Corpora for Music Information Research in Indian Art Music

Ajay Srinivasamurthy
ajays.murthy@upf.edu

Gopala Krishna Koduri
gopala.koduri@upf.edu

Sankalp Gulati
sankalp.gulati@upf.edu

Vignesh Ishwar
vignesh.ishwar@upf.edu

Xavier Serra
xavier.serra@upf.edu

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Research corpora are representative collections of data and are essential to develop data-driven approaches in Music Information Research (MIR). We address the problem of building research corpora for MIR in Indian art music traditions of Hindustani and Carnatic music, considering several relevant criteria for building such corpora. We also discuss a methodology to assess the corpora based on these criteria and present an evaluation of the corpora in their coverage and completeness. In addition to the corpora, we briefly describe the test datasets that we have built for use in many research tasks. In specific, we describe the tonic dataset, the Carnatic rhythm dataset, the Carnatic varṇaṁ dataset, and the Mridangam stroke dataset. The criteria and the evaluation methodology discussed in this article can be used to systematically build a representative and comprehensive research corpus. The corpora and the datasets are accessible to the research community from a central online repository.

1. INTRODUCTION

Computational approaches in Music Information Research (MIR) need data for developing algorithms and for testing approaches. A carefully designed data collection is critical for the success of these approaches. To develop such MIR approaches and advance knowledge, there is a need for research corpora that can be considered authentic and representative of the real world.

A research corpus is an evolving collection of data that is representative of the domain under study and can be used for relevant research problems. A good data corpus includes data from multiple sources and can even be community driven. In the context of MIR, since its practically infeasible to work with the whole universe of music, a research corpus acts as a representative subset for research. Hence, algorithms and approaches developed and technologies demonstrated on the research corpus can be assumed to generalize to real world scenarios.

A test corpus or a test dataset is often a subset of the research corpus, possibly with additional metadata for use in a specific research task. In experiments, test corpora are used to develop tools, and to evaluate and improve their performance. Computational approaches are developed using these datasets and then extended to the research corpus. Hence test corpora can even consist of synthetic data that can be used for testing. Unlike the research corpus, a test corpus is fixed for use in a specific experiment. A test corpus can evolve, but each version of the dataset used in a specific experiment is retained for better reproducibility of research results.

Building a research corpus itself is a research problem and has been studied in many fields such as linguistics, speech and biomedical language processing[1, 2, 3]. There are also many central repositories of corpora such as the Linguistic Data Consortium [1] for language resources and PhysioBank [2] for physiological signals.

There have been efforts to compile large collections of music related data, e.g. the Million Song Dataset [4], which is a good research corpus for several MIR tasks on contemporary popular music. However, despite the importance of a good research corpus in MIR, the problem of building it has received little attention by the research community. There have been no studies on a systematic way to compile and curate a research corpus. Recently, Peeters [5] presented a unified way to describe annotated MIR test datasets. Serra [6] elucidated a set of design principles to build and compile a research corpus, based on a set of primary considerations such as Purpose, Coverage, Completeness, Quality and Reusability. We use these primary considerations in this article to develop a corpus for MIR in Indian Art Music.

Musics of the world might share some basic concepts such as melody and rhythm, but some salient aspects can be described completely only by considering the specificities of that music culture. For such studies, in the context of the CompMusic project [7], Serra emphasized the need for culture specific research corpora to develop approaches that utilize the important aspects of the music culture.

The primary aim of CompMusic is to build culture specific computational methodologies for better exploration of music collections through meaningful music concepts and automatically extracted melody, rhythm and semantic descriptors. Working with five music traditions of the world, the data driven methodologies in CompMusic primarily involve signal processing, machine learning and semantic web technologies. Hence, there has been a significant ef-

[1] https://www.ldc.upenn.edu/

[2] http://www.physionet.org/physiobank/

fort towards the design and compilation of research corpora for relevant problems in the music cultures being studied. In this article we present the research corpora for Carnatic and Hindustani music, the art music traditions from India. There have also been several test datasets created from the research corpora to develop and test algorithms for specific research tasks. We also present some of the test datasets and specific tasks in which they can be used.

Each of the aforementioned music cultures can be described in terms of musical concepts, music content and the music community. The elements of the corpora can be associated with one or more of these categories and hence useful for computational tasks in these three aspects. Central to both the corpora is an audio music recording with its metadata. All audio in both the corpora are stereo recordings sampled at 44.1 kHz and stored as 160 kbps mp3 files for ease of transmission and storage.

An important concern in research is the reproducibility of the experiments, which necessitates a corpus accessible to the research community. When possible, we emphasize the use of open repositories of information such as MusicBrainz [3] and Wikipedia. The releases in the Carnatic [4] and Hindustani [5] corpora have been organized into collections in MusicBrainz. For audio, since there are no open repositories of quality audio, we use easily accessible commercial recordings. Further, the test datasets and the derived information such as annotations and extracted features are openly available [6] . In CompMusic, we are developing a tool for navigating through music collections called *Dunya*[8], which also acts as the central permanent online repository to store the metadata, audio, annotations and research results. Dunya is open source and provides an API for accessing these data.

As we described earlier, the research corpora are growing entities through continued efforts. Hence, the numbers presented in this article are only indicative and are of secondary importance. We primarily emphasize on presenting a scientific approach to develop a corpus and evaluate its suitability for a particular set of research tasks. We emphasize on methodologies that can be used to evaluate a corpus on the aspects of coverage and completeness. Apart from the description of the corpora, a methodology for evaluation of the corpus is an important contribution of this article. We further note that in addition to the sources described in this article, there are several other sources that can be used for computational research in Indian Art Music, and eventually could be a part of the corpus.

2. CARNATIC MUSIC RESEARCH CORPUS

The Carnatic music research corpus mainly comprises of audio recordings, its associated editorial metadata, lyrics, scores, contextual information on music concepts, and community (social) information from online music forums and other sources. Audio recordings, editorial metadata, scores, and lyrics are the content used by signal processing and machine learning approaches. Contextual information and the forum discussions form the music concepts and community information used for semantic analysis.

There are several considerations in collecting a corpus of Carnatic music. A concert, called a kutcheri (Kachēri), is the natural unit of Carnatic music and used as the main unit of music distribution. A concert has one or more lead artists (mainly vocal, veena, violin, or flute), melodic accompaniment (mainly violin), and one or more percussion accompaniments (mainly Mridangam). Carnatic music is predominantly composition based and most commercial releases are concerts, comprising of several pieces that are improvised renderings of compositions. Vocal music is predominant in Carnatic music and most of the compositions are to be sung. Even in instrumental music, the lead artist aims to mimic vocal singing [9]. The melody and rhythm is organized based on the frameworks of rāga and tāḷa [7] . The rāga and tāḷa are the most important metadata associated with a composition and hence a recording of the composition. Each composition is composed in one or more rāgas and tāḷas.

Based on these considerations, we consulted expert musicians and musicologists, such as T M Krishna [8] to arrive at a representative collection of Carnatic music audio. The main institutional reference for Carnatic music is the Madras Music Academy (MMA) [9] , which is a premier institution dedicated to Carnatic music and organizes the annual music conference in Chennai, India. The annual Carnatic music festival is one of the largest music festivals in the world, with a significant part of the Carnatic music community taking part in it. The MMA has been driving scholarly research and opinion in Carnatic music. The MMA has a panel of experts that formulates the procedure and standards for the selection of artists for the music festival. The MMA has been recording concerts and its archive can be considered a standard repository of Carnatic music. However, the archive is not openly available online. We thus followed the musical criteria followed by the MMA and procured the audio from commercially available releases. Though Carnatic music is spread across South India, the choice of MMA as an institutional reference will have an influence on the research corpus introducing a bias towards the music scene in Chennai.

We wished to compile concerts over several generations of musicians. We started with the artists that have been performing at the MMA in the last five years, and then expanded the collections to include their teachers, and popular musicians of their era. The record label Charsur [10] specializes in Carnatic music and the core of our audio collection is from their catalog of music concerts. Hence, the corpus consists of audio from commercially available releases from Charsur and other music labels. The corpus presently consists of 248 releases(concerts) with 1650 audio recordings (346 hours) spanning 1068 compositions. The number

[3] http://musicbrainz.org/

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

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

[6] http://compmusic.upf.edu/corpora

[7] Some audio examples at http://compmusic.upf.edu/examples-taala-carnatic

[8] http://www.tmkrishna.com/

[9] http://musicacademymadras.in/

[10] http://www.charsur.com/

of other relevant music entities in the corpus is described in Table 1 (column 2). Though we focus on concerts with vocalist leads, we also have instrumental music releases (mainly with Veena, Violin, Flute, Saxophone, and Mridangam in lead). The whole audio collection is commercial and easily accessible, but is not open and distributable. However, efforts are underway to compile a freely available open collection of Carnatic music.

The editorial metadata associated with each release has been stored and organized in MusicBrainz. The primary metadata associated with each concert is the name of the release, the lead and the accompanying artists, and the musical instruments in the concert. For each audio recording contained in the release, the relevant metadata are the artists performed on the track, the name of the composition/s and the composer, rāga/s, tāḷa/s, musical form/s. MusicBrainz assigns a unique identifier (MBID) for each entity in MusicBrainz, such as the artist, composer, instrument, recording, work, and a release. This helps to organize the metadata in an effective way. All the editorial metadata was entered using Roman alphabet and a roman transliteration was used when the language of the release was not English. The rāga and tāḷa information was added as tags, though a recent version of MusicBrainz supports work attributes with which this information can be stored and accessed better. Efforts are underway to convert the existing tags into work attributes.

Since Carnatic music is predominantly a vocal music tradition, lyrics play an important role. A significant part of the rendition of a composition is improvised and hence the scores associated with a composition are of limited use, nonetheless important. The lyrics and scores, even though not time aligned to audio recordings, are useful for computational analysis and hence we compiled them. The primary languages in which Carnatic music is composed are Telugu, Tamil, Kannada, Sanskrit, and Malayalam. There are several published compilations of lyrics and scores for most of the currently performed compositions, such as the ones of the three most popular composers in Carnatic music: Tyāgarāja, Śyāmā śāstri and Muttusvāmi dīkṣitar (e.g. [10]). However, these compilations are not machine readable and hence not accessible for computational analysis.

There are good online open repositories for lyrics, such as sahityam.net [11] , which is a wiki of lyrics of Carnatic compositions. Sahityam.net is our primary source for machine readable lyrics. It uses a uniform scheme for transliteration to Roman script and hence has minimal ambiguity. In some cases, it provides additional commentary, references, and example renditions. Sahityam.net currently hosts lyrics for about 1820 compositions of Carnatic music. Machine readable scores are more difficult to access, with no comprehensive machine readable score compilations available. A set of machine readable (HTML, Word) scores compiled by Dr. Shivkumar Kalyanaraman [12] is the main source of scores.

	Corpus	Raaga.com	Kutcheris	Charsur
Rāgas	246	489 (42%)	N/A	301 (68%)
Tāḷas	18	16 (100%)	N/A	21 (85%)
Composers	131	598 (17%)	N/A	256 (42%)
Artists	233	501	2978	264 (48%)

Table 1: Coverage of the Carnatic music corpus. The number in parentheses is the *overlap* measure in percentage. N/A indicates data not available.

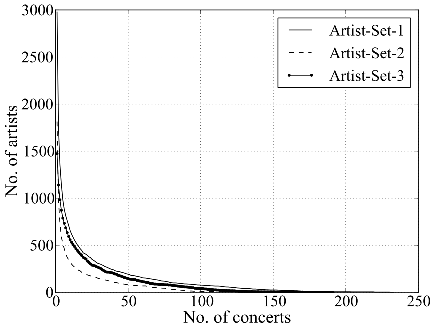


Figure 1: The number of artists by the number of their performances.

The music community and music concepts related information in the corpus form the primary source of information for semantic analysis, and come from various sources from the Internet. Kutcheris.com [13] is an up-to-date directory of artist biographies, music venues, concerts and events. The category of Carnatic music on Wikipedia [14] is a source of contextual information including music concepts. We have added a lot of information and contributed to Wikipedia with the help of experts. While Wikipedia acts as an encyclopedia of music concepts providing linked information, online music forums with discussions provide opinions from which some of these links can be inferred. The rasikas.org [15] Carnatic music forum is an active forum of Carnatic music listener community with useful discussions about Carnatic music concepts, concerts, and performances. It is an important source of data useful for community profiling.

2.1 Coverage

A research corpus needs to be representative of the real world in the concepts that are primary to the music culture. The aim of a coverage analysis is to estimate the comprehensiveness of the corpus with respect to another representative reference source. For Carnatic music, a coverage analysis is presented for artists, rāgas, tāḷas, and composers. For artist coverage, we chose to use Kutcheris.com as the primary reference since it is up-to-date with current artists and their performances. We use the last five years of their concert listings. Many of the artists and the concerts listed on Kutcheris.com are from Chennai. Charsur's release catalog provides information about rāgas, tāḷas, composers and artists. Raaga.com [16] is an Indian music streaming service and its Carnatic channel is another reference for

[11] http://www.sahityam.net

[12] http://www.shivkumar.org

[13] http://www.kutcheris.com

[14] http://en.wikipedia.org/wiki/Category:Carnatic_music

[15] http://www.rasikas.org/

[16] http://www.raaga.com

rāgas, tāḷas, composers and artists. However, Raaga.com has many light music forms included in its Carnatic channel, some of which we have consciously excluded from our corpus. Hence it is to be noted that numbers and the analysis with Raaga.com will have an adverse influence from these other included music forms. The data from each of these reference sources was crawled from their online catalogues. The data from raaga.com was crawled in March, 2012 (in the current version of raaga.com (2014), data about some entities is missing) and from the others in March, 2014. We observed that nearly every source had duplicate entities mostly arising due to spelling variations (e.g. Tyagaraja, Tyaagaraaja). We merged the duplicates by matching the longest common subsequence in the strings and by using Damerau-Levenshtein distance.

Table 1 shows the coverage of the Carnatic corpus in comparison to the references. For each music entity e, we define a coverage measure called the *overlap* (O) as,

$$O_e^r = \frac{|S_e^c \cap S_e^r|}{|S_e^r|} \qquad (1)$$

where O_e^r is the *overlap* measure of the entity e with reference r, S_e^c is the set of entities in the corpus, S_e^r is the set of entities in the reference, and $|S|$ denotes the cardinality of a set S. An *overlap* of 100% is achieved if all the elements in the reference set are present in the corpus. Table 1 shows the *overlap* measure for rāgas, tāḷas and composers for both Raaga.com and Charsur. We can see that there is a good coverage of tāḷas and a satisfactory coverage of rāgas in the corpus. The composer coverage with respect to Raaga.com is poor since it includes the light music composers in its set of composers.

Among the 233 artists who have at least one recording in the corpus, 74 are lead artists (lead vocal or lead instrumental). Further, we have 28 violin accompanying artists and 48 unique percussion artists in the corpus. The concerts listed by Kutcheris.com span the whole year and all through the day. However, the evening concerts are more recognized, and we took it to be a measure of popularity of the artists. Moreover, the evening concerts during the music season lasting from November to January are ticketed. For a coverage analysis, we thus consider three categories of artists: Artists-Set-1 (all the artists), Artists-Set-2 (artists who have performed in the evening concerts, through the year) and Artists-Set-3 (artists who have performed in evening concerts between November and January). Of the 2978 total artists present in Set-1 on Kutcheris.com concert listings, there are 1814 artists in Set-2 and 1472 artists in Set-3.

The number of concerts performed by each artist is also an indicator of popularity. Though there are a large number of artists in Kutcheris, we see that the distribution of the number of concerts they have performed is exponential (Fig. 1), e.g. there are only about 200 artists who have over 50 concerts. Hence to capture this fact, we used the set of artists in the corpus and computed the *overlap* as defined in Eq. 1 through different subsets of artists in Kutcheris.com, sweeping over the number of concerts (at least) they have performed.

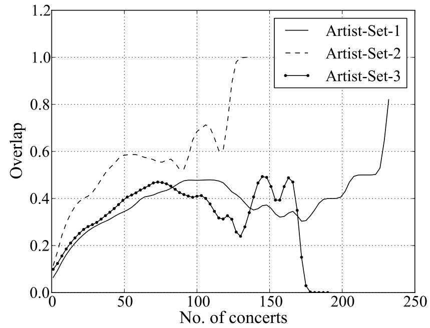


Figure 2: Coverage of Carnatic artists. The ordinate is the overlap value of the set of artists in corpus, compared against a set of artists in Kutcheris.com who have performed in at least as many concerts as the abscissa.

Accompanying metadata	# Recordings	% of total
Lead artist	1650	100
Accompanying artists	1221	74.00
Rāga	959	58.12
Tāḷa	917	55.58
Work (Composition)	989	59.94

Table 2: Completeness of the Carnatic music corpus, showing the number of recordings in which the corresponding metadata is available.

Fig. 2 shows the *overlap*, using a set of artists that have performed at least as many concerts as the number shown on the abscissa. The *overlap* is also shown for the three categories of artists we discussed before. We can see that the *overlap* increases as we consider more frequently performing artists and becomes almost constant. The artists who have performed the most concerts are often the accompanying artists, and are few in number, which explains why the *overlap* becomes a constant, when we discount the *overlap* for more than 150 concerts. When we consider a large number of concerts, the *overlap* values are unreliable since the number of artists is less. In general, we can see that the *overlap* is better for Artists-Set-2 than Artists-Set-1 and Artists-Set-3, showing that the corpus has more representation of artists from evening concerts round the year.

2.2 Completeness

In the context of this article, completeness of the corpus refers mainly to the completeness of the associated metadata for each recording, primarily from MusicBrainz. Even though carefully built, the editorial metadata associated with a release and its recordings can be incomplete. There are three possible reasons for incomplete metadata. Many releases do not provide all the required metadata on the CD. In many releases, only the lead artist is listed, without the accompanying artists. It is seen very often that the composition information is also absent on the CD cover. The second reason is that the editorial metadata was not completely entered into MusicBrainz. This is sometimes seen with release and recording relationships that were left incomplete by the person who added the metadata. Further, since all the metadata, including the rāga/tāḷa tags, are imported and linked automatically, there can be import errors due to variations in transliterations and spelling. Multiplicity of languages used in Carnatic music further adds to these inconsistencies. These import errors are the third reason for incomplete metadata.

Missing metadata in MusicBrainz can only be completed by manually adding the missing fields to MusicBrainz. However, we are also exploring automatic metadata completion based on other relations on the release or the recording, using semantic web approaches. The missing data due to transliteration errors have been addressed to an extent by making curated lists of entities such as rāgas and tāḷas, and using robust algorithms for matching and linking metadata. Despite significant efforts, there are many recordings and releases that have incomplete metadata.

Table 2 shows the completeness of the recordings in the corpus, including all the three factors that result in incomplete metadata. All the recordings have a lead artist, but about a quarter of the recordings (429/1650) do not have accompanying artist information. Rāga, tāḷa and work (composition) are listed for about half the recordings. It is to be noted that these numbers reflect only the recordings for which we were completely sure of the editorial metadata. There are several recordings that have the required metadata but deemed incomplete since we could not accurately match it to a related entity in the curated lists.

3. HINDUSTANI MUSIC RESEARCH CORPUS

Similar to Carnatic music, Rāg and tāl[17] are the fundamental music concepts in Hindustani music and hence the main theme around which the corpus has been built. Hindustani music tradition is much more diverse and heterogeneous and thus presents a significant challenge to compile a good research corpus. Though vocal music is predominant, instrumental music in Hindustani music is also popular. The main focus in Hindustani music is on improvisation and compositions are short. For Hindustani music corpus we focus on two important vocal music styles - Dhrupad and Khyāl. A typical khyāl performance has lead vocals/instrument, with a melodic accompaniment - harmonium or a sāraṅgi, and a rhythmic accompaniment - Tabla. In dhrupad style, pakhāvaj is the main rhythmic accompaniment.

There are many institutions that have compiled huge audio archives of Hindustani music. The primary of them are the ITC Sangeet Research Academy (ITC-SRA), Sangeet Natak Academy, and the All India Radio (AIR). Each of these institutions own thousands of hours of expert curated music recordings that represent the real world performance practice. ITC-SRA is a premier music academy of Hindustani music and has taken up major efforts in the archival of music. Sangeet Natak Academy is India's national academy for music, drama and dance. AIR is the largest public broadcaster in India and has a huge archive of Hindustani music curated over many decades. AIR awards grades to musicians and its archives can be considered as a reference. None of these archives are publicly available and we compiled the audio in our corpus using these collections as a reference. We consulted expert musicians and musicologists, such as Dr. Suvarnalata Rao at the National Centre for the Performing Arts (NCPA), Mumbai, India to curate the audio collection in the corpus.

	Corpus	ITC-SRA	Swarganga
Artists	360	240 (19%)	629 (14%)
Rāgs	176	185 (48%)	534 (13%)
Tāls	32	N/A	59 (37%)
Works	685	N/A	1957

Table 3: Coverage of the Hindustani music corpus. The number in parentheses is the *overlap* measure in percentage. N/A indicates data not available.

Accompanying metadata	# Recordings	% of total
Lead Artist	1096	100
Accompanying artist	658	39.88
Rāg	960	58.18
Tāl	627	38.00
Work (Bandish)	576	34.91

Table 4: Completeness of the Hindustani music corpus showing the number of recordings in which the corresponding metadata is available.

The audio collection in the corpus comprises of commercially available music releases from several music labels. It mainly consists of khyāl and dhrupad vocal music releases, though a significant number of instrumental music releases are present. The corpus presently has 233 releases with a total of 1096 recordings (300 hours). As with Carnatic music, the editorial metadata associated with each release is stored in MusicBrainz. The metadata associated with each release is the name of the release, the lead and the accompanying artists, and the musical instruments in the concert. For each audio recording in the release, the relevant metadata are the artists performed on the track, the name of the composition/s (*bandish*) and the composer/s (if composed), rāg/s, tāl/s, lay/s (tempo class), form/s, and section/s. All the editorial metadata was entered using Roman alphabet, following a uniform transliteration scheme for a better consistency.

Hindustani music is mainly improvised and hence lyrics and scores are not very relevant for computational analysis. Bhatkhande [11] and Ramashray Jha [12] compiled lyrics and scores of bandishes using a standardized notation for Hindustani music. However, they are not available in a machine readable form. Swarganga Music Foundation [18] has a good archive of rāgs, tāls and bandishes. The category of Hindustani music on Wikipedia [19] is a source of contextual information including music concepts of Hindustani music.

3.1 Coverage

The methodology followed for the coverage analysis of Hindustani music is the same as followed for Carnatic music. We present the coverage analysis for artists, rāgs, tāls

[17] Some audio examples at http://compmusic.upf.edu/examples-taal-hindustani

[18] http://www.swarganga.org/

[19] http://en.wikipedia.org/wiki/Category:Hindustani_music

and compositions. The coverage analysis for Hindustani music is more complex than Carnatic music. This can be attributed to the heterogenous nature of the music repertoire, and to the lack of dedicated recording labels like Charsur in the case of Carnatic music. For each of these entities we choose two main references, ITC-SRA and Swarganga.

Unlike Carnatic music, the unit of music distribution in Hindustani music is not often a concert. Further, it is geographically spread over the Indian sub-continent and hence there is no single repository of Hindustani music performances, such as Kutcheris.com for Carnatic music. Therefore, it is challenging to do a comprehensive artist coverage analysis like the one presented for Carnatic music.

Table 3 shows the coverage of the Hindustani corpus. We see that the corpus and the chosen references have comparable number of entities, but the *overlap* is less. This is primarily because we mainly focused on recordings made in last 20-30 years to ensure good recording quality and to reflect current performance practices. On the other hand both the references focus primarily on archiving Hindustani music and hence consist of several generations of artists, infrequent rāgs and tāls, and a more comprehensive list of compositions. Further, the Hindustani corpus is mainly composed of vocal music recordings with a focus on only two styles, khyāl and dhrupad. The reference archives additionally include instrumental music and several other styles of Hindustani music.

3.2 Completeness

The completeness of the editorial metadata for Hindustani music is shown in Table 4. We see that the editorial metadata for all the recordings at least includes the lead artist, and for more than half of the collection, the accompanying artists (658/1096). Roughly 90% of the corpora is annotated with the rāg label and more than half with the tāl label. Work (bandish) labels are present for nearly half of the collection (576/1096). Ālāp performances in Hindustani music are not compositional works, and hence should be discounted while assessing the completeness of work metadata. But due to the unavailability of such an information (ālāp labels), ālāp performances are also included in assessment and hence work completeness is an underestimate.

4. TEST DATASETS

The test datasets were designed for specific tasks and contain additional information such as annotations and derived data [20]. They are useful for various melody and rhythm analysis tasks. We describe each dataset briefly emphasizing the primary research task where they can be used.

4.1 Indian Art Music Tonic Dataset

Estimating the tonic of the lead vocals/instrument is a primary task and forms the basis for many melodic analysis tasks in both Hindustani and Carnatic Music. We built the Indian Art Music Tonic dataset [21] for the development of tonic identification approaches [13, 14, 15]. The dataset is a subset of recordings in the Hindustani and the Carnatic research corpora, each manually annotated with the tonic of the lead artist. Table 5 shows a numerical breakdown of the dataset.

	Pieces	Vocal	Male/Female	Instrumental
Hindustani	245	193 (79)	137/56 (56/23)	52 (21)
Carnatic	352	235 (67)	170/65 (48/19)	117 (33)
Total	597	428 (72)	307/121 (51/21)	169 (28)

Table 5: The Indian Art Music Tonic dataset. Numbers in parentheses in columns 3-5 show the values in percentage of the value in second column.

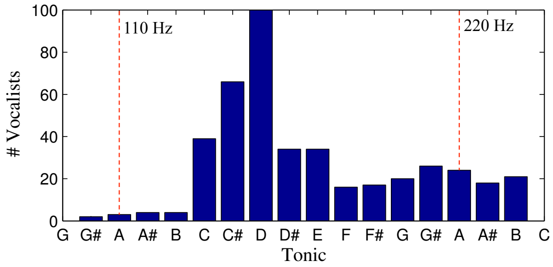


Figure 3: A histogram of vocal tonic values in the dataset.

The collection consists of recordings of both Hindustani and Carnatic music recordings, selected such that it has a balanced mix of vocal and instrumental music, male and female singers, old and new recordings, and different styles within these two music traditions. Each recording can be uniquely identified using the MBID of the recording. The annotation for each recording in the dataset is the tonic pitch for vocal performances and tonic pitch-class for instrumental performances, and was manually annotated and verified by a professional Carnatic musician [13]. Fig. 3 shows the distribution of tonic (only for vocal recordings) in the dataset. The distribution peaks around the tonic values of C# and D owing to the larger fraction of male vocalists in the dataset.

4.2 Carnatic Rhythm Dataset

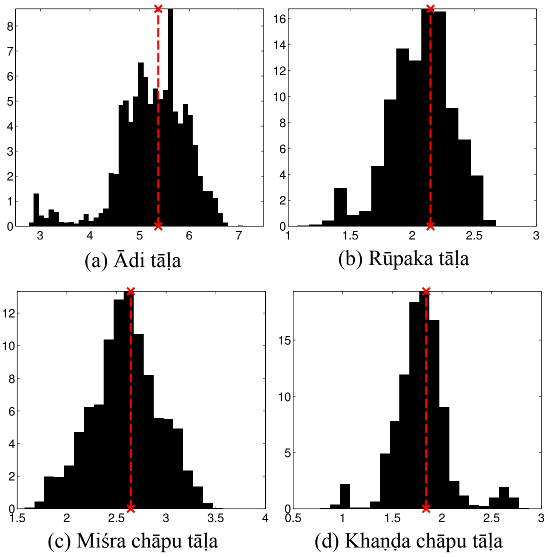
The Carnatic Rhythm Dataset [22] is a rhythm annotated test corpus for many automatic rhythm analysis tasks in Carnatic Music. The collection consists of audio excerpts from the Carnatic research corpus, manually annotated time aligned markers indicating the progression through the tāḷa cycle, and the associated tāḷa related metadata. The dataset has pieces in four popular tāḷas (Table 6) that encompass a majority of Carnatic music. The pieces include a mix of vocal and instrumental recordings, recent and old recordings, and span a wide variety of forms. All pieces have percussion accompaniment, predominantly Mridangam. Of the 176 excerpts, 120 are full length pieces.

There are several annotations that accompany each excerpt in the dataset. The primary annotations are audio synchronized time-stamps indicating the different metrical positions in the tāḷa cycle - the sama (downbeat) and other beats. The annotations were created using Sonic Visual-

[20] http://compmusic.upf.edu/datasets
[21] http://compmusic.upf.edu/iam-tonic-dataset
[22] http://compmusic.upf.edu/carnatic-rhythm-dataset

Tāḷa	# Pieces	$\overline{\text{LEN}}$	Size	# Samas	# Ann.
Ādi	50	4.85	252.78	2882	22793
Rūpaka	50	4.62	267.45	7582	22668
M. Chāpu	48	6.59	342.13	7795	31055
K. Chāpu	28	4.41	134.62	4387	13111
Total	**176**	**5.06**	**996.98**	**22646**	**89627**

Table 6: The CompMusic Carnatic Music Rhythm Dataset showing the number of pieces, median length of each piece $\overline{LEN}$, the total size of the dataset (in minutes), and the number of annotations. M. Chāpu and K. Chāpu refer to Miśra chāpu and Khaṇḍa chāpu tāḷas respectively. # Samas is number of sama annotations, #Ann. refers to all beat annotations (including sama).



(a) Ādi tāḷa

(b) Rūpaka tāḷa

(c) Miśra chāpu tāḷa

(d) Khaṇḍa chāpu tāḷa

Figure 4: The histogram of median tāḷa cycle length for different tāḷas in the dataset. The abscissa is length in seconds, the ordinate is the number of recordings.

izer [16] by tapping to music and manually correcting the taps. Each annotation has a time-stamp and an associated numeric label that indicates the position of the beat marker in the tāḷa cycle. In addition, for each excerpt, the tāḷa of the piece and eḍupu (offset of the start of the piece, relative to the sama) are recorded. The possibly time varying tempo of a piece can be obtained using the beat and sama annotations. The distribution of the median length of the tāḷa cycle for each of the four tāḷas in the dataset is shown in Fig. 4. The length of the tāḷa cycle is indicative of the tempo of the piece. Though there is no notated tempo for a composition and the musician is free to choose a tempo, we empirically observe that musicians tend to choose a narrow range of tempo.

The dataset is intended to be a test corpus for several computational rhythm analysis tasks in Carnatic music [17, 18]. Possible tasks include tāḷa, sama and beat tracking, tempo estimation and tracking, tāḷa recognition, rhythm based segmentation of musical audio, structural segmentation, audio to score/lyrics alignment, and rhythmic pattern discovery.

Rāga	# Recordings	# Duration (min)
Ābhōgi	5	29
Bēgaḍa	3	27
Kalyāṇi	4	27
Mōhanaṁ	4	24
Sahāna	4	28
Sāvēri	5	36
Śrī	3	26
Total	**28**	**197**

Table 7: The Carnatic Varṇaṁ Dataset

4.3 Carnatic Varṇaṁ Dataset

Carnatic varṇaṁ dataset [23] is a collection of 28 solo vocal recordings, recorded for our work on intonation analysis of Carnatic rāgas [19]. The collection has the audio recordings, tāḷa cycle annotations and music score in a machine readable format. The dataset consists of seven different varṇaṁs in seven rāgas sung by five young professional Carnatic vocalists with a music training of more than 15 years. They are all set to ādi tāḷa. Since intonation analysis requires clean pitch tracks, the varṇaṁs were recorded without any accompanying instruments, except the drone. The dataset is described in Table 7.

The recordings were manually annotated with the sama (downbeat) of the tāḷa cycles using Sonic Visualizer, and each cycle was later divided into the 8 beats of ādi tāḷa. The music scores for the seven varṇaṁs were procured from the archive curated by Dr. Shivkumar Kalyanaraman [24] and manually converted to a machine readable format (yaml).

The distinct advantage of this dataset is the free availability of the audio content. Along with the annotations, it can be used for melodic analyses such as characterizing intonation, motif discovery and tonic identification. The availability of a machine readable scores allows the dataset to be used for audio-score alignment.

4.4 Mridangam Stroke Dataset

The Mridangam Stroke dataset [25] is a collection of 7162 audio examples of individual strokes of the Mridangam in various tonics. The dataset can be used for training models for each Mridangam stroke [20]. The dataset comprises of ten different strokes played on Mridangams with six different tonic values. The dataset is described in Table 8, with stroke labels along rows and tonic values along columns. The audio examples were recorded from a professional Carnatic percussionist in semi-anechoic studio conditions using SM-58 microphones and an H4n ZOOM recorder. The audio was sampled at 44.1 kHz and stored as 16 bit wav files.

5. SUMMARY

We presented the research corpora and the associated test datasets for MIR in Hindustani and Carnatic music. We discussed the considerations in building such culture specific

[23] `http://compmusic.upf.edu/carnatic-varnam-dataset`
[24] `http://www.shivkumar.org/music/varnams/index.html`
[25] `http://compmusic.upf.edu/mridangam-stroke-dataset`

	B	C	C#	D	D#	E	Total
Bheem	5	3	1	0	15	25	**49**
Cha	57	50	54	67	49	53	**330**
Dheem	127	86	78	12	111	54	**468**
Dhin	48	48	63	12	198	113	**482**
Num	81	98	97	18	143	60	**497**
Ta	145	165	217	180	119	105	**931**
Tha	200	185	211	224	196	160	**1176**
Tham	88	80	35	29	92	50	**374**
Thi	438	334	369	283	444	345	**2213**
Thom	136	80	72	91	128	135	**642**
Total	**1325**	**1129**	**1197**	**916**	**1495**	**1100**	**7162**

Table 8: The Mridangam Stroke Dataset. The row and column headers are the stroke labels and the tonic values, respectively.

corpora and provided an evaluation of their coverage and completeness. The corpora and the test datasets discussed in this article would be useful to build computational algorithms and approaches for a better computational description of Indian Art Music. The methodology used for evaluation of the corpora can be extended to other research corpora. Both the corpora are available for research through the web application Dunya.

Acknowledgments

The compilation of the corpora and the datasets presented in this paper has been a collective effort of the members of the CompMusic project and a number of collaborators. 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] M. Wynne, Ed., *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005.

[2] S. Pan and W. Weng, "Designing a speech corpus for instance-based spoken language generation," in *Proc. of the 2nd International Conference on Natural Language Generation*, 2002, pp. 49–56.

[3] K. B. Cohen, P. V. Ogren, L. Fox, and L. Hunter, "Empirical data on corpus design and usage in biomedical natural language processing," in *AMIA Annual Symposium Proceedings*, 2005, pp. 156–160.

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *Proc. of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, Miami, USA, Oct. 2011, pp. 591–596.

[5] G. Peeters and K. Fort, "Towards a (Better) Definition of the Description of Annotated MIR Corpora," in *Proc. of 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, Oct. 2012, pp. 25 – 30.

[6] X. Serra, "Creating research corpora for the computational study of music: the case of the compmusic project," in *Proc. of the 53rd AES International Conference on Semantic Audio*, London, Jan. 2014.

[7] X. Serra, "A multicultural approach in Music Information Research," in *Proc. of 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, USA, Oct. 2011, pp. 151–156.

[8] A. Porter, M. Sordo, and X. Serra, "Dunya: A system for browsing audio music collections exploiting cultural context," in *Proc. of 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, Nov. 2013, pp. 101–106.

[9] T. Viswanathan and M. H. Allen, *Music in South India*. Oxford University Press, 2004.

[10] T. K. Govinda Rao, *Compositions of Tyāgarāja*. Ganamandir Publications, 2009.

[11] V. N. Bhatkhande, *Hindustani Sangeet Paddhati: Kramik Pustak Maalika Vol. I-VI*. Sangeet Karyalaya, 1990.

[12] R. Jha, *Abhinav Geetanjali Vol. I-V*. Sangeet Sadan Prakashan, 2001.

[13] S. Gulati, "A Tonic Identification Approach for Indian Art Music," Master's Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2011.

[14] J. Salamon, S. Gulati, and X. Serra, "A Multipitch Approach to Tonic Identification in Indian Classical Music," in *Proc. of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, Oct. 2012, pp. 499–504.

[15] S. Gulati, A. Bellur, J. Salamon, H. G. Ranjani, V. Ishwar, H. A. Murthy, and X. Serra, "Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation," *Journal of New Music Research*, vol. 43, no. 1, pp. 55–73, 2014.

[16] C. Cannam, C. Landone, and M. Sandler, "Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files," in *Proc. of the ACM Multimedia 2010 International Conference*, Florence, Italy, Oct. 2010, pp. 1467–1468.

[17] A. Srinivasamurthy, A. Holzapfel, and X. Serra, "In Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music," *Journal of New Music Research*, vol. 43, no. 1, pp. 97–117, 2014.

[18] A. Srinivasamurthy and X. Serra, "A supervised approach to hierarchical metrical cycle tracking from audio music recordings," in *Proc. of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014, pp. 5237–5241.

[19] G. Koduri, V. Ishwar, J. Serrá, and X. Serra, "Intonation analysis of ragas in Carnatic music," *Journal of New Music Research*, vol. 43, no. 1, pp. 73–94, 2014.

[20] A. Anantapadmanabhan, A. Bellur, and H. A. Murthy, "Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization," in *Proc. of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 2013, pp. 181–185.