Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project

Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona

Correspondence should be addressed to Xavier Serra (xavier.serra@upf.edu)

ABSTRACT

A fundamental concern in music information research is the use of appropriate data sets, research corpora, from which to perform the needed data processing tasks. These corpora have to be suited for the specific research problems to be addressed and the design criteria with which to create them is a research task to which not much attention has been paid. In the CompMusic project we are studying several non-western art music traditions and a major effort has been the creation of appropriate data collections with which to study and characterise the melodic and rhythmic aspects of these traditions. In this article we go over the criteria used to create these collections and we describe the specificities of each of the collections gathered.

1. INTRODUCTION

In the CompMusic project[1] we work on computational approaches to describe music recordings by emphasising the use of domain knowledge of particular music traditions. We are focusing on five music cultures: Arab-Andalusian (Maghreb), Beijing Opera (China), Turkish-makam (Turkey), Hindustani (North-India) and Carnatic (South-India). A target application for this work is a system with which to browse through audio music collections of the chosen cultures; being able to discover specific characteristics of the music and relationships between different musical concepts. The target user of the system is a music lover of the particular music tradition.

There are many great traditions that we could have chosen for our work, but we had to limit ourselves and make decisions. The main criteria used for selecting the specific traditions were: (1) they had to have musical personalities contrasting with the current popular western music, (2) they had to have alive performance practices and strong social and cultural relevance, (3) there had to be musicological and cultural studies available about them, (4) it had to be feasible to put together large enough research corpora of machine-readable music data from existing sources, and (5) the selected traditions had to provide a diverse set of music repertoires with which to study new and diverse MIR problems.

In this paper we focus on the issue of the research corpora, on the criteria used to create them and on describing the specificities of each of the collections gathered.

2. CULTURE SPECIFIC CORPORA

Isn't music a universal language? Why then focus on particular music traditions and on culture specific corpora?

The discussion of the universality of music is a long one and it is completely biased by the discipline within which the discussion takes place. For example ethnomusicologists tend to focus on identifying musical specificities of particular musical cultures while in cognitive sciences a major interest is in finding musical capacities that might be universal.

Among ethnomusicologists there is a wide consensus that all musics are not alike, and that the approaches used to understand Western musics are not appropriate for other types of musics. Bruno Nettl [1] offers a good discussion on some of these topics. He supports the idea that music is not a universal language but that the diverse musics that exist are not as mutually unintelligible as languages. In his article he tries to identify some music universals, but he only succeeds in identifying very simple and broad ones. For example he points out that all musical cultures have singing and some instrumental music and that in most vocal music the chief melodic interval is around the major second. He also identifies that most musical utterances tend to descend at the end and that

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

most musics have a rhythmic structure based on a distinction among note lengths and among dynamic stresses.

In the fields of neuroscience and cognition most of the studies have been focusing on finding universals but more recently some studies are also concerned with the topic of cultural diversity, thus trying to identify cultural differences in music cognition. Morrison and Demorest [2] write that there is a great and often culture-specific diversity of musical practices differentiated in part by form, timbre, pitch, rhythm, and other structural elements. They claim that musical interactions situated within a given cultural context begin to influence human responses to music as early as one year of age.

An excellent attempt in finding music universals is the work by Brown and Jordania [3]. In their paper they argue in favour of studying music universals, mainly for comparative analysis, and by building on top of Nettl's work they give an excellent typology for their proposed music universals.

For our research approach we start from the idea that there are universal musical concepts, like melody and rhythm, but we want to emphasise, and show, that many important aspects of a particular music recording can be better understood by considering cultural specificities. To emphasise this view we have focused on a few art music traditions, thus on types of music that have been formalised and for which theoretical frameworks have been proposed for their understanding. We have explicitly excluded Western classical music from our selected cultures. We wanted to promote a fresh approach to our research, without the bias that current MIR research has towards western pop music and thus influenced by western classical music concepts.

3. CRITERIA FOR CREATING CORPORA

The audio music collections of each music culture plus the accompanying information configure the core of the CompMusic project. They constitute our research corpora. But what constitutes a good research corpora?

The design of research corpora is a research problem in itself. These corpora have to be appropriate for all the data driven research that wants to be carried out, but very little has been written about design principles of corpora from the perspective of information science [4], there are a few references in the use of standardise corpora in fields such as speech processing and linguistics, but basically nothing in the field of MIR.

We should not confuse the idea of research corpora with the one of test corpora (also known as test collections). Research corpora are collections of authentic data used to perform experiments to advance knowledge. The test corpora are the ground truths, collections of authentic or invented data used for testing, evaluating performance, and calibrating the tools used in experimentation. A recent paper by Peeters [5] referred to the problem of test corpora in MIR. Here we are concern with research corpora, thus on putting together corpora that capture the essence of a particular music culture for performing our research work.

For the case of CompMusic we have been putting together research corpora for the five music traditions being studied. The types of data collected are audio recordings and editorial metadata, which are complemented with descriptive information about the items we have, and in some cases with music scores and/or lyrics. The core unit in a corpus is an audio recording and the set of information items that accompany it.

The main criteria that we have taken into account in the creation of the corpora have been: Purpose, Coverage, Completeness, Quality, and Reusability.

Purpose: The first step in the design of a corpus is to define the research problem that wants to be addressed and the research approach that will be used. In our case we want to develop methodologies with which to extract musically meaningful features from audio music recordings, mainly related to melody and rhythm. The approaches are based on signal processing and machine learning techniques; thus the corpus has to be aligned with this purpose.

Coverage: A corpus has to include data representative of all the concepts to be studied and given our quantitative approach, there has to be enough samples of each instance for the data to be statistically significant. For our research we need to have audio recordings, plus appropriate accompanying information, covering the varieties of melodies and rhythms present in each musical culture.

Completeness: In each corpus every audio recording is complemented with a set of data fields and the idea of completeness relates to the percentage of fields filled, thus how complete the corpus is. For our corpora this mainly refers to the completeness of the editorial metadata and of the descriptive information accompanying each audio recording.

Quality: The data has to be of good quality. The audio has to be well recorded and the accompanying information has to be accurate. We have used well produced recordings whenever possible and the accompanying information has been obtained from reliable sources and validated by experts.

Reusability: The research results have to be reproducible and that means that the corpus has to be available for the research community to use. In our case we have emphasised the use of specific open repositories that are either already suitable or than can be adapted to our needs. For example, for editorial metadata we use MusicBrainz,[2] for other complementary information we use repositories like Wikipedia, and for the audio, for which it is practically impossible to use open repositories, we use easily available commercial recordings.

4. ARAB-ANDALUSIAN MUSIC CORPUS

The main concept of Arab-Andalusian music is *ṭab'*, which is the andalusian term for musical mode that also relates to the emotional state produced by the melodies of this mode in the listener [6]. All the melodies of a particular ṭab' constitute an homogeneous set called *nawba*. Each nawba is structured as a sequence of five *myazen* (plural of *mizán*), each one on a particular rhythmic pattern. Each mizán begins with an instrumental prelude and is followed by a number of sung poems, *sana'i*, which include instrumental accompaniment and short instrumental interludes. Additionally, each mizán has 3 versions depending on its cadence: *muassa'* (slow), *mahzúz* (intermediate) and *muṣarraf* (fast). A complete nawba in the Arab-Andalusian music tradition of Morocco can last up to six or seven hours, which makes it practically unfeasible for orchestras to perform a full nawba in a concert. Usually an orchestra, in a concert, will perform one mizán of one nawba. The naming convention among musicians for the mizán played is to use the name of its characteristic rhythm followed by the name of the nawba it belongs to. For example, *quddam msharqi* refers to the rhythm quddam of ṭab' msharqi. Sometimes, the recordings only include a part of a mizán, thus only one version of the rhythm. In this case, the rhythm version name will precede the name of the rhythm. For example, *insiráf quddam msharqi* refers to the fast cadence version of the rhythm quddam of ṭab' msharqi.

So far we have gathered 56 hours of recordings, performed by 3 different orchestras from Morocco (Orchestra Bríhí of the city of Fes, Orchestra of the Tetouan Music Conservatory, and Orchestra of Tetouan) and a big ensemble that combined 2 orchestras (Orchestra Bríhí of Fes and Orchestra of the Tetouan Music Conservatory). Table 1 shows the number of instances of each rhythm/nawba pair in the recordings. Our goal is to increase the collection size up to 200 hours of recordings, filling the absent rhythm/nawba pairs, which we consider a sufficiently big corpus to carry out both our intended qualitative and quantitative research.

All the recordings have been selected by the musician and musicologist Amin Chaachoo [6]. Most of them were recorded in the 1960s and 1970s and they mainly come from radio programs and personal recordings. The reason to use these older recordings is because of the high musical quality of the performing orchestras at that time, which included some of the most recognised maestros of Arab-andalusian music in Morocco. One problem is that the recording quality of many of them is not as good as we would have liked, but we cannot do much about that since there are very few well produced commercial recordings of this music.

Given our focus on reusability, we are making all of these recordings available on Internet Archive,[3] and all the metadata associated to them are being stored and organised in MusicBrainz.[4] All the metadata has been entered in its original language and script, Arab.

About music scores we have not yet decided what strategy to follow. We have a well published collection of the lyrics of the whole repertoire [7] and a good set of music transcriptions [8], but not in machine readable form. It should be feasible to convert a representative part of it into a digital format like MusicXML.[5]

5. BEIJING OPERA MUSIC CORPUS

The music in Beijing opera is mainly structured according to two basic principles, *shengqiang* and *banshi*, which in a broad sense define respectively its melodic and rhythmic components [9]. On top of these two structural principles, the system of role-types (formalised performative categories according to which all fictional

[2] `http://musicbrainz.org/user/compmusic/collections`

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

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

[5] `http://www.musicxml.com/`

Nawba/Mizán	Basít	Qaïm Wa Nisf	Btayhí	Darj	Quddám
Raml al-Máya	3	0	3	0	2
al-Isbahán	1	2	2	0	2
al-Máya	2	0	4	1	2
Rasd al-Dayl	5	1	1	0	1
al-Istihlál	2	4	4	3	1
al-Rasd	1	0	3	0	2
Garíbat al-Husayn	3	0	2	1	2
al-Hijaz al-Kabir	1	1	3	1	0
al-Hijaz al-Msharqi	0	0	4	1	3
Iráq al-Ajam	1	0	1	0	4
al-Ussáq	1	2	1	0	1

Table 1: Numbers of the Arab-Andalusian audio collection.

characters are constructed) impose particular constrains to the execution of shengqiang and banshi. The interaction of these three components, hence, offers a substantial account of Beijing opera music. Our goal in Beijing opera is to characterise these components and analyse their interaction. For this, the corpus being collected for Beijing opera is composed of audio recordings, music scores, lyrics, and the accompanying metadata.

Beijing opera is an eclectic art form in which music is a core element, but not the only one, since the mime, dance and acrobatics are also very important. For our research we have decided to use the arias as unit of analysis; in them the musical aspect of the opera is most fully expressed and for that we have purchased well produced commercial CDs, albums, containing compilations of arias. Our current collection includes 48 albums, which contain 510 recordings (tracks) featuring 381 arias and over 46 hours of audio. In terms of coverage of the corpus, some observations must be mentioned regarding the delimitations of the research scope in terms of the three basic principles defined previously. Since most of the arias are formed by two or more banshi, the corpus has a representative sample of the most common ones. As for shengqiang, for historical reasons, Beijing opera uses many different ones, although two of them, *xipi* and *erhuang*, conjointly known as *pihuang*, are considered to be the core ones. Considering the numbers of the audio collection in table 2, although we have samples of 11 different shengqiang, recordings of arias sung in xipi and erhuang sum 413, representing 81% of the collection. Therefore, due to the importance of these two shengqiang as stated in the literature and supported by the data in the corpus, research will focus on them. Finally, the system of role types is a complex one with many subdivisions and subtleties that we are not yet considering. The five categories shown in the table have been selected according to their specificities regarding singing and consequently musical relevance. From these observations and our defining criteria we are aware that the corpus is still to be improved. From table 2 it can be argued that the sample of singers, especially for some role-types, is still not big enough for some of the proposed computational analysis. In terms of representativeness, there are some considerations not yet taken into account that should be regarded in the future. These are related with singing schools covered by the corpus, which might have a direct influence in the musical performance, the generations of singers, and the historic period in which the arias were created.

The editorial metadata of the albums is stored and organised in MusicBrainz.[6] All the metadata has been entered in its original language and script, that is Chinese and Chinese simplified characters, but transliteration according to the pinyin system are also offered. In the case of releases and recordings the transliterations are stored as related pseudo-releases and for artists and works they appear as aliases. Other extra information, such as role-type, shengqiang, banshi and school cannot yet be properly stored in the MusicBrainz repository but we are working for making it possible.

Besides the audio recordings and the metadata describing it, we plan to include music scores. We have access to

[6] `http://musicbrainz.org/collection/40d0978b-0796-4734-9fd4-2b3ebe0f664c`

two published collections of scores [10] [11] which include the scores of 151 of the arias for which we have recordings, 40%; proportion that ought to be improved. However the main problem is that we do not have access to machine readable versions of these scores, so we will have to input a relevant part of these scores into a digital format. We also plan to include the lyrics of all the arias for which we have recordings, which we can easily gather from publicly accessible websites.[7]

6. TURKISH MAKAM MUSIC CORPUS

This corpus is centred on the Ottoman/Turkish traditional/classical/art music (geleneksel/klasik Osmanlı/Türk (sanat) musikisi') and for this music, *makam* is the main defining musical concept. Makam can be best understood within the context of modal practice and it lies in between the concept of a scale and of a tune. The rhythmic counterpart to the melodic concept of makam is *usul*. An usul is a rhythmic pattern of a certain length that contains a sequence of strokes with varying accents. Within CompMusic we are studying diverse aspects of this music tradition related to these concepts, such as intonation, tonic detection, melodic characterisation and rhythmic characterisation [13].

For the rhythmic and melodic study of the makam music tradition, audio recordings together with music scores are fundamental data sources. These two sources allow us to analyse the music from different points of view. However, in order to take the most advantage of this complementarity we need to time align them. This is a research task that needs to be done specifically for this repertoire and that we have already started [14].

In term of audio recordings, the Turkish Radio and Television (TRT) probably has the most representative and better produced collection of makam music recordings. Unfortunately their archives are not publicly accessible and only a small part of the collection has been commercially released as CDs. In this respect our efforts have been focused on putting together a collection from commercially available CDs, albums. We currently have 225 albums, which include close to 2000 different compositions on all the common makams (Table 3) and it covers a variety of genres (e.g. classical, folk, religious) forms (e.g. ilahi, şarkı, peşrev) and instrumentations. We have stored and organised the editorial information in MusicBrainz.[8]

With respect to the scores, we have gathered a collection that was curated by M. K. Karaosmanoğlu [12]. This collection, named symbTr, includes 1700 musical pieces featuring 155 different makams, 100 usuls and 48 forms. It covers compositions from both Turkish classical and folk music traditions spanning different historical eras. The vocal pieces also include the lyrics aligned to the music. Currently the representation used is a simple text-based format but we are in the process of converting the whole collection to MusicXML. This score collection is the most comprehensive collection of publicly available machine readable scores of Turkish makam music.

We have already obtained some research results by using part of this corpus [15] [14] [16].

7. HINDUSTANI MUSIC CORPUS

For the case of the Indian art music traditions, both Hindustani and Carnatic, *rags* and *tals* are the basic concepts with which to describe melody and rhythm. These concepts have been the main concern around which this corpus has been compiled. Raga eludes a simple and concise definition, but technically we can say that a raga is a musical entity in which the choice of notes; their order and hierarchy, the manner of intonation of individual notes, relative duration and their specific melodic approach, are clearly defined. Tala is the rhythmic framework governing the temporal aspect of the music and it can be described as having a certain number of time units or beats (*matra-s*) and more importantly sections into which these beats are grouped and stressed.

Given the musical prominence of the voice, we have emphasised the recordings in which the voice is the lead instrument. In order to make our decisions we have used the help of musicologists, like Suvarnalata Rao [17], and of expert musicians.

The Hindustani music tradition is very heterogeneous and thus gathering a representative corpus is a difficult task. Because of this, we also decided to concentrate on two of the major traditional styles: *dhrupad* and *khyal*. We identified three institutions that have already compiled audio archives that could be used for our purposes: ITC Sangeet Research Academy,[9] Sangeet Natak

[7] such as http://www.xikao.com and http://www.jingju.com

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

[9] http://www.itcsra.org

Role Type	Recordings	Singers	Albums	Xipi recordings	Erhuang recordings	*Duration*
Laosheng	189	14	16	93	80	16h 8m
Jing	56	5	5	22	28	4h 13m
Laodan	20	3	3	6	7	2h 23m
Dan	186	17	18	85	45	18h 12m
Xiaosheng	59	9	6	40	7	5h 11m
Total	510	48	48	246	167	46h 7m

Table 2: Numbers of the Beijing Opera audio collection.

Recordings	3841
Albums	225
Works	1980
Artists	584
Composers	422
Lyricists	316
Makams	129
Usuls	67
Total duration	258h 43m

Table 3: Numbers of the Turkish Makam audio collection.

Academy,[10] and All India Radio.[11] These audio archives have been curated by experts and represent a real-world scenario, each consisting of several thousands hours of music. ITC Sangeet Research Academy is one of the premier music academies dedicated to Hindustani music and it has done a major effort in archiving audio recordings. The Sangeet Natak Academy is India's national academy for music, dance and drama, and it has a huge collection of Hindustani music recordings. Finally, the All India Radio is the biggest broadcaster in India, it has served as a musical reference when talking about Hindustani music and also hosts a large archive of Hindustani music recordings. However these archives are not publicly available; we had to create our collection from well produced commercial CDs but we used these recognised institutions as reference.

Table 4 gives some useful numbers on the current audio music collection. For example, the number of rags and tals gives an indicative measure of the diversity/coverage of our recordings and the number of lead artists show the different stylistic variations and nuances of our repertoire. However we still need to develop appropriate statistics for measuring the coverage and completeness of the corpus.

The editorial metadata of the CDs has been stored and organised in MusicBrainz.[12] All the text information has been entered in english given that this is the language that most of the CD albums use. We have helped develop MusicBrainz in order for it to support the specific concepts used in Hindustani music.

The music scores and the lyrics in Hindustani music are not as relevant [17] as in the other traditions and we have decided to not include them in our corpus. However the *bandish-s* are short melodic compositions used as basis of improvisation for which there are online repositories[13] and that eventually could be added to the corpus.

We have done some computational work to study the melody and rhythm of this music by using part of the collection [18] [23] [16].

8. CARNATIC MUSIC CORPUS

Like in Hindustani music, *ragas* and *thalas* are the two basic concepts with which to describe the melody and

[10] http://sangeetnatak.gov.in

[11] http://allindiaradio.gov.in

[12] http://musicbrainz.org/collection/213347a9-e786-4297-8551-d61788c85c80

[13] https://www.swarganga.org/bandishbase.php

Recordings	970
Albums	233
Lead artists	98
rags	278
tals	42
Total duration	271h

Table 4: Numbers of the Hindustani audio collection.

rhythm in Carnatic music. Even though there are substantial differences between these two Indian art music repertoires, the general description of raga and thala given in the previous section also apply here.

The voice has a major prominence, even more than in Hindustani music, so in most of the compiled recordings singers are the lead artists. For our selection of recordings we have consulted with expert musicians and musicologists, like T M Krishna [19].

The main institutional reference for gathering the collection has been the Madras Music Academy (MMA).[14] This is an institution that has been hosting an annual music festival since 1935, a festival considered the main model for the Carnatic music tradition. The MMA has also been hosting scholarly discussions for a long time and these have shaped the evolution of the musical concepts being used. The Academy has an expert committee of veteran and contemporary artists, who formulate and follow a procedure by which artists are selected for the festival, criteria that has a strong influence for the reputation of Carnatic musicians. The MMA has been recording their concerts for a long time and its archive is a main reference for this music tradition. However, this archive, like most others, is not openly available online and thus we had to rely on gathering commercial CDs, even while following the musical criteria used by the MMA. The main record company specialised in Carnatic music is Charsur[15]. They have been publishing high quality CDs for 15 years and thus the core of our collection is their catalog of music concerts.

In Carnatic music, the concert, *kutcheri*, is a natural musical unit that is also used as the unit for music distribution (typically a concert is recorded in two CDs). We have use it as the main entity in the corpus. Also in gathering the audio collections we wanted to cover several generations of musicians. For this, we started from the artists that have been performing at the MMA during the last 5 years, and then expanded the collection by including their teachers, *gurus*. Table 5 shows some numbers of the corpus. The corpus is organised in terms of concerts, recordings (songs), artists, ragas, thalas, and composers.

The editorial metadata has been stored and organised in MusicBrainz.[16] All the text information has been entered in english, given that this is the language that most of the published recordings use. We have gathered much information about the artists in the collection, mainly from Wikipedia and Kutcheris.com.[17] We also added a lot of information by ourselves with the help of experts.

The Carnatic tradition is very much composition based, which is one of the main differences with Hindustani music. Because of this, we plan to gather scores and lyrics in order to be able to carry out some computational analysis using them. There exists published compilations of scores from most of the currently performed compositions, for example the ones by the three most recognised composers; Tyagaraja [20], Syama Sastri [21] and Dikshitar [22]. Again the problem is to have them in a machine readable form. Just for the lyrics there are good online repositories, like shatyam.net.[18]

The Carnatic corpus has been the most worked on so far in CompMusic, being the most complete one with respect to the criteria identified. We are getting close to have a corpus that can be used for many research and application tasks. We have already done some computational work to study the melody and rhythm of this music by using part of this collection [18] [25] [23] [16] and we have been able to develop a preliminary version of a web discovery system that exploits this corpus in a practical application [24].

[14] `http://http://www.musicacademymadras.in`

[15] `http://www.charsur.com`

[16] `http://musicbrainz.org/collection/f96e7215-b2bd-4962-b8c9-2b40c17a1ec6`

[17] `http://kutcheris.com`

[18] `http://www.shatyam.net`

Concerts (releases)	248
Recordings	1889
Lead artists	80
Violin artists	31
mridangam artists	32
ragas	249
thalas	20
Composers	152
Total duration	397h 10m

Table 5: Numbers of the Carnatic audio collection.

9. CONCLUSSIONS

In this article we have addressed the issue of the development of research corpora for music information research, using the CompMusic project as an example.

Putting together corpora is a demanding and costly effort that cannot be easily accomplished within most research projects. In CompMusic we have been fortunate to have the resources for doing it. The work is not finished yet, but we believe that the collections already compiled should be valuable for the research community and for tasks in music information research that go beyond our project's specific tasks.

In the process of putting together the different corpora it became clear that there were no stablished criteria from which to develop the corpora. Thus, in parallel to creating the corpora, we have been identifying these criteria and in this paper we have tried to make them explicit and exemplify them. There is still a lot to do, both in terms of developing more concrete criteria for developing research and test corpora and in terms of having quantitative ways for measuring how well a given corpus fulfils those criteria. From our first analysis we know that we have to further develop our corpora, some more than others, specially in terms of coverage and completeness.

In parallel to the research corpora we have also been putting together test corpora for some of the particular research problems we are working on. We have not mentioned anything about them in this article. The criteria used to create these test corpora is more related with the evaluation carried out in the particular research task [26]. For example we have created a corpus for tonic identification in both Carnatic and Hindustani music [25] and one for rhythm analysis in Carnatic music [16]. We are in the process of defining the best way to organise and distribute these test corpora and the ones we will develop.

10. ACKNOWLEDGEMENTS

The gathering of the corpora presented here has been a collective effort in which all the members[19] of the CompMusic project have participated plus a number of collaborators. The CompMusic project is funded by the European Research Council under the European Unions Seventh Framework Program (FP7/2007-2013) / ERC grant agreement 267583.

11. REFERENCES

[1] B. Nettl, "The Universal Language: Universals of Music", *The Study of Ethnomusicology: Thirty-one Issues and Concepts* (2005) 2nd edition, Urbana and Chicago: University of Illinois Press, pp. 42–49.

[2] S. J. Morrison and S. M. Demorest, "Cultural Constraints on Music Perception and Cognition", *Progress in brain research* (2009) Elsevier, vol. 178, pp. 67–77.

[3] S. Brown and J. Jordania, "Universals in the Worlds Musics", *Psychology of Music* (2011) 41(2), pp. 229–248.

[4] W. J. MacMullen, *Requirements Definition and Design Criteria for test Corpora in Information Science* (2003) SILS Technical Report 2003-03, School if Information and Library Science, University of North Carolina at Chapel Hill.

[5] G. Peeters and K. Fort, "Towards a (better) Definition of the Description of Annotated m.i.r. Corpora", *Proceedings of the International Society for Music Information Retrieval Conference* (2012) Porto, Portugal.

[19] `http://compmusic.upf.edu/node/3`

[6] A. Chaachoo, *La Música Andalusí: Al-Ala. Historia, Conceptos y Teoria Musical* (2011) Almuzara.

[7] M. Chaachoo, *Diwan al-Ala* (2009)

[8] Y. Chami, *La nawba al-Isbahan* (1956) Ministerio de Instrucción pública y Bellas Artes (delegación Zona Norte) y el Instituto General Franco de Estudios e Investigación Hispano-Árabe, Tetuán.

[9] E. Wichmann, *Listening to Theatre: The Aural Dimension of Beijing opera* (1991) University of Hawaii Press.

[10] *Jingju qupu jicheng* (Collection of Beijing opera arias scores) Shanghai, Shanghai wenyi chubanshe (1998) 10 vols.

[11] *Jingju qupu jingxuan* (Selected music form Beijing opera of China) Shanghai, Shanghai yinyue chubanshe, (1998-2005) 2 vols.

[12] M. K. Karaosmanoğlu, "A Turkish Makam Music Symbolic Database for Music Information Retrieval: SymbTr", *Proceedings of the International Society for Music Information Retrieval Conference* (2012) Porto, Portugal.

[13] B. Bozkurt, R. Ayangil, A. Holzapfel, "Computational Analysis of Turkish Makam Music: Review of the State-of-the-Art and Challenges" (2014) *Journal of New Music Research*.

[14] S. Şentürk, A. Holzapfel and X. Serra, "Linking Scores and Audio Recordings in Makam Music of Turkey" (2014) *Journal of New Music Research*.

[15] E. Ünal, B. Bozkurt and M. K. Karaosmanoğlu, "A Hierarchical Approach to Makam Classification of Turkish Makam Music, using Symbolic Data" (2014) *Journal of New Music Research*.

[16] A. Srinivasamurthy, A. Holzapfel and X. Serra, "In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music" (2014) *Journal of New Music Research*.

[17] S. Rao and P. Rao,"Hindustani Music: A Musicological and Computational Perspective" (2014) *Journal of New Music Research*.

[18] P. Rao, J. Ch. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, S. Gulati and H. Murthy, "Classification of Melodic Motifs in Raga Music with Time-series Matching" (2014) *Journal of New Music Research*.

[19] T. M. Krishna and V. Ishwar, "Carnatic Music: Svara, Gamaka, Motif and Raga Identity", 2nd CompMusic Workshop (2012) Istanbul (Turkey).

[20] T. K. G. Rao, *Compositions of Tyagaraja* (1995) Chennai: Ganamandir Publications.

[21] T. K. G. Rao, *Compositions of Syama Sastri* (1997) Chennai: Ganamandir Publications.

[22] T. K. G. Rao, *Compositions of Muddusvami Dikshitar* (1997) Chennai: Ganamandir Publications.

[23] S. Gulati, A. Bellur, J. Salamon, Ranjani H. G., V. Ishwar and X. Serra, "Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation" (2014) *Journal of New Music Research*.

[24] A. Porter, M. Sordo and X. Serra, "Dunya: A System for Browsing Audio Music Collections Exploiting Cultural Context", *Proceedings of the International Society for Music Information Retrieval Conference* (2013) pp. 101–106, Curitiba, Brasil.

[25] G. K. Koduri, J. Serrà and X. Serra, "Intonation Analysis of ragas in Carnatic Music" (2014) *Journal of New Music Research*.

[26] J. Urbano, M. Schedl and X. Serra, "Evaluation in Music Information Retrieval" (2014) *Journal of Intelligent Information Systems*. vol. 3, pp. 345–369.