

A Tonic Identification Approach for Indian Art Music

Sankalp Gulati

MASTER THESIS UPF 2012

Master in Sound and Music Computing

Master thesis supervisor:

Xavier Serra

Co-supervisor:

Justin Salamon

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona



Copyright © 2012 by Sankalp Gulati

This is an open-access article distributed under the terms of the *Creative Commons Attribution 3.0 Unported License*, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

In this thesis we address a fundamental aspect of the computational analysis of Indian art music, the automatic identification of the tonic of the lead performer. We propose two approaches for tonic identification in Indian art music, which take advantage of the characteristic feature of this music tradition by performing a multi-pitch analysis.

We provide a short introduction to Indian art music, explaining the concept of tonic in the context of this music tradition. We review the main audio features, techniques and works relevant to the computational analysis of the tonal aspect of music and present a critique of previous approaches to tonic identification in Indian art music.

A detailed description and the implementation steps for both the proposed methods are presented. The audio signal is transformed using a multi-pitch representation, which is then used to construct the pitch histograms. The tonic is identified using the prominent peaks of a pitch histogram. Following a classification based approach the system automatically learns the best set of rules to select the peak of the histogram that represents the tonic. In addition to the multi-pitch representation, the second method also analyzes the predominant melody pitches to estimate the tonic octave. Further, we also present a proposal for a complete iterative system for tonic identification which aims to use both audio and metadata.

The methods are evaluated on a sizable diverse database of Indian art music, compiled as a part of the CompMusic¹ project. The obtained results are good and demonstrate the advantage of using a multi-pitch approach. A detailed error analysis is performed and the plausible reasons for errors are discussed. The thesis is concluded with a summary of the work, highlighting the main conclusions and the contributions made.

¹<http://compmusic.upf.edu/>

Acknowledgements

First and foremost, I would like to thank my supervisor, Xavier Serra, for giving me an opportunity to join the Music Technology Group and letting me contribute to the CompMusic project. I am grateful to him for supporting my research throughout the year. His guidance has been of immense value to the current thesis.

I extend my gratitude to Justin Salamon, who has given me plenty of support, ideas and advice. I thank him for all his efforts, specially for diligently reviewing this document and providing me with the melody extraction code.

Next, I would like to thank Emilia Gómez for her suggestions, advice and her valuable time for our discussions. The guidance, ideas and insights provided by Perfecto Herrera and Joan Serrà regarding the machine learning related issues have been instrumental in the experiments. The music technology courses at UPF, that provided me with a gamut of knowledge and insights, has been a delightful experience and I thank my professors for the same.

I am grateful to members of 55.308, Mohammed Sordo, Gōpāla K. Kōḍūri and Sertan Şentürk for their help and support in not only the thesis but also for the enthusiasm they brought into the group. This year has been made memorable by the entire batch of SMC and I thank them from the bottom of my heart.

I am obliged to Cristina Garrido, Alba Rosado, Sonia Espí, Vanessa Jimenez and Lydia García for assisting me with all the legal formalities and making my stay in Barcelona essentially effortless.

Thanks to Prof. Preeti Rao for providing me guidance, advice and initiating me into the area of music information retrieval. I am also grateful to Amruta Vidwans, Kaustuv Kanti Ganguli, Ashwin Bellur, Vignesh Ishwar to help me at various stages of this research work.

This research was funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 267583 (CompMusic).

Last but not the least, I am grateful to my family and friends who inspire and support me to continue doing what I love. I want to thank Mrs. Aparna & Vishal Kapoor and Shefali Bajpai for playing a crucial role in my life.

Contents

1	Introduction	6
1.1	Indian Art Music Traditions	6
1.2	Tonic in Indian Art music	8
1.3	Tonic in Western music	10
1.4	Motivation and Goals	11
2	Scientific Background and Related Work	13
2.1	Feature Extraction	13
2.1.1	Fundamental Frequency Estimation	14
2.1.2	Multi-pitch Analysis	16
2.1.3	Predominant Melody Extraction	18
2.1.4	Pitch Class Distribution Features	19
2.2	Key Estimation in Western Music	21
2.3	Tonal Structure of the Tānpūrā	22
2.4	Tonic Identification in Indian Art Music	26
3	Multi-pitch Approach to Tonic Identification	29
3.1	Overview of the Methods	29
3.2	Method 1	33
3.2.1	Multi-pitch Analysis	33
3.2.2	Candidate Selection	40
3.3	Method 2	43
3.3.1	Tonic Pitch-class Identification	43
3.3.2	Tonic Octave Estimation	44
3.4	Tonic identification system	48

3.5	Evaluation Methodology	50
3.5.1	Database	50
3.5.2	Annotations	52
3.5.3	Performance Evaluation	53
4	Results and Discussion	55
4.1	Results	55
4.2	Discussion	58
5	Conclusions and Future work	68
5.1	Summary of the work	68
5.2	Main conclusions	69
5.3	Open issues and future work	70
5.4	Contributions	72
	References	73

List of Figures

2.1	Steps of the fundamental frequency estimation process	15
2.2	General block diagram for the PCD computation from audio	20
2.3	Image of a tānpūrā showing different tunings	24
2.4	Spectrogram of a section of solo tānpūrā field recording.	24
2.5	Spectrum of a tānpūrā recording	25
2.6	Diagram showing cross-section view of tānpūrā bridge	25
3.1	Spectrogram of an excerpt of Hindustani music	30
3.2	Block diagram of the proposed methods	32
3.3	Detailed block diagram of multi-pitch analysis module	34
3.4	Peaks of the salience function computed for a Hindustani music excerpt	38
3.5	Comparison of the histograms constructed using different methods .	39
3.6	Features extracted from a pitch histogram for training the classifier	41
3.7	Example of predominant melody histogram	45
3.8	Block diagram of the proposed iterative tonic identification system .	50
3.9	Visualization of the three datasets S1, S2 and S3 used in the eval- uation of proposed approaches.	51
3.10	Distribution of tonic frequency for male and female vocal perfor- mances in our music collection.	52
3.11	Screenshot of the MATLAB GUI used for tonic annotations	53
4.1	An example of the decision tree obtained for Method 1	61
4.2	Demonstration of the effect of frequency range selected for the multi- pitch analysis.	62

4.3 Melody histogram of three excerpts on which the rule-based approach fails	67
---	----

List of Tables

3.1	Database description	51
4.1	Performance accuracy of Method 1 with instance normalization . . .	56
4.2	Performance accuracy of Method 1 without instance normalization .	56
4.3	Performance accuracy of first stage of Method 2 with instance normalization	57
4.4	Performance accuracy of first stage of Method 2 without instance normalization	57
4.5	Performance accuracy of second stage of method 2 for rule-based approach	57
4.6	Performance accuracy of second stage of method 2 for classification based approach	58

Chapter 1

Introduction

This thesis addresses a fundamental aspect of the computational analysis of Indian art music, the automatic identification of the tonic. Automatic tonic identification is a crucial first step for a more detailed tonal and melodic analysis of this music tradition, such as intonation analysis, motivic analysis and automatic rāg identification. Tonic identification therefore is a fundamental research problem for the computational analysis of Indian art music, that calls for a cultural specific approach which exploits the specificities present in this music tradition.

Subsequent sections in this chapter provide a brief introduction to the Indian art music, describing the meaning of the tonic in the context of this music tradition and highlighting the importance and precise goal of this research work.

1.1 Indian Art Music Traditions

In this work, Indian art music refers to both the art music traditions of the Indian subcontinent: *Hindustani*¹ (also known as North Indian) music (Bor, Delvoye, Harvey, & Nijenhuis, 2010; Danielou, 2010), prominent in the northern regions of India, Pakistan, Nepal, Afghanistan and Bangladesh; and *Carnatic*² music, widespread in the southern regions of the peninsular (Singh, 1995; Viswanathan & Allen, 2004). In this thesis we use the word “Art” instead of “Classical” to refer

¹http://en.wikipedia.org/wiki/Hindustani_classical_music

²http://en.wikipedia.org/wiki/Carnatic_music

to these music traditions. Raja presents interesting arguments emphasizing the appropriateness of such a terminology (Raja, 2012, Page 1).

The roots of the Indian art music can be traced back to *Sāmved*³⁴, which is one of the four *Vedas*⁵ (ancient core Hindu scriptures) that describes music at length (Trivedi, 2008; Singh, 1995). The *Sāmved* dates back to around 1000 BC, consists of a collection of religious hymns (taken from *R̥gved*⁶, the oldest ved), to be sung using specifically indicated melodies called as *Sāmagan*⁷ (Griffith, 2004). However, the contemporary form of Indian art music is a confluence resulting from the cultural interactions between the Persian, Greek, Arabic, Iranian and Indian people (Kaul, 2007; Saraf, 2011; Singh, 1995).

In both Hindustani and Carnatic music, the *rāg*⁸ (Bagchee, 1998; Danielou, 2010; Viswanathan & Allen, 2004) is the fundamental melodic framework around which the whole music is built upon and the *tāl*⁹ (Clayton, 2000; Sen, 2008) provides the rhythmic framework. Though the Hindustani and Carnatic music traditions share the fundamental musical concepts, the music is significantly different in each tradition (see (Narmada, 2001) for a comparative study of rāgs). Each tradition imbibes its own cultural specificities and has a different approach to music. Moreover, there are several sub-forms (sub-genres), within each tradition, classified based on the different singing styles and instrumentation (Viswanathan & Allen, 2004; Bor et al., 2010).

The seven solfege symbols (Sa, Re, Ga, Ma, Pa, Dha and Nī in short-form) used in Indian art music are termed as *svaras*¹⁰ (Danielou, 2010; Bagchee, 1998). Except ‘Sa’ and ‘Pa’ (the fifth with respect to Sa), every other svar has two or three variations, where each of them has a specific function in a given rāg rendition (Viswanathan & Allen, 2004).

Over the centuries these music traditions have been orally transmitted from one generation to the next, following a hierarchical model of music education such

³<http://en.wikipedia.org/wiki/Samaveda>

⁴<http://www.sacred-texts.com/hin/sv.htm>

⁵<http://en.wikipedia.org/wiki/Vedas>

⁶<http://en.wikipedia.org/wiki/Rigveda>

⁷<http://en.wikipedia.org/wiki/Samagana>

⁸<http://en.wikipedia.org/wiki/Raga>

⁹[http://en.wikipedia.org/wiki/Tala_\(music\)](http://en.wikipedia.org/wiki/Tala_(music))

¹⁰<http://en.wikipedia.org/wiki/Swaras>

as *gharānā*¹¹ in Hindustani music (Saraf, 2011; Mehta, 2008). Each *gharānā* or school of music has its own ideology and characteristic style of music performance (Deshpande, 1989).

Indian art music is essentially heterophonic¹² with a main melody being sung or played by the lead artist (Bagchee, 1998). Many times an instrument provides a melodic accompaniment and follows the lead performer like a shadow (Viswanathan & Allen, 2004). A typical arrangement in a performance of Indian art music consists of a lead performer (occasionally a duo), a rhythm accompaniment generally provided by *tablā*¹³ in Hindustani music and *mṛdangam*¹⁴ in Carnatic music, a constantly sounding drone in the background and frequently a melodic accompaniment using *harmonium*¹⁵ in Hindustani and violin in Carnatic music. The drone sound which is mainly produced by the *tānpūrā*¹⁶ is the only component that adds a harmonic element to the performance (Bagchee, 1998). Generally there are different sets of instruments used in Carnatic and Hindustani music, with an exception of the *tānpūrā*.

1.2 Tonic in Indian Art music

Tonic is one of the basic concepts in any tonal music across the world (Stevens, 2004; Castellano, Bharucha, & Krumhansl, 1984). Broadly, it refers to a particular tone that acts as a focus around which the other tones are organized. However, its precise meaning, function and significance might vary a lot and therefore it should be understood within a given cultural context. This section highlights these issues from the perspective of Indian art music.

Tonic is the foundation of melodic structures in both Hindustani and Carnatic music (Viswanathan & Allen, 2004; Danielou, 2010). It is the base pitch of a performer, carefully chosen in order to explore the full pitch range effectively in a given *rāg* rendition. The tonic acts a reference and the foundation for the melodic

¹¹<http://en.wikipedia.org/wiki/Gharana>

¹²<http://en.wikipedia.org/wiki/Heterophony>

¹³<http://en.wikipedia.org/wiki/Tabla>

¹⁴<http://en.wikipedia.org/wiki/Mridangam>

¹⁵<http://en.wikipedia.org/wiki/Harmonium>

¹⁶<http://en.wikipedia.org/wiki/Tambura>

integration throughout the performance (Deva, 1980). That is, all the tones in the musical progression are constantly referred and related to the tonic pitch. All the accompanying instruments such as *tablā*, violin and *tānpūrā* are tuned using the tonic of the lead performer. It should be carefully noted that tonic in Indian art music refers to a particular pitch value not to a pitch-class. The frequency range of the tonic pitch for male and female singers spans more than one octave (roughly 110-260 Hz) (Sengupta, Dey, Nag, Datta, & Mukerjee, 2005). Specific cases¹⁷ where the singer has the tonic pitch at the extremes are, M D Ramanathan¹⁸ (tonic around 105 Hz) and Veena Sahasrabuddhe¹⁹ having the tonic pitch around 238 Hz.

In any performance of Indian art music (in both Hindustani and Carnatic), the tonic is the Sa (also referred as *Ṣadja*) svar around which the whole rāg is built upon (Danielou, 2010; Bagchee, 1998). Other set of svaras used in the performance derive their meaning and purpose in relation to this reference and to the specific tonal context established by the given rāg (Deva, 1980). The importance of the tonic in Indian art music means identifying the tonic is crucial for many other types of tonal analyses such as intonation analysis (Serrà, Koduri, Miron, & Serra, 2011; Koduri, Serrà, & Serra, 2012), motif analysis (Ross, Vinutha, & Rao, 2012) and rāg recognition (Chordia & Rae, 2007; Koduri, Gulati, & Serra, in press).

Both the performer and the audience need to hear the tonic pitch throughout the concert. This is accomplished by a constantly sounding drone instrument in the background of the performance, which reinforces the tonic. Along with tonic, the drone also emphasises other notes like the fifth, fourth or sometimes the seventh, depending on the choice of the rāg. Essentially, the drone is the reference sound that establishes all the harmonic and melodic relationships between the pitches used during a given performance. Typically the drone is produced by either the *tānpūrā*²⁰, electronic *tānpūrā*²¹ or śruti box²² for the case of vocal music and by the

¹⁷The cases which we have encountered in music collection of CompMusic project, there might be other cases with more extreme tonic values.

¹⁸<http://musicbrainz.org/release/7dda9bb7-81f6-45c4-888e-b924b23613cc>

¹⁹<http://musicbrainz.org/recording/22e45ddb-9b88-406e-996a-2136730d72d4>

²⁰<http://en.wikipedia.org/wiki/Tambura>

²¹http://en.wikipedia.org/wiki/Electronic_tanpura

²²http://en.wikipedia.org/wiki/Shruti_box

sympathetic strings of instruments such as sitār²³, sārangi²⁴ and vīṇā²⁵ for the case of instrumental performances. A detailed description of the different tunings used in the tānpūrā (the most commonly used drone instrument) and a brief discussion on its acoustical properties is provided in Section 2.3.

1.3 Tonic in Western music

In the current work there is no intention of comparing the concept of tonic in Indian art music against the concept of tonic in Western music. However, to better appreciate the differences, it is worthwhile to briefly go over the definitions of the musical concepts related to the tonic in Western music.

In the context of Western music the tonic is associated with the idea of key and tonality. These two terms are often used synonymously but are different as is mentioned in *The New GROVE Dictionary of Music and Musicians* (Grove & Stanley, 1980). These two concepts are defined as follows:

Key: The quality of a musical composition or passage that causes it to be sensed as gravitating towards a particular note, called the key note or the tonic. One therefore speaks of a piece as being in the key of C major or minor, etc. The key of a movement commonly changes during its course through the process of modulation, returning to the home key before the end (Grove & Stanley, 1980, Vol. 10)

Tonality: While the word key is linked with the idea of a diatonic scale in which the notes, intervals and chords are contained, a tonality reaches further than the note content of a major or minor scale, through chromaticism, passing reference to other key areas, or wholesale modulation: the decisive factor in the tonal effect is the functional association with the tonic chord (emphasized by functional theory), not the link with a scale (which is regarded as the basic determinant of key in theory of fundamental progressions). A tonality is thus an expanded key (Grove & Stanley, 1980, Vol. 19)

Without going deep into a musicological discussion, we see that at a surface

²³<http://en.wikipedia.org/wiki/Sitar>

²⁴<http://en.wikipedia.org/wiki/Sarangi>

²⁵http://en.wikipedia.org/wiki/Saraswati_veena

level the tonic in Indian art music is more of an attribute of the lead performer, whereas in Western music it is related to a musical piece. Also, in the former case it refers to a specific pitch value and in the later to a pitch class.

1.4 Motivation and Goals

Advancements in the information technologies has significantly changed the way we interact with the music, be it generation, distribution, storage, browsing & discovery or listening to it. Considering the enormous volume of available music collections, automatic description of musical material becomes a pre-requisite in many situations for developing intelligent systems to be able to perform the aforementioned tasks. Moreover, a thorough understanding of the musical concepts within a given cultural context is the first step towards developing automatic description systems.

Over the past couple of decades the Music Information Retrieval (MIR) community has made significant advancements in the automatic description of music. However, the main focus of the research in MIR has been centered around Western popular music. The technologies developed for this music type do not always respond well to the multicultural reality of the diverse world musics (Serra, 2011). Methodologies developed are not always directly applicable to the other rich music traditions of the world, such as Indian art music, Makam music in Turkey or Chinese Han music. In fact, the research problems themselves are different for each of these music traditions.

These factors create a need for identifying the research problems and developing methodologies which are specific to music cultures (Serra, 2011). This would make our understanding of the musical concepts more universal and should help in bridging the semantic gap (Wiggins, 2009). In this thesis, we present our work on Indian art music, which as described in section 1.1 is a rich old tradition with profound literature and musicological studies.

We focus on the identification of the tonic of the lead performer as this information is fundamental to the description of Indian art music, needed for many melodic analyses such as intonation analysis, motivic analysis and automatic rāg recognition. Our main goals can be summarized as:

- Devise an approach for the automatic labelling of large databases of Indian art music with the tonic pitch for vocal music and tonic pitch-class for instrumental music.
- Utilize culture specific characteristics of Indian art music for automatic identification of the tonic.
- Evaluate proposed approach on a sizable database.
- Review relevant past work and highlight the scientific background of similar tasks to identify meaningful audio features.

Note that we make a distinction between the identification of the tonic pitch (for vocal music) and the tonic pitch-class (for instrumental music). For many melodic analyses such as intonation analysis, the information regarding the tonic octave becomes crucial. This is because the precise intonation and timbral characteristics of a particular svar might be different in different octaves. We found that for the vocalists the concept of tonic pitch is well defined (pitch of the middle register Sa), whereas, for the instrumentalists it is not very clear. Therefore, for the instrumental music we only aim to identify the tonic pitch-class.

Chapter 2

Scientific Background and Related Work

This chapter presents a review of the main audio features, techniques and works relevant to the task of tonic identification. As very little work has been done on tonic identification for Indian art music in the past, the review primarily focuses on the tonal features and signal representations that could be useful for this work. A detailed summary of the tonal structure of the *tānpūrā* is also presented (section 2.3), followed by a summary of the existing work on tonic identification in Indian art music (section 2.4).

2.1 Feature Extraction

In this section we present some of the relevant audio features and signal representations that have been used for the tonal analysis in tasks similar to the tonic identification. Feature extraction is typically the first crucial step in automated analysis of a real-world data. It involves extracting features (sometimes called as descriptors) which are numeric values representing a specific characteristic or attribute of the data. In our given context of the automatic tonal analysis, these features are extracted from the audio data. Essentially the idea is to transform the raw audio data so that the information (perceptual in nature) crucial to a specific task is easily and reliably available. This often results in a drastic reduction in the

overall amount of data.

We review literature on F0-estimation (Section 2.1.1), multi-pitch analysis (Section 2.1.2), predominant melody extraction (Section 2.1.3) and pitch-class distribution features (Section 2.1.4). The following paragraphs describe each of them.

2.1.1 Fundamental Frequency Estimation

The most basic low-level feature that relates to the tonal aspect of a sound is the frequency and its perceptual analog pitch. It is one of the fundamental dimensions of the sound, other dimensions being loudness, duration, and timbre. Most of the real-world pitched sounds can be represented as a superposition of complex sinusoids, meaning that they consist of many sinusoids. These sinusoids are referred as partials or harmonics, as they are in harmonic relationship with the lowest frequency sinusoid, which is called the fundamental frequency (F0). The pitch of a complex tone is generally related to its fundamental frequency. Therefore the problem of fundamental frequency estimation is also sometimes referred as pitch estimation (however, the differences should be kept in mind). Sometimes, despite the fundamental frequency component being absent from the complex tone, the perceived pitch still remains as the fundamental frequency. This phenomenon is referred as missing fundamental (Schmuckler, 2004).

Due to the importance of the pitch information in the speech and music domain, there exists a wide body of research on this topic. However, we notice that many of the proposed algorithms deal only with monophonic audio data, i.e. they can only reliably estimate the pitch of a single sound source and it must be the only sound source present in the audio signal. There are many ways in which the pitch estimation algorithms are classified in the literature. A common way is to divide them into those working in the time domain and those working in the frequency domain, though some algorithms can be expressed in both. A comprehensive discussion on these algorithms is beyond the scope of this thesis. Here, we provide a short overview of the most commonly used algorithms, and refer the readers to the appropriate source for more detailed information.

The fundamental frequency estimation process is typically divided into three sub-processes; pre-processing, F0-extraction and the post-processing as shown in

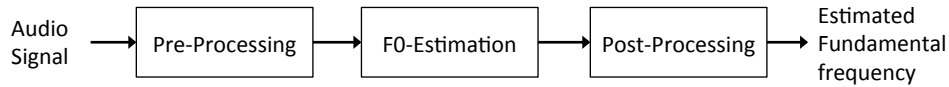


Figure 2.1: Steps of the fundamental frequency estimation process

Figure 2.1 (McKinney, cited in (Hess, 1984)). The goal of the pre-processing step is to apply transformations to the signal which will improve the accuracy of the F0-extraction in the following step. This is typically achieved by performing de-noising, normalization or spectral whitening. The main F0-extraction process generates a sequence of fundamental frequency estimates, computed from overlapping chunks of the input data (frames). Finally, the post-processing block tries to detect errors in the F0-estimation (exploiting continuity constraints) and subsequently performs a correction or smoothing of the obtained pitch contour. In subsequent paragraphs we review only the algorithms for performing the main F0-estimation. For other processes, we refer the reader to (Klapuri, 2003a; Gómez, 2006; V. Rao & Rao, 2010).

The algorithms which have been used most frequently belong to the time domain (lag-based) category. The Auto-correlation function (ACF) based method, which was originally proposed by Rabiner for pitch detection has been one of the key algorithms (Rabiner, 1977). Later, modifications in the ACF based method proposed by Boersma improved the performance, and made this approach robust to additive noise and strong formants (Boersma, 1993). Issues related to the resolution of the pitch estimate due to the finite sampling rate and analysis window size have been addressed by Medan et al. (Medan & Yair, 1991). Another key method for F0-estimation that closely relates to the ACF based methods uses the average mean difference function (AMDF). Proposed by Cheveigné et al. (YIN), this algorithm significantly reduces the error rates found in the ACF based methods (De Cheveigné & Kawahara, 2002).

Besides the time domain algorithms, methods based on spectrum auto-correlation (Lahat & Niederjohn, 1987), subharmonic summation (Hermes, 1988) and harmonic matching (Maher & Beauchamp, 1994) are often used in the music analyses. Some of these algorithms are also applied to obtain the F0-estimation in

polyphonic music scenarios, i.e. when multiple pitched sources are present in the audio. Two-way mismatch criterion (TWM) (Maher & Beauchamp, 1994), which is the foundation of harmonic matching algorithms, has proven to be robust against the sparse tonal interferences in polyphonic music (V. Rao & Rao, 2010). An attempt to combine the advantages of different periodicity estimation algorithms (both from time and spectral domain) was made by A. Klapuri (Klapuri, 2000). The author computes the periodicity function for different frequency bands and then combine them later, making the method robust to inharmonicities and corrupted frequency bands.

For a detailed review of fundamental frequency estimation methods, the reader may refer to (Gómez, 2006; Klapuri, 2003a; V. Rao & Rao, 2010).

2.1.2 Multi-pitch Analysis

Most of the fundamental frequency estimation algorithms mentioned above are designed to work on monophonic audio (section 2.1.1). It is not appropriate to use them in the polyphonic or heterophonic music scenario. First, because they are affected by the presence of other tonal components in the audio, which could lead to erroneous results. And second, we might be interested in extracting more than one pitch value, as there are multiple pitched sources present in the music. In few cases the approaches proposed for F0-estimation in monophonic audio signals are applied to polyphonic audio as well. For example, the TWM procedure mentioned above is extended to extract multi pitch information (Maher & Beauchamp, 1994; V. Rao & Rao, 2010).

(Klapuri, 2003b) proposes an iterative approach based on the concept of source separation (Benaroya, Bimbot, & Gribonval, 2006). In every iteration the pitch of the most prominent sound component is estimated and that source is subtracted from the signal. This algorithm works reasonably well for a wide range of fundamental frequencies and for different kinds of sources. A brief discussion on some of the existing multi-pitch extraction algorithms can be found in (Klapuri, 2003a).

In the subsequent paragraphs we describe the generic structure of the multi-pitch analysis systems proposed recently (Salamon & Gómez, 2012; V. Rao & Rao, 2010; Klapuri, 2006). Typically, the multi-pitch analysis part of these systems have

three main modules; 1) Signal representation, 2) Saliency function computation, 3) F0-candidate extraction. Following paragraphs describe these modules, presenting commonly used state-of-the-art approaches for each of these tasks.

The signal representation module transforms the audio signal to a sparse representation, typically comprising of sinusoids with their respective amplitudes. Some of the challenges at this step include; extracting sinusoids as accurately as possible (both in terms of their frequency and amplitude), filtering out the true sinusoids from the sidelobes of the applied window and handling the non-stationarity of the audio signal. Many techniques have been proposed to handle these issues. Using the parabolic interpolation or the phase vocoder based methods to accurately estimate the frequency and amplitude of the sinusoids (Dressler, 2006), techniques like mainlobe matching to filter out side-lobes of the window (Griffin & Lim, 1988), using variable length windows or a multi-resolution FFT to improve the frequency resolution are some techniques frequently found in the literature (Klapuri, 2000).

Once the sinusoids are extracted, a saliency function is constructed using them, which represents the saliency of different pitch values over time. A frequently used approach for the saliency function computation is harmonic summation (Klapuri, 2006). In this method the saliency of a given frequency is computed as the weighted summation of the energies of all the sinusoids which are found at integral multiples of the given frequency (i.e. at its harmonic locations). A two-way-mismatch (TWM) method as mentioned previously has also been used for the computation of the saliency function (Maher & Beauchamp, 1994). In (Goto, 2004), M. Goto estimates the relative predominance of every possible F0 candidate (represented as a probability density function) by using maximum a posteriori probability estimation and F0 temporal continuity criteria. A comparative study of these approaches can be found in (Poliner, Ellis, & Ehmann, 2007; Salamon, Gómez, & Bonada, 2011; V. Rao & Rao, 2010)

Finally, reliable F0 candidates are extracted from the saliency function, either by selecting the peaks of this function or by applying some heuristic techniques to enhance the selection of the candidates belonging to a particular source, as done in (V. Rao & Rao, 2010). The precise approach to select F0 candidates depends upon the end task, whether it is the multi-F0 trajectory extraction or the predominant melody extraction. The later is discussed in the upcoming paragraphs.

2.1.3 Predominant Melody Extraction

As the name suggests, this task aims at estimating the main or predominant melody line from a polyphonic or heterophonic audio music recording. By this, we mean a scenario where multiple pitched sources or multiple melody lines exist simultaneously at a given point in time. Here the predominant melody is considered as the time varying pitch trajectory of the lead or dominant musical source. In (Poliner et al., 2007), the authors roughly define it as “the melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the ‘essence’ of that music when heard in comparison”. However, providing a strict definition of the melody is not an easy task, as the concept of melody is based on the judgement of human listeners and should be defined within a given cultural context.

A large amount of algorithms proposed for this task can be grouped as “salience based” methods as described in (Salamon & Gómez, 2012). These are also referred to as ‘understanding without separation’ paradigm as described by Scheirer (Scheirer, Vercoe, Benton, & Scheirer, 2000). Here, a pitch salience function is computed from the audio signal (as described in previous paragraphs). Using this pitch salience function, potential F0 candidates are extracted based on their prominence in the salience function. Later, applying tracking rules the system identifies the F0 trajectory that best represents the melody line. One of the motivations behind performing this kind of task was to track both the prominent melody and bass-line, which could later be used for applications like music transcription (Ryynänen & Klapuri, 2008) or music scene description (Goto, 2004). Perceptually motivated approaches combining heuristic rules and melodic characteristics are also proposed to identify the melodic contours which belong to the main melody and the ones that do not (Salamon & Gómez, 2012; Paiva, Mendes, & Cardoso, 2006). The concept of main and secondary melody is culture specific, which from an Indian music perspective is well described and discussed by V. Rao (V. M. Rao, 2011).

Another approach to predominant melody extraction could be first separating the predominant melodic source from polyphonic or heterophonic audio and subse-

quently applying F0-estimation techniques designed to handle monophonic audio scenarios. In (Lagrange, Martins, Murdoch, & Tzanetakis, 2008), the authors propose a method to extract the predominant melodic source (frequently the singing voice) inspired from computational auditory scene analysis (CASA).

For a detailed review of the state-of-the-art in predominant melody extraction we refer to (Poliner et al., 2007; Salamon & Gómez, 2012; V. M. Rao, 2011).

2.1.4 Pitch Class Distribution Features

One of the most common sets of descriptors used for analyzing the tonal content of a music material (tonality) are frequently referred to as pitch-class distributions (PCD), pitch-class profiles (PCP), Harmonic pitch-class profiles (HPCP) or chroma features in general. Though these features are used within similar context (tonal analysis of music), they rely on very different implementations; essentially a vector of features describing the salience of the different tones or pitches present in an audio signal. Typically a 12 dimensional vector (also referred to as bins) with values indicating the amount of energy present in the audio corresponding to each of the 12 semitones. Many applications often demand a finer analysis, in which each semitone is further divided into 2-3 bins, leading to a 24 or 36 bin PCD.

According to Gómez (Gómez, 2006) well computed pitch class distribution features should fulfill these requirements:

1. Represent the pitch class distribution of both monophonic and polyphonic signals.
2. Consider the presence of harmonic frequencies.
3. Robustness to noise that sound at the same time: ambient noise (e.g. live recordings), percussive sounds, etc.
4. Independence of timbre and played instrument, so that the same piece played with different instruments has the same tonal description.
5. Independence of loudness and dynamics.

- Independence of tuning, so that the reference frequency can be different from the standard A 440 Hz.

These properties make HPCPs an ideal candidate for the features to be used in tasks like cover song identification (Gómez & Herrera, 2006; Ellis & Poliner, 2007; Serra & Gomez, 2008), chord recognition and tonality analysis (Peeters, 2006; Gómez, 2006) and audio matching (Muller, Kurth, & Clausen, 2005).

Fine grained pitch class distributions have been used in tonal analysis of makam music, specifically for the problem of tonic and makam recognition (Ioannidis, 2010; Bozkurt, 2008; Gedik & Bozkurt, 2010).

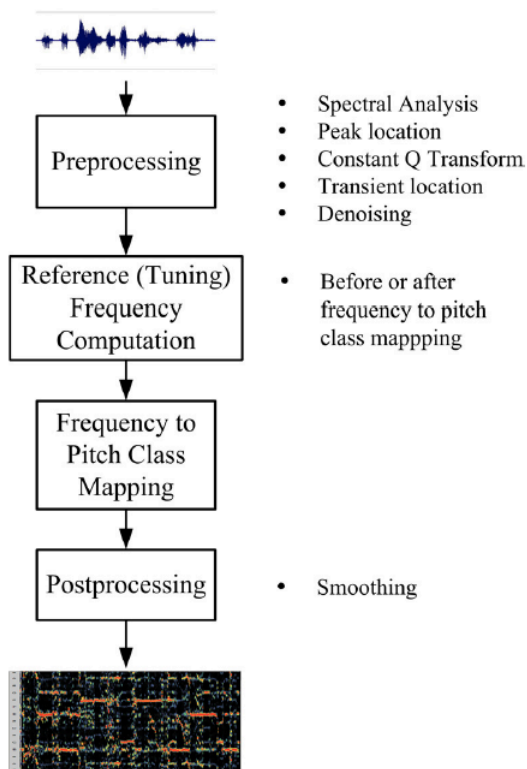


Figure 2.2: General block diagram for methods for pitch class distribution computation from audio. Figure extracted from [(Gómez, 2006)] with the permission of the author.

Gómez also outlines the generic schema for approaches computing an instantaneous evolution of pitch class distributions (Gómez, 2006) (Figure 2.2). Every

module in this diagram is carefully adapted to suit the application and the context in which these features are used. A variety of different implementations for the Chroma features is studied by Muller et al. in (Muller, Ewert, & Kreuzer, 2009; Muller & Ewert, 2010). For a comprehensive review of various approaches to PCD computation and tonality analysis, we refer to (Gómez, 2006).

2.2 Key Estimation in Western Music

In the last decade, key estimation and tonality modeling in western music has been one of the comprehensively studied topic in the MIR community. At the same time, the problem is also studied by researchers from musicological, music cognition & psychology domain. The sheer amount of undergoing research in this direction reflects the complexities involved in this task of key perception by human beings. In this subsection we review some of the relevant work done in Western music.

Two basic ways of modeling human key perception have been proposed by Brown H., namely, the structural and functional approach (Brown, 1988). Both these approaches have received empirical support and they are often regarded as complementary. According to the distinction made by Brown, the structural approach to tonality considers that the listeners derive the key perception or tonal centers by using aggregated pitch content of a musical piece and by deciding which key profile best correlates with the distribution of pitch classes. In this approach, the prevalence of a pitch class primarily depends upon its cumulative duration. However there are secondary factors affecting the perceptual salience of a pitch class, like its repetition, or metrical position. Structural approach pays relatively less importance to the local ordering of the pitch values. On the other hand, the functional approach is quite complementary to this approach. It pays less importance to the pitch class material and more emphasis on the sequential intervallic implications.

One of the seminal approaches for automatic key estimation is the key profile method proposed by Krumhansl and Kessler (Krumhansl & Kessler, 1982; Krumhansl, 1990). The approach is based on the correlation of the established tonal hierarchies that are experimentally determined using the probe-tone method

with the pitch class distribution (PCD) of the musical material. Pitch saliences in PCD are affected by the pitch repetition and their relative duration in a musical piece. Many recent approaches to key estimation are motivated by this methodology and use these experimentally determined key profiles to match with the PCDs extracted from the audio (Peeters, 2006). In these methods, the pitch class distributions derived from the audio are correlated with each of these templates (12 for major and 12 for minor scale), and the one resulting in the highest correlation is selected as the correct key note and the mode. However the extraction of the PCDs from the audio is not a trivial task. Some of the challenges involved are the presence of the multiple pitch sources in the audio and their corresponding harmonics (Gómez, 2006). One of the methods to tackle this difficulty is proposed by Gómez (Gómez, 2006). In her approach instead of the pitch class profile a harmonic pitch class profile is built directly from the audio taking appropriate contributions from all the harmonics. A quite comprehensive overview of the main studies pertaining to tonality and key estimation in western music can be found in her dissertation (Gómez, 2006).

An alternate model to represent the tonality in Western music is the spiral model, proposed by Chew. (Chew, 2002). In this model the pitches are represented in a three-dimensional space (3D spiral), and every key has its particular place in that space.

2.3 Tonal Structure of the Tānpūrā

The presence of the drone in the background is a characteristic feature of Indian art music and plays a very crucial role. The emergence of this concept dates back to AD 1600 (Bagchee, 1998). The drone acts as a reference of the music to a tonal background, reinforcing all the harmonic and melodic relationships. The presence of the drone brings out the issues of intonation and consonance more than it otherwise would have been. As described by Deva (Deva, 1980) without a drone the intonation and the tonality of the music are governed by the tonal memory (a matter of retrospect and post relation of tones). But with the employment of the drone, a musician is forced to constantly refer his tones to this tonal background both for the intonation and consonance resolution.

The tonal structure of the drone is thus a very important aspect of this music tradition. The current section briefly describes the tonal structure of the tānpūrā, which is the main drone instrument used to accompany the lead performer. In particular, we discuss the different types of tuning, its timbral characteristics, different playing styles and finally highlight the characteristic feature of this instrument, i.e. the rounded (buzzing) sound created by the special configuration of the bridge.

Tānpūrā is a long-necked plucked lute, which comes in different sizes that corresponds to the different pitch ranges it can produce. The largest one is for the male singers, a smaller size for the female singers and the smallest one to accompany the instrumentalists. It usually has 4 strings (5 or 6 in rare cases) which are plucked serially in a regular pattern to create a rounded resonant sound.

Figure 2.3 shows a 4 string tānpūrā with its important body parts duly labelled and the frequently used tuning configurations listed on the side. The pegs corresponding to 4 strings are marked with the numbers 1, 2, 3 and 4 respectively. The two middle strings of the tānpūrā (corresponding to pegs 2 and 3) are tuned to the tonic pitch of the lead performer (Sa), while the fourth string (corresponding to peg 4) is tuned an octave below the tonic pitch (sa). In addition to reinforcing the tonic pitch, tānpūrā also produces secondary pitch classes. The first string of the tānpūrā (attached to peg 1) is frequently tuned to the fifth (pa) with respect to the tonic pitch, in the lower octave, resulting in pa-Sa-Sa-sa type of tuning. For rāgs which omit pa (fifth), the first string is tuned to the natural fourth (ma) as ma-Sa-Sa-sa. Furthermore, for some rāgs it is desirable to tune the first string to seventh (nī) as nī-Sa-Sa-sa, where ‘nī’ is one semitone below the tonic pitch (Sa).

The tānpūrā sound is composed of rich overtones, with the higher harmonics adding energy to various pitch classes. Deva presents a detailed analysis of the spectral characteristics of the tānpūrā sound (Deva, 1980). The author also provides an interesting historical perspective on the emergence of the tānpūrā and its significance in Indian art music. Figure 2.4 & 2.5 show the spectrogram and spectrum of a short audio excerpt taken from a solo tānpūrā field recording ¹. We notice from these figures that the tānpūrā sound is quite bright with a low spectral roll-off and has a dense spectrum.

¹<http://www.freesound.org/people/sankalp/sounds/153263/>

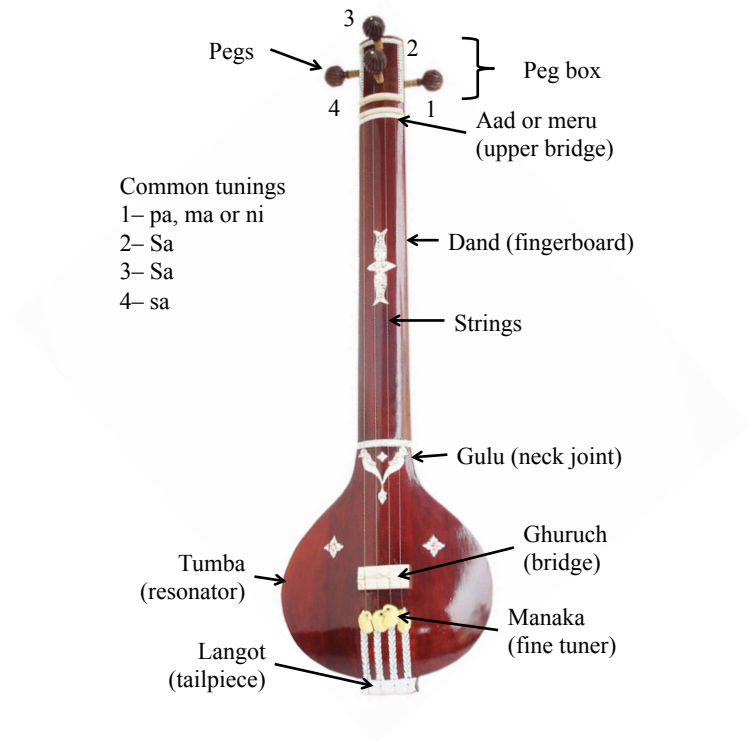


Figure 2.3: Image of a tãnpūrã with all its important body components duly labelled. Common tuning configurations are also shown.

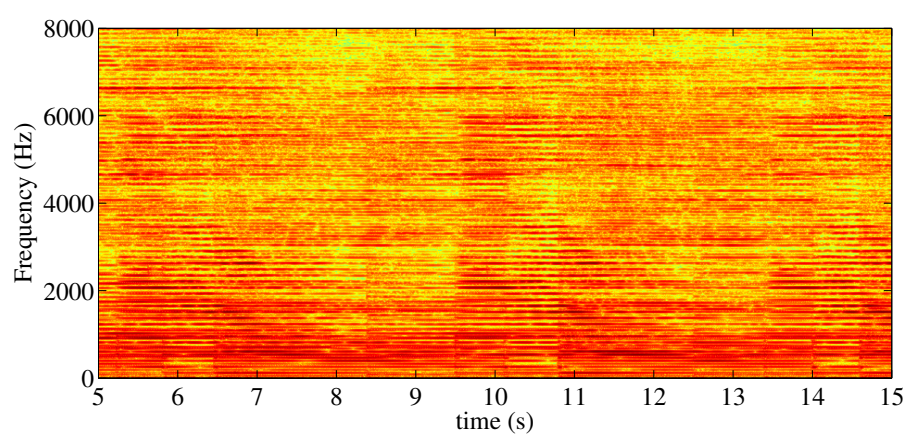


Figure 2.4: Spectrogram of a section of solo tãnpūrã field recording.

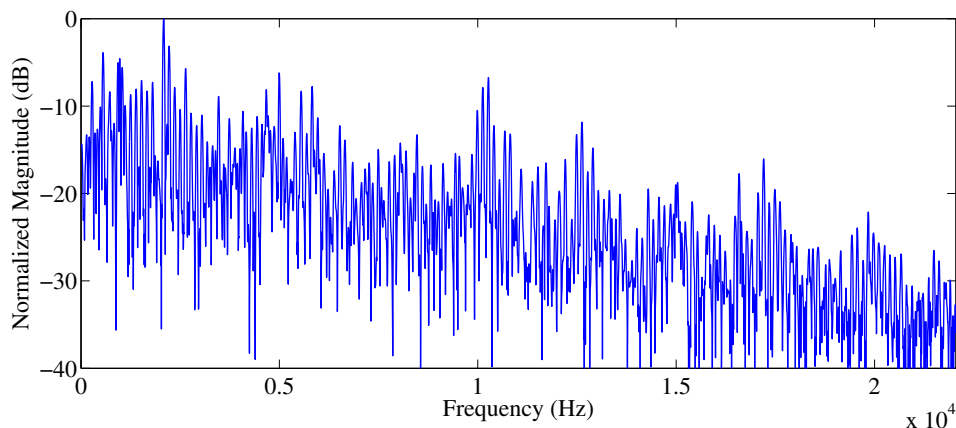


Figure 2.5: Spectrum of a tānpūrā recording.

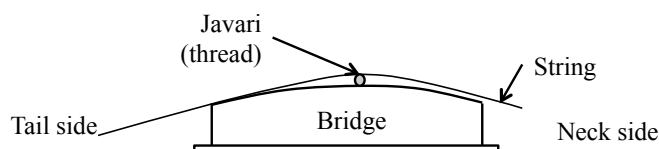


Figure 2.6: Diagram showing cross-section view of tānpūrā bridge, marking the position of javārī (thread) between the bridge and the strings.

The tānpūrā strings are gently plucked by the fingers to avoid discreet transients. The fingers are repeatedly rolled over the strings to create a constant slow rhythmic pattern (which bear no relation with the speed of the song). The playing style slightly differs in Hindustani and Carnatic music. In Carnatic music one complete cycle of the rhythm pattern consists of 6 matrās (beats for simplicity), whereas in Hindustani music it contains 5 matrās.

The unique resonant sound of the tānpūrā is due to the special type of the bridge, and because of the *javārī* (special thread) that is inserted between the bridge and the strings at a specific point. Raman studied this Phenomenon in depth; describing how the javārī makes the vibration modes in strings violate the helmholtz law (Raman, 1921). Detailed explanations of the conducted experiments to observe this effect are summarized in (Bagchee, 1998). Figure 2.6 shows a cross section view of the bridge, indicating the position of the javārī (thread) between the bridge and the strings.

Several recorded samples of the acoustic and electronic tānpūrā with different

tuning settings and tonic values can be obtained from the Freesound website ^{2 3} .

2.4 Tonic Identification in Indian Art Music

Despite the importance of automatic tonic identification in the description of various aspects of Indian art music, the problem has not received much attention in the past. This section describes briefly the approaches found in the literature and summarizes their shortcomings.

One of the first approaches for tonic identification was proposed by Sengupta et al. which uses only the *ālāp*⁴ (opening section of a typical performance of Indian art music) sections of 118 solo vocal recordings to get the clean pitch tracks (Sengupta et al., 2005). The author considers only the stable regions (60 ms) in the pitch envelope for the analysis. The proposed approach is quite brute force and begins by considering a large number of frequency values as the candidates for the tonic pitch. The candidate that best represents the distribution of stable pitch regions according to a defined error minimization criterion is selected as the tonic. In addition to the audio data, information regarding the gender of the singer (male or female) is also used to restrict the tonic pitch range. The range of the tonic pitch considered for the male singers is 95-175 Hz and for the female singers it is 185-255 Hz. The evaluation methodology used by the author is somewhat problematic. For instance, the final results are provided in terms of the average error in tonic frequency in Hz. First, it does not reflect for how many files the tonic was correctly identified. And second, as the error is reported in terms of the frequency, it is perceptually not very meaningful. The perceptual error (error in cent scale) for the same frequency error will be different for different tonic frequencies.

The approach proposed by Ranjini et al. takes advantage of the peculiarities found in the melodies of Carnatic music (Ranjani, Arthi, & Sreenivas, 2011). This method is also based on prominent pitch extraction and uses the pitch distribution of the predominant melody as a representation for the tonal structure. In Carnatic

²<http://www.freesound.org/people/sankalp/packs/9600/>

³<http://www.freesound.org/people/sankalp/packs/9571/>

⁴<http://en.wikipedia.org/wiki/Alap>

music the extent of *gamak*⁵ (the amount of deviation from the central frequency) on the Sa and Pa svar of the rāg is minimal and so the histogram peaks corresponding to these svaras are expected to be narrower (more peaky) as compared to the other svaras. These narrow peaks are identified by analysing the parameters of fitted semi-continuous GMMs. The authors only mention that the Praat software is used for the pitch extraction, without describing in detail the parameters used or indicating its performance accuracy on the chosen dataset. These chosen parameters can make a lot of difference in the pitch extraction task. Also, Praat’s pitch extraction algorithm is designed to work in the monophonic case, but it has been used for a polyphonic case (that has lots of harmonic content as tānpūrā is usually present along with the voice) in this work. The database used to evaluate this approach is confined only to the performances of the sampūrṇa rāg (specific category of rāgs comprising of all the seven svaras). Moreover, the size of the database is also quite small, comprising only 55 ālāpnās.

The research in the area of automatic tonic identification for Indian art music is still in the nascent stages. Some of the shortcomings in the existing approaches can be summarised as:

1. Approaches are solely based on the analysis of the predominant melody. No efforts have been made to exploit the presence of the drone sound in the background, which is an important cue for the identification of the tonic.
2. Researchers have used monophonic pitch trackers, whereas the music material under investigation has polyphonic elements (i.e. multiple pitched sources present in the audio).
3. The approaches are evaluated on a limited database, both in terms of the number of recordings and the diversity present in the selected musical material.
4. All the previous approaches aim to identify the Sa (Ṣaḍja), which is the tonic pitch-class. It lacks the information about the correct tonic octave that might be crucial in tasks such as intonation analysis.

⁵[http://en.wikipedia.org/wiki/Gamaka_\(music\)](http://en.wikipedia.org/wiki/Gamaka_(music))

As we see, there is a wide scope for improvement in these approaches and in devising new methods that explore the cultural specificities present in Indian art music.

Chapter 3

Multi-pitch Approach to Tonic Identification

This chapter describes in detail the proposed approach for tonic identification using multi-pitch analysis of the audio excerpts. We present two different methods for tonic identification, to be able to work on both vocal and instrumental music. Section 3.1 gives an overview of the proposed methods and presents the motivation and scope of application for both of them. Sections 3.2 and 3.3 explain the implementation steps for the methods (Method 1 and Method 2), describing each step in detail and the evaluation methodology is presented in Section 3.5.

In addition to the proposed methods, a proposal for a complete iterative system for tonic identification in Indian art music is presented in Section 3.4. The system aims to incrementally utilize all the available data (audio data and metadata) to identify the tonic and also estimate a confidence measure for each output.

3.1 Overview of the Methods

The proposed methods use a multi-pitch analysis of the audio signal to identify the tonic in both Hindustani and Carnatic music. The motivation for adapting a multi-pitch analysis is twofold: first, the music material under investigation is non monophonic (includes many instruments playing simultaneously). Second, we know that the tonic is continuously reinforced by the drone sound, an important

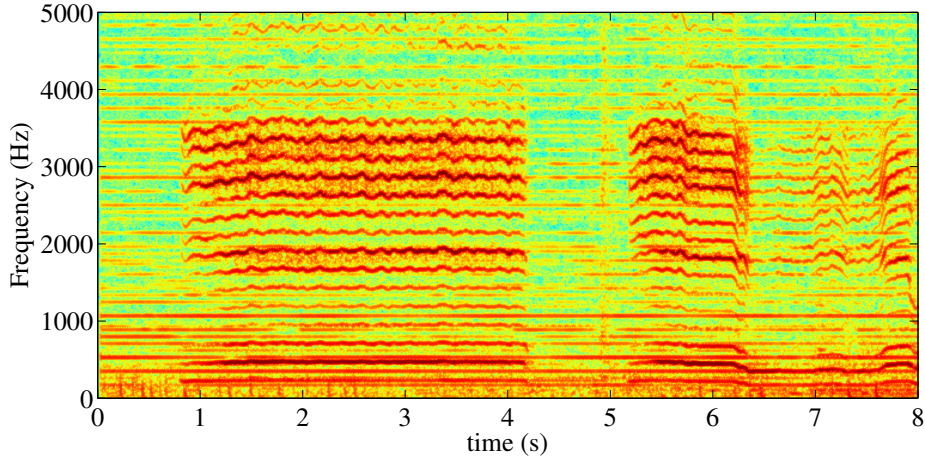


Figure 3.1: Spectrogram of an excerpt of Hindustani music with two clearly visible types of harmonic series, one belonging to the drone and the other to the lead voice.

cue that can not be exploited if we extract a single pitch value for each frame of the audio recording. To illustrate this point in Figure 3.1 we display the spectrogram of a short audio excerpt of Hindustani music. It clearly shows two types of harmonic series; first consists of nearly flat lines and corresponds to the drone instrument (playing Sa and Pa). The second harmonic series (which start at around 1 second) corresponds to the voice of the lead performer. If we only consider the pitch of the lead performer (which is quite dominant) in our analysis, we loose the information in the drone sound, which in this case is a important indicator of the tonic pitch.

In this thesis we propose two methods, namely, Method 1 and Method 2. Method 1 is solely based on a multi-pitch analysis (Salamon, Gulati, & Serra, 2012) whereas Method 2 uses a multi-pitch analysis and predominant melody information to perform tonic identification. As the tonic pitch range of male and female singers spans more than one octave (Section 1.2) and we require the tonic octave information for many melodic analyses (Section 1.4), the method should not only identify the tonic pitch-class but also the octave in which it lies (tonic pitch identification covers both these tasks). Method 1 identifies the tonic pitch-class and the tonic octave in a combined way, directly by performing a multi-pitch analysis of the audio signal. However, Method 2 divides this task into two stages; first, identifying the tonic pitch-class by performing a multi-pitch analysis and second, using the predominant melody pitches to establish the tonic octave. The Following

points highlight the motivation and scope of application of these methods.

- Method 1: The two middle strings of the tānpūrā are tuned to the tonic of the lead performer (Section 1.2 and 2.3) and the rest of the strings with respect to the tonic. Therefore, the drone sound contains sufficient information to extract the tonic pitch. Method 1 capitalizes on this logic and identifies the tonic pitch directly by performing a multi-pitch analysis of the audio.

This method is designed to work for vocal music. Since we are interested in identifying only the tonic pitch-class for instrumental music (Section 1.4) and this method performs tonic pitch identification in a single stage, it is not applicable to instrumental music. Instrumental music pieces are not annotated with the tonic pitch (but with the tonic pitch-class), which is required for training in Method 1.

- Method 2: The motivation behind proposing the second method is:
 1. To enable the method to be used for both vocal and instrumental excerpts.
 2. While annotating the excerpts with the tonic pitch it was observed that the decision of the tonic octave is primarily based on the pitch range of the sung melody.
 3. We found that in few cases when musicians go on tour, they take the liberty to use a smaller tānpūrā instead of a big one. In such situations, the middle two strings of the tānpūrā generate a pitch sensation of an octave above to the tonic. This argument got some support when we carefully examined an electronic tānpūrā. We observed that the range of the tonic frequencies which an electronic tānpūrā is capable of producing is nearly one octave. Additionally, it has a knob to change the equalization depending upon whether the singer is male or female. However, we know that the combined tonic pitch range for male and female singers spans more than one octave (110-260 Hz, Section 1.2). Therefore, there is a chance that in some rare circumstances the singers might be using the drone instrument as a support for only the tonic-pitch class. Note that there is no consolidated evidence found for this

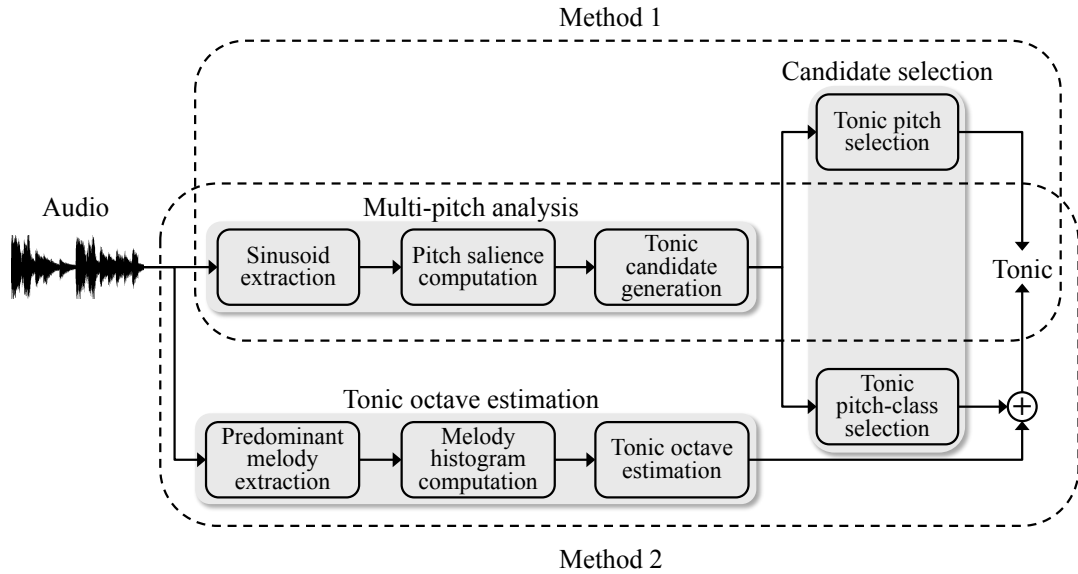


Figure 3.2: Block diagram of the proposed methods. There are three main processing units, indicated by the gray regions, namely, multi-pitch analysis, candidate selection and tonic octave estimation.

argument. This possibility came up during our discussions with some of the musicians.

These factors suggest that it might be beneficial to include the melody information in addition to the drone in the analysis. Therefore, Method 2 divides this process into two stages and identifies the tonic pitch-class by performing a multi-pitch analysis and subsequently analyzes the predominant melody pitches to obtain the tonic octave. Thus, in addition to vocal music, Method 2 is also applicable to instrumental music, where we only identify the tonic pitch-class and omit the stage of the tonic octave estimation.

Figure 3.2 shows a coarse block diagram of both the proposed methods, demonstrating the differences and shared processing blocks between the two. As we see in the figure, there are three main processing units (regions with gray background); first, the multi-pitch analysis, which takes the audio signal as input and generates tonic pitch (or pitch-class) candidates. This processing block is exactly the same

for Method 1 and Method 2. Second, the candidate selection, which identifies the correct tonic candidate using a learned template based on a classification approach (different for both the methods). And the third, tonic octave estimation unit, which as the name suggests identifies the octave in which the tonic pitch lies. This block is only included in Method 2.

Detailed implementation steps for both the methods are provided in the subsequent sections. Section 3.2 describes Method 1 in detail, explaining each of the individual processing blocks, the multi-pitch analysis (Section 3.2.1) and the candidate selection (Section 3.2.2). Section 3.3 describes Method 2, explaining both the stages, the tonic pitch-class identification (Section 3.3.1) and the tonic octave estimation (Section 3.3.2). Note that the multi-pitch analysis module is same for both the methods and therefore it is explained only while describing Method 1.

3.2 Method 1

This method involves two main processing steps, the multi-pitch analysis and the candidate selection (Figure 3.2) (Salamon et al., 2012). Both these steps are explained below.

3.2.1 Multi-pitch Analysis

The multi-pitch analysis used in this thesis is taken from the first block of the melody extraction algorithm proposed by Salamon and Gómez in (Salamon & Gómez, 2012) (Sinusoid extraction and Saliency function computation blocks). The input to the multi-pitch analysis module is audio signal and it generates tonic pitch candidates. Figure 3.3 shows a detailed block diagram of this module, highlighting various sub-tasks involved. Each of the constituent blocks is elaborated below in subsequent sections, providing all the necessary implementation details.

Sinusoid Extraction

We start off by extracting the sinusoidal components of the audio signal, as done in most of the tonal analysis. This process is divided into three parts (Figure 3.3); spectral transform, peak picking and sinusoid frequency and amplitude correction.

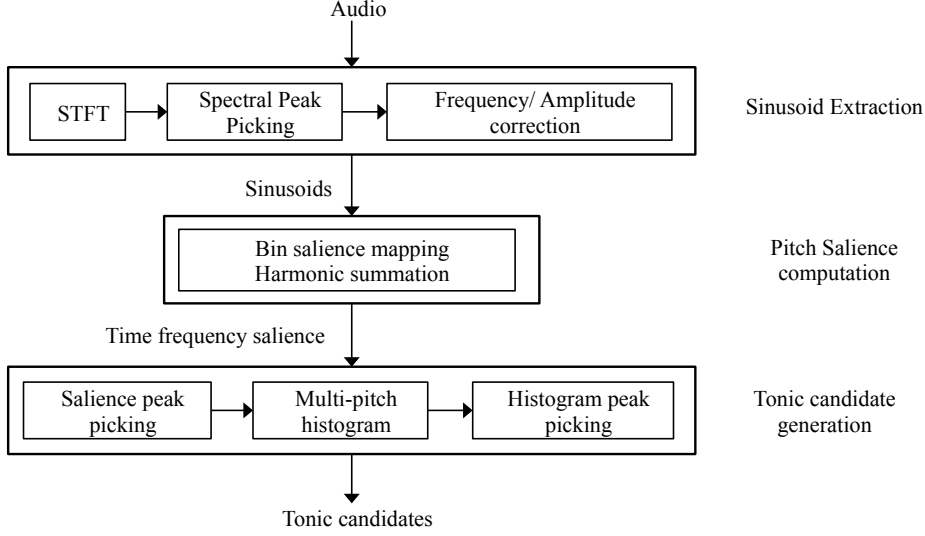


Figure 3.3: Detailed block diagram of multi-pitch analysis module

We use Short-Time Fourier Transform (STFT) to transform the audio signal from a time domain to a time-frequency domain representation. STFT is given by:

$$X_l(k) = \sum_{n=0}^{M-1} w(n) \cdot x(n + lH) e^{-j \frac{2\pi}{N} kn}, \quad (3.1)$$

$$l = 0, 1, \dots \text{ and } k = 0, 1, \dots, N - 1$$

where $x(n)$ is the time domain signal, $w(n)$ the windowing function, l the frame number, M the window length, N the FFT length and H the hop size. We use the Hamming windowing function with a window size of 46.4 ms, a hop size of 11.6 ms and a $\times 4$ zero padding factor, which for data sampled at $f_S = 44.1$ kHz gives $M = 2048$, $N = 8192$ and $H = 512$ (Salamon & Gómez, 2012). Given the FFT of a single frame $X_l(k)$, spectral peaks p_i are selected by finding all the local maxima k_i of the magnitude spectrum $|X_l(k)|$.

Not all the spectral peaks correspond to valid sinusoids, there are many spurious peaks (relatively low energy) generated as a result of the windowing. Numerous techniques are proposed to filter out this noise and to extract true sinusoids

(mainlobe matching technique (Griffin & Lim, 1988) etc). In the current implementation, we apply simple energy threshold to discard the spurious spectral peaks. The energy threshold (T_s) is calculated as follows:

$$\begin{aligned} T_s &= \max(T_r, \alpha), \\ T_r &= E_m + \beta \end{aligned} \tag{3.2}$$

where T_r is the relative threshold w.r.t the maximum spectral peak (E_m) for each frame, α is the an absolute threshold and β is a relative threshold parameter. We use $\alpha = -70$ dB and $\beta = -40$ dB.

The frequency resolution in STFT is limited by the spectral resolution (bin frequencies). So, the frequencies of the sinusoids extracted by performing spectral peak picking will be quantized to the bin frequencies. A Simple way to improve the resolution is to increase the FFT size at an expense of high computational cost. Current implementation already has FFT size of 8192 points, which is reasonably high. Alternatively, one could resort to techniques such as parabolic interpolation or phase vocoder based methods to correct the frequency and amplitude of sinusoids. We apply three-point parabolic interpolation, given by following equation:

$$\begin{aligned} f &= \frac{\alpha - \gamma}{2(\alpha - 2\beta + \gamma)}, \\ y &= \beta - \frac{1}{4}(\alpha - \gamma)f \end{aligned} \tag{3.3}$$

where f and y are the interpolated frequency and amplitude values of the sinusoid, α , β and γ are the amplitudes (in logarithmic domain, dB) of the three highest samples around the spectral peak (β).

Pitch Salience Computation

The extracted sinusoids are used to compute a salience function, a time-frequency representation indicating the salience of different pitches over time. In this work we use a salience function proposed by Salamon and Gómez in (Salamon et al., 2011). The method is based on harmonic summation (Klapuri, 2006). In short, the salience of a given frequency is computed as a weighted summation of energy found at all the integer multiples (harmonics) of that frequency. This brings out

the fundamental frequency component of the complex sinusoidal mixture, as it receives contributions from all its harmonics. The peaks of the salience function at a given time instance represent the prominent pitches present in that frame. Note that though the two concepts; pitch (which is perceptual) and fundamental frequency (which is a physical measurement) are not identical, for simplicity we use these two terms interchangeably.

The constructed salience function spans a pitch range of 5 octaves, starting from 55 Hz to 1.76 kHz. To reduce the computational cost, the frequency values are quantized into a total of 600 bins on a cent scale, such that the resolution of each bin is 10 cents (sufficient for our analysis). The mapping between a given frequency value f_i in Hz to its corresponding bin index $b(f_i)$ is given by:

$$b(f_i) = 1200 \frac{\log_2(f_i/f_r)}{\eta} + 1 \quad (3.4)$$

where f_r is the reference frequency, η is the bin resolution in cents. We use $f_r = 55$ Hz and $\eta = 10$, which is sufficient for our analysis.

At each frame, the salience of a pitch $S(j)$ (at j^{th} bin) is computed using N_p number of extracted sinusoids with frequencies \hat{f}_i and magnitudes \hat{a}_i . The computation is done as follows:

$$S(j) = \sum_{h=1}^{N_h} \sum_{i=1}^{N_p} g(j, h, \hat{f}_i) \cdot (\hat{a}_i)^\beta \quad (3.5)$$

where N_h is the number of harmonics considered (a crucial parameter), β is a magnitude compression factor and $g(j, h, \hat{f}_i)$ is the function that defines the weighting scheme. We use $N_h = 20$ and $\beta = 1$ in the current implementation.

Note that in (Salamon & Gómez, 2012), a relative magnitude threshold ($e(\hat{a}_i)$) is also applied at this step, to discard the low-energy sinusoids with respect to the maximum energy sinusoid in each frame. In current implementation we perform this operation during the sinusoid extraction step (section 3.2.1, equation 3.2). Another critical component of the harmonic summation is the weighting function ($g(j, h, \hat{f}_i)$), which defines the weight given to a sinusoid when it is considered as the h^{th} harmonic of the bin j . We use the weighting scheme as follows:

$$g(j, h, \hat{f}_i) = \begin{cases} \cos^2(\delta \cdot \frac{\pi}{2}) \cdot \alpha^{h-1} & \text{if } |\delta| \leq 1 \\ 0 & \text{if } |\delta| > 1 \end{cases} \quad (3.6)$$

where $\delta = |b(\hat{f}_i/h) - j|/10$ is the distance in semitone between the folded frequency \hat{f}_i/h and the center frequency of the bin j , and α is the harmonic weighting parameter (we use $\alpha = 0.8$). The non zero values for $|\delta| < 1$ means that each sinusoid not just contributes to a single bin of the salience function (i.e. $b(\hat{f}_i/h)$) but also to the neighboring bins with a \cos^2 weighting. Performing this smoothed weighting avoids potential problems that may arise due to the quantization of salience function into bins and inharmonicities present in the audio. (Salamon & Gómez, 2012) addresses the issue of finding the optimal values of the aforementioned parameters for predominant melody extraction task.

To have a better understanding, we present an example showing the peaks of the salience function (Figure 3.4) for the same audio excerpt whose spectrogram is shown in Figure 3.1. We notice that the tonic pitch (Sa) and fifth (Pa) played by the *tānpūrā* are clearly visible along with the peaks corresponding to the voice. However, we observe that the salience of the pitch values corresponding to the voice is much higher than the *tānpūrā* sound. It is because of the arrangement that the drone always function in the background. We are interested in exploiting the drone signal as much as possible and therefore the prominence of pitched content of the lead performer needs to be normalized. This issue is handled in the next stage while obtaining the tonic candidates from salience function.

Tonic Candidate Generation

We proceed to extract the potential tonic pitch candidates using the salience function computed in the previous step. Each candidate is represented by a frequency and an amplitude value. The process of generating the tonic candidates includes three sub-tasks (Figure 3.3); detecting peaks of the salience function, computing a pitch histogram and extracting candidates as the peaks of this histogram.

Peaks of the salience function represent the prominent pitches of the lead instrument, voice and other predominant accompanying instruments present in the audio recording at every point in time. A histogram of the pitch values corre-

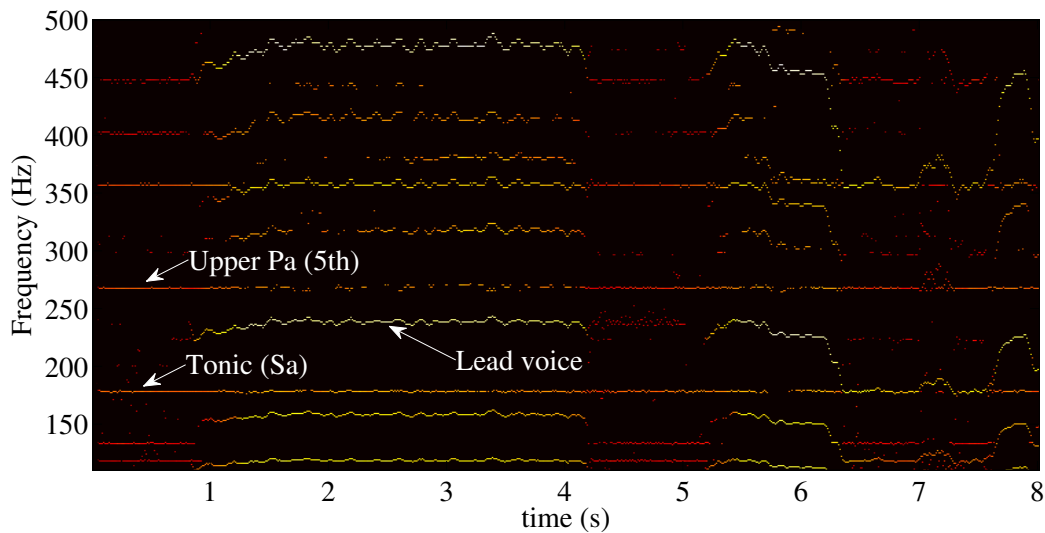


Figure 3.4: Peaks of the salience function for the same excerpt shown in Figure 3.1. The top 10 peaks of the salience function are selected for each frame. Magnitude of a peak is in logarithmic scale (dB)

sponding to the peaks in the salience function for the entire excerpt would indicate which pitches are repeated most often throughout the excerpt. Though the pitch histograms have been used previously for tonic identification (Ranjani et al., 2011), they were constructed using only the predominant melody. Therefore, in many cases the tonal information provided by the drone instrument is not taken into consideration.

We start by selecting the peaks of the salience function at each frame. As the frequency range of the tonic pitches chosen by the singers in Indian art music is within a finite range, between 110-260 Hz (Figure 3.10), we can limit the range from which the salient pitches are selected. We chose a lenient frequency range of 110-370 Hz to select the peaks from the salience function. This ensures that the range of the tonic pitch for both male and female singers is covered. Moreover, the range spans nearly 2 octaves, and therefore the system must be able to identify not only the correct tonic pitch-class but also the octave in which it is played. For each frame, we select the 10 most salient pitch values within the frequency range of 110-370 Hz. The selected peaks are used to construct a multi-pitch histogram, which represents the cumulative occurrences of different pitches at the level of the whole audio excerpt.

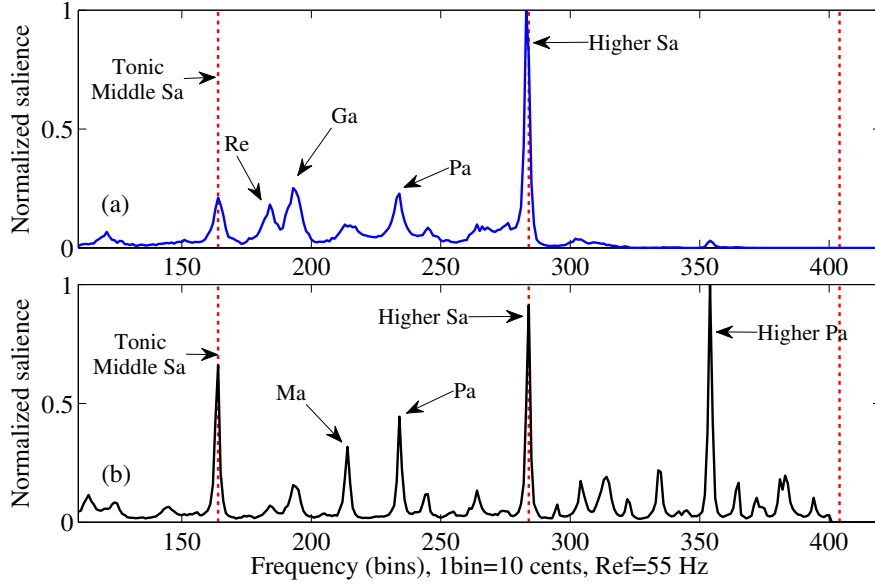


Figure 3.5: Histograms constructed using a) predominant melody (blue) and b) multi-pitch salience function (black). The tonic pitch-class locations are indicated by the red dotted lines.

However, We notice that generally the lead voice/instrument is much louder than the drone sound (Figure 3.4). To normalize this bias towards the dominant source, we drop the saliences of the peaks and consider only their frequency of occurrence. This way a peak that corresponds to the voice has equal weight in the histogram as the peak which corresponds to the drone.

Since we consider only the frequency of occurrence of pitches while constructing the pitch histogram, the pitches produced by the drone instrument (tonic and either Pa, Ma or $N\bar{i}$) have high saliences in the histogram, as the drone is constantly playing in the background. In such cases the pitch distribution depends heavily on the tuning of the drone instrument. This would not be the case if we only considered the predominant melody for histogram computation, as the pitch distribution then would depend on the specific rāg, making this task a joint estimation of the tonic and the rāg, eventually adding more complexity to the problem. To demonstrate this, Lets make a small comparison of the pitch distributions computed using different methods. Figure 3.5 shows two pitch histograms

computed using (a) predominant melody and (b) multi-pitch salience function. These histograms are computed using a three minute long audio excerpt from our database. The pitch axis is plotted in cents, and the histogram is normalised by the magnitude of its highest peak. We see that in the pitch histogram computed using predominant melody (a), the top three peaks correspond to svar Sa, Ga and Re (prominent svaras of rāg *Sindh Bhairavī*), whereas for the later case (b), the top three peaks correspond to Sa (in two octaves) and Pa, which are the prominent svaras produced by the drone instrument.

The tonic pitch will not always be the highest peak of the pitch histogram . We therefore consider top 10 peaks of the histogram $p_i(i = 1 \dots 10)$, one of which corresponds to the tonic pitch. We call them tonic pitch candidates and store both frequency and amplitude of each of these candidates for every audio excerpt. The next section describes the process of selecting the correct tonic candidate using a template learned with a classification based method.

3.2.2 Candidate Selection

This section describes the process of selecting the correct tonic candidate, using both frequency and amplitude of each of the 10 extracted candidates. We perform tonic candidate selection using an automatically learned set of rules (can be treated as a template) based on classification. Extracting features from the pitch histograms we train a classifier to predict the class of the instance, which is then used to infer the correct tonic candidate. Here, a crucial part of the process is to select the class labels appropriately, so that together with the peaks of the pitch histograms (candidates), it can be used to select the candidate corresponding to tonic pitch. Subsequent paragraphs describe this process in detail.

As seen in Figure 3.2, at this step, Method 1 and Method 2 differ from each other. In Method 1 we aim to identify the true tonic pitch candidate whereas in Method 2 the target is to select a candidate that corresponds to the tonic pitch-class. This section explains the candidate selection procedure followed in Method 1.

The pitches used in a musical piece are in relation to the tonic of the lead performer. Taking this into account, we hypothesize that the tonic can be identified

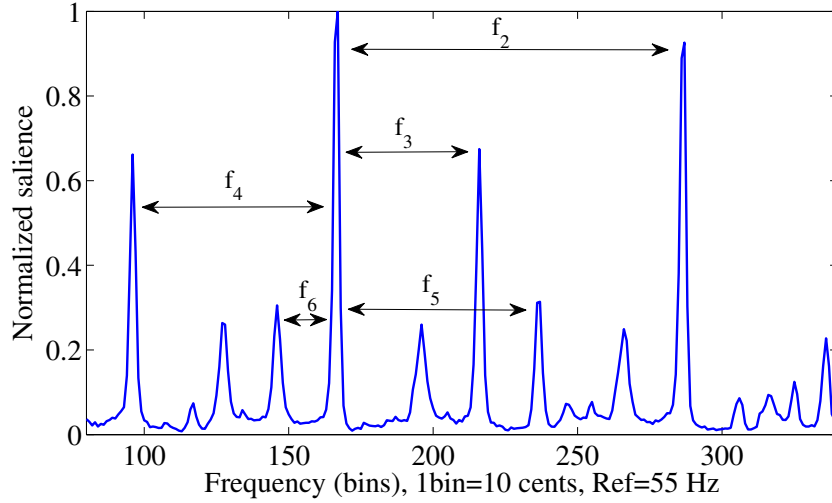


Figure 3.6: An example of multi-pitch histogram displaying five pitch-interval features $f_i (i = 2 \dots 6)$ of the 10.

based on the relationship of the pitch intervals between the most dominant pitches in the recording and their rate of occurrence. We propose a classification based approach to automatically learn the best set of rules to select the correct tonic candidate based on these relationships.

The first step is to encode the pitch intervals between the candidates in a meaningful way such that it can be used as a feature set to train a classifier. We compute the distance between every tonic candidate (p_i) and the most salient candidate in the histogram (p_1). This gives us a set of features $f_i (i = 1 \dots 10)$ (pitch-interval features), where f_i is distance in semitone between p_i and p_1 . Another set of features $a_i (i = 1 \dots 10)$ (amplitude features) include the amplitude ratios of all the candidates with respect to the highest candidate. To visualize the extracted features for a better understanding, in Figure 3.6 we show the pitch-interval features f_i . Only five of the ten pitch-interval features are shown to keep the figure clean. The other set of features $a_i (i = 1 \dots 10)$ are simply the amplitude values of the peaks (p_i) as the histogram is already normalized by the highest peak.

We annotate each audio excerpt with a class label (as explained below) and use 20 features (f_i, a_i) to train a classifier in order to predict the class label. In this way the system automatically learns the best set of rules that maximise the

class prediction. The strategy for labelling an instance with a class should be such that it allows us to uniquely associate the tonic pitch with it, given all the 10 candidates.

In this method the class label of an instance is assigned as the rank of the tonic pitch in the ordered list of all the candidates, arranged in descending order of their peak magnitude. For example, if the candidate corresponding to the tonic pitch is the second highest peak of the histogram, we assign a label “Second”. Theoretically it is a 10 class problem, as the allowed tonic rank can go as low as tenth. But after analysing the training data we found that the lowest tonic rank was fifth and hence only 5 classes are used in the experiment. Moreover, 96% of the instances are labelled with one of the top three classes (first, second, third).

At this point we have each instance annotated with a class label and a set of 20 extracted features. Next, we proceed to select the relevant features for our specific task. We use the WEKA data-mining software for all the classification related steps (Hall et al., 2009). We perform attribute selection using the *CfsSubsetEval* attribute evaluator and BestFirst search method (Hall, 1999) with a 10-fold cross validation option set. We select the features which are used in at least 80% of the folds.

Subsequently, a C4.5 (J48) decision tree is trained using WEKA to learn best set of rules to reliably identify the correct tonic candidate (Quinlan, 1993). Note that we also tried other classifiers, namely support vector machine (Sequential Minimal Optimization (SMO) with polynomial kernel) and an instance based classifier K^* (Witten, Frank, & Hall, 2011). However the accuracy obtained by the J48 decision tree was considerably higher and so for the rest of the thesis we present our results based on this classifier. Additionally, the advantage of using a decision tree is that the resulting classification rules can be easily interpreted and visualized. It is crucial to understand the classification rules, especially because in our proposed approach it is a part of the methodology and not just used for the evaluation.

We noticed that the number of instances belonging to each class in our training dataset was highly uneven. Training a classifier in this way might result into a biased learning, favoring the class which has more number of instances. To mitigate this effect due to uneven number of instances per class, we also performed experiments with instance normalization by repeating the number of instances in

minority class. We used the ‘supervised.instance.Resample’ filter in WEKA with ‘biastoUniformClass’ option set to 1 to normalize the number of instances per class (Witten et al., 2011).

3.3 Method 2

This method divides the task of tonic identification into two stages, the tonic pitch-class identification, performed using a multi-pitch analysis, similar to Method 1 and the tonic octave identification using the predominant melody information (Figure 3.1). This enables Method 2 to be applicable to both vocal and instrumental music. Recall that for the instrumental excerpts we only require the tonic pitch-class (Section 1.4 and 3.1) and therefore, we omit the second stage of octave estimation in such cases. The following paragraphs describe both these stages in detail.

3.3.1 Tonic Pitch-class Identification

The tonic pitch-class identification in Method 2 is performed in a similar way as the tonic pitch identification in Method 1. It involves two main processing steps, the multi-pitch analysis and the candidate selection. The multi-pitch analysis module used in this method is same as described for Method 1 (Section 3.2.1). Both these methods differ at candidate selection step, which for Method 2 is described below.

Candidate Selection

The difference between the candidate selection module used in this method and in Method 1 is in the class labelling strategy that is followed to train the classifier. In this method, the class labels assigned to each instance while learning the model are derived from the pitch histograms using the tonic pitch-class information. Recall that in Method 1, labelling an instance with a class also requires the tonic octave information in addition to the pitch-class, which is unavailable for instrumental excerpts. Apart from this, the rest of the procedure followed for selecting the correct tonic pitch-class candidate is same as the procedure described for tonic candidate selection in Method 1 (Section 3.2.2).

The class labels assigned to each instance in this method is the best rank of the tonic pitch-class amongst all the candidates. Note that we use the term ‘best’ to highlight that we select the highest rank of all the candidates corresponding to the tonic pitch-class and since we considered a frequency range of more than one octave, we may have multiple peaks, representing the same pitch class but at different octaves. This task is also theoretically a 10 class classification problem. However, as we have relieved a constraint (peak as tonic pitch-class not an exact pitch) there are greater number of instances (98.7%) labelled with one of the top three classes (first, second, third) as compared to the Method 1.

Since in this method we use only the tonic pitch-class information to train the classifier, it can be applied to instrumental music as well (we only have the tonic pitch-class annotations for instrumental music).

3.3.2 Tonic Octave Estimation

The octave in which the tonic of the singer lies is an important information crucial for many melodic analyses, as seen earlier (Section 1.4). This section describes the process of estimating the tonic octave (i.e. the second stage of Method 2) using the tonic pitch-class and extracted predominant melody contour.

The pitch range for the majority of singers lies within three octaves, where the tonic chosen by them is the middle register Sa. The tonic is thus the lowest Sa svar sung by the vocalist (with an exception of the madhyam-śruti case, which is rarely witnessed (Section 4.2)). This motivates us to analyze the predominant melody contour in order to automatically estimate the tonic octave.

The process of estimating the tonic octave is divided into three steps (see Figure 3.2), namely, predominant melody extraction, melody histogram computation and finally octave estimation using the constructed histogram.

For the predominant melody extraction we use the algorithm proposed by Salamon and Gómez (Salamon & Gómez, 2012), who kindly provided us with an implementation. Their system is specifically designed to extract the pitch contour of the dominant melodic source (lead performer in our case) in a situation where multiple pitched components exist simultaneously in the audio signal. The key ideas behind the system are; using multi-pitch analysis to handle polyphonic mu-

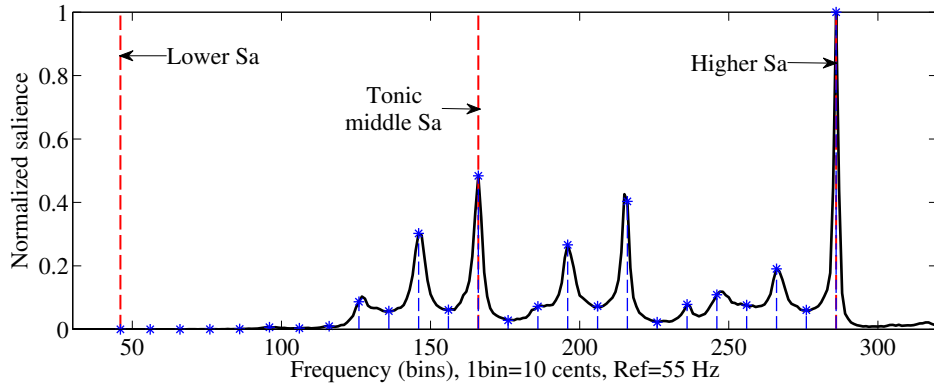


Figure 3.7: An example of the predominant melody histogram extracted from a song in our database. The red lines mark the tonic pitch-class locations

sic and exploiting the characteristics of the melodic contours to filter out erroneous pitch contours. We use this system with the default parameter values, which we found works quite well for our analysis. A Vamp plugin to use this method in the SonicVisualizer ¹ can be obtained from one of the authors website ².

The extracted pitch contour is used to construct the melody histogram. The histogram summarizes all the pitches used in the melody indicating their frequency of usage. It thus becomes an ideal representation to be used for our analysis. Before computing the histogram the pitch values are converted into a cent scale and quantized into 600 bins with a resolution of 10 cents per bin (Equation 3.4). An example of a melody histogram is shown in Figure 3.7. The red lines mark the pitch values (in bins) corresponding to tonic pitch-class (Sa) in different octaves. As can be seen, the tonic pitch corresponds to bin 166 which is the lowest Sa that has non-zero salience in the histogram. We propose two different approaches to select the tonic octave from the melody histogram; a rule-based approach (RB) and classification-based approach (CB).

Rule-based Approach

In this approach the tonic octave is estimated by applying a simple rule on the melody histogram. Ideally, the tonic pitch is the lowest tonic pitch-class used in

¹<http://www.sonicvisualiser.org>

²<http://www.justinsalomon.com/melody-extraction.html>

the melody. Therefore, it would be sufficient to select the lowest tonic pitch-class in the melody histogram, which has a non-zero value. An example of the melody histogram is shown in Figure 3.7, where all the bins corresponding to the tonic pitch-class in different octaves are marked by the red lines. It can be seen from the figure that the middle tonic pitch-class (middle Sa) is the lowest tonic pitch-class which has a non-zero value. However, we observed in some rare cases that the melody extraction algorithm makes octave errors and estimates pitches which are sub-multiples of the true pitch values. This results into a non-zero value in melody histogram at a sub-multiple of the bin corresponding to the tonic pitch, which eventually leads to an error. A solution to this would be to take ratios of histogram values at tonic pitch-class locations in adjacent octaves. As the octave errors are very rare, this ratio would still be maximum at the tonic octave. We calculate the ratio $R(i)$ at every bin corresponding to the tonic pitch-class in different octaves ($i = 1, 2, \dots, N$) as shown below:

$$R(i) = \frac{h(j_i)}{h(j_{i-1}) + \epsilon}, \quad (3.7)$$

$$j_i = \text{mod}(\eta, 120) + 120 \cdot (i - 1), i = 1, 2, 3, 4, 5$$

where i is the octave index, h is the histogram value, j_i is the bin index of the tonic pitch-class in the octave i , η is the bin index of the tonic pitch-class (input given by previous stage), ϵ is a very small number (minimum floating point value) to avoid division by zero.

The correct tonic octave is given by the index $i = I$ at which the Ratio $R(i)$ is maximum.

$$I = \arg \max_i R(i) \quad (3.8)$$

Classification Based Approach

We noticed certain cases where the rule-based method is bound to fail. An example of such a case is the madhyam-śruti songs in which the singer may not sing the tonic pitch at all (explained in Section 4.2). In such cases the natural fourth (Ma) with respect to the tonic pitch is considered as the Sa svar of the rāg. Therefore, analysing melody histograms at only the tonic pitch class locations won't be sufficient to estimate the tonic octave. Another problem is that many times low

frequency pitches are not tracked by the melody extraction algorithm. In such situations salience of the lowest tonic pitch-class used in the melody would be low in the melody histogram, which leads to an error. A detailed discussion on the challenges in the rule-based approach and analysis of erroneous cases is presented in the discussion section 4.2.

To handle the specific cases mentioned above, we adapt a classification based approach. The key idea here is not to rely on only the tonic pitch-class locations in the melody histogram but to parametrize the whole histogram and model the lowest octave of the sung melody. The system would automatically learn the best set of rules and pitch classes in the melody histogram which are crucial for identifying the tonic octave.

For every tonic pitch-class in different octaves we extract a set of 25 features. These features are the values of melody histogram at 25 equidistant locations spanning two octaves, centered around itself. Basically, these are sampled values of melody histogram, 12 for each semitone below the tonic pitch class and 12 for each semitone above the pitch class and one itself. This gives us a set of 25 features $h_i(i = 1 \dots 25)$. An example is shown in Figure 3.7 for a tonic pitch-class at bin number 166 which actually corresponds to correct tonic octave. The sampled histogram at 25 equidistant locations centered around 166th bin is marked by blue stars. Next, we assign a class label to each instance in our dataset, which are essentially all the possible tonic pitch-classes in different octaves for all the histograms. We assign a class ‘TonicOctave’ if the instance (tonic pitch-class) is in the tonic octave, else ‘NonTonicOctave’. The ground-truth tonic annotations are used for labelling the classes. We then train a classifier using the 25 extracted features to learn the best set of rules in order to predict the class label for every tonic pitch-class. By predicting the class (‘TonicOctave’ or ‘NonTonicOctave’) of every possible tonic pitch-class in different octaves, we can identify the correct tonic octave. We are also interested in knowing what features (essentially the histogram samples at 25 equidistant points) are useful for this classification task, as these features will reflect roughly which svaras in lower register are crucial factors in deciding the tonic octave.

We use the WEKA data-mining software for this classification task too. We perform the attribute selection in the same way as did before, using the *Cfs*-

SubsetEval attribute evaluator and BestFirst search method with a 10-fold cross validation option set (Hall, 1999; Witten et al., 2011). We select the features which are used in at least 80% of the folds. Subsequently, a C4.5 (J48) decision tree is trained using WEKA to learn the best set of rules to predict the class labels.

Note that for computing the melody histogram we have used the whole audio file. This is justified, because if our aim is to find out the lowest tonic pitch-class used by the singer in the melody, we need to listen to all of it. Otherwise, we would have to incorporate the knowledge regarding the tonic pitch range for male and female singers. We know that the tonic pitch range for male and female singers is typically between 110-260 Hz (Section 1.2). Therefore, for the pitch values which correspond to the pitch-classes between 130-220 Hz, there exists only one possibility of the tonic pitch. In a such situation we do not even need to apply any algorithm to estimate the tonic octave. Note that this holds true for majority of the data in our database. We also conduct experiments to see the effect of including the information regarding the possible tonic pitch range in the system. Furthermore, if the gender of the singer is known, there is only one possibility of the tonic pitch given the tonic pitch-class, as the individual tonic pitch ranges for both male and female singers are contained within one octave. Therefore if the gender of the singer is known this stage of tonic octave estimation can be omitted.

3.4 Tonic identification system

This section presents an overview of the proposed practical system for tonic identification which aims at recursively utilizing all the available data (audio and relevant metadata) and obtaining results with maximum confidence. The motivations behind such a system are:

1. Prevalent methodologies in MIR primarily focus on using only a single type of data source (Barrington, Turnbull, & Yazdani, 2009). Most of the approaches either use the available audio data, music scores or the contextual metadata to accomplish certain tasks. Recent efforts towards semantic music discovery combine audio content analysis with social contextual data and metadata (Barrington et al., 2009). However, there should be more attempts

specifically in the area of automatic music description to explore the potential of combining the complementary type of data, to achieve practical solutions with better accuracies.

2. The concept of a confidence measure is rarely seen in the existing systems. This issue particularly becomes important in situations where a method is used as a building block in another system. In such situations, we might want to compromise the overall accuracy of the method in exchange for a high confidence value, to avoid error propagation. One might argue that the overall accuracy of a method reflects its statistical confidence value, but at the same time we should consider that the method could have been developed for achieving an overall high accuracy, rather than obtaining results with a high reliability. Moreover the concept of confidence measure can allow us to iteratively utilize the available data, as will be described while explaining the proposed system.

Motivated by the aforementioned ideas, the proposed system combines the audio data and the available metadata for the identification of the tonic. Based on the derived confidence measure, the system tries to combine these two data sources to maximise the accuracy in an iterative manner. Figure 3.8 shows the block diagram of the complete iterative system.

As we notice in the figure, all the available data is fed to a data selection module, which decides what fraction of the data and which type of data is to be supplied to the automatic tonic identification module in each iteration. The data selection module has a predefined preferential order of the data to be fed into the system. The order is such that the audio data is utilized fully before using the metadata (as for Indian art music metadata in an organised and machine readable form is harder to obtain than the audio). The system can be started with a fraction of a minute of the audio data (the duration which is enough for a human listener to identify tonic). Based on the derived confidence measure more and more audio data would be pumped into the system. The iterative process will be terminated when the confidence reaches a threshold for it to be safely considered as 100% accurate estimation. In case we couldn't reach the desired confidence value even after utilizing the full audio data, metadata regarding the rāg, artist, gender of

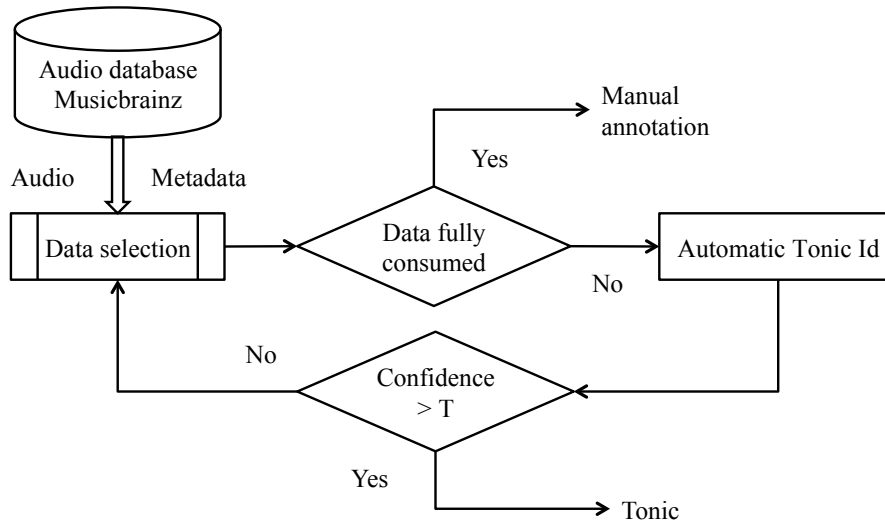


Figure 3.8: Block diagram of the iterative tonic identification system; both the audio and the metadata is input to the system. The system returns a tonic frequency along with the associated confidence value

the singer will be fed to the system; such that using the minimum amount of extra information we achieve the desired confidence value while maximising the accuracy at the same time.

3.5 Evaluation Methodology

3.5.1 Database

The music collection used to evaluate the proposed approaches is a subset of the musical material compiled as part of CompMusic project (Serra, 2011). The core database used in this work is comprised of 352 full length audio songs, containing both vocal (237) and instrumental (115) musical pieces. However, for the evaluation of specific methods/approaches at various stages, multiple short excerpts are extracted from the full audio songs. In addition to the audio data we also possess the relevant metadata corresponding to each song, uploaded to Musicbrainz³. Ev-

³<http://musicbrainz.org/>

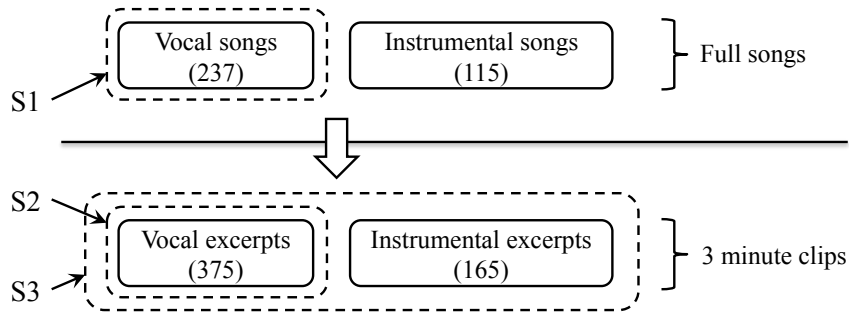


Figure 3.9: Visualization of the three datasets S1, S2 and S3 used in the evaluation of proposed approaches.

Dataset	Size	Len.	Hi.(%)	Ca.(%)	M(%)	F(%)	#U song	#U artists
S1	237	full	27.4	72.6	75.2	24.8	237	34
S2	375	3 min	37	63	77.8	22.2	237	34
S3	540	3 min	36	64	NA	NA	352	54

Table 3.1: Database description; summary in terms of different constituting components; Hindustani (Hi), Carnatic (Ca), male (M), female (F), number of unique songs (U song), number of unique artists (U artists)

ery song in our database is a part of the Hindustani and Carnatic music collections in Musicbrainz.

Typically the length of the audio songs in Indian art music can vary from 3-4 minutes to more than 1 hour. The characteristics of the musical content (both in terms of musical concepts and acoustical characteristics) may also vary a lot, from slow and unmetred *ālāp* in one song to fast *tān* in another. Therefore, for the evaluation of the methods in which only a few minutes of the audio data is used (to show that only a small amount of data is sufficient to perform the task), it is very important that multiple excerpts are extracted from different sections of the song, to preserve the diversity. In the current work, the short excerpts which are used for the evaluation are 3 minutes long and for the songs more than 12 minutes in length, 3 such excerpts are extracted from the start, middle and the end of the song. Otherwise, only one excerpt per song is extracted from the beginning.

We use 3 different datasets derived from our core-database. Figure 3.9 shows these different sets (S1, S2 and S3) indicating the constituent musical material (vocal or instrumental) and the amount of mutual overlap in them. To better

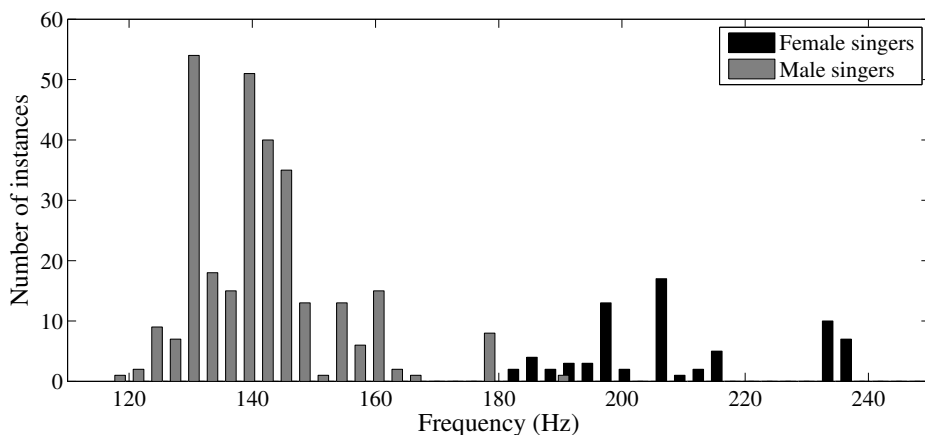


Figure 3.10: Distribution of tonic frequency for male and female vocal performances in our music collection.

understand the datasets, general statistics are provided in Table 3.1. In this table Hi. stands for the number of songs belonging to Hindustani music (in percentage), Ca. is for Carnatic music, M and F denotes the percentage of the songs sung by male and female singers respectively, #U_{song} and #U_{artist} is the number of unique songs and artists present in the database. In Figure 3.10 we display the distribution of tonic frequencies in our music collection for both male and female singers (only for vocal excerpts). This reflect another dimension of the diversity present in the database.

3.5.2 Annotations

The tonic pitch (pitch-class for instrumental music) for each excerpt was manually annotated by the author, which was later verified and corrected by a professional Carnatic musician (the number of discrepancies was very small). To assist the tiresome process of annotation, we used the candidate generation part of the proposed approach. Using the multi-pitch histogram we extracted the top 10 candidate frequencies for the tonic in the range of 110 to 300 Hz. Notice that the frequency range of the tonic is kept quite lenient for doing annotations. To further accelerate the process, we also designed a simple MATLAB[®] GUI. A screenshot of the GUI can be seen in Figure 3.11.

Using this GUI the user can load at-once the list of files which are to be anno-

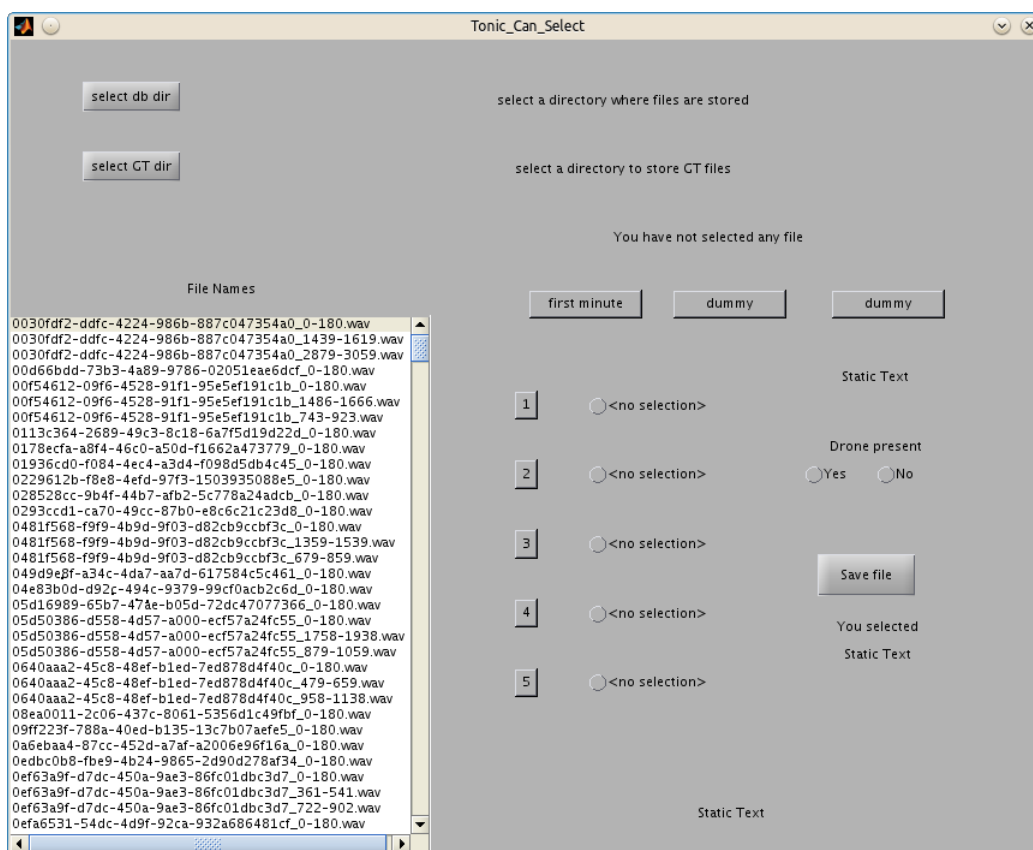


Figure 3.11: Screenshot of the MATLAB GUI used for tonic annotations

tated. Selecting a file loads all the 10 tonic candidate frequencies along with the corresponding audio. The annotator can then listen to the candidate frequencies (tones) one by one together with the original audio in order to identify the tonic frequency. This process was followed for the tonic annotations of all the songs in our database.

3.5.3 Performance Evaluation

We introduced two methods, Method 1 (M1) and Method 2 (M2) for the tonic identification task. M1 identifies the tonic pitch and is applicable to vocal music pieces whereas M2 identifies the tonic pitch-class and caters to both the vocal and instrumental excerpts. M2 also include another stage of tonic octave estimation which is only required for the vocal excerpts and not for the instrumental pieces.

The evaluation for vocal excerpts is done in terms of the percentage of the excerpts for which the tonic pitch is correctly identified, whereas for the instrumental music it is the tonic pitch-class. An output is considered as correct if it is within a bracket of 25 cents from the ground-truth value.

For a detailed performance evaluation and deeper analysis of the approaches, we evaluate both the methods at individual stages. M1 is evaluated for the tonic pitch identification accuracy on the dataset S2, which contains only vocal excerpts. The first stage of M2 is evaluated for the tonic pitch-class identification task on the dataset S3 containing both vocal and instrumental excerpts. The second stage of M2, for the tonic octave estimation is evaluated on dataset S1 containing full vocal recordings. For the evaluation of the tonic octave estimation, we calculate the accuracy in terms of the percentage of the excerpts for which tonic octave is correctly identified. The annotated tonic frequencies by the author (section 3.5.2) served as the ground-truth for calculating all the accuracies.

Chapter 4

Results and Discussion

This chapter presents the results obtained from the performance evaluation of both the proposed methods. We also present an analysis of the accuracies obtained for different tasks and provide plausible explanations for the erroneous cases wherever possible.

4.1 Results

In this section we present the evaluation results for our proposed approach for tonic identification. Recall that the evaluation strategy used for calculating performance accuracies was detailed in Section 3.5.3.

The performance accuracies for the tonic pitch identification task using Method 1 (M1) on the dataset S2 for both with and without normalization are summarized in Table 4.1 and 4.2. These tables show the performance of M1 on the whole dataset ('full'), as well as the obtained accuracies as a function of different attributes such as Hindustani and Carnatic music, for male and female singers. Note that even when the results are shown for a particular attribute (such as Hindustani music), the classifier is trained using the whole database. It also shows a breakdown of the total errors made by the system in terms of different types of errors. The erroneous cases can be classified into four categories; first, octave errors (Oct.Err) is when the tonic pitch-class is correctly identified but not the tonic-octave. Second, the 'Pa' or fifth type errors (Pa Err.) is when instead of identifying the tonic pitch the

Filter	Accuracy(%)	Oct. Err.(%)	Pa Err.(%)	Ma. Err(%)	Others
Full	84.27	2.13	4	4	5.6
Hind.	88.49	2.16	4.32	1.44	13.6
Carn.	81.8	2.12	3.81	5.51	6.8
Male	87.67	1.37	2.73	2.39	15.82
Female	72.28	4.82	8.43	9.64	4.82

Table 4.1: Performance accuracy of M1 on dataset S2 with instance normalization. Results shown for the full dataset and also a breakup of the performance is presented for the songs belonging to Hindustani (Hind.) & Carnatic (Carn.) music and to male & female singers categories.

Filter	Accuracy(%)	Oct. Err.(%)	Pa Err.(%)	Ma. Err(%)	Others
Full	93.05	1.3	2.7	1.87	1.1
Hind.	96.4	1.44	0	0.72	1.44
Carn.	91.1	1.27	4.24	2.54	0.85
Male	93.83	1.37	3.08	0.68	1.03
Female	90.36	1.20	1.2	6.02	1.20

Table 4.2: Performance accuracy of M1 on dataset S2 without instance normalization. Results shown for the full dataset and also a breakup of the performance is presented for the songs belonging to Hindustani (Hind.) & Carnatic (Carn.) music and to male & female singers categories.

system picks its fifth (Pa) either on the higher or the lower octave. Similar to this is the third category, the ‘Ma’ or fourth type errors (Ma Err.) when the selected pitch is the fourth of the tonic pitch in any octave. All other kinds of errors belong to the ‘Others’ category.

The performance accuracies for tonic pitch-class identification task using the first stage of Method 2 on database S3 for both with and without instance normalization are summarized in Table 4.3 and 4.4. These tables also show a break up of total errors made by the system in terms of different types of errors, as explained above.

The results for the tonic octave estimation using the second stage of Method 2 for both the approaches (rule-based and classification-based) are shown in Table 4.5 and 4.6. The evaluation is done both with and without imposing a constraint on the tonic pitch range. In the former case, the allowed frequency range for the

Filter	Accuracy(%)	Pa Err.(%)	Ma. Err(%)	Others
Full	76.67	10.37	6.29	6.67
Vocal	76.53	10.933	5.6	6.93
Inst.	76.96	9.09	7.88	6.06
Hind.	84.69	6.63	2.55	6.12
Carn.	72.1	12.5	8.43	6.97

Table 4.3: Performance accuracy of M2 (tonic pitch-class identification) on dataset S3 with instance normalization. Results shown for the full dataset and also a breakup of the performance is presented for the songs belonging to Hindustani (Hind.) & Carnatic (Carn.) music and to vocal & instrumental (Inst.) music .

Filter	Accuracy(%)	Pa Err.(%)	Ma. Err(%)	Others
Full	92.96	2.59	2.96	1.48
Vocal	94.13	2.67	1.87	1.33
Inst.	90.3	2.42	5.45	1.82
Hind.	94.39	1.53	2.04	2
Carn.	92.15	3.2	3.49	1.16

Table 4.4: Performance accuracy of M2 (tonic pitch-class identification) on dataset S3 without instance normalization. Results shown for the full dataset and also a breakup of the performance is presented for the songs belonging to Hindustani (Hind.) & Carnatic (Carn.) music and to vocal & instrumental (Inst.) music.

Filter	Accuracy (no tonic-limit)(%)	Accuracy (tonic-limit)(%)
Full	89.5	96.2
Male	89.32	95.5
Female	89.83	98.3
Hind.	96.92	98.46
Carn.	86.62	95.35

Table 4.5: Performance accuracy of M2 (tonic octave estimation) on dataset S1 for rule-based approach. Results shown for both the cases; without imposing any limit on allowed tonic pitch range and constraining it to a limit of 110-260 Hz

Filter	Accuracy (no tonic-limit)(%)	Accuracy (tonic-limit)(%)
Full	96.62	98.73
Male	98.88	100
Female	89.83	94.91
Hind.	92.31	95.38
Carn.	98.26	100

Table 4.6: Performance accuracy of M2 (tonic octave estimation) on dataset S1 for classification based approach. Results shown for both the cases; without imposing any limit on allowed tonic pitch range and constraining it to a limit of 110-260 Hz

tonic pitch was restricted to 110-260 Hz. Note that the results shown are only for the tonic octave estimation stage, evaluated individually using the ground-truth tonic pitch-class information.

4.2 Discussion

In this section we analyse the obtained results, comment on performance of the proposed methods while scrutinizing the erroneous cases and providing plausible explanations wherever possible. The discussion is organized on the basis of the evaluated tasks, namely, tonic pitch identification, tonic pitch-class identification and tonic octaves estimation.

Tonic pitch identification

We see in Table 4.2 that Method 1 obtains a good accuracy of 93.05% for the tonic pitch identification task on the complete dataset S2 (without instance normalization). More importantly, since the allowed frequency range for tonic pitch was more than one octave (110-370 Hz), it means that the system is able to correctly identify not just the tonic pitch-class but also the corresponding octave. This is already a significant advancement, as past approaches only targeted the identification of the tonic pitch-class.

The method is evaluated for both cases of with and without performing instance normalization while training the classifier. As can be seen, normalizing the instances per class results in an inferior performance compared to leaving the

number of instances per class as in the original database. This can be attributed to the fact that some classes contain a very small number of instances (in absolute terms) and the normalization is performed by repeating the instances of the minority class. After instance normalization the classifier predicts the minority classes with better accuracy, but at the same time the prediction accuracy for the majority classes drops down by a small amount. The increased accuracy for predicting the minority classes does not improve the overall accuracy because a slight decrease in prediction accuracy of the majority classes causes a greater drop in the performance. It appears that if our goal is to achieve maximum overall accuracy, it is better to ignore the specific rare cases than try to learn them. Because in an attempt to learn the rules for specific rare cases, the system starts having more confusions in the prediction of majority classes. In the remaining part of the discussion we present all the analysis for the case of without instance normalization.

Analysing the obtained results as a function of musical style (Hindustani or Carnatic) and gender of the singer (male or female) revealed interesting insights. We observe that the performance for the excerpts belonging to Hindustani music (96.4%) is better compared to the performance obtained for Carnatic music (91.1%) (Table 4.2). Examining the data on a broad level, we noticed that in many Carnatic excerpts the loudness level of the drone sound in relation to the lead performer was quite low. Consequently, this results in frames where all the prominent peaks in the salience function correspond to the lead voice (note that the salience function contains many more peaks than the true F0 due to the harmonic summation). This means that the peaks in the pitch histogram corresponding to the drone sound have quite low magnitude, resulting in a low tonic rank and eventually leading to incorrect tonic identification.

Considering the performance accuracy as a function of gender of the singer (Table 4.2), we observe that the system works better for the excerpts performed by male singers (93.83%) as compared to those sung by female singers (90.36%). A plausible reason for this could be the uneven amount of male and female performances in our dataset. Since the dataset is substantially populated by male singer performances (77.8%, Table 3.1), the classification rules are better learned for these excerpts. Also, we notice that the frequency range chosen for the con-

struction of the pitch histograms is well tuned for the tonic pitch range of male singers. The frequency range for the computation of pitch histograms was selected based on the overall high accuracy and therefore, for the same reasons (dominance of male performances), the selected frequency range appears to be biased towards male singers. An analysis of how the chosen frequency range during multi-pitch analysis affects the accuracy of the system, particularly the performances of female singers is provided in subsequent paragraphs.

Further analysing the results, we examined the type of errors commonly made by the system. The most frequent error types were selecting the fifth (Pa) or the fourth (Ma) as the tonic or identifying the tonic in another octave. These type of errors are understandable, as Pa or Ma is the secondary pitch-class that is often produced by the drone instrument in addition to the tonic. Moreover, for the male singers the errors were selecting the higher Pa or Ma as tonic, whilst for female singers it was selecting the lower Pa or Ma. This can be attributed jointly to the differences in typical tonic frequencies for male and female singers, together with the frequency range chosen for constructing the pitch histograms. Errors apart from these three types, are quite rare. This error analysis together with a close examination of the pitch histograms suggests that some of these errors could be avoided if the secondary pitch-class produced by the drone instrument is known. In these cases a fundamental cause for an error is the confusion between the Ma and Pa tuning cases, which arises due to similar pattern of the histogram peaks. As a result, the system applies rules which are learned for handling Pa tuning cases onto Ma tuning cases and vice versa, leading to error.

Thus, if the tuning configuration of the drone instrument is known, then some of these errors arising because of the Ma-Pa confusion can be avoided. As a matter of fact, the tuning configuration of the drone instrument is related to the rāg, and so, incorporating rāg information could lead to an improvement. This work is left for future investigation.

The resulting decision tree exhibits musically meaningful relationships. An example of a decision tree obtained for Method 1 is shown in Figure 4.1. We see that the pitch intervals used by the tree to make the decisions correspond very well to the typical intervals between the prominent pitches in the drone sound. The distance between tonic Sa to lower Pa or tonic Sa to higher Ma is rounded

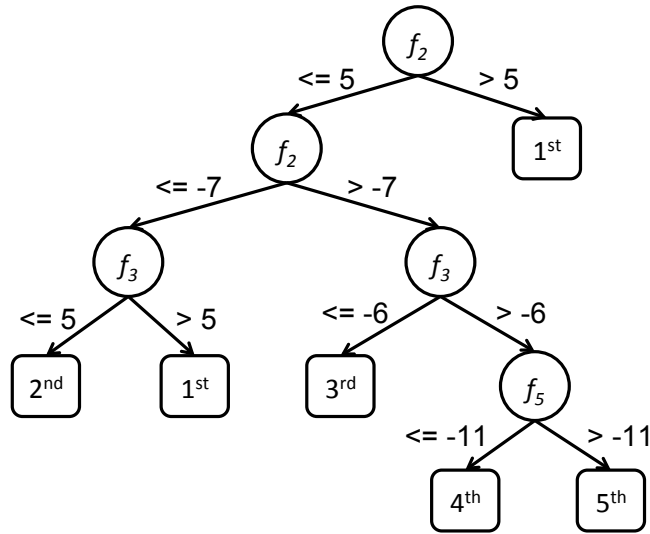


Figure 4.1: An example of the decision tree obtained for Method 1

off to 5 semitones and between tonic Sa to lower Ma which is same as tonic Sa or higher Pa is 7 semitones. The branch of the tree where the distances are 6 and 11 semitone contains very few instances.

Examining the rules of the same decision tree (Figure 4.1) in depth we observe that one of the most useful pieces information is the relationship between the top two peaks of the histogram (f_2). Whenever the second highest peak is more than 5 semitones apart, the highest peak is chosen as the tonic. This condition always corresponds to one of the two cases; either the second peak is Pa (in a Pa tuning) leading to $f_2 = 7$ or it is the higher Sa resulting in $f_2 = 12$. In both these conditions the first peak corresponds to the tonic Sa and is correctly identified every time. The case of $f_2 = \mp 12$ is quite a common scenario, where both the peaks correspond to Sa in different octaves. Branching left, the tree checks whether the highest peak corresponds to Pa (7 semitone above the tonic, $f_2 = -7$). To validate this hypothesis it checks the relationship with the third peak. If third peak is found at 5 semitone above the highest peak (thus corresponding to Sa one octave above the tonic), the system confirms that first peak is Pa and correctly selects the second peak which corresponds to the tonic pitch. Otherwise the hypothesis is rejected, and the pattern corresponds to a case of Ma tuning where the highest peak is Ma. Similar interpretations can be made for all the rules.

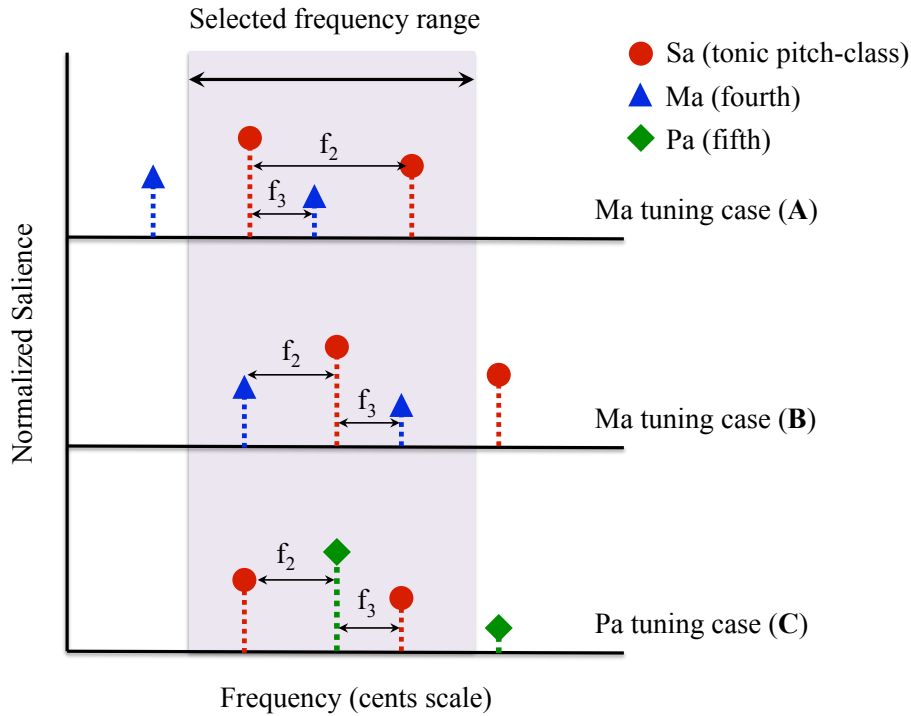


Figure 4.2: Demonstration of the effect of frequency range selected for the multi-pitch analysis, showing a specific scenario where Ma tuning cases are confused with Pa tuning cases, leading to errors.

As a final step in our analysis of the results obtained for the tonic pitch identification task, we investigate an interesting observation. In Table 4.2 we notice that for the songs performed by female singers, the percentage of Ma type errors is considerably high (6.02%) compared to the rest. Examining this specific scenario exposed one of the effects of the frequency range selected for constructing pitch histograms on the system's performance. A highly common pattern observed in the peaks of the pitch histograms is shown in Figure 4.2 (pattern-A). The third highest peak can either be Ma or Pa, depending upon the tuning of the drone instrument. We consider the case of Ma tuning for this case-study. For this pattern (A), the system easily identifies the true tonic pitch as the highest peak in the histogram. This can be easily deduced from the decision-tree shown in Figure 4.1, as $f_2 = 12$. A plausible reason for this pattern being very common is the fact that the tonic pitch-class is reinforced much more than the secondary pitch-class produced

by the drone instrument. For example, in *tānpūrā*, three strings are producing the Sa, whereas only one string produces secondary pitch-class (Pa, Ma or Nī). So, the magnitude of the peaks corresponding to the Sa pitch-class is typically higher.

The tonic frequencies chosen by female singers are considerably higher than those by male singers (Figure 3.10). So, for the case of a female singer, the pitch-histogram with the same pattern (A) of the peaks, as seen above, would look like its transposed version as shown in Figure 4.2 (B). But, because the frequency range which is used to construct the pitch histogram remains the same for both male and female singers (110-370 Hz), the peak corresponding to the higher Sa is now outside this range and will not be considered in the analysis. So effectively, in such a case, the section of the pitch histogram which is within the frequency range exhibits another type of pattern (B). This pattern (B) now resembles another pattern (C) shown in the same figure, which is a typical pattern found in the excerpts where the Pa tuning is used by the drone instrument and are sung by male singers. The two patterns (B and C) look very similar but notice the relative location of the tonic pitch with respect to the highest peak is different in both the patterns. This implies that the system will apply the same set of rules to identify the tonic pitch in both situations (as the patterns are identical) and would eventually select a wrong tonic pitch in at least one of the cases. Because the excerpts with Pa tuning of the drone instrument and sung by male singers are more common in the database compared to excerpts with Ma tuning and sung by female singers, the learned rules are such that they (rules) favor the former case of Pa tuning. As a result, we select the lower Ma as the tonic pitch (in pattern B), applying the same logic used when we identify the correct tonic pitch when the highest peak is Pa (in pattern A). As we see here, the chosen frequency range for computing the pitch histograms plays a crucial role and affects the performance of the system. We realize that the chosen frequency range in our implementation is more suited for the performances of male singers. This is understandable because the frequency range is selected on the basis of the overall accuracy of the system and the male performances in our database are significantly higher in numbers.

Tonic pitch-class identification

The results obtained for the tonic pitch-class identification task using the first stage of Method 2 for both with and without instance normalization are shown in Table 4.3 and 4.4. The method performs well, achieving an accuracy of 92.96% on dataset S3, without instance normalization. We see that this method works well for not only the vocal excerpts but also for the instrumental excerpts .

When instance normalization is performed the obtained results are inferior compared to the case when number of instances are not normalized. A possible explanation for this is already discussed in the previous section 4.2. In the remaining part of the discussion we present all the analysis for the case when the number of instances are not normalized.

We also analyse the performance accuracy as a function of different attributes such as for vocal, instrumental, Hindustani and Carnatic excerpts, similar to the analysis performed in the previous section. Table 4.4 shows the obtained accuracy for the whole database (92.96%), vocal excerpts (94.13%), instrumental pieces (90.3%), excerpts belonging to Hindustani music (94.39%) and Carnatic music (92.15%). We notice that the performance on the vocal excerpts is better compared to the instrument excerpts. A plausible reason for this difference in performance could be the presence of drone instrument as an accompaniment. For vocal music, there is always a drone instrument accompanying the lead performer, whereas for the instrumental songs a dedicated drone instrument might be absent in some cases. In many Indian instruments such as sitār, vīṇā and sārangi the sympathetic strings of the instrument reinforce the tonic pitch and other important pitch classes. Therefore in some instrumental performances an external drone instrument is not used. However, the loudness level of the sound produced by the sympathetic strings in relation to the sound produced by the main strings is considerably low in many cases, leading to incorrect tonic identification.

Examining the type of errors made by the system, we noticed that the commonly made errors were of the type Ma or Pa. Moreover, we found that the reasons behind these errors are similar to the ones which caused errors in tonic pitch estimation task in Method 1, as explained earlier. We see that both Method 1 and Method 2 make identical errors, which is comprehensible as they share the core

methodology.

Tonic octave estimation

Table 4.5 and 4.6 show the results obtained for the tonic octave estimation task using rule-based and classification based approach of the second stage of Method 2 on dataset S1. The rule-based approach achieves an accuracy of 89.5%, whereas the classification based approach yields 96.62% accuracy (without applying tonic-range constraint). As we can see, the difference in the performance accuracy is quite considerable.

Examining the erroneous cases for the rule-based approach, we found multiple scenarios where this approach would fall short of estimating the correct tonic octave. However, these situations are not very common and occur only in few songs, as is reflected from the obtained overall accuracy of 89.5%, which is reasonably good. The two main reasons for the rule-based approach to select the wrong tonic octave are as follow:

- In few songs the tonic pitch of the singer does not correspond to the Sa svar of the rāg. For these songs the root svar or Sa of the rāg in the melody corresponds to the higher fourth of the tonic pitch (Ma). This concept is termed as madhyam-śruti. It appears in the performances of the rāgs where the pitch range to render *āroh* and *avroh* (ascending and descending) is truncated. As the pitch range used in the exposition of these rāgs is low as compared to the other rāgs, the singer transposes the Sa of the rāg to fourth of the tonic instead of choosing Sa to be the tonic pitch.

As we can imagine, in these situations our hypothesis that the lowest Sa sung by the singer corresponds to the tonic pitch, i.e. in the correct octave does not holds true. Consequently, applying the rule-based approach leads to an octave error.

- For some excerpts sung by the male singers we noticed that the lower frequencies are not well tracked by the melody extraction algorithm. This may be due to the fact that we did not change any parameter of the melody extraction algorithm. The parameters of this algorithm are tuned for the

melodies in Western music, which generally do not span very low frequency regions (100-140 Hz) as the melodies sung by the male singers in Indian art music do. As a result, the salience of the lowest Sa in the melody histogram is either null or extremely low, leading to an octave error if we use the rule-based approach.

Figure 4.3 shows section of three melody histograms computed from the excerpts belonging to cases mentioned above. The figure shows two octaves of each histogram, centered around the tonic pitch. The case (a) is corresponding to a song with madhyam-śruti, (b) a female vocalist of Hindustani music singing as low as 600 cents below the tonic and (c) a male Carnatic singer with a tonic of 129 Hz, a case where F0-estimation algorithm did not track low pitch values in the melody. The histograms in this figure give an idea about the limitations of the rule-based approach in handling these specific cases.

As seen earlier, the classification based approach performs better with an accuracy of 96.62% as compared to the rule-based method. This is because the later is based on a simple rule, which only considers the value of the melody histogram at the locations of tonic pitch-class, whereas the classification based approach uses the sampled form of the whole melody histogram capturing much more information. The classification based approach basically models the section of the melody histogram which corresponds to the lower register of the singer.

We also evaluate both the approaches for tonic octave estimation incorporating the knowledge of possible frequency range of tonic pitch. We apply a constraint that the tonic pitch can only lie between 110-260 Hz. This considerably improves the performance of the rule-based approach which now achieves an accuracy of 96.2%. The accuracy for the classification based approach also increases to 98.73% but not as significantly as for the former case. We observe that the performance of the classification based approach does not depend a lot on the selected frequency range. Examining the errors for the classification based approach, we found that some of the erroneous cases are the songs performed by the female singers in which the artist sings up to 600 cents below the tonic (Figure 4.3, (b)). Singing that low from the tonic pitch will make the lower octave of the melody histograms resemble the lower octave of the histograms obtained from the madhyam-śruti cases, with

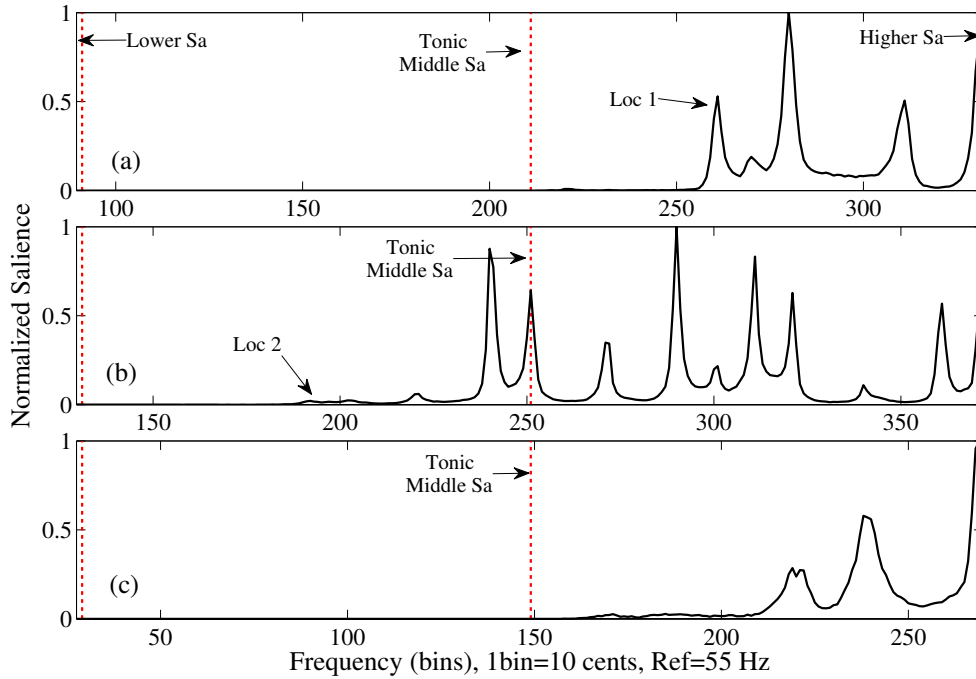


Figure 4.3: Melody histogram of three excerpts, centered around the tonic pitch. The case (a) is corresponding to a song with madhyam- \acute{s} rutu, (b) a female vocalist of Hindustani music singing as low as 600 cents below the tonic and (c) a male Carnatic singer with a tonic of 129 Hz, a case where F0-estimation algorithm did not track low pitch values in the melody.

respect to the tonic pitch-class (Figure 4.3, (a)). As a result, the classifier treats it like a madhyam- \acute{s} rutu case and selects the octave which is one below the true octave. This is because in the madhayama- \acute{s} rutu cases the lower octave of the melody histogram lies above the tonic pitch (See Figure 4.3).

Another interesting observation is that the rule-based approach performs equally good for the performances of male and female singers, and better for Hindustani music compared to Carnatic music. However, the classification based approach performs better for the performances of male singers compared to female singers and for Carnatic music than the Hindustani music. This can be attributed to the predominance of male singers and Carnatic music cases in our database (Table 3.1).

Chapter 5

Conclusions and Future work

In the last chapter of this dissertation we provide an overview of the work carried out and summarize the achievements. We also discuss issues which are still unresolved and need further investigation, as well as present some suggestions for future work.

5.1 Summary of the work

We started with an introduction of Indian art music and briefly presented an overview of this music tradition, highlighting the important musical concepts that define Indian art music. Later, we described the characteristics of this music material to get an idea about the kind of music under investigation. Subsequently, we explained the meaning of tonic in the context of Indian art music and underlined the importance of this concept. We also presented the definition of key and tonality, two concepts that relate to tonic in Western music in order to stress on the differences in the contexts in which the term tonic is used in both these music traditions. Finally, we set forward clear goals for this research work and our motivation for the computational analysis of Indian art music for the task of automatic tonic identification.

We have reviewed main audio features, techniques and works relevant to the computational analysis of the tonal aspect of music. We discussed state-of-the-art approaches for F0-estimation, multi-F0 estimation, predominant melody extrac-

tion and for the computation of pitch-class distributions. To have a little background on the methodologies adapted for tonal analyses in Western music, we briefly presented commonly used techniques for the key estimation task. A review of the relevant work done for tonic identification in Indian art music was presented, which revealed that there is a wide scope for improvement in these approaches. Later, we described the tonal structure of the *tānpūrā*, explaining different types of tunings, its spectral characteristics and the reason for its unique sound texture.

We proposed and implemented two methods for the tonic identification in Indian art music based on multi-pitch analysis. We have given a detailed explanation on both the methods in terms of the approach followed, motivation behind each of the approaches and included all the necessary implementation details. We also present a proposal for a complete iterative system for tonic identification in Indian art music. We evaluated both the methods on a sizable database and showed that we achieve a good accuracy following this approach. We finally presented a detailed analysis of the erroneous cases, examining closely each type of error that the system made and gave plausible explanations for most of them.

5.2 Main conclusions

The problem of automatic tonic identification in Indian art music has not received deserved attention from the MIR community in the past, even though it is one of the fundamental problem in computational analysis of Indian art music. Moreover, past approaches to this problem were confined to tonic pitch-class identification, doing which we completely discard an important information of the tonic octave.

We proposed two methods to identify tonic of the lead artist in a performance of Indian art music, which are capable of identifying not just the tonic pitch-class but also the corresponding tonic octave. The methods are based on multi-pitch analysis of the audio signal, in which multiple pitches present in the audio recording at a given instance of time are used to construct a pitch histogram. The peaks of the histogram represents the dominant pitches used in the excerpt. This analysis enables us to take into account not just the pitches used by the lead performer but also the ones produced by the drone instrument. Using the pitch histograms and a classification based approach, we were able to learn optimal set of rules for the

identification of the tonic. Additionally, the learned rules are easily to interpret and are musically coherent.

We also analyze the predominant melody to obtain the tonic octave information as a part of the second method. We used a rule-based and a classification based approach to identify the tonic octave from the melody histogram. By using a simple set of rules on melody histogram at tonic pitch-class locations, we were able to identify the tonic octave quite reliably. Furthermore, using the melody histogram values at each semitone and a classification base approach, the automatically learned set of rules performed much better. Both the methods were evaluated on a sizable collection of excerpts consisting of a wide variety of music pieces, artists, arrangement and recording conditions. More importantly, the approach adapted is suitable for both Hindustani and Carnatic music, male and female performances, vocal and instrumental music and it uses only a short excerpt of the full performance (exception being only the tonic octave estimation task which requires full performance).

The work presented in this thesis is primarily from a computational perspective. We plan to include more knowledge about the perception and cognition of a human listener for identifying the tonic pitch in our future work.

5.3 Open issues and future work

While we achieve our set goals for this work, there are many issues which are still unresolved and need further investigation. In this section we list out some of those issues which we found are interesting and should be addressed in the future work.

- **Source separation:** We proved that the drone sound can be successfully utilized using a multi-pitch approach to identify the tonic pitch in Indian art music. However, there are cases where the approach fails to identify the correct tonic. Some of the reasons behind such mistakes are: the loudness level of the drone sound in some excerpts is too low as compared to the lead performer. Also, we discarded the saliences of F0 candidates before constructing pitch histogram (to normalize the bias towards lead performer). As a result of this, the candidate corresponding to tonic pitch has equal weight in the

histogram as compared to the one which corresponds to the secondary pitch class of the drone instrument, even though the tonic pitch candidate is typically more salient. A solution to these problems can be performing source separation upfront, to extract the signal component corresponding to the drone sound and use only that component to identify the tonic pitch.

- **Utilization of metadata** : Our proposed approach make use of only the audio data while not utilizing the available metadata corresponding to the songs in our database. However, metadata can be a potential source of information, particularly for some culture specific knowledge that can aid the identification of tonic.

We observed that the type of errors commonly made by the system were either the octave errors, fifth errors (Pa) or the fourth errors (Ma). In such cases, information regarding the gender of the singer or the rāg corresponding to the performance might help in resolving the confusions, leading to improvement in the system.

A proposal for a complete system for tonic identification utilizing both the audio data and relevant metadata is presented in Section 3.4.

- **Perceptual and cognitive studies**: Studies pertaining to how an human listener identifies the tonic pitch in Indian art music and the duration of the audio data sufficient for this task should be done. This might help in improving the methodologies used by the automatic tonic identification approaches and it also provides a baseline for the amount of audio data needed to match the human performance. Furthermore, perceptual aspects related to this task such as the appropriate pitch resolution that should be considered in a computational analysis should also be studied in future.
- **Cultural influence**: An interesting study is be to analyse the effect of cultural background of the human listener in the task of tonic identification. We observed that for some melodies even a non-listener of Indian art music was able to identify the tonic, whereas some melodies demanded a thorough knowledge and understanding of the rāg.

5.4 Contributions

This section summarizes the relevant contributions associated with the work done in this thesis.

- Scientific background and review of relevant works and techniques pertaining to tonic identification in Indian art music.
- A new approach to tonic identification in Indian art music based on multi-pitch analysis of audio data.
- Compilation of Hindustani music collection ¹ comprising of around 125 CD releases, carried out as a part of CompMusic project ² ³. Uploading all the metadata corresponding to the collection into Musicbrainz ⁴. Tonic annotations for the songs used in the evaluation of the proposed approach.
- Compilation of solo recordings of both acoustic (#12) and electric (#20) tānpūrā, with different tuning configurations and tonic pitches. Sounds uploaded with all the relevant details on Freesound ⁵ ⁶.
- Code: An optimized C++ implementation for the computation of pitch salience function using harmonic summation. The code can be obtained from Github⁷.
- Relevant publications:

J. Salamon, S. Gulati and X. Serra, *A Multipitch Approach to Tonic Identification in Indian Classical Music*. In Proc. of ISMIR 2012, October, Porto, Portugal.

S. Gulati, J. Salamon and X. Serra, *A Two-stage Approach for Tonic Identification in Indian Art Music*. 2nd CompMusic Workshop, Istanbul 2012

¹<http://musicbrainz.org/collection/5d9b5dc6-507b-4f1a-abc4-fefd14f5e84c>

²<http://musicbrainz.org/user/compmusic/collections>

³<http://compmusic.upf.edu/>

⁴<http://musicbrainz.org/>

⁵<http://www.freesound.org/people/sankalp/packs/9571/>

⁶<http://www.freesound.org/people/sankalp/packs/9600/>

⁷<https://github.com/sankalpg/HarmonicSummation.git>

References

- Bagchee, S. (1998). *NAD Understanding Raga Music*. Business Publications Inc.
- Barrington, L., Turnbull, D., & Yazdani, M. (2009). Combining audio content and social context for semantic music discovery. In *Proc. 32nd ACM SIGIR*.
- Benaroya, L., Bimbot, F., & Gribonval, R. (2006, jan.). Audio source separation with a single sensor. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1), 191 - 199.
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings 17* (pp. 97–110).
- Bor, J., Delvoye, F. N., Harvey, J., & Nijenhuis, E. T. (Eds.). (2010). *Hindustani Music: Thirteenth to Twentieth Centuries* (First ed.). New Delhi: Manohar Publishers and Distributors.
- Bozkurt, B. (2008). An automatic pitch analysis method for turkish maqam music. *Journal of New Music Research*, 37(1), 1–13.
- Brown, H. (1988). The interplay of set content and temporal context in a functional theory of tonality perception. *Music Perception*, 5(3), 219–249.
- Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of north India. *Journal of experimental psychology: General*, 113(3), 394–412.
- Chew, E. (2002). The spiral array: An algorithm for determining key boundaries. In *Proceedings of the Second International Conference, ICMAI 2002* (pp. 18–31). Springer.
- Chordia, P., & Rae, A. (2007). Raag recognition using pitch-class and pitch-class dyad distributions. In *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*.

- Clayton, M. R. L. (2000). *Time in Indian music: rhythm, metre, and form in North Indian rag performance*. Oxford University Press.
- Danielou, A. (2010). *The Ragas of Northern Indian Music*. New Delhi: Munshiram Manoharlal Publishers.
- De Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am*, 111(4), 1917–1930.
- Deshpande, V. (1989). *Indian Musical Traditions: An Aesthetic Study of the Gharanas in Hindustani Music* (Second ed.). Popular Prakashan.
- Deva, B. C. (1980). *The Music of India: A Scientific Study*. Delhi: Munshiram Manoharlal Publishers.
- Dressler, K. (2006). Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06) . . .*, 247–252.
- Ellis, D. P., & Poliner, G. E. (2007). Identifying ‘Cover Songs’ with Chroma Features and Dynamic Programming Beat Tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* (pp. IV–1429–IV–1432). IEEE.
- Gedik, A., & Bozkurt, B. (2010). Pitch-frequency histogram-based music information retrieval for turkish music. *Signal Processing*, 90(4), 1049–1063.
- Gómez, E. (2006). *Tonal Description of Music Audio Signals*. Unpublished doctoral dissertation, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.
- Gómez, E., & Herrera, P. (2006). The song remains the same identifying versions of the same piece using tonal descriptors. Available from [files/publications/8f564f-ISMIR-2006-Egomez-Pherrera.pdf](#)
- Goto, M. (2004, September). A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4), 311–329.
- Griffin, D., & Lim, J. (1988). Multiband excitation vocoder. In *IEEE Transactions on Acoustics, Speech and Signal Processing* (Vol. 36, pp. 1123–1235).
- Griffith, R. T. H. (2004). *Hymns of the Samaveda*. Kessinger Publishing.
- Grove, G., & Stanley, S. (1980). *The New Grove dictionary of music and musicians* (First ed.). Macmillan Publishers.

- Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning*. Unpublished doctoral dissertation, University of Waikato.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009, November). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1), 10–18.
- Hermes, D. (1988, 1988). Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83, 257 - 264.
- Hess, W. (1984). Pitch Determination of Speech Signals: Algorithms and Devices. *The Journal of the Acoustical Society of America*, 76(4), 1277.
- Ioannidis, L. (2010). *Estimating the makam of polyphonic music signals : template-matching vs . class-modeling*.
- Kaul, D. M. (2007). *Hindustani and Persio-Arabian Music* (First ed.). Kanishka Publishers, Distributors.
- Klapuri, A. (2000). Qualitative and Quantitative Aspects in the Design of Periodicity Estimation Algorithms. In *Proceedings of the European Signal Processing Conference*.
- Klapuri, A. (2003a). Melody description and extraction in the context of music content processing. *Journal of New Music Research*.
- Klapuri, A. (2003b, November). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6), 804–816.
- Klapuri, A. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*.
- Koduri, G. K., Gulati, S., & Serra, X. (in press). Survey and Evaluation of Pitch-distribution Based Raaga Recognition Techniques. *Journal of New Music Research*.
- Koduri, G. K., Serrà, J., & Serra, X. (2012, Oct.). Characterization of intonation in carnatic music by parametrizing pitch histograms. In *13th int. soc. for music info. retrieval conf.* Porto, Portugal.
- Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch (Oxford Psychology Series, No 17)*. Oxford University Press, USA. Hardcover.
- Krumhansl, C. L., & Kessler, E. J. (1982, July). Tracing the dynamic changes

- in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, 89(4), 334–368.
- Lagrange, M., Martins, L., Murdoch, J., & Tzanetakis, G. (2008, feb.). Normalized cuts for predominant melodic source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2), 278–290.
- Lahat, M., & Niederjohn, R. (1987). A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(6), 741–750.
- Maher, R., & Beauchamp, J. W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *The Journal of the Acoustical Society of*, 95(April), 2254–2263.
- Medan, Y., & Yair, E. (1991). Super resolution pitch determination of speech signals. *IEEE transactions on signal processing*, 39(1), 40–48.
- Mehta, R. (2008). *Indian Classical Music and Gharana Tradition* (First ed.). Readworthy Publications Pvt. Ltd.
- Muller, M., & Ewert, S. (2010, March). Towards Timbre-Invariant Audio Features for Harmony-Based Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3), 649–662.
- Muller, M., Ewert, S., & Kreuzer, S. (2009, April). Making chroma features more robust to timbre changes. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1877–1880). IEEE.
- Muller, M., Kurth, F., & Clausen, M. (2005). Chroma-based statistical audio features for audio matching. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*. (pp. 275–278). IEEE.
- Narmada, M. (2001). *Indian Music and Sancharas in Raagas*. Delhi: Somnath Dhall, Sanjay Prakashan.
- Paiva, R. P., Mendes, T., & Cardoso, A. (2006, December). Melody Detection in Polyphonic Musical Signals: Exploiting Perceptual Rules, Note Saliency, and Melodic Smoothness. *Computer Music Journal*, 30(4), 80–98.
- Peeters, G. (2006). Chroma-based estimation of musical key from audio-signal analysis. In *Proc. 7th International Conference on Music Information Retrieval (ISMIR)* (pp. 155–120).
- Poliner, G., Ellis, D., & Ehmann, A. (2007). Melody transcription from mu-

- sic audio: Approaches and evaluation. *IEEE Trans. on Audio, Speech and Language Process*, 15(4), 1247–1256.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Rabiner, L. (1977, February). On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(1), 24–33.
- Raja, D. S. (2012). *Hindustani Music Today* (First ed.). D.K. Printworld (P) Ltd.
- Raman, C. V. (1921). On some Indian stringed instruments. In *Indian Association for the Cultivation of Science* (Vol. 33, pp. 29–33).
- Ranjani, H., Arthi, S., & Sreenivas, T. (2011). Carnatic music analysis: Shadja, swara identification and rAga verification in AlApana using stochastic models. *Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE Workshop*, 29–32.
- Rao, V., & Rao, P. (2010). Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8), 2145–2154.
- Rao, V. M. (2011). *Vocal Melody Extraction from Polyphonic Audio with Pitched Accompaniment*. Unpublished doctoral dissertation, Indian Institute of Technology Bombay.
- Ross, J. C., Vinutha, T. P., & Rao, P. (2012, Oct.). Detecting melodic motifs from audio for Hindustani classical music. In *Proc. 13th International Conference on Music Information Retrieval (ISMIR)*. Porto, Portugal.
- Ryynänen, M. P., & Klapuri, A. P. (2008, September). Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music. *Computer Music Journal*, 32(3), 72–86.
- Salamon, J., & Gómez, E. (2012, August). Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6), 1759–1770.
- Salamon, J., Gómez, E., & Bonada, J. (2011). Sinusoid extraction and salience function design for predominant melody estimation. In *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11), Paris, France* (pp. 73–80).

- Salamon, J., Gulati, S., & Serra, X. (2012, October). A Multipitch Approach to Tonic Identification in Indian Classical Music. In *Proc. 13th International Conference on Music Information Retrieval (ISMIR)*. Porto, Portugal.
- Saraf, R. (2011). *Development of Hindustani Classical Music (19th & 20th centuries)* (First ed.). Vidyanidhi Prakashan.
- Scheirer, E. D., Vercoe, B. L., Benton, S. A., & Scheirer, E. D. (2000). *Music-listening systems* (Tech. Rep.). Massachusetts Inst. of Tech.
- Schmuckler, M. A. (2004). Pitch and pitch structures. *Neuhoff, J., editor, Ecological Psychoacoustics*, 271–315.
- Sen, A. K. (2008). *Indian Concept of Rhythm* (Second ed.). New Delhi: Kanishka Publishers, Distributors.
- Sengupta, R., Dey, N., Nag, D., Datta, A. K., & Mukerjee, A. (2005). Automatic Tonic (SA) Detection Algorithm in Indian Classical Vocal Music. In *National Symposium on Acoustics* (pp. 1–5).
- Serra, J., & Gomez, E. (2008, March). Audio cover song identification based on tonal sequence alignment. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 61–64). IEEE.
- Serrà, J., Koduri, G. K., Miron, M., & Serra, X. (2011). Assessing the tuning of sung indian classical music. In *Proc. 12th International Conference on Music Information Retrieval (ISMIR)*.
- Serra, X. (2011). A Multicultural Approach to Music Information Research. In *Proc. 12th International Conference on Music Information Retrieval (ISMIR)*.
- Singh, J. (1995). *Indian Music* (First ed.). Munshiram Manoharlal Publishers Pvt Ltd.
- Stevens, C. (2004). Cross-cultural studies of musical pitch and time. *Acoustical Science and Technology*, 25(6), 433–438.
- Trivedi, R. (Ed.). (2008). *Bharatiya Shastriya Sangit: Shastra, Shikshan Va Prayog* (First ed.). Sahitya Sangam, New 100, Lookerganj, Allahabad. INDIA.
- Viswanathan, T., & Allen, M. H. (2004). *Music in South India*. Oxford University Press.
- Wiggins, G. a. (2009). Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. *2009 11th IEEE Interna-*

tional Symposium on Multimedia, 477–482.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining : practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann. Paperback.