Creating Corpora for Computational Research in Arab-Andalusian Music

Mohamed Sordo
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
mohamed.sordo@upf.edu

Amin Chaachoo
Asmir Center for Musicological Research
Tetouan, Morocco
chaachooamin@gmail.com

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

ABSTRACT

Research corpora are fundamental for the computational study of music. The design criteria with which to create them is a research task in itself. These corpora need to be well suited for the specific research problems to be addressed. Since these research problems are also shaped by musical, cultural and other specific aspects of the music traditions to be studied, the research corpora should take these specificities into account. In this paper we address the problems of creating corpora for computational research on Arab-Andalusian music, considering several relevant criteria for creating such corpora. We focus on the problems raised during the annotation process of the corpora, specifically the language issues surrounding this art music tradition. Following the criteria, we created a research corpus consisting of audio recordings with their corresponding metadata, lyrics and music scores. So far we have gathered 338 recordings from 3 different Arab-Andalusian music schools of Morocco, covering most of the musical modes, rhythms and forms of this art music tradition. The Arab-Andalusian corpus is accessible to the research community from a central online repository. Moreover, the audio recordings of this corpora are freely available through the Internet Archive repository. The Arab-Andalusian corpus can be used to generate test datasets, which can be used as ground truth to test several computational research tasks.

Categories and Subject Descriptors

H.3.9 [**Information systems**]: Information systems applications—*Digital libraries and archives*

General Terms

Digital Libraries, Research Corpora

1. INTRODUCTION

In the context of the CompMusic [13] project we aim to study several non-western art music traditions. A major effort has been dedicated to the creation of appropriate data collections for the study and characterization of the cultural specific aspects of these traditions. Building a research corpus is a research problem by itself and it has been studied in other disciplines (see for instance [9, 16]). There have been efforts to compile large collections of music related data such as the Million Song Dataset [1]. However, few efforts have been dedicated to study systematic ways to compile a research corpus for music [10, 15]. Recently, Serra [14] provided a set of criteria to compile and curate research corpora. These criteria include concepts such as purpose, coverage, completeness, quality and reusability. In this paper we refer (directly or indirectly) to these concepts while providing a specific criteria for Arab-Andalusian music. Following this criteria we created a research corpus for Arab-Andalusian music consisting of audio recordings with their corresponding metadata, lyrics and music scores. The corpus is openly available for research[1]. When possible, we emphasize the use of open repositories of information such as MusicBrainz and Internet Archive. In CompMusic, we are developing a tool for navigating through music collections called Dunya [12], which also acts as the central online repository to store the metadata, audio, annotations, lyrics, scores and research results. Dunya provides an API[2] for accessing these data.

DLfM '14, September 12 2014, London, United Kingdom.

ACM 978-1-4503-3002-2/14/09$15.00
http://dx.doi.org/10.1145/2660168.2660182 .

2. ARAB-ANDALUSIAN MUSIC

The Arab-Andalusian music (from now on AAM) is a musical tradition that can be traced back to the 9th century in Al-Andalus, to the muslims and christians living in Moorish Spain [2]. AAM is the result of many influences, including Middle-East Arabic classical music, the Hispanic music traditions of the Iberian peninsula, and other classical traditions such as the Gregorian and Byzantine. The Andalusian tradition is maintained in quite a few north African regions [5, 11], mainly in Morocco, Algeria, and Tunisia. In this paper we focus on the Andalusian music of Morocco. AAM has been preserved and kept alive as an oral tradition. It is typically associated with certain social occasions and circumstances, such as weddings, festivals, etc. The modern study of the AAM theory started in the colonial period (20th century), but most of these studies did not consider the plurality and the diverse influences of AAM [2].

AAM is a patrimony music, which means its compositions are fixed, although ornamentations and controlled improvisations (i.e., improvisations which respect the modes and characteristics of AAM) are allowed. The main concept of AAM is *ṭabʿ*, which is the andalusian term for musical mode, and at the same time means the emotional state produced by the melodies of this mode in both the performer and listener [2]. All the melodies of a particular *ṭabʿ* constitute an homogeneous set called *nawba*. A *nawba* is structured as a sequence of five myazen (plural of *mizān*), each one on a particular rhythmic pattern. Each *mizān* begins with an instrumental

[1] `http://compmusic.upf.edu/corpora`

[2] `https://github.com/MTG/dunya/blob/master/andalusian/API_README.md`

Form	Recordings
Andalusian music	165
Popular music	16
Taqsīm	24
Mshāliya	33
Tawshiya	33
Inshād	22
Tawshiya Sanʿa	33
Mawwāl	12
TOTAL	338

Table 1: Corpus recordings

Nawba/Mizān	Basit	Qāʾim wa Nisf	Btayhi	Darj	Quddām
Raml al-Māya	3	2	5	0	4
al-Isbahān	1	3	4	0	5
al-Māya	4	1	3	1	10
Rasd al-Dayl	6	3	3	0	6
al-Istihlāl	3	6	5	3	4
al-Rasd	3	0	3	0	6
Gharībat al-Husayn	5	3	1	0	3
al-Hijaz al-Kabir	3	1	3	1	2
al-Hijaz al-Msharqi	4	0	5	0	5
Irāq al-Ajam	1	0	4	0	7
al-Uššāq	3	3	1	0	4

Table 2: Distribution of the Andalusian music recordings.

prelude and is followed by a number of sung poems, *sanāʾiʿ*, which include instrumental accompaniment and short instrumental interludes. Additionally, each *mizān* has three versions depending on its cadence: *muassaʿ* (slow), *mahzúz* (intermediate) and *muṣarraf* (fast). A complete *nawba* in AAM of Morocco can last up to six or seven hours, which makes it practically unfeasible for music orchestras to perform a full *nawba*. Usually an orchestra will perform a certain *mizān* of a given *nawba*. The naming convention among musicians is to use the name of the *mizān* followed by the name of the *nawba*. For example, *quddām msharqi* refers to the rhythm *quddām* of nawba *msharqi*. There are 26 *ṭabʿs* in AAM, but only 11 *nawbas* have been preserved [2]. Given its oral tradition many of the *ṭabʿs*' melodies were lost so coherent *nawbas* could not be formed with the remaining melodies. For this reason, some Moroccan music experts annexed these remaining melodies to existing *nawbas* with similar *ṭabʿ*. Apart from the concepts of *ṭabʿ*, *nawba* and *mizān*, there are other forms or genres in the AAM tradition, such as *tawshiya* (measured instrumental prelude), *mshāliya* (free instrumental prelude), *taqsīm* (instrumental improvisation), *inshād* (a sung poem with 2 verses), *mawwāl* (free vocal improvisation), etc. which are mainly used as preludes or interludes between the different sections of a *nawba* [2].

3. RESEARCH CORPORA

The AAM research corpus mainly consists of audio recordings with their corresponding metadata, lyrics (or *sanāʾiʿ*) and music scores. In the following subsections we describe their characteristics along with the issues we found while gathering them.

3.1 Recordings

So far we have gathered 112 hours of recordings, performed by four different orchestras from the three most important music schools in Morocco (Fes, Rabat and Tetouan). Table 1 shows the distribution of the corpus recordings into the different genres of the tradition, whilst Table 2 provides the distribution of the AAM recordings into their corresponding *nawba* and *mizān*. The AAM recordings cover most of the different combinations of *mizān* and *nawba*, except for *mizān darj*, where only few recordings could be gathered. Most of these pieces were recorded in the 1960s and 1970s and mainly come from radio programs and personal recordings. When compared to more contemporary recordings, the production quality of the chosen recordings is somehow poor. However, the reason to use these older recordings is because of the high musicality and virtuosity of the performing orchestras at that time, which included some of the most recognized masters of AAM in Morocco. Additionally, unlike most of the contemporary recordings, the selected recordings can be freely distributed. Given our focus on reproducibility, we are making all of these recordings available on Internet Archive, more specifically on the Community Audio collection. In order to ease the access to the recordings, we annotate and tag each recording, including the name of the performance (*mizān* + *nawba*), the name of the orchestra, the date of the performance, the performance type, the style of music and the country of origin. The recordings then can be easily accessed via a search query[3]. Even though the collection size might seem small for other musics, we consider it to be sufficiently representative and complete for a patrimony music like AAM.

The editorial metadata associated with each recording has been stored and organized in MusicBrainz[4]. This includes the title of the release (which typically refers to the event where the recording was performed), the title of the recording (which contains the *mizān* and *nawba* of the recording), the name of the orchestra, the director of the orchestra and the artists who performed in that recording, along with their instrument. For each one of these "entities", a unique identifier (an MBID) is assigned in MusicBrainz. This provides an effective way to organize metadata. In order to avoid mismatches in metadata due to different transliterations, we decided to enter the metadata in its original Arabic script. Nonetheless, we also provide a roman transliteration of the metadata, which is currently stored in our own dataset, and also available through the previously mentioned Dunya API. Roman transliteration of Arabic is not a trivial process. Apart from the different existing standards, the difficulty resides in the fact that Arabic is normally not vocalized. Arabic is a type of Abjad language [4]. In other words, fluent readers of Arabic do not need vowels (and other diacritics) in order to solve all the probable ambiguities and to read it properly. This however causes a serious problem from a computational perspective. In the last few years there has been a growing interest in vocalizing Arabic text [6, 18]. For this corpus we have developed our own code for transliterating Arabic metadata. The code first vocalizes the text using the implementation by Zarrouki [17], and then implements the American Library Association - Library of Congress (ALA-LC) standard for transliterating Arabic[5]. This code (which still has some issues, especially the vocalization module) is freely available[6] for researchers and developers to use, modify and distribute.

Finally, we also annotated the starting and ending time stamps of each section in the audio recordings. Each section may correspond to one of the three main sections of a *nawba* (*muassaʿ*, *mahzúz* and *muṣarraf*) or other forms of AAM. The annotations can be very

[3] http://archive.org/search.php?query=arab-andalusian

[4] http://musicbrainz.org/collection/142ea0d7-7fdf-4ea5-9b04-219f68023d01

[5] http://www.loc.gov/catdir/cpso/roman.html

[6] https://github.com/CompMusic/ArabicTransliterator

useful for tasks such as audio segmentation, music discovery, etc.

3.2 Lyrics

AAM is in fact a succession of sung poems called *sanā'i'* (plural of *san'a*), which concur with the mode and rhythm of the *nawba*. These poems typically follow the classical Arabic poetry patterns. Furthermore, there are two types of poems in AAM: muwaššaḥa and zajal [8]. The succession of *sanā'i'* can vary in number or order depending on the music school, the context (e.g., festivals, weddings, etc.) and the director of the orchestra.

At the end of the 18th century, Mohammed ibn al-Hussein al-Haik, a renowned Moroccan poet and musician, published a songbook called Kunnaš Al-haik [7] compiling all the 11 *nawbas* with their corresponding *sanā'i'*. This songbook is still considered nowadays as one of the most important references for AAM. Nevertheless, despite this effort, due to lack of communication, the work by al-Haik could not reach all the schools of Morocco. Many schools used only those poems that already existed in their context to create the songs. This resulted in recordings of the same AAM songs by different schools to differ from each other.

For this corpus we use the *sanā'i'* as defined and compiled by Mehdi Chaachoo [3], which take the school differences into account. Each recording of the corpus has its corresponding succession of *sanā'i'*, manually selected from Chaachoo's songbook by listening to all of the recordings. At the time of writing these manuscript we have gathered lyrics of the recordings corresponding to 9 *nawbas* (out of the existing 11 *nawbas*). Furthermore, we have also segmented the lyrics by sections (see Section 3.1). This segmentation can be useful for computational tasks such as audio to lyrics alignment or score to lyrics alignment, and also for more musicological studies. Finally, the lyrics are also romanized using the ALA-LC standard as explained in Section 3.1.

3.3 Music scores

Even though the melodies in AAM are fixed, they are subject to numerous ornamentations and improvisations by the musician, depending on his emotion or state of mind. For this reason, two performances of the same music piece by the same orchestra might have substantial melodic and structural differences. Hence, it is more appropriate to have transcriptions at a recording level rather than a composition level. In many cases, due to the large amount of ornamentations, it might be possible that the original melody is lost [2]. Moreover, unlike more rational musics, AAM cannot be performed only from the score, but also from the tradition. Thus, the central element of AAM is the recording (performance). Scores are complimentary to the recordings for the study of AAM.

For the research corpus, we asked an Arab-Andalusian musicologist to transcribe all the recordings of our collection. So far we have managed to gather the transcriptions of the recordings of 3 *nawbas* (out of a total of 11 *nawbas*). The transcriptions are stored in `MusicXML` format.Currently the transcriptions show the score, as well as the tempo and the name of each performed *san'a*. As future work, we would also like to align the scores with their corresponding lyrics.

4. SUMMARY

We presented a criteria for creating research corpora for computational research on Arab-Andalusian music, putting an emphasis on the cultural specificities of this art music tradition. Following this criteria we created a research corpus of Arab-Andalusian music consisting of audio recordings with their corresponding metadata, lyrics and music scores. We discussed various issues surrounding the creation of the corpus, such as language issues and other cultural specific aspects of this art music tradition. We are continuously putting effort into growing and completing this research corpus, especially the lyrics and the music scores. The research corpus (including the audio recordings) is freely available through open repositories and our project's central repository.

5. ACKNOWLEDGMENTS

The compilation of the corpora presented in this paper has been a collective effort of the members of the CompMusic project. The CompMusic project is funded by the European Research Council under the European Union's Seventh Framework Program (ERC grant agreement 267583).

6. REFERENCES

[1] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. of the 12th ISMIR*, pages 591–596, 2011.

[2] A. Chaachoo. *La Música Andalusí. Al-Ála. Historia, conceptos y teoría musical.* Almuzara, Córdoba, 2011.

[3] M. Chaachoo. *Diwan Al-Ala*. Imprimerie Al Khalij Al Arabi, Tetouan, Morocco, 2009.

[4] P. T. Daniels and W. Bright. *The world's writing systems*. Oxford University Press, 1996.

[5] M. Guettat. *La musique arabo-andalouse. L'empreinte du Maghreb*. El Ouns/Fleurs sociales, Paris/Montreal, 2000.

[6] N. Habash and O. Rambow. Arabic diacritization through full morphological tagging. In *Proc. of NAACL-HLT 2007*, pages 53–56, 2007.

[7] I. A. M. ibn al-Hussein al-Haik (ed. Malik Bennouna). *Kunnash al-Haik*. Rabat, Morocco, 1999.

[8] J. T. Monroe. Zajal and muwashshaha: Hispano-arabic poetry and the romance tradition. In S. K. Jayyush, editor, *The Legacy of Muslim Spain*, pages 398–419. EJ Brill, 1992.

[9] S. Pan and W. Weng. Designing a speech corpus for instance-based spoken language generation. In *Proc. of Int. Conf. on Natural Language Generation*, 2002.

[10] G. Peeters, K. Fort, et al. Towards a (better) definition of annotated mir corpora. In *Proc. of the 13th ISMIR*, 2012.

[11] C. Poché. *La musique arabo-andalouse*. Cité de la Musique / Actes Sud, Paris/Arles, 1995.

[12] A. Porter, M. Sordo, and X. Serra. Dunya: A system for browsing audio music collections exploiting cultural context. In *Proc. of 14th ISMIR*, pages 101–106, 2013.

[13] X. Serra. A multicultural approach in music information research. In *Proc. of the 12th ISMIR*, pages 151–156, 2011.

[14] 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, 2014.

[15] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra. Corpora for music information research in indian art music. In *Proc. of the Sound and Music Computing Conference*, 2014.

[16] M. Wynne, Arts, and H. D. Service. *Developing linguistic corpora: a guide to good practice*, volume 92. Oxbow Books Oxford, 2005.

[17] T. Zarrouki. mishkal. `https://github.com/linuxscout/mishkal`, 2014.

[18] I. Zitouni, J. S. Sorensen, and R. Sarikaya. Maximum entropy based restoration of arabic diacritics. In *Proc. of the 21st Int. Conf. on Computational Linguistics*, pages 577–584, 2006.