AUTOMATIC TONIC IDENTIFICATION IN INDIAN CLASSICAL MUSIC

A THESIS

submitted by

ASHWIN BELLUR

for the award of the degree

of

MASTER OF SCIENCE
(by Research)

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

July 2013

THESIS CERTIFICATE

This is to certify that the thesis titled **AUTOMATIC TONIC IDENTIFICATION IN INDIAN CLASSICAL MUSIC** , submitted by **Ashwin Bellur**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Science**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. R. Aravind
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date:

Prof. Hema. A. Murthy
Research Guide
Professor
Dept. of Computer Science and Engineering
IIT-Madras, 600 036

Place: Chennai
Date:

ACKNOWLEDGEMENTS

The completion of this thesis has been possible because of the contribution of several people. I would like to express my gratitude to my advisor Prof. Hema A. Murthy for her guidance and unwavering support. She has been a constant source of encouragement and perspective and has played a major role in instilling confidence in me as a researcher. My interactions with her over the last four years have not only shaped my outlook on research, but on life as well. Her unabated energy and enthusiasm for research is something I truly admire and can only aspire to emulate.

I am grateful to my co-advisor Prof. R. Aravind and the members of the GTC Committee, Prof. C. S. Ramalingam and Prof. V. Kamakoti for their insightful comments and suggestions with respect to my thesis. I would also like to express my gratitude to Prof. Xavier Serra for letting me be a part of the remarkable Compmusic project. It has been an absolute pleasure working with him and the other members of the consortium. I wish my very best for this fantastic project.

I would like to thank Badri, Raghav, Srikanth and other members of Donlab for their support and encouragement over the years. I would like to thank Vignesh Ishwar, in particular, for having played an invaluable part in the writing of this thesis. I deeply appreciate the long hours he invested in vetting and validating the work done in this thesis. I would also like acknowledge the contribution of Gopal and Sankalp from MTG Barcelona. They have been instrumental in helping me set up the database used in this work. The many deliberations with Sankalp played a crucial role in helping me compile and interpret the results reported in this work.

I would like to acknowledge the European Research Council for funding the research under the European Unions Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

Lastly, I would like to thank my parents, my sister Sharmila and Cousin Prasad

Ram for their unreserved support and encouragement. Knowing that they have my back has always made my life so much easier and has given me the strength and courage to pursue any path I wish to choose.

ABSTRACT

KEYWORDS: Indian classical music ; Tonic identification; Pitch histograms; Group delay processing; Low level audio descriptors; Non-negative matrix factorization.

Music Information Retrieval (MIR) involves study of musical repertoires and musical concepts using computational methodologies. Use of computation methodologies on music have found real world applications like searching music in large repositories, archival of music, music recommendation and music synthesis. MIR studies for Indian classical music are gaining importance given that the music repertoire has an active listening community and is increasingly being consumed online. Indian classical music with its distinct characteristics offers a number of interesting challenges from the perspective of search, retrieval and recommendation. This thesis addresses the task of automatic tonic identification, a task of fundamental importance in the context of MIR for Indian classical music.

In Indian classical music, tonic pitch is the reference note chosen by the lead artist for a performance. Knowledge of the tonic pitch is essential in Indian classical music in order to perform melodic analysis across artists and performances. Features extracted from different performances need to be normalized using the respective tonic before further analysis. Hence automatic tonic identification serves as a prerequisite for studying melodic concepts using large data driven techniques.

In this work an attempt is made to utilize the various cues the music offers to the listener to help identify the tonic. We propose processing two types of cues to automatically identify the tonic pitch: 1) melodic characteristics and 2) tuning of the drone. The thesis first explains the cues in greater detail and suggests appropriate features that need to be extracted from the audio to capture the cues. It then details several knowledge based signal processing techniques that can be adopted to further enhance the tonic information in the extracted features.

In order to process the melodic characteristics, pitch histograms are used as the basic representation. It is then shown that the pitch histograms can be processed using the group delay function to emphasize certain characteristics of the music that manifest in a pitch histogram, to aid in accurate tonic identification. To justify such a processing technique it is illustrated that the pitch histogram can be characterized as the squared magnitude response of a set of resonators in parallel that do not have constant Q. Interesting properties of the group delay response of such a setup are then illustrated and are shown to be useful in identifying the tonic.

To process the drone, a cepstrum based pitch extraction technique is proposed. It is also shown that by estimating pitch of low energy frames, tonic can be identified with greater speed and higher accuracy. In order to further enhance the performance and also illustrate the ubiquitous nature of the drone, a non-negative matrix factorization technique based method is developed to identify tonic.

The proposed techniques are tested on a large and varied database. It is shown that very high accuracies of automatic tonic identification ($\approx$ 95%) can be attained using the methods proposed.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **i**

ABSTRACT **iii**

LIST OF TABLES **viii**

LIST OF FIGURES **xi**

ABBREVIATIONS **xii**

1 Overview of the thesis **1**

1.1 Introduction 1

1.2 Overview 2

1.3 Organization of the thesis 3

1.4 Contribution of the thesis 3

2 Tonic Identification in Indian Classical music **5**

2.1 Introduction to tonic identification 5

2.2 Previous work 8

2.2.1 Steady state based approach 8

2.2.2 Semi-Continuous GMM based approach 8

2.2.3 Multi pitch approach 9

2.3 Database 10

2.4 Summary 12

3 Tonic identification using melodic cues **13**

3.1 Introduction 13

3.1.1 Śruti and svara in Indian classical music 14

3.1.2 Analysis of pitch histograms 15

3.2 Group delay function 18

3.2.1 Minimum phase group delay function 19

3.2.2 Properties of minimum phase group delay function 22

3.3 Processing pitch histograms using group delay function 23

3.4 Methods to identify the tonic using GD histograms 29

3.4.1 Method 1 - concert based method 29

3.4.2 Method 2 - template matching 30

3.4.3 Method 3 - segmented histograms 33

3.5 Experiments and results 34

3.5.1 Performance of CBM 35

3.5.2 Performance of TMM and SHM 36

3.5.3 Error analysis 39

3.6 Summary 42

4 Processing the drone for tonic identification 44

4.1 Motivation and context 44

4.2 Drone and its characteristics in Indian classical music 45

4.3 Cepstrum based pitch extraction for tonic identification 48

4.3.1 Tonic identification using cepstral pitch method and pitch histograms 51

4.3.2 Audio descriptors to select frames 51

4.3.3 Non-negative matrix factorization based tonic identification 55

4.4 Experiments and Results 58

4.4.1 Method 1 - pitch histogram with cepstrum based pitch extraction 58

4.4.2 Method 2 - audio descriptors 60

4.4.3 Method 3 - non-negative matrix factorization 63

4.4.4 Error Analysis 66

4.5 Summary 67

5 Summary and Conclusion 69

5.1 Salient Points 69

5.2 Criticism of the work 70

5.3 Future work 70

LIST OF TABLES

2.1 Details of the database. H – Hindustani, C – Carnatic, M – Male, F – Female, I – Instrumental 11

3.1 Accuracy in %. TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic pitch class 36

3.2 Performance of TMM (% accuracy). TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic Pitch Class 36

3.3 Performance of SHM (% accuracy). TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic Pitch Class 36

4.1 Low level audio descriptors 52

LIST OF FIGURES

2.1 Pitch Histogram of three performances of *rāga Kāmbōji* by three different artists. The solid red lines denote the *svara Sa* 7

2.2 Pitch Histogram of the three performances on the cent scale, after normalizing with respect to the tonic. *Svara Sa* can be seen at -1200, 0 and 1200 cents 7

3.1 Pitch histogram of an item of Carnatic Music (plot a) and an item of Hindustani music (plot b). The red stem indicates the tonic pitch . . 16

3.2 Pitch histograms for each of the manually transcribed svaras along with the histogram for the complete item 17

3.3 Block digram for deriving minimum phase group delay function from pitch histograms 21

3.4 Pitch and Group Delay Histograms (a) Pitch histogram of a 3 minute Carnatic music item. (b) The GD histogram. Red stems indicate the bin values of *svara Sa* and blue stems indicate the *svara Pa* 23

3.5 Column 1 - Squared magnitude response, Column 2 - Pole zero plot, Column 3 - Group delay plot, for a two pole system with $\alpha_1 = 1$ and varying α_2 25

3.6 Column 1 - Squared magnitude response, Column 2 - Pole zero plot, Column 3 - Group delay plot, for a three pole system with $\alpha_1 = \alpha_3 = 1$ and varying α_2 27

3.7 Pitch histograms for each of the manually transcribed svaras along with the histogram for the complete item and the group delay processed histogram. The group delay processed histogram is normalized to have a maximum height of 1. Red stems indicate the *Sa* and the blue stem indicates the *Pa* 28

3.8 Concert method - plots a,b,c and d correspond to the group delay histograms for 4 items of a concert. Plot e is the bin-wise product of the above 4 histograms 30

3.9 Illustration of the template matching procedure. Plot 1 shows the local peaks in Gd histograms. Plot 2 and 3 show the template matching procedure for two different cases. The black strip represents the local peak assumed as the tonic and the blue strip represents the corresponding template. Plot 2 being the correct estimate of the *Sa*, a better template match is obtained. 32

3.10 Histogram and GD histogram of a single four minute Carnatic piece. The histogram bin marked with a “*” on the X axis is the most frequented note, whereas the bin marked “+” is the *svara Sa* 34

3.11 Plots 1-4 show segment-wise histogram and GD histogram. Plot 5 is the product of GD histograms in plot 1-4, with *Sa* as the global peak 34

3.12 The performance of TMM and SHM on database DB3 37

3.13 The performance of TMM and SHM on database DB4 38

3.14 The performance of TMM and SHM on Male vocal Female vocal and Instrumental music. M – Male, F – Female, I –Instrumental 40

3.15 The performance of TMM and SHM on Male vocal Female vocal and Instrumental music. TP – Tonic Pitch, TPM – Tonic Pitch with Metadata, TPC – Tonic Pitch class. Last letter H – Histogram, G – GD Histogram 41

3.16 The performance of TMM and SHM using GD histograms on database DB4. TP – Tonic Pitch, TPM – Tonic Pitch with Metadata, TPC – Tonic Pitch class 43

4.1 Spectogram of the *tambura* in isolation 46

4.2 Figure a is the pitch extracted from a *tambura* excerpt and Figure b is the corresponding histogram 47

4.3 Spectogram of an excerpt of Carnatic music 47

4.4 Procedure for cepstrum based pitch extraction 49

4.5 Cepstrum for three frames of data. Plots a, d and g are the cepstrum plots for the *tambura* in isolation. Plots b, e and h are cepstrum plots for male vocal in isolation. Plots c, f and i are cepstrum plots for the mixed signal 49

4.6 Block diagram for method 1 51

4.7 Low level audio descriptors and their relative values for prominent drone frames 53

4.8 Block diagram for method 2 53

4.9 Waveform of 60s duration and short term energy 54

4.10 Plot (a) is the pitch histogram computed from pitch extracted from all the frames. Plot (b) is the pitch histogram computed using pitch from short energy frames 54

4.11 Block diagram for method 3 56

4.12 The performance on determining pitch in the lower and middle octave performance range, using varying quantities of data. LO - Lower Octave, MO- Middle Octave 59

4.13 The performance on determining pitch in the lower octave range, using varying quantities of data. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument, FI - Full Item 60

4.14 The performance on using various audio descriptors. ZCR - Zero crossing rate, STE - Short term energy. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument 61

4.15 The performance on selecting drone prominent frames using short term energy. FI – Full Item 62

4.16 The performance on using Random and Constant activation matrix on DB4. FI – Full Item 64

4.17 The performance of NMF based method in tonic identification. FI – Full Item 65

4.18 Error analysis - pitch histogram on low energy frames based method . H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument 66

4.19 Error analysis - NMF based method. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument 67

ABBREVIATIONS

MIR	Music Information Retrieval
GMM	Gaussian Mixture Models
SC-GMM	Semi Continuous Gaussian Mixture Models
NMF	Non-negative Matrix Factorization
DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
GD	Group Delay
CBM	Concert Based Method
TMM	Template Matching Method
SHM	Segmented Histogram Method
TP	Tonic Pitch
TPC	Tonic Pitch Class

CHAPTER 1

Overview of the thesis

1.1 Introduction

Music Information Retrieval (MIR) deals with study of musical repertoires and concepts using computational methodologies. Given that much of music and the related industry and products are going online, the field of MIR has begun to gain significant importance. There has been a major paradigm shift in the manner in which music is discovered, listened and created over the last few years. The accumulation of vast quantities of data online provides major challenges in organizing, searching and retrieving music. Methodologies that are based on a combination of musicological studies, signal processing and machine learning techniques have sought to solve some of these interesting challenges.

Research in MIR has primarily focussed on studying (Serra, 2011) commercial Western repertoires and concepts related to it. The availability of a rich consumer base as well large amounts of high quality data are some the important factors for such a bias. The methodologies developed are therefore dictated by the musical repertoires analysed. Most of the MIR techniques that have been developed are designed for Western music and cannot be subjected to non-Western musical repertoires.

Indian classical music, with the presence of a large active performing and listening community, well developed musicological studies and availability of good quality data renders itself viable to MIR treatment. The two main sub-genres of Indian classical music, Hindustani and Carnatic music are increasingly being distributed and consumed online, presenting a large number of challenges in terms of search, archival and recommendation.

In this work, the task of automatic tonic identification, one of the fundamental tasks in both Hindustani and Carnatic genres of Indian classical music, is addressed. As will be seen in the following sections and chapters, automatic tonic identification is of primary importance and is in fact a prerequisite for any kind of data driven melodic analysis. In this work, an attempt is made to combine domain knowledge with signal processing techniques, to develop expert systems which, given an audio excerpt, can identify the tonic pitch automatically.

1.2 Overview

In Indian classical music, the tonic is the reference pitch with respect to which all the other notes in a melody are defined. The tonic is decided by the lead performer, usually depending upon the performer's vocal/instrumental range. All other instruments that are part of the ensemble, also tune to the tonic or the reference pitch. The objective of this work is to automatically identify the tonic pitch when given an audio excerpt. The information of the tonic is essential, as it will allow normalizing of melodic features across musicians and the study of invariant melodic concepts across musicians using these normalized features.

In this work, we attempt to utilize the various cues the music offers to the listener to help identify the tonic. The cues used in this work can be classified into two broad classes 1) Melodic characteristics 2) Tuning of the drone. The drone refers to the sound produced by an instrument or an electronic device in the background serving as the reference throughout a performance. In order to use these cues, appropriate features that need to be extracted from the audio are first analyzed. The features can seldom be used directly for tonic estimation. Signal processing algorithms are then employed to enhance tonic information in the extracted features. The features extracted and the processing techniques proposed to identify tonic are different for each of these cues being analyzed.

Several strategies are then proposed for using these processed representations for identifying the tonic pitch. The strategies/methods proposed stem from the domain knowledge of the cues and manner in which they manifest in their respective representations. The performance of the techniques is evaluated on a large and varied database. The results indicate that the tonic pitch can be obtained fairly accurately by employing the methods proposed in this thesis.

1.3 Organization of the thesis

The organization of the thesis is as follows: In Chapter 2, the role of the tonic in Indian classical music and the importance of its identity for melodic analysis is discussed. Then the various musical cues on offer to identify tonic pitch are illustrated. The two main cues 1) Melodic characteristics 2) Tuning of the drone, that have been used in this work for identifying the tonic are elaborated upon. Some of the previous efforts to utilize these cues for tonic identification are also discussed. Chapter 2, then details the database that has been compiled for analyzing the characteristics of the music, as well as for testing the methods suggested in this work.

Chapter 3 and Chapter 4 are dedicated to the two broad classes of cues being analyzed in this work for tonic identification. While Chapter 3 deals with processing the cues arising from the melodic characteristics of the art form, Chapter 4 deals with processing the drone to identify the tonic pitch. Several knowledge based signal processing techniques are developed to process these cues in the respective chapters.

The performance of the methods proposed in Chapters 3 and 4 are tested on the database and are reported in the respective chapters. Chapter 5 reviews the salient points presented in the thesis. Possible other applications of the developed strategies as well as some of the shortcomings are also discussed in Chapter 5. The work concludes by discussing some of the possible future efforts for tonic identification

1.4 Contribution of the thesis

The following are the main contributions of the thesis:

1. Two complementary cues are processed from the audio and utilized for tonic identification.
2. Several knowledge based signal processing methods are proposed to exploit each of these cues.

3. Pitch histogram based approaches are developed for utilizing the melodic cues for tonic identification. A novel group delay based processing technique is developed to accentuate the tonic information present in pitch histograms. Interesting properties of the group delay function are illustrated though this novel processing technique.

4. To process and to determine the tuning of the drone in Indian music, a cepstrum based pitch extraction method is shown to suffice. The properties of the drone in terms of low level audio descriptors are illustrated. These features characterizing the drone are then shown to useful in enhancing the performance of tonic identification.

5. A non-negative matrix factorization based technique is developed to determine the tuning of the drone and to thereby identify tonic with minimal data.

6. Through developing several strategies to use these cues, the thesis also sheds light upon some of the interesting properties of the music and the manner in which they manifest on signal processing.

CHAPTER 2

Tonic Identification in Indian Classical music

2.1 Introduction to tonic identification

A fundamental concept in Indian Classical music is the one of tonic. The tonic is the base pitch chosen by a performer, which serves as a reference throughout a performance. Melodies are defined relative to the tonic. Accompanying instruments also tune to the tonic of the lead performer. The tonic is generally preserved across the various items which are presented in a concert[1].

The seven basic notes or *svaras* as they are referred to in Indian classical music (Indian equivalent to the solfège) are, *Ṣaḍja, Ṛṣabha, Gāndhāra, Madhyama, Pañcama, Dhaivata* and *Niṣāda*. These are generally represented using the syllables *Sa, Ri, Ga, Ma, Pa, Da* and *Ni* respectively. The *ṣaḍja* or the *svara Sa* serves as the base note with all the other notes of the melody, depending on the position of the *Sa* on the frequency scale. In particular, the tonic pitch is the *svara Sa* in the middle octave range of a performance. As the lower, middle and higher octaves of a performance are referred to as *mandra stāyi*, *madhya stāyi* and *tāra stāyi* both in Hindustani and Carnatic music, the tonic pitch is also referred to *madhya ṣaḍja*.

A simplified view of the difference from Western classical music would be that Indian music uses a fixed tonic, while Western classical music uses a movable tonic i.e for each key a different reference tonic is used. While in Western music, a fixed frequency is used as reference for tuning, invariably A4 (440 Hz), in Indian classical music, the root note is first chosen by a musician and all the melodies that the musician performs in a concert are defined relative to this root note, referred to as the tonic.

There have been various efforts to apply computational methods to analyze different aspects of Indian music using pitch (Krishnaswamy, 2003) as the basic feature. In

[1] except in the rare event where the *madhyama śruti* is used in Carnatic music for performing specific items.

Chordia and Rae (2007); Chordia *et al.* (2009), pitch class distribution and pitch dyads are employed for automatic *rāga* recognition. Pandey *et al.* (2003) employ a form of hidden markov models with pitch as the basic feature to do the same. In Krishnaswamy (2004) an attempt is made to study inflections/*gamakas* using pitch contours. Serra *et al.* (2011) address the tuning issue in Indian music using pitch histograms. In Ross *et al.* (2012) and Ishwar *et al.* (2013) attempts are made to spot recurring melodic motifs using pitch contours as the feature. All these works indicate that pitch and pitch contours can be used to study a variety of melodic concepts of Indian classical music. Given the relevance of pitch and pitch contours for studying melodic concepts, the knowledge of the tonic pitch becomes imperative for applying large data driven techniques across musicians. Using the tonic pitch, the pitch extracted from performances of numerous artists can be normalized with respect to the tonic and then studied. In works that have used data driven techniques to study melodic concepts, either the tonic has been manually identified (Chordia and Rae, 2007; Chordia *et al.*, 2009; Ishwar *et al.*, 2013), or other means like interval histograms (Serra *et al.*, 2011) were developed to overcome the effect of varying tonic.

In order to illustrate the effect and the need for normalizing pitch with respect to the tonic, histograms computed using the pitch extracted from three Carnatic items in *rāga Kāmbōji*, rendered by three different artists with different tonic pitch are shown in Figure 2.1. The histograms has been normalized to have a maximum height of 1. As each item has been rendered using a different tonic, the pitch histogram of the same *rāga* can be seen to occupy different pitch ranges (Figure 2.1) with barely discernible similarities across the pitch histograms.

Figure 2.2 shows the histogram on the cent scale, computed after normalizing each of the pitch contours with its respective tonic. Cent is an unit of measure used for musical intervals on the logarithmic scale as shown in Equation 2.1.

$$c = 1200(log_2(\frac{f_2}{f_1})) \tag{2.1}$$

The variable c is the cent value of frequency f_2 with respect to the tonic f_1.

Given that the pieces are of the same *rāga*, similarities across histograms are evident in Figure 2.2 when compared to Figure 2.1. These normalized histograms on the cent scale can now be used for studying melodic concepts, like the characteristics of the melody *Kāmbōji* across artists, which would not be possible with just pitch contours and pitch histograms.

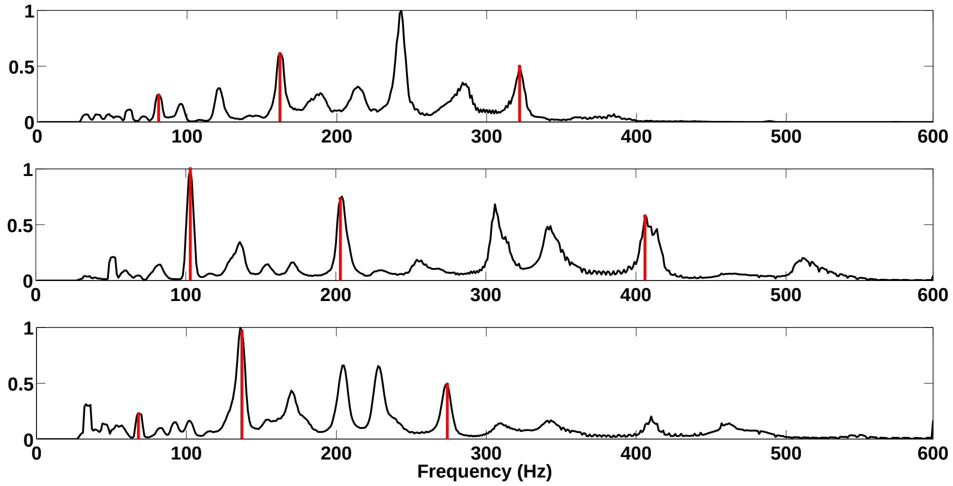

Figure 2.1: Pitch Histogram of three performances of *rāga Kāmbōji* by three different artists. The solid red lines denote the *svara Sa*

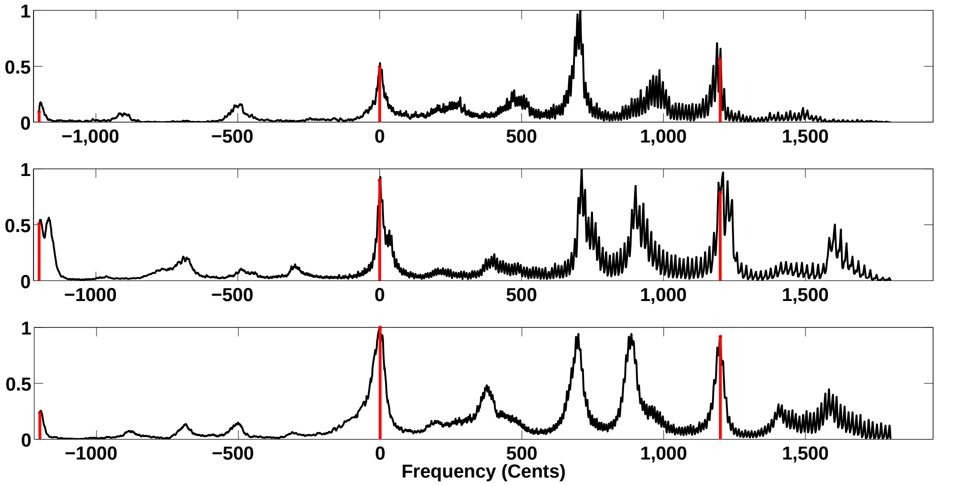

Figure 2.2: Pitch Histogram of the three performances on the cent scale, after normalizing with respect to the tonic. *Svara Sa* can be seen at -1200, 0 and 1200 cents

2.2 Previous work

There have been a number of efforts (Sengupta *et al.*, 2005; Ranjani *et al.*, 2011; Salamon *et al.*, 2012; Gulati *et al.*, 2012) to automatically identify the tonic pitch, given an audio excerpt. In Sections 2.2.1, 2.2.2 and 2.2.3 a brief review of these works are presented.

2.2.1 Steady state based approach

Sengupta *et al.* (2005) is one of the earliest works to attempt automatic tonic identification. The method uses pitch contours as the basic feature. It proposes that given the pitch contour of an item of Hindustani music, the "steady regions" must be first detected. It hypothesizes that these regions denote the notes of the melody. A large number of frequency values within a certain range are then treated as candidate tonic pitch values. The method proposes that for the right candidate tonic, the steady regions detected should be at certain specified ratios to the tonic as prescribed by the tuning system. The candidate tonic for which the steady regions best follow the prescribed ratios is deemed to be the tonic pitch. The method however, is tested on small database of Hindustani music which had been created specifically for testing the method proposed.

2.2.2 Semi-Continuous GMM based approach

In Ranjani *et al.* (2011), an attempt is made to use the melodic characteristics of Carnatic music to identify the tonic. In Indian classical music, more so in Carnatic music as will be illustrated in detail in Chapter 3, the *svaras* do not indicate a specific pitch value, rather indicates a pitch region. Out of all the *svaras* that can comprise a melody, the music dictates that the *svara Sa* and *Pa* are allowed span a narrower range when compared to the other notes. In Ranjani *et al.* (2011), an effort is made to capture this trait of the music by first computing Parzen distribution of the pitch extracted from the audio. For each of the candidate peaks in the distribution, a 36 mixture Semi-Continuous Gaussian Mixture Model is computed (SC-GMM), by fixing the means of the GMM at the expected theoretical positions of the *svaras* of the melody, with respect to the candidate peak. The variance and mixture weights of the SC-GMM are iteratively estimated. It is hypothesized that for the actual tonic peak, the mixtures denoting the *svaras Sa* and *Pa* in all the 3 octaves will have the least variance. A number of ways are suggested for using the variance values and mixture weights in order to identify the tonic pitch.

2.2.3 Multi pitch approach

In Salamon *et al.* (2012); Gulati *et al.* (2012), a multi pitch based approach is explored with the intention of tracking the pitch that is produced by the drone in the background. The drone is used in Indian classical music to establish the tonic and it serves as reference throughout a performance. The fact that the drone is omnipresent in the background in all the frames is exploited by constructing a histogram using the multi-pitch extracted. It is hypothesized that the peaks of the pitch corresponding to the drone should dominate, given its omnipresent nature. But the drone, along with the *svara Sa*, also produces pitch corresponding to the *svara Pa Ma* or *Ni*, due to which dominant peaks other than the tonic pitch are also present. In order to identify the peak corresponding to the tonic pitch amongst these peaks, a supervised approach is adopted. The relation between the peaks are learnt during the training phase using decision trees. In the testing phase, given a new excerpt of music, the pitch histogram is first computed using multi-pitch and the decision trees are used to determine the peak corresponding to that of the tonic pitch.

These previous efforts to automatically identify tonic pitch primarily differ in the musical cues they attempt to use. The cues can be broadly divided into two classes

- Melodic Characteristics
- Tuning of the drone

In this thesis, an effort is made to develop methods to use both the melodic characteristics as well as characteristics of the drone, to identify the tonic.

For determining the tonic pitch using melodic characteristics, an attempt is made to process cues similar to the one proposed in Ranjani *et al.* (2011). The fundamental difference between the approach used by Ranjani *et al.* (2011) and the work reported here is that, while in Ranjani *et al.* (2011) a SC-GMM based technique is proposed, in this work modeling of the pitch histograms by an all-pole model and its subsequent processing using group delay function is proposed. The possible advantages of such a processing technique over SC-GMM is explained in detail in Chapter 3.

In Chapter 4, techniques are developed to process the audio to determine the tuning of the drone and hence the tonic pitch. While in Salamon *et al.* (2012); Gulati *et al.* (2012) a multi-pitch with decision trees based approach is developed, in this work a simple cepstral pitch method with the use of the domain knowledge is shown to suffice. To further improve the performance a NMF based technique is also developed. While the methods proposed in Salamon *et al.* (2012); Gulati *et al.* (2012), require a training phase due to use of decision trees, the methods proposed in the work do not have a learning phase.

2.3 Database

To effectively study the performance of the methods proposed in this work, a large varied database is used. Due to absence of a standard database, multiple databases that have been compiled by different groups (Salamon *et al.*, 2012; Gulati *et al.*, 2012; Bellur *et al.*, 2012) for the purpose testing their respective tonic identification techniques are used in this work. Though these databases were initially created with the intention of highlighting the importance of particular cues for tonic identification, these databases combined, factor in all the different types of data that might be provided for tonic identification. This is of pertinence as the nature of data available influences the kind of musical cues that can be processed for identifying the tonic pitch.

Two main factors that were considered in organizing the database. The first factor that was considered was the quality and type of recording. In order to study melodic concepts of Indian music, it is imperative to analyze data across the decades considering the evolving nature of Indian classical music. This is essential, so as to ensure that one does not arrive at conclusions about the melodic concepts by just analyzing the modern day performance styles, as one might due to the ease of availability of good quality data. Use of audio data from across the decades implies that the quality of data available would vary. A large amount of recordings of yesteryear artists available are not professionally recorded, but rather are amateur recordings part of personal collections. Even the professionally recorded audio are many a time cassette recordings which have been converted to digital audio. One of the important challenges with such recordings would be the processing of the drone for identifying the tonic. In these recordings, due to the amateur setup, the drone might be extremely faint, masked by the noisy conditions or it may sometimes be even absent. In order to test the methods proposed in these works on such recordings, a large number of concerts were picked at random from a personal collection of amateur recordings of live concerts. However the amateur recording collection was restricted to Carnatic music due to lack of availability of similar data for Hindustani music. The details are enumerated in Table 2.1. As also can be seen in Table 2.1, to factor in good quality audio, a large collection of Audio CDs, of both Hindustani and Carnatic music, released by popular labels was also compiled.

The second factor that was considered was the duration of music used for identifying the tonic. Sometimes shorter excerpts of the original item might be presented for tonic identification. Many of the aggregate melodic cues based methods described in Chapter 3 might not be present when short excerpts are used. A different approach might be required for such data. The excerpts used in this work are of varying duration, ranging from 3 minute excerpts to entire concerts (Table 2.1).

Table 2.1: Details of the database. H – Hindustani, C – Carnatic, M – Male, F – Female, I – Instrumental

Database	Dur	Type	Items	Artists	H	C	M	F	I
DB1	Full concert	Amateur recording + audio CD	78	22	0	100	72	20	8
DB2	Full Song	Amateur recording	485	22	0	100	69	23	8
DB3	Full Song	Audio CD	428	71	45	55	72	28	0
DB4	3 min excerpt	Audio CD	1206	114	43	57	45	28	27

As can be seen in Table 2.1, the database has 4 components, each comprising of a large number of artists across male vocal, female vocal and instrumental artists. While database DB1 and DB2 consist of only Carnatic music (compiled at IIT Madras), database DB3 and DB4 consist of both Hindustani and Carnatic music (compiled with MTG Barcelona). The reasons for compiling DB1 where the tonic is identified for a complete concert will be detailed in Chapter 3.

In order the generate the ground truth for testing the methods, the tonic was manually identified for the whole database by professional musicians. Given an item of music, the professional musician first played the different tones on a pitch pipe to determine the nearest semitone to the tonic pitch of the item. To arrive at the precise estimate of the tonic, the musician then played tones, varying them $1Hz$ at a time, within the semitone, along with the classical piece using the *Audacity* software. The pitch at which the tone sounded same as that of the *svara Sa* in the middle octave of the item was identified as the tonic pitch.

For the complete database, the tonic pitch was found to be in the following range:

- 100 Hz - 180 Hz for items with lead as male vocal
- 140 Hz - 250 Hz for items with lead as female vocal
- 120 Hz - 210 Hz for items with instrumental lead

2.4 Summary

The concept of tonic pitch in Indian classical music was first discussed. The importance of automatic tonic identification was motivated by analyzing previous efforts to perform computational studies on Indian music, in which the varying tonic between artists proved to be a bottleneck. Using cent histograms as an example, the importance of the knowledge of the tonic for performing melodic studies was further highlighted. Some of the previous efforts to automatically identify the tonic pitch were subsequently discussed. The basic difference in these prior approaches was discussed, in particular the nature of cues picked from the music to identify the tonic pitch. It was inferred that these cues used can be broadly divided into two classes, melodic cues and tuning of the drone.

CHAPTER 3

Tonic identification using melodic cues

3.1 Introduction

Melody is a fundamental element in most music traditions. Although melody is a common term that is used to categorize certain musical elements, each tradition has specific differences. Indian classical music is an example of a tradition with melodic traits very different from that of Western classical music. In Western classical music, a melody is normally defined as a succession of discrete tones, tones that belong to a given scale and tonality context. Melodic studies use symbolic representation of music and use concepts such as notes, scales, octaves, tonality and key signatures. Also, Western classical music uses equal temperament tuning, due to which melodic analysis of a Western piece of music is normally based on a quantized representation of pitches and durations, within a well defined framework of possible relationships (Serra, 2011).

Melody in Indian classical music relates to the concept of *Rāga*. This has very little correspondence to the tones and scales of Western music. A *Rāga* prescribes the way a set of notes/*svaras* are to be inflected and ordered. Indian classical music tradition has been preserved and has evolved as an oral tradition in which notation plays a very little role. When used, the written notations are more for archival and reference purposes, than as documents which can be referred to during a performance.

As mentioned earlier, notes in Indian music are referred to as *svaras*. A *svara* is more a pitch region than a specific pitch frequency. The concept of discrete tones is not quite appropriate for Indian Music. A melody is better defined as the pitch-centered temporal variations in a musical piece. The characteristics of these pitch regions denoting *svaras* are decided by its scale, melodic pattern and specific melodic treatment. In this chapter, an attempt is made to derive a relationship between this signature characteristic of Indian classical music and the features extracted from the audio signal. These features are then processed to identify the tonic pitch.

3.1.1 Śruti and svara in Indian classical music

Intonation in Indian classical music has been a topic of much theoretical discussion and debate. In particular the concept of *śruti* that deals with tuning has received widespread attention. In Rao and Meer (2010), an effort is made to trace the historical development of the concept of *śruti* and detail some of the recent findings of computational analysis.

Typically, in Indian classical music, seven *svaras* denoted as *Sa, Ri, Ga, Ma, Pa, Da, Ni* are used. Except the tonic and the fifth, that is *Sa* and *Pa,* all the other notes have two variations, which accounts for twelve notes in an octave, referred to as *svarasthānas*. The term *śruti* here refers to the subtle division of the octave in Indian classical music rather than 12 discrete *svaras*. The concept of *śruti* is a much contested topic primarily due to the fact that unlike Western music, the note or the *svara* does not indicate a specific pitch frequency, rather a pitch area (Deva, 1965). The presence of *gamakas* and *mīnḋs* which are the core characteristics of the art form, as has been illustrated in Krishnaswamy (2003, 2004); Subramanian (2007), implies that the notes are not rendered flat, but are inflected. Due to this characteristic of the music, as various recent works have indicated (Levy, 1982), there are no fixed number of divisions of an octave as described by many differing theories over the centuries, rather the divisions are decided by the scale, melodic pattern and specific melodic treatment (Meer, 1980; Ranade, 1957). In fact in Deva (1952), it is argued that there are "micro-distinctions" between pitches of the *svara* due to *gamakas* and *mīnḋs* which are impossible to measure and quantify.

In Serra *et al.* (2011), these characteristics of Indian music are studied through constructing pitch histograms on a large data set of Hindustani and Carnatic music. It is suggested that the tuning for Hindustani and Carnatic music tended towards equal temperament and just intonation forms of tuning respectively, though not exactly at the theoretical positions. Further it is illustrated through using pitch histograms that there are many more divisions of the octave than the standard 12 semitones found in Western classical music with pitch distributions more spread out in Carnatic music when compared to Hindustani music.

3.1.2 Analysis of pitch histograms

Pitch histograms, as has been illustrated already, serve as useful representation for studying melodic concepts in Indian classical. These histograms have been used for studying intonation (Serra *et al.*, 2011) as well as in *rāga* recognition systems (Koduri *et al.*, 2012) in the context of Indian classical music. Pitch histograms are used as the primary representation in this chapter. To compute the pitch histogram, given that Indian classical music is heterophonic in nature, Yin (Cheveigne and Kawahara, 2002.) – autocorrelation based mono pitch extraction algorithm is used to extract pitch from an item of music. Pitch is extracted using Yin on all the frames of the given audio. The resulting pitch contour indicates the pitch of the most prominent source in each of the frames. Figure 3.1 shows two histograms computed on the pitch extracted from an item each of Carnatic (*Rāga Bēgaḍa*) and Hindustani music (*Rāga Yaman*) of 3 minute duration. The pitch was extracted using Yin with a hop size of $10ms$ and window size of $133ms$. Bin width of $1Hz$ was used to compute the histogram. The same configuration has been used for computing pitch histograms in the rest of the chapter.

In order to develop techniques to identify tonic pitch with pitch histograms as the primary form of representation, a large number of pitch histograms were studied. The following observations were made:

- A peak is always seen at the tonic pitch value in the histogram. The peak is not necessarily the most prominent. A peak denoting the *svara Sa* is present in any excerpt of music, even when the *svara Sa* is not being rendered by the lead performer in the excerpt. This is because, the drone (if present) is tuned to the tonic pitch, along with the percussion and accompanying instruments, resulting in a peak at the tonic pitch.

- As expected, the distribution of the pitch frequency appears to be continuous in nature in the pitch histogram, even though a melody is based on a scale of finite *svaras*. Peaks are seldom observed at expected theoretical positions of the *svaras* of a given melody. This property of Indian music, in particular Carnatic music, renders the peak picking process difficult.

- With the *svara Sa* as the tonic, the *svara Pa* is the fifth i.e exactly at 1.5×(pitch of *Sa*) (Deva, 1952). The same cannot be assumed for the other *svaras* of the melody as one might not find peaks indicating the *svaras* at the theoretical positions (Serra *et al.*, 2011)[1].

[1] Deva et al. in Deva (1952) suggests that *Sa* and *Pa* were in fact variable notes earlier with its position depending on the *grāma*. The concept of *grāma* is not relevant in the present day practice

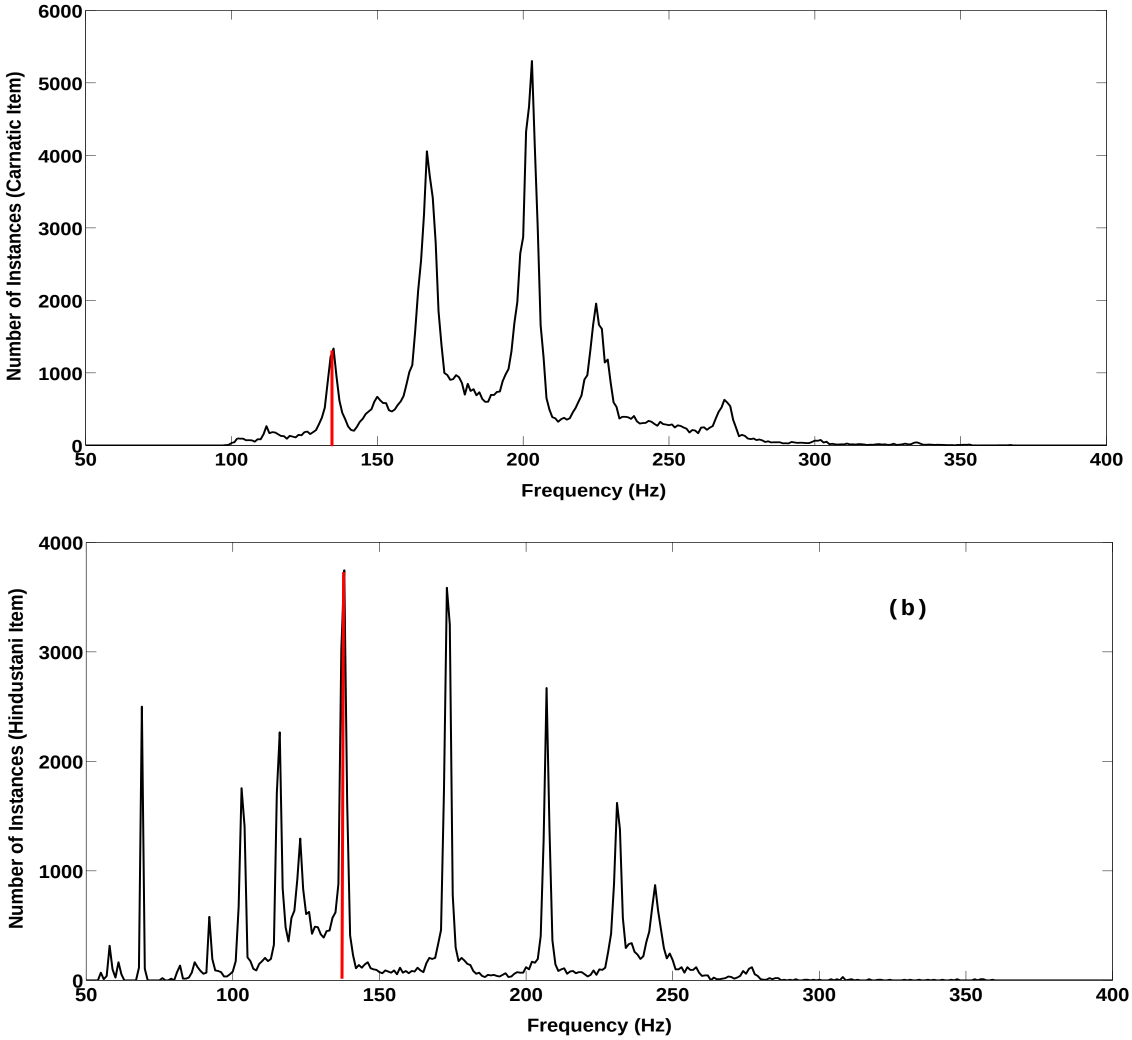

Figure 3.1: Pitch histogram of an item of Carnatic Music (plot a) and an item of Hindustani music (plot b). The red stem indicates the tonic pitch

- Another characteristic of Indian music, is that *svaras* that constitute a melody have dissimilar inflections, which manifest as peaks with varying bandwidth in the pitch histogram. This characteristic is even more prominent in Carnatic music than in Hindustani music as the melody is replete with *gamakas*. This is clearly evident from Figure 3.1. It was observed that, amongst the inflected *svaras*, the tonic as well as the fifth, in all the octaves are inflected less. The peaks corresponding to these notes have narrower bandwidths in the pitch histogram relative to the other *svaras*. In fact these *svaras* are referred to as *achala svaras* as they admit less inflection and have a fixed position.

In order to further illustrate the above mentioned properties, the Carnatic music item for which the pitch histogram is depicted in Figure 3.1, was manually transcribed. In order to obtain the pitch values corresponding to each *svara* from the audio and analyse the distribution of these pitch values, the following annotation procedure was followed.

- Pitch was extracted for the entire item.
- *Tāla* cycles were manually annotated.
- The *tāla* cycles were then synchronized with the notation semi-automatically.
- Pitch values corresponding to each *svara* were obtained and respective histograms were computed.

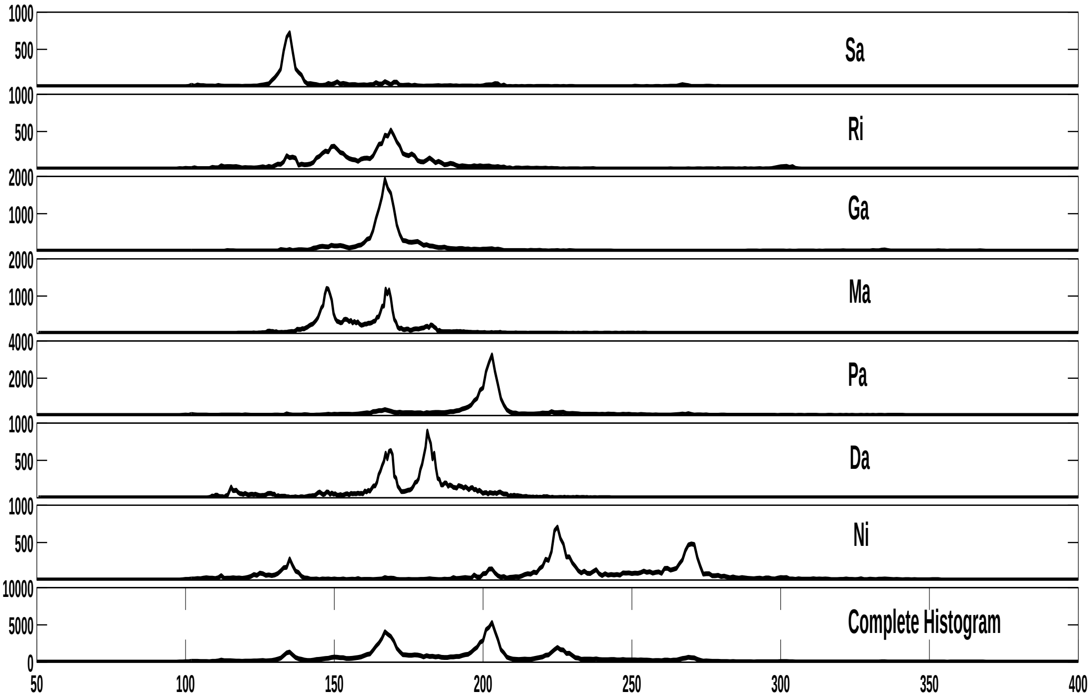

Figure 3.2: Pitch histograms for each of the manually transcribed svaras along with the histogram for the complete item

As can be seen in Figure 3.2, the *svaras* exist as pitch regions, with *svaras Sa* and *Pa* less inflected and with peaks at the theoretical pitch value. It can also be seen that the peaks indicating the *svaras* of the melody as well as the bandwidth information of the *svaras* are not clearly visible in the pitch histogram of the entire item. These observations imply that in order to identify the tonic pitch value, it is essential to resolve and pick peaks from a pitch histogram that appears continuous. Given that the tonic will invariably have a prominent peak, further analysis of peaks and their bandwidths might suffice to identify tonic.

In Ranjani *et al.* (2011) an attempt was made to use these characteristics to identify tonic for Carnatic music. A Semi-Continuous Gaussian Mixture Modeling (SC-GMM) based technique was employed to quantify the bandwidth information. In order to apply SC-GMMs, it is assumed that Carnatic music follows just intonation system of tuning and peaks corresponding to the notes are presumed to be at theoretical positions as suggested by the just intonation system. It was shown in Serra *et al.* (2011) that though Carnatic music tends towards just intonation, the peaks are not necessarily at the theoretical positions. It is also shown that Hindustani music tends more towards equi-temperament tuning system.

In this work, we propose a knowledge based signal processing technique to process pitch histograms, in order to identify the tonic pitch. We propose processing the pitch histogram using the group delay function in order to resolve peaks as well as preserve bandwidth information. No assumptions are made regarding the tuning system followed by the music repertoires. Several strategies are developed to use the group delay processed histograms for identifying the tonic pitch. Given that the relative differences in inflections is more prominent in Carnatic music, histogram processing techniques and strategies to identify the tonic were first experimented and developed using items of Carnatic music. These techniques and strategies were then used for Hindustani music also.

In the following Section, we review the group delay function and its properties before illustrating the benefits of processing pitch histograms using the group delay function.

3.2 Group delay function

The group delay function is defined as the negative derivative of the phase spectrum of a given signal. Using group delay functions, it has been shown that the phase spectrum, which generally appears to be noisy owing to being wrapped, can be gainfully processed to extract useful information. Group delay functions (Yegnanarayana *et al.*, 1984) have been extensively studied in the context of speech processing, for both formant and pitch estimation (Murthy and Yegnanarayana, 1991; Rajan and Murthy, 2013). Group delay based features have also found application in speaker and speech recognition systems (Hegde *et al.*, 2007; Bozkurt *et al.*, 2007; Padmanabhan and Murthy, 2009).

One of the flavors of group delay processing, is the minimum phase group delay function derived from the magnitude spectrum of a speech signal. The minimum phase group delay function has been shown to be useful in applications such as spectral estimation and formant extraction (Yegnanarayana and Murthy, 1992; Murthy and Yegnanarayana, 1991). It has also been shown that the minimum phase group delay function can be used to process any positive function (Nagarajan *et al.*, 2003*a*). In Nagarajan *et al.* (2003*a*) it was shown that such a processing technique can be used on the short term energy of a speech utterance for the purpose of segmenting continuous speech into syllable like units, leading to its usage in applications like speech recognition and synthesis systems (Rao *et al.*, 2005).

It is the minimum phase group delay function that is used for processing the pitch histograms in this work. Before going into the details of such a processing technique, a short review of deriving minimum phase group delay function from the magnitude spectrum is presented in Section 3.2.1. Some of the known properties of the minimum group delay function so obtained are illustrated. The procedure to derive the minimum phase group delay function for a pitch histogram is then detailed.

3.2.1 Minimum phase group delay function

Let $|X(e^{j\omega})|$ be the magnitude spectrum of a real signal and $|X(e^{j\omega})|^2$ be the squared magnitude spectrum. The magnitude spectrum and the squared magnitude spectrum of any real signal satisfies the symmetry property, i.e.,

$$\begin{aligned} |X(e^{j\omega})| &= |X(e^{-j\omega})| \\ |X(e^{j\omega})|^2 &= |X(e^{-j\omega})|^2 \end{aligned} \tag{3.1}$$

Inverse Fourier transform of such a symmetric positive real function will yield a two sided real signal. The causal portion of the resulting signal is a minimum phase signal (Nagarajan *et al.*, 2003*b*). The reasoning behind the causal portion being a minimum phase signal, as has been illustrated in detail in Nagarajan *et al.* (2003*b*), is as follows:

The squared magnitude response $|X(e^{j\omega})|^2$ can be written as,

$$\begin{aligned}|X(e^{j\omega})|^2 &= X(e^{j\omega})X^*(e^{j\omega}) \\ &= X(z)X^*(1/z^*)|_{z=e^{j\omega}}\end{aligned} \tag{3.2}$$

let

$$C(Z) = X(z)X^*(1/z^*) \tag{3.3}$$

From Equation 3.3, it is evident that for a pole or a zero outside the unit circle, $C(z)$ will have a conjugate reciprocal pair inside the unit circle. Applying inverse Z-transform on $C(z)$ will result in a symmetric signal with the causal part corresponding to the minimum phase signal represented by poles and zeros inside the unit circle in the Z-domain.

The same result is obtained on performing the IDFT of $|X(e^{j\omega})|^2$, i.e

$$\begin{aligned}IDFT(|X(e^{j\omega})|^2) &= IDFT(X(e^{j\omega})X^*(e^{j\omega})) \\ y[n] &= conv(x[n], x[-n])\end{aligned} \tag{3.4}$$

where $conv$ in Equation 3.4 stands for the convolution operation between $x[n]$ and $x[-n]$. Thus $|X(e^{j\omega})|^2$ can be represented as the Fourier transform of the autocorrelation of some causal sequence. The causal portion of $y[n]$ is a minimum phase signal. The derivative of the Fourier transform phase of the causal portion results is the minimum phase group delay function. If the windowed causal portion of $y[n]$ is $y^{'}[n]$, group delay function $\tau(e^{j\omega})$ is given by

$$\tau(e^{j\omega}) = -\partial(arg(DFT(y^{'}[n])))/\partial\omega \tag{3.5}$$

Deriving minimum phase group delay response for any positive function

The magnitude spectrum is a positive symmetric function. Therefore, techniques applied for deriving the minimum phase group delay from the magnitude spectrum can be

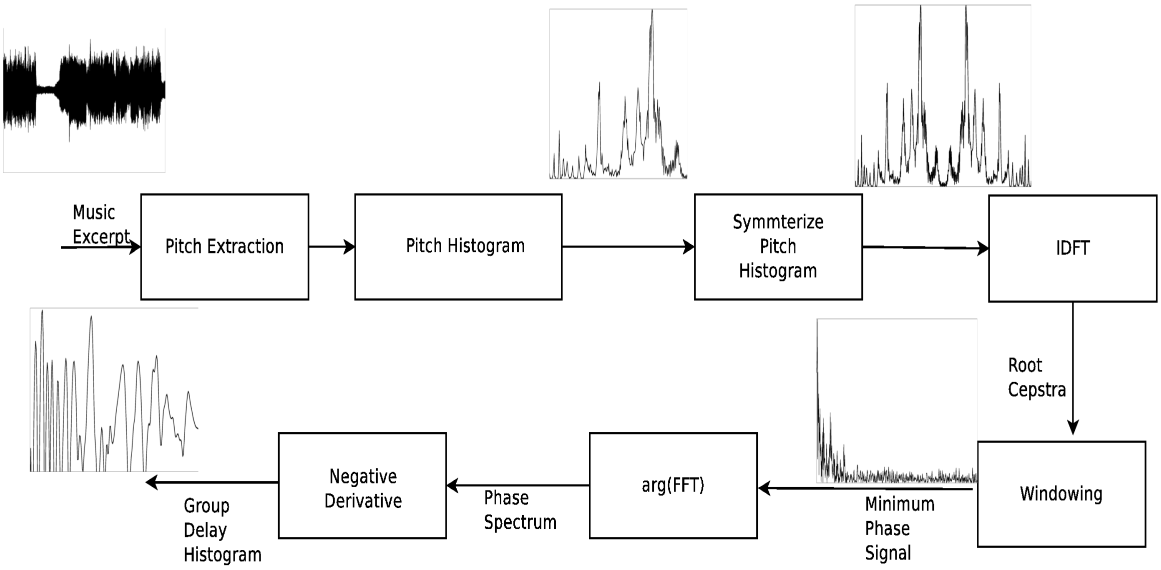

Figure 3.3: Block digram for deriving minimum phase group delay function from pitch histograms

applied to any positive symmetric function. For example, the short term energy contour was processed using the group delay function in Nagarajan *et al.* (2003*a*) . A procedure which is similar to the algorithm used for syllable segmentation in Nagarajan *et al.* (2003*a*) is used in this work. The exact procedure adopted is as follows (block diagram in Figure 3.3)

1. In order to treat the pitch histogram as a squared magnitude spectrum, a symmetric histogram generated by lateral inversion about the Y-axis is appended to the pitch histogram. If the original pitch histogram was made up of N bins, it is now extended to 2N-1 bins, with its symmetric counterpart from bin N+1 to 2N-1. Let the symmetric histogram be $P[k]$.

2. The causal portion of the inverse Fourier transform of a power function has minimum phase properties (Nagarajan *et al.*, 2003*b*). In order to extract the minimum phase signal, inverse discrete Fourier transform $IDFT(P[k])$ is computed. Let the resultant sequence be $p[n]$.

3. Minimum phase signal is then extracted by windowing the causal portion of $p[n]$ with a Hamming window of size N.

4. If $c[n]$ is the windowed causal portion of $p[n]$, the minimum phase group delay $\tau[k]$ is estimated as given in Equation 3.6. Δ denotes first difference and arg denotes unwrapped phase.

$$\tau[k] = -\Delta arg[DFT(c[n])] \tag{3.6}$$

5. $\tau[k]$ is the group delay processed histogram, henceforth referred to as the GD histogram.

3.2.2 Properties of minimum phase group delay function

In speech analysis, the group delay function of this minimum phase signal has found several applications. The additive properties of the group delay function as will be explained using the next set of equations, has been extensively exploited in these works.

For a given system $H(e^{j\omega})$, let the magnitude spectrum $|H(e^{j\omega})|$ be given, where $H(e^{j\omega})$ is defined as

$$H(e^{j\omega}) = H_1(e^{j\omega})H_2(e^{j\omega}) \tag{3.7}$$

$$|H(e^{j\omega})| = |H_1(e^{j\omega})||H_2(e^{j\omega}| \tag{3.8}$$

$$arg(H(e^{j\omega})) = arg(H_1(e^{j\omega})) + arg(H_2(e^{j\omega})) \tag{3.9}$$

Then the group delay function, which is defined as the negative derivative of phase is given by

$$\begin{aligned} \tau_h(e^{j\omega}) &= -\partial(arg(H(e^{j\omega})))/\partial\omega \\ &= -\partial(arg(H_1(e^{j\omega})))/\partial\omega - \partial(arg(H_2(e^{j\omega})))/\partial\omega \\ &= \tau_{h1}(e^{j\omega}) + \tau_{h2}(e^{j\omega}) \end{aligned} \tag{3.10}$$

where $\tau_{h1}(e^{j\omega})$ and $\tau_{h1}(e^{j\omega})$ correspond to the group delay of $H_1(e^{j\omega})$ and $H_2(e^{j\omega})$ respectively. From Equations 3.7 and 3.10, we can see that multiplication in the magnitude spectrum domain becomes additive in the group delay domain. This property of the group delay function was shown to be very useful in speech processing, especially in spectral estimation and formant tracking. In Yegnanarayana (1979), it was first shown that by modeling the vocal tract as a cascade of resonators, group delay spectrum of such a setup behaves like the squared magnitude response in the vicinity of the resonances. At the same time, due to the additive nature of the group delay spectrum, formants are shown to be better resolved. This resolving ability of the group delay function is also used to obtain syllable boundaries by processing the short term energy function in Nagarajan *et al.* (2003*a*). In the following section we process the pitch histograms using group delay functions. We illustrate some of the advantages of such a processing technique from the perspective of tonic identification.

3.3 Processing pitch histograms using group delay function

In this work we propose processing the pitch histogram, a positive function, using group delay function. Group delay processing the pitch histogram was initially explored with the intention of using the resolving ability of the group delay function to resolve peaks that are in close vicinity in the pitch histogram. On processing the pitch histograms using the group delay function, it was observed that the group delay function along with resolving the peaks also accentuated peaks with narrow bandwidth, generally peaks corresponding to *svaras Sa* and *Pa*, a favourable trait from the perspective of tonic identification. Figure 3.4 shows a typical outcome on group delay processing a pitch histogram. Figure 3.4a is the pitch histogram constructed on an item of Carnatic music and Figure 3.4b is the GD histogram. The *svaras Sa* and *Pa* are indicated using red and blue stems respectively.

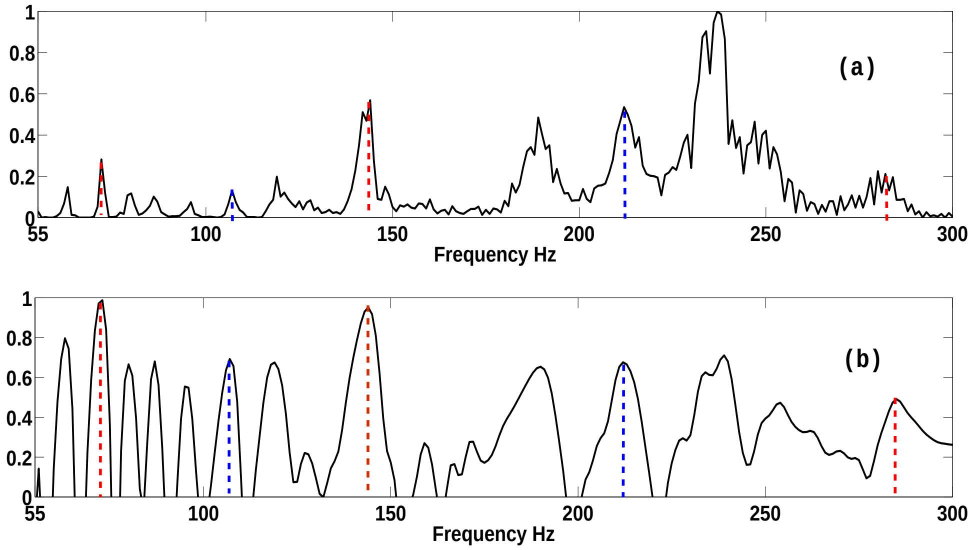

Figure 3.4: Pitch and Group Delay Histograms (a) Pitch histogram of a 3 minute Carnatic music item. (b) The GD histogram. Red stems indicate the bin values of *svara Sa* and blue stems indicate the *svara Pa*

In order to explain these interesting and useful characteristics of group delay processing, we propose characterizing pitch histograms as the response to a set of resonators in parallel. Similar to short term energy in Nagarajan *et al.* (2003*a*), the pitch histogram can be assumed to be the squared magnitude $|H(e^{j\omega})|^2$ response of a system $H(e^{j\omega})$. At the same time, unlike the magnitude spectrum in speech (Yegnanarayana, 1979; Murthy and Yegnanarayana, 1991) or short term energy Nagarajan *et al.* (2003*a*), where all formants/peaks are generally thought of as responses to constant Q filters from a cascade of resonators, some of the peaks in the pitch histogram can be characterized by both a large gain and a large bandwidth. This is primarily because the gain corresponds to the frequency of occurrence of a particular note in a given melody. To mimic the pitch histogram and to achieve a particular peak-gain at resonance, $H(e^{j\omega})$ can be thought of as the squared magnitude response of resonators connected in parallel with different gains, rather than a cascade of resonators:

$$H(z) = \sum_{k=1}^{M} \frac{\alpha_k}{(1 - d_k z^{-1})} \tag{3.11}$$

with $d_i = r_i e^{(j\omega_i)}$, r_i is the radius of the pole and ω_i is the angle of the pole.

To justify this model, we estimate the squared magnitude response of two resonators in parallel given by Equation 3.12.

$$H(z) = \frac{\alpha_1}{(1 - d_1 z^{-1})} + \frac{\alpha_2}{(1 - d_2 z^{-1})} \tag{3.12}$$

with poles at $d_1 = 0.9e^{(j\pi/4)}$, $d_2 = 0.7e^{(j\pi/3)}$ and different values of α_1 and α_2.

The angular frequency of the pole corresponds to the location of the peak, while the radius of the pole relates to the bandwidth of the pole. As the poles are close to each other, it can be seen in column 1 of Figure 3.5, that the two peaks are not resolved.

In this parallel setup, the width of the peak does not necessarily imply a decrease in gain. A gain term is included in Equation 3.12 to account for the height of the peak.

The Z-transform of the parallel connection becomes:

$$H(z) = (\alpha_1 + \alpha_2) \frac{1 - c_1 z^{-1}}{(1 - d_1 z^{-1})(1 - d_2 z^{-1})} \tag{3.13}$$

with $c_1 = \dfrac{\alpha_2 d_1 + \alpha_1 d_2}{\alpha_1 + \alpha_2}$.

The first two columns of Figure 3.5 show the squared magnitude spectrum and pole-zero plot respectively, for the system given by Equation 3.13. In this system, α_1 is set to 1, while α_2 is varied. Since the model assumed is a parallel connection, zeroes are introduced as indicated in Equation 3.13. It can be seen that as α_2 increases, the zero moves towards the pole of the other resonator, thus annihilating the pole of the second resonator.

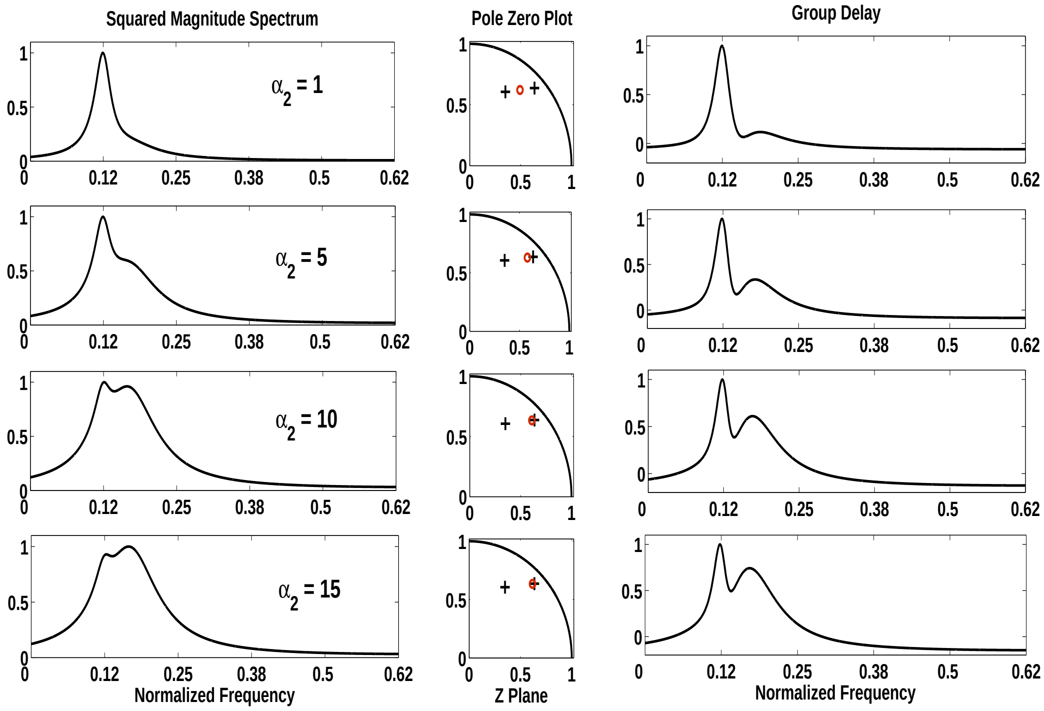

Figure 3.5: Column 1 - Squared magnitude response, Column 2 - Pole zero plot, Column 3 - Group delay plot, for a two pole system with $\alpha_1 = 1$ and varying α_2

Pitch histograms exhibit similar behavior, i.e., peaks at various bin values can have varying heights and bandwidths. This primarily depends on the frequency of the note and the inflection that is associated with the note. Pitch histograms can therefore be characterized as a set of resonators in parallel with an appropriate choice of pole angles (ω_i), gains (α_i) and pole radii (r_i) respectively.

Clearly, estimating position of peaks and their bandwidths even for the synthetic spectrum of Figure 3.5 is non-trivial. Given that pitch histograms in Carnatic music show similar traits, tonic identification from pitch histograms of Carnatic music is difficult. We conjecture that processing such histograms using the group delay function will aid in tonic identification. Having shown that the histogram can be assumed to be the squared magnitude response of the system in Equation 3.13, we will now illustrate some of the properties of the group delay response of such a setup which will explain the outcomes of group delay processing observed in Figure 3.4.

The system in Equation 3.13 modeling a 2 peak histogram can be generalized as (Oppenheim and Schafer, 1990),

$$H(z) = G\frac{\prod_{k=1}^{M-1}(1 - c_k z^{-1})}{\prod_{k=1}^{M}(1 - d_k z^{-1})} \tag{3.14}$$

where M corresponds to the number of peaks in the histogram and G is the gain. The squared magnitude spectrum of this system evaluated on the unit circle is given by:

$$|H(e^{j\omega})|^2 = (G)^2\frac{\prod_{k=1}^{M-1}(1 + |c_k|^2 - 2Re\{c_k e^{-j\omega}\})}{\prod_{k=1}^{M}(1 + |d_k|^2 - 2Re\{d_k e^{-j\omega}\})} \tag{3.15}$$

For the system given in Equation 3.14, the group delay function is defined as the negative derivative of the continuous phase function of the system (arg$[H(e^{j\omega})]$), i.e

$$\tau(\omega) = -\frac{d}{d\omega}arg[H(e^{(j\omega)})] \tag{3.16}$$

The group delay function of the system given in Equation 3.14 will take the form (Oppenheim and Schafer, 1990)

$$\begin{aligned}\tau(\omega) = &\sum_{k=1}^{M}\frac{|d_k|^2 - Re\{d_k e^{-j\omega}\}}{1 + |d_k|^2 - 2Re\{d_k e^{-j\omega}\}} \\ &- \sum_{k=1}^{M-1}\frac{|c_k|^2 - Re\{c_k e^{-j\omega}\}}{1 + |c_k|^2 - 2Re\{c_k e^{-j\omega}\}}\end{aligned} \tag{3.17}$$

where Re indicates the real part.

For any pole with $d_i = r_i e^{j\omega_i}$, as $\omega \to \omega_i$, the group delay function takes the form,

$$\frac{|d_i|^2 - Re\{d_i e^{-j\omega}\}}{1 + |d_i|^2 - 2Re\{d_i e^{-j\omega}\}} \approx \frac{K_i}{1 + |d_i|^2 - 2Re\{d_i e^{-j\omega}\}} \tag{3.18}$$

where K_i is some constant.

It can be seen from Equations 3.15, 3.17, and 3.18 that the group delay behaves like the squared magnitude response of the resonators at the resonance frequencies (Yegnanarayana, 1979). The additive property of group delay is evident from the third column of Figure 3.5. While in the squared magnitude spectrum (column 1), the peaks (pole angles) are not resolvable for small values of α_2 and bandwidths (pole radius) are difficult to discern at high values of α_2, resolution of the peaks and bandwidth information are better preserved in the group delay extracted (Column 3 of Figure 3.5) .

To further illustrate the characteristics of the group delay function, the squared magnitude, pole-zero plot and group delay response for three resonators in parallel are shown in Figure 3.6. α_1 and α_3 are fixed at $\alpha_1 = \alpha_3 = 1$, with $d_1 = 0.9e^{(j\pi/4)}$, $d_2 = 0.7e^{(j\pi/3)}$ and $d_3 = 0.9e^{(j4\pi/3)}$. Square magnitude and group delay response are shown for various values of α_2. It is interesting to see that in the group delay column, the peak due to the pole with a larger pole radius of 0.9 at d_3, dominates even at large values of α_2. This ability of the group delay function to accentuate peaks with narrow bandwidths is indeed crucial for tonic identification.

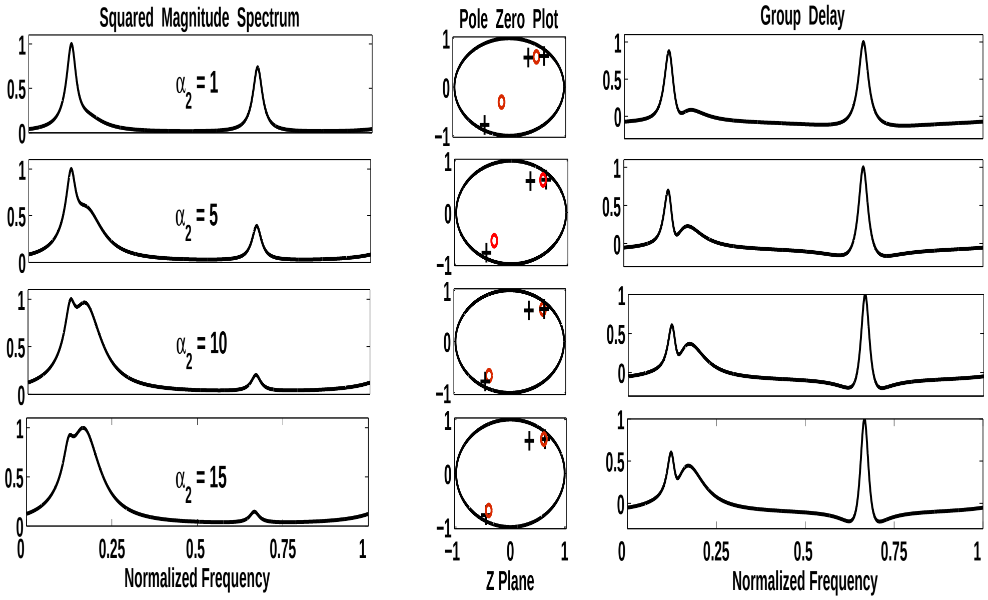

Figure 3.6: Column 1 - Squared magnitude response, Column 2 - Pole zero plot, Column 3 - Group delay plot, for a three pole system with $\alpha_1 = \alpha_3 = 1$ and varying α_2

Figure 3.7 shows the group delay processed histogram. The histogram is normalized to have a maximum height of 1. The pitch histogram of the transcribed *svaras* and the pitch histogram for the whole item is also shown in Figure 3.7. That is, the first seven plots are the histograms computed using the pitch values of the transcribed *svaras* and the eigth plot is the histogram computed using the histogram of the complete item. We can observe that the group delay processed histogram is not only smooth but also the peaks are sharper. Also, peaks corresponding to that of the tonic and fifths (peaks with narrow bandwidth), are emphasized in the GD histogram. The accentuation however was observed to be sharper in the case of Carnatic music when compared with Hindustani music, owing to larger variations in the bandwidths due to the presence of *gamakas*. Methods to identify tonic utilizing these features of group delay processing are proposed in the following section.

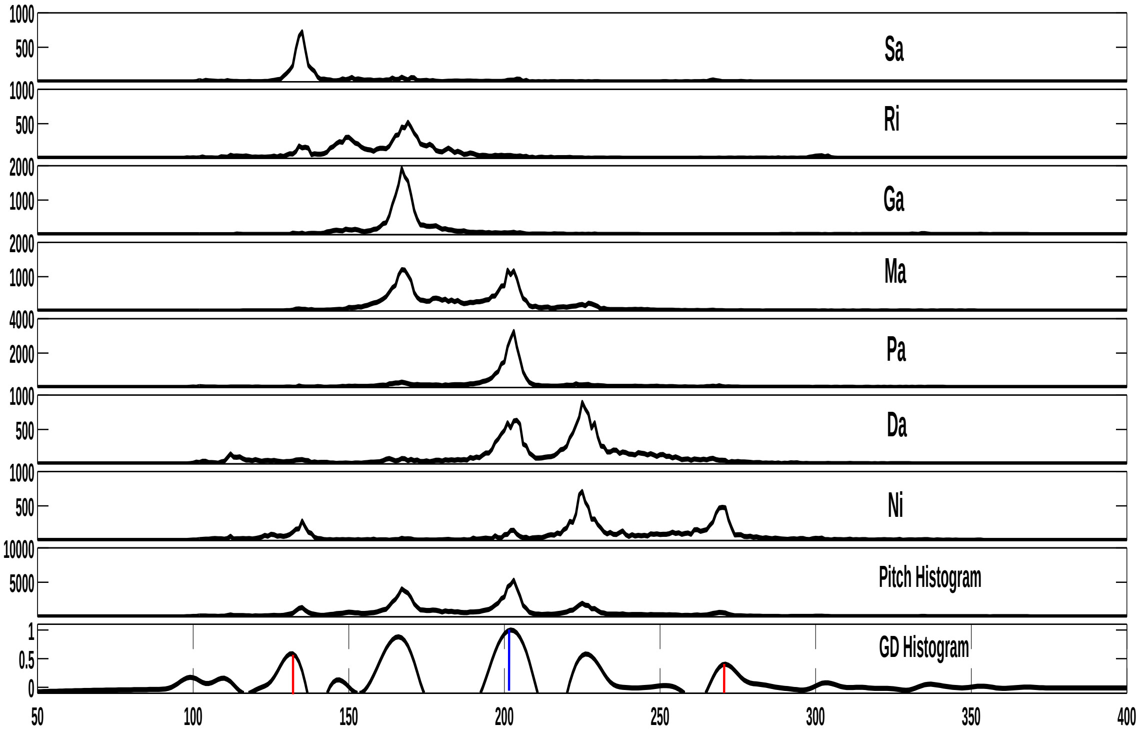

Figure 3.7: Pitch histograms for each of the manually transcribed svaras along with the histogram for the complete item and the group delay processed histogram. The group delay processed histogram is normalized to have a maximum height of 1. Red stems indicate the *Sa* and the blue stem indicates the *Pa*

3.4 Methods to identify the tonic using GD histograms

In this section, we attempt to use the pitch histogram, and the GD histogram representation of an item of music to identify the tonic. Three methods are proposed in this section. Each method attempts to utilize some specific properties of the music, as is manifested in a histogram. While the first method is for Carnatic music alone, the second and third methods are for both Carnatic and Hindustani music.

3.4.1 Method 1 - concert based method

The database for Indian music is generally in the form of audio CDs or recordings of concerts. A concert or an audio CD can be considered as a unit by itself. A Carnatic music concert or an audio CD, consists of a number of items in different *rāgas*. The number of items can range from 5 to as many as 20 items. The *rāgas* are seldom repeated. Although the *rāgas* are different, the tonic in which they are rendered is kept constant. In addition to this, every *rāga*, contains the *Sa*, along with a subset of the 12 semitones which make up the *rāga*. The basic idea behind the approach proposed in this method, is to identify the tonic of every concert. To detect the tonic of the concert the following algorithm is used:

1. Compute the GD histograms of all individual items, namely, $GDP_i, 1 \leq i \leq n$ in a concert, where n corresponds to the number of items in a concert.
2. Compute $\prod_{i=1}^{n} GDP_i$. That is, the histograms computed using smaller segments are multiplied bin-wise one single histogram.

Since the rāgas of each item are different, with *svara Sa* being present across all items, the peak corresponding to that of the tonic must dominate in Step 2. As an example, Figure 3.8 shows the GD histograms for four items (plots *a*, *b*, *c*, *d*) of the same concert. The fifth row, i.e plot *e* in Figure 3.8 is the product of the four GDP_is evaluated on the four items performed in the concert. The dominant peak is the tonic used in the concert. It must be noted, that for an individual item, the most dominant peak might not be the tonic (first row in Figure 3.8), but other notes may dominate the individual histogram. But with the *Sa* present in the percussion and drone, and with

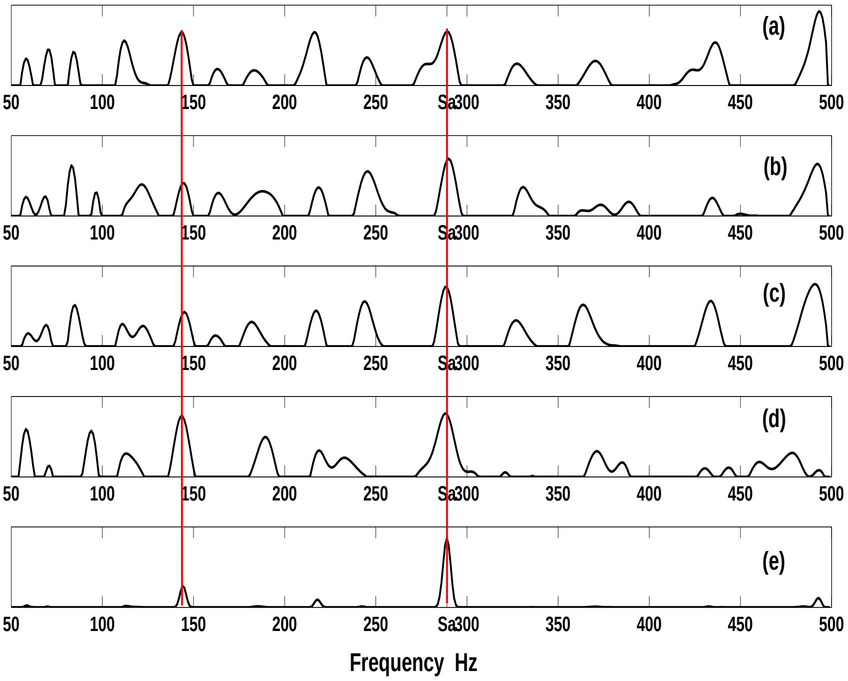

Figure 3.8: Concert method - plots a,b,c and d correspond to the group delay histograms for 4 items of a concert. Plot e is the bin-wise product of the above 4 histograms

every *rāga* having the *Sa*, a prominent peak at *Sa* in the histogram and GD histogram is guaranteed. Tonic identification is thus reduced to determining the frequency of the peak that has the maximum value.

On extending the proposed method to Hindustani music, it was observed that the method was not very accurate. On analyzing the concerts/audio CDs, it was observed that even though the tonic is kept constant in a concert/audio CD, the number of pieces in a concert are far fewer (2 - 4 items). In the event of one of the *svaras*, other than the *Sa*, dominating a particular item of the concert, resulted in erroneous identification of tonic for the complete concert/audio CD.

3.4.2 Method 2 - template matching

Although the previous method can be used for normalizing pitch values for a large number of items in a concert, it will not work when only individual items are available. The objective in this method as that of the next method is to perform tonic identification when provided with individual items.

In this method, the less inflected nature and the fixed ratio between *Sa* and *Pa* are exploited ($Pa = 1.5 \times Sa$). This method is comparable to that of Ranjani *et al.* (2011), where they attempt to exploit the same characteristics using SC-GMM. While in Ranjani *et al.* (2011) five different rules are explored, in this work, *Sa* - *Pa* - *Sa* template is used on the histogram and GD histograms.

The procedure to detect tonic is as follows:

- Compute histograms and GD histograms.
- Let f_i, $i \in [1, N]$, correspond to the frequencies of the N peaks of the histogram.
- Let L be a vector such that:
  $L[k] = v_i$ for $k \in [f_i ... f_N]$, v_i is the height of the peak at f_i and $L[k] = 0, elsewhere$
- Each peak location is a candidate *Sa*. Now let f_j be the frequency of a candidate *Sa*, say S_j. $j \in [1, N]$.
- Given the frequency of S_j, the expected frequencies of *Sa* and *Pa* across the 3 octaves under consideration are

$$\begin{bmatrix} 0.5(f_j) & 0.75(f_j) & f_j & 1.5(f_j) & 2(f_j) & 3(f_j) \\ S_{j_{lower}} & P_{j_{lower}} & S_j & P_j & S_{j_{higher}} & P_{j_{higher}} \end{bmatrix}$$

$$E = [0.5\ 0.75\ 1\ 1.5\ 2\ 3]f_j$$

- Let T_j be the template vector for a test piece such that: $T_j[k - \delta : k + \delta] = 1$ for $k \in E$; δ allows for a leeway of δ around the expected peak. $T_j[k] = 0$ elsewhere.
- $C_j = L^T T_j$
- $tonic = \underset{j}{argmax}\ \ C_j, j \in [1, N]$

The procedure detailed above attempts to fit a template for every peak of the histogram, each of which is treated as a candidate *Sa*. With *svaras Sa and Pa* at fixed ratios, indicated by accentuated peaks due to group delay processing, the template prescribed would best match the tonic pitch. This feature is illustrated in Figure 3.9. The first plot is the GD histogram. The second and third plot show the template matching procedure for two candidate peaks, indicated by the black strip. The tonic pitch being 200 Hz in this case, a better template match is obtained in the second plot when compared to the third plot. The template is indicated using blue strips.

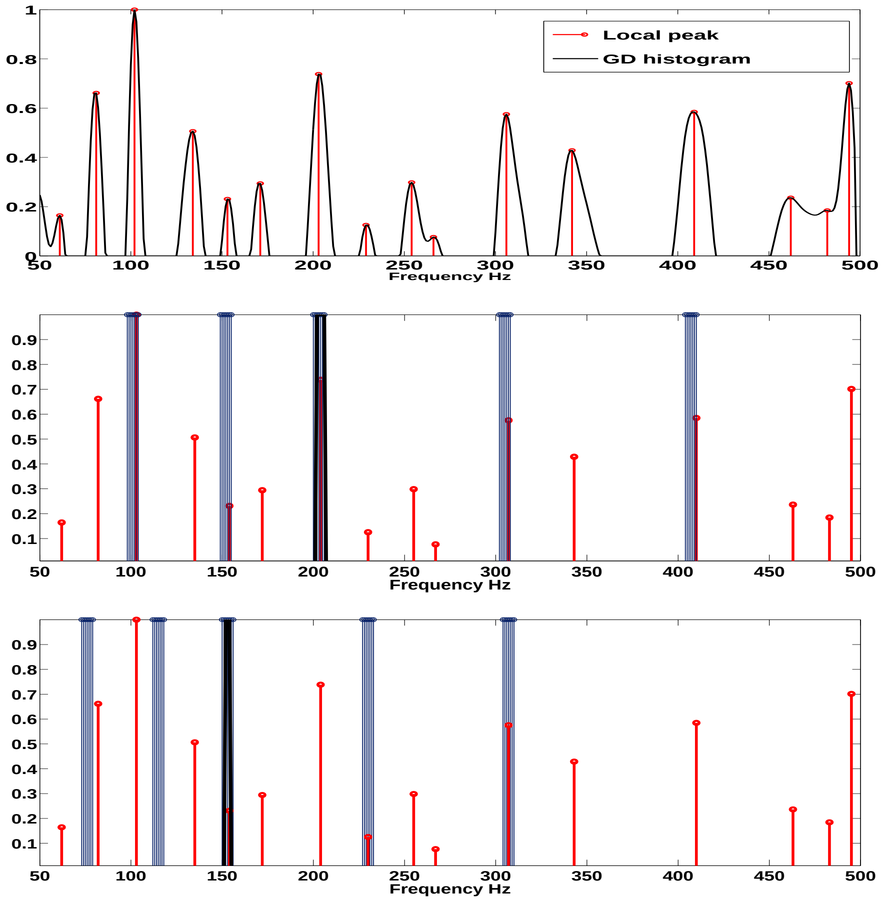

Figure 3.9: Illustration of the template matching procedure. Plot 1 shows the local peaks in Gd histograms. Plot 2 and 3 show the template matching procedure for two different cases. The black strip represents the local peak assumed as the tonic and the blue strip represents the corresponding template. Plot 2 being the correct estimate of the *Sa*, a better template match is obtained.

Vādi - Samvādi template

An attempt was made in Bellur *et al.* (2012) to extend the template matching method to factor in the *rāga* information for tonic identification. It was shown that instead of using the *Sa-Pa-Sa* template, a customized template created using the *rāga* information can be used to for tonic identification. It was observed in Bellur *et al.* (2012), for Hindustani music in particular, that every *rāga* has *vādi* and *samvādi svaras*, which are essentially *svaras* that are most dominant in a given *rāga*. In the event of the *rāga* identity being known, it was proposed in Bellur *et al.* (2012), that *Sa-vādi-samvādi-Sa* template can be used to obtain higher accuracies.

The use of the *vādi-samvādi* has not been further explored in this thesis owing to the nature of the database being used in this work. While in Bellur *et al.* (2012), the results are reported on a relatively small database of Hindustani music, it was found that compiling the required metadata for the database described in Chapter 2 is non-trivial. In particular, compiling the *vādi* and *samvādi* information for every *rāga* in the database was found to be a difficult task.

3.4.3 Method 3 - segmented histograms

In the template matching method, the assumption is that the peak at which the template fits best is the *Sa*. There are a few drawbacks in this method. Since the template is basically using the fact that the *Pa* is $1.5\times$ *Sa* with respect to the *Sa*, there might be a perfect template fit for another set of notes with the same template. It is also possible that for excerpts with a faint drone and *svara Pa* not being part of the melody, there might not be a peak at *svara Pa*, leading to erroneous tonic identification.

As an alternate to the template matching method, another method for tonic identification was devised using *piece-wise* histograms. Figure 3.10, shows the histogram and GD histogram of a single four minute Carnatic piece. The note marked "*" on the X axis is most frequented, whereas "+" is the tonic. Global peak picking would have resulted in an erroneous identification of tonic. As an attempt to identify tonic, even in the case of it not being the most dominant note even in the GD histogram, a given music item is segmented into units of duration of $\approx$ 1 minute. The histograms and GD histograms are calculated on the pitch extracted from the segmented items. As mentioned before, the presence of the drone and the *mṛdaṅgaṁ* or the *tablā* ensure that the segmented histograms will always show a local peak at the tonic, which might not be the case for the other svaras of the Rāga. In Figure 3.11, plots 1-4 show the histograms and GD histograms computed on the segmented pieces. It can be seen that a local peak at the *Sa* is always present. Figure 3.11 also illustrates the ability of the group delay function to accentuate peaks with narrow bandwidth. The *Sa* peak gets emphasized in each of the segmented GD histograms. Similar to the procedure adopted in the concert based method, the bin-wise product of the segmented GD histograms is computed. This is followed by picking the global peak to determine the tonic.

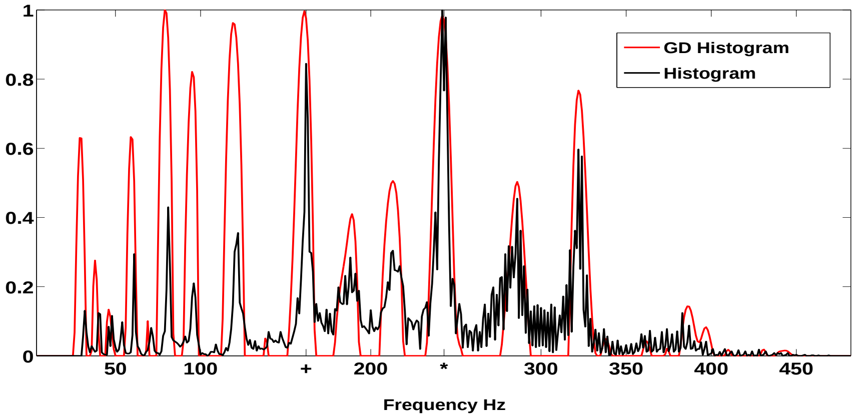

Figure 3.10: Histogram and GD histogram of a single four minute Carnatic piece. The histogram bin marked with a "*" on the X axis is the most frequented note, whereas the bin marked "+" is the *svara Sa*

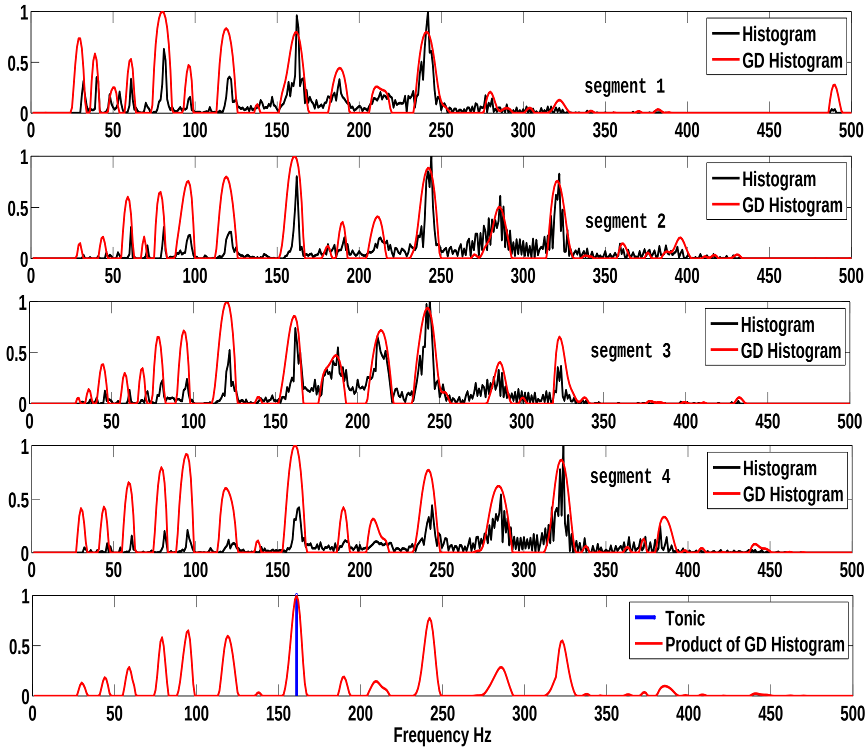

Figure 3.11: Plots 1-4 show segment-wise histogram and GD histogram. Plot 5 is the product of GD histograms in plot 1-4, with *Sa* as the global peak

3.5 Experiments and results

The methods proposed in Section 3.4 were tested on the database detailed in Section 2.3. Even though the methods in this chapter were primarily devised to use melodic cues that manifest in an item of Carnatic music, they are tested on the complete varied database. This would ensure the usefulness of the cues and the efficacy of the methods proposed for different types of data. Method 1 i.e Concert based method (CBM) is only tested on DB1, while Method 2 and 3 i.e the Template Matching Method (TMM) and the Segmented Histogram Method (SHM) are tested on DB2, DB3 and DB4. During testing, the following variations were experimented with for each of the methods.

- Determining the middle octave Sa i.e the tonic pitch: Given that the tonic for the whole databases across performers ranges from 100 Hz to 250 Hz, only the peaks within this range are considered as candidate tonic pitch values.
- Determining the *svara Sa* in any of the octaves: This is termed as tonic pitch class accuracy as defined in Salamon *et al.* (2012). This is an important criteria to test, given that pitch halving and doubling is invariably prevalent in pitch extraction algorithms. Peaks in the range 50 Hz to 400 Hz are considered as candidate *svara Sa* values.
- Determining the tonic using the metadata of the lead performer, i.e if the lead artist is male vocal, female vocal or instrumentalist. On using metadata the candidate tonic pitch is further restricted for each of the classes as follows:
  - Male artists - 100 Hz to 180 Hz
  - Female artists - 140 Hz to 250 Hz
  - Instrumental music - 120 Hz to 210 Hz

3.5.1 Performance of CBM

Table 3.1 shows the performance of CBM on using histograms and GD histograms. The tonic identified is deemed to be accurate if it is within the $2Hz$ range of the ground truth. Using CBM on the database DB1, it can be seen that close to 100% result can be obtained. Further analyzing the concerts for which the method was erroneous, it was observed that due to poor quality of the recording, the tonic pitch had changed between items. Bin-wise multiplication of the histograms nullified the tonic pitch. Hence, it can be assumed that on availability of the audio for a complete concert with a constant tonic pitch, this method can be adopted to identify the tonic pitch with very high accuracy.

It should also be noted that there is no added benefit of using the GD histogram in this particular method. It can also be seen that for this particular method, metadata regarding the type of artist is also not required and high performance is achieved without any metadata.

Table 3.1: Accuracy in %. TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic pitch class

Type	TP	TPWM	TPC
Histogram	96.15	97.44	97.44
GD Histogram	96.15	97.44	97.44

3.5.2 Performance of TMM and SHM

The Template Matching Method (TMM) and the Segmented Histogram Method (SHM) were tested on three databases DB2, DB3 and DB4. Tables 3.2 and 3.3 report the accuracy of identifying tonic using TMM and SHM respectively for each of the databases. The values are accuracy in % on estimating tonic pitch, tonic pitch class and tonic pitch with metadata are reported.

Table 3.2: Performance of TMM (% accuracy). TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic Pitch Class

Database	Histogram			GD Histogram		
	TP	TPWM	TPC	TP	TPWM	TPC
DB2	78.35	87.22	79.38	86.57	91.75	89.07
DB3	68.46	88.79	76.40	79.67	94.39	84.58
DB4	70.48	87.73	77.36	80.10	91.87	86.73

Table 3.3: Performance of SHM (% accuracy). TP – Tonic Pitch, TPWM – Tonic Pitch With Metadata, TPC – Tonic Pitch Class

Database	Histogram			GD Histogram		
	TP	TPWM	TPC	TP	TPWM	TPC
DB2	74.23	82.89	77.32	84.95	89.90	86.6
DB3	60.96	87.62	75.23	67.76	90.88	80.84
DB4	64.18	83.17	74.54	70.15	88.23	82.01

From analyzing the results in Table 3.2 and 3.3, it can be seen that TMM performs better than SHM across all databases. It can also be clearly inferred from Table 3.2 and 3.3 that processing of the histograms using the group delay, does give a considerable boost to the performance of the methods across databases.

When it comes to the three categories of accuracy measurement, it can be seen that both methods perform better in estimating the tonic pitch with the metadata, especially for databases DB3 and DB4. Given that DB2 has only Carnatic music while DB3 and DB4 comprise of Hindustani music too, the performance of the methods were analyzed individually for Carnatic and Hindustani music. Figure 3.12 and 3.13 show the results of such an analysis on DB3 and DB4 respectively.

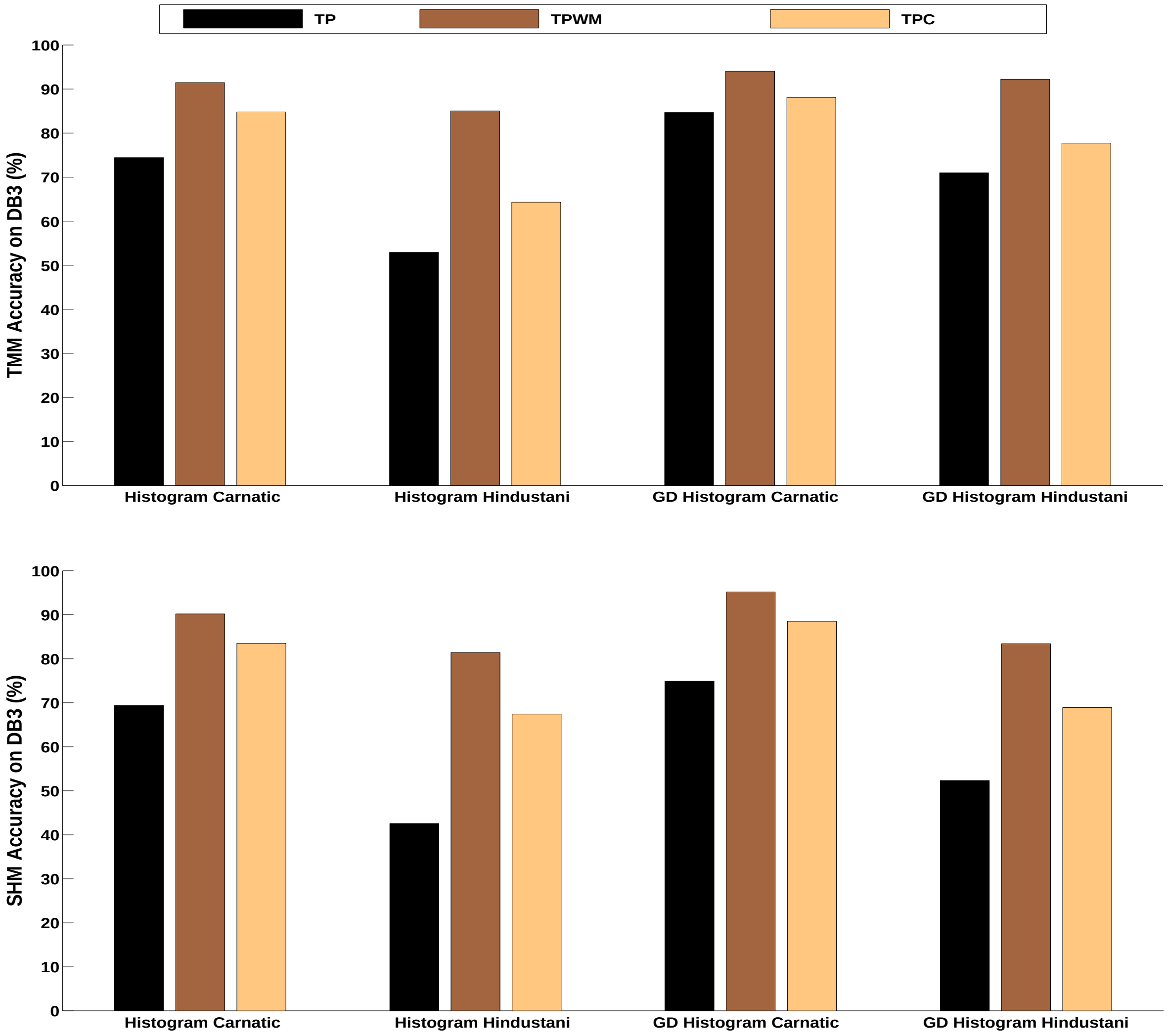

Figure 3.12: The performance of TMM and SHM on database DB3

It can be seen that the methods perform better on Carnatic music when compared with Hindustani music. As suspected the dip in performance on not using metadata is considerable for Hindustani music and relatively less for Carnatic music. Though group delay processing seems to aid both forms of music, poorer performance of the methods for Hindustani music can be attributed to less prominence of *svara Sa* when compared with Carnatic music. This can be reasoned by looking at the performance of methods on plain pitch histograms, where the success of the methods rely only on the prominence

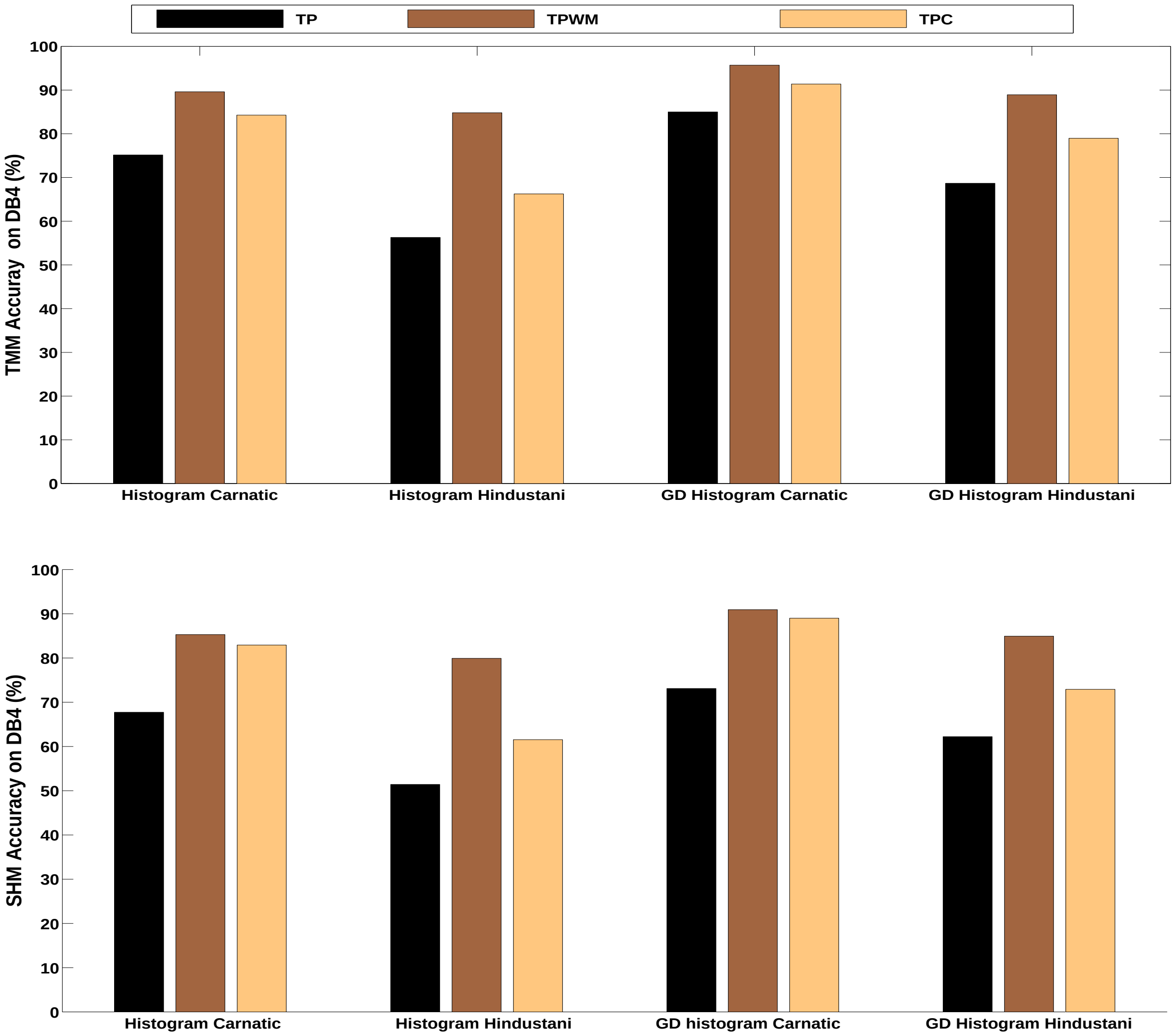

Figure 3.13: The performance of TMM and SHM on database DB4

of *Sa*. Across both databases on just using the pitch histogram, Carnatic music is better suited when compared with Hindustani music.

Another interesting observation is that the tonic pitch class (TPC) accuracy is higher when compared with tonic pitch (TP) across databases. On comparing TPC and TP accuracies for both methods, it can be observed that there is a marked difference between TP and TPC accuracies on using SHM. This can attributed to the fact that SHM relies on the simple tallest peak picking procedure. On occasions, when due to pitch halving/doubling or due to the melody being rendered, the lower/upper octave *Sa* might dominate the histogram. Given that the *Sa* is less inflected, they get emphasized by the group delay function leading to erroneous tonic pitch estimation although the tonic pitch class is accurately estimated. The usefulness of group delay is once again validated through the bar plots with accuracies around 90% achieved for all cases using GD histogram.

In order to further analyze the performances of the method, accuracies on using male vocal, female vocal and instrumental excerpts were estimated. Figure 3.14 shows the performance of the methods in identifying the tonic pitch with and without metadata using GD Histograms. It is interesting to observe that the metadata seems to be particularly important for accurate identification of the tonic pitch for female vocal excerpts. Analyzing the ground truth for female vocal, it was seen that a large number of the tonic values were above 200 Hz. Without the use of metadata, all the peaks in the range 100 - 250 Hz are considered as candidate tonic value. This results in taller peaks corresponding to *svaras* other than the tonic in the lower octave regions, arising due to melodies rendered by the accompanying instruments or pitch halving get erroneously identified as the tonic pitch. When meta data is used, the middle octave range is only considered. The performance thus improves considerably. It also seems to be clear from the bar graphs in Figure 3.14 that the performance of the methods on DB2 and DB3 is better compared to DB4. This can be attributed to the nature of database DB4. While DB2 and DB3 are made up of complete items, DB4 consists of three minute excerpts randomly picked from an item. Given that the methods rely on melodic characteristics that manifest in a complete item, performance of the methods is marginally better for databases DB2 and DB3.

3.5.3 Error analysis

In this section, an attempt is made to analyze the nature of the errors that occur in the methods proposed. In order to perform these analysis the tonic pitch estimated by the methods is quantized into one of the 12 semitones. If the tonic pitch estimated is erroneous and corresponds to the fifth, it is *svara Pa*. If the tonic pitch estimated is $\approx 1.33 \times Tonic$ it is deemed to be the *svara Ma*. Rest of the errors are categorized as 'others'. The reason for such a categorization is as follows: The drone that is used as the reference, produces along with *svara Sa*, the *svara Pa* or *Ma* depending on the melody being performed. Given that the drone is always present there are prominent peaks

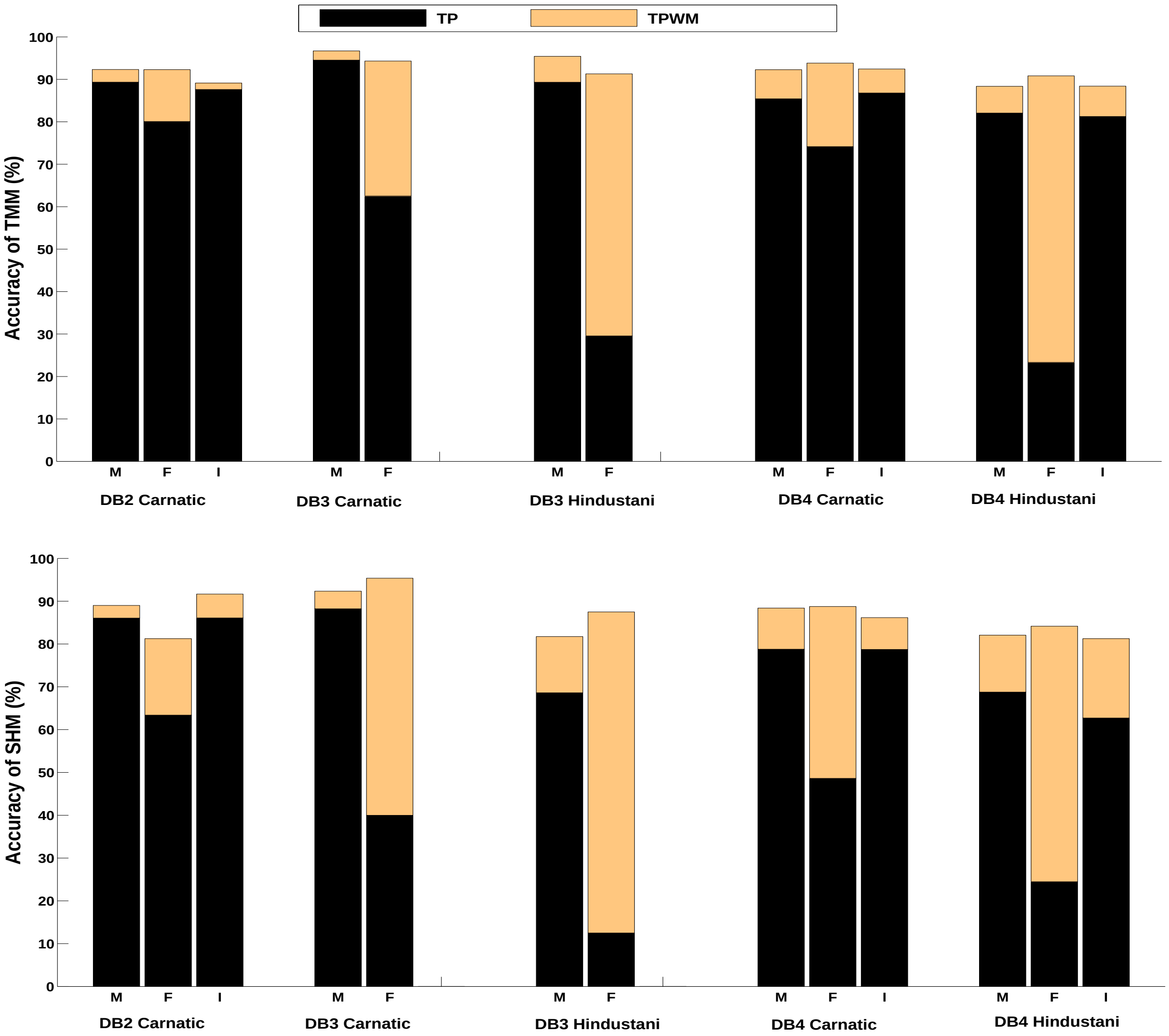

Figure 3.14: The performance of TMM and SHM on Male vocal Female vocal and Instrumental music. M – Male, F – Female, I –Instrumental

found at the bin values that correspond to *Pa* or *Ma*, which in the event of dominating a histogram (as will be seen in the next Chapter) end up getting picked as the tonic pitch. Also, as has been noted earlier in the Chapter, along with *Sa*, the *svara Pa* is also less inflected. This implies that group delay histogram ends up accentuating the peak corresponding to *Pa* also, which might lead to erroneous identification.

Another reason for categorizing *Ma* errors is that in certain cases, where the Ma has peak at the theoretical value i.e $1.33 \times Tonic$, the *Sa-Pa-Sa* template used in method TMM will also fit the *svaras Ma-Sa-Ma*. This is because where Ma is taken as the reference, *Sa* behaves like the fifth, thereby leading to erroneous identification.

It can be observed in Figure 3.15 that there is considerable reduction in the number

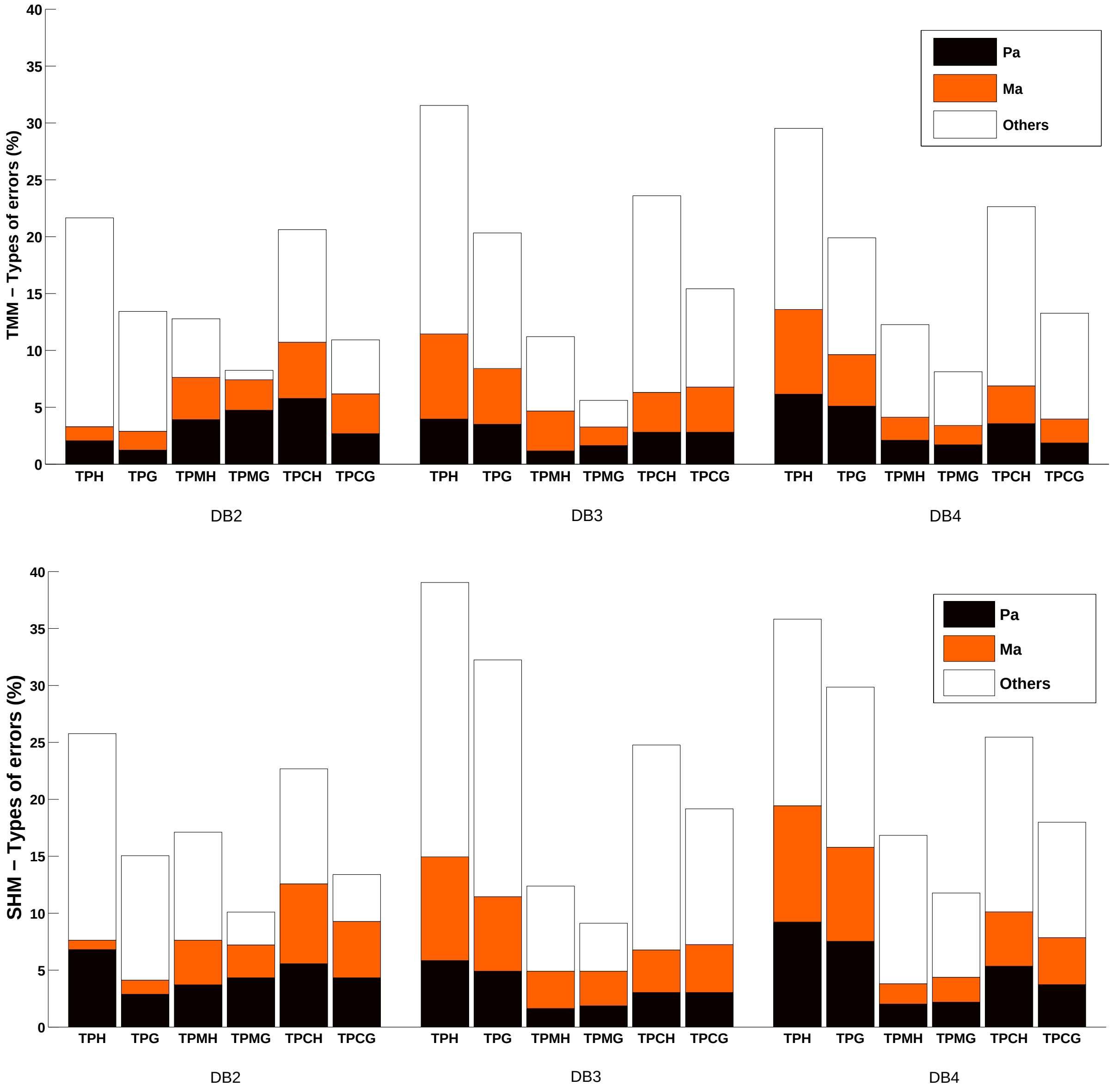

Figure 3.15: The performance of TMM and SHM on Male vocal Female vocal and Instrumental music. TP – Tonic Pitch, TPM – Tonic Pitch with Metadata, TPC – Tonic Pitch class. Last letter H – Histogram, G – GD Histogram

of errors that belong to the ’others’ category on group delay processing. Also, the number of errors that belong to the *Pa* category actually increase on group delay processing. These observations are on the expected lines given that group delay processing will accentuate the peak corresponding to *svara Pa* and while suppressing those *svaras* belonging to the ’others’ category. It can also be observed that for both methods, using the metadata reduces the number of errors that fall into the ’others’ category considerably.

In order to further illustrate the nature of errors even after group delay processing, the performance on the methods on different classes of data from database DB4 is shown in Figure 3.16. Database DB4 is chosen as it has a large number of instances for each of the subclasses.

Both the methods, as can be observed in Figure 3.16, perform better for Carnatic music when compared to Hindustani music. This might be attributed to the fact that the group delay processing is more helpful in the case of Carnatic music owing to the fact that the *svaras* are a lot more inflected when compared to Hindustani music. It can also be observed that for items with female lead vocal, errors in the ’others’ category seem to be extremely dominant on estimating the tonic pitch without metadata. Though, these errors seem to reduce on restricting the peak picking range using the metadata. Another interesting observation is that the *Ma* errors seem to be extremely prevalent in Hindustani music, especially in the case of SHM method, suggesting peaks corresponding to the *svara Ma* dominate pitch histograms in Hindustani music.

3.6 Summary

Melodic characteristics of Indian classical music, in particular the inflected nature of *svaras* was illustrated. It was proposed that the varying inflections of *svaras* and the fixed ratio between *svara Sa* and *svara Pa* can be used for tonic identification. The manner in which these traits manifest in a pitch histogram was then illustrated. It was proposed that the traits required for tonic identification can be enhanced using group delay processing. Pitch histograms were modeled as magnitude response of system of resonators in parallel and properties of the group delay function for such a system was explained. It was shown that group delay histograms were indeed beneficial for tonic identification and several strategies were subsequently developed to utilize it. Three methods were proposed in this Chapter. The concert based method was shown to be highly accurate when presented with a whole concert. Amongst the template matching and segmented histograms methods, template matching method was found to perform better. Using metadata regarding the artists, accuracies in the range of 90 to 95% was attained for the different databases.

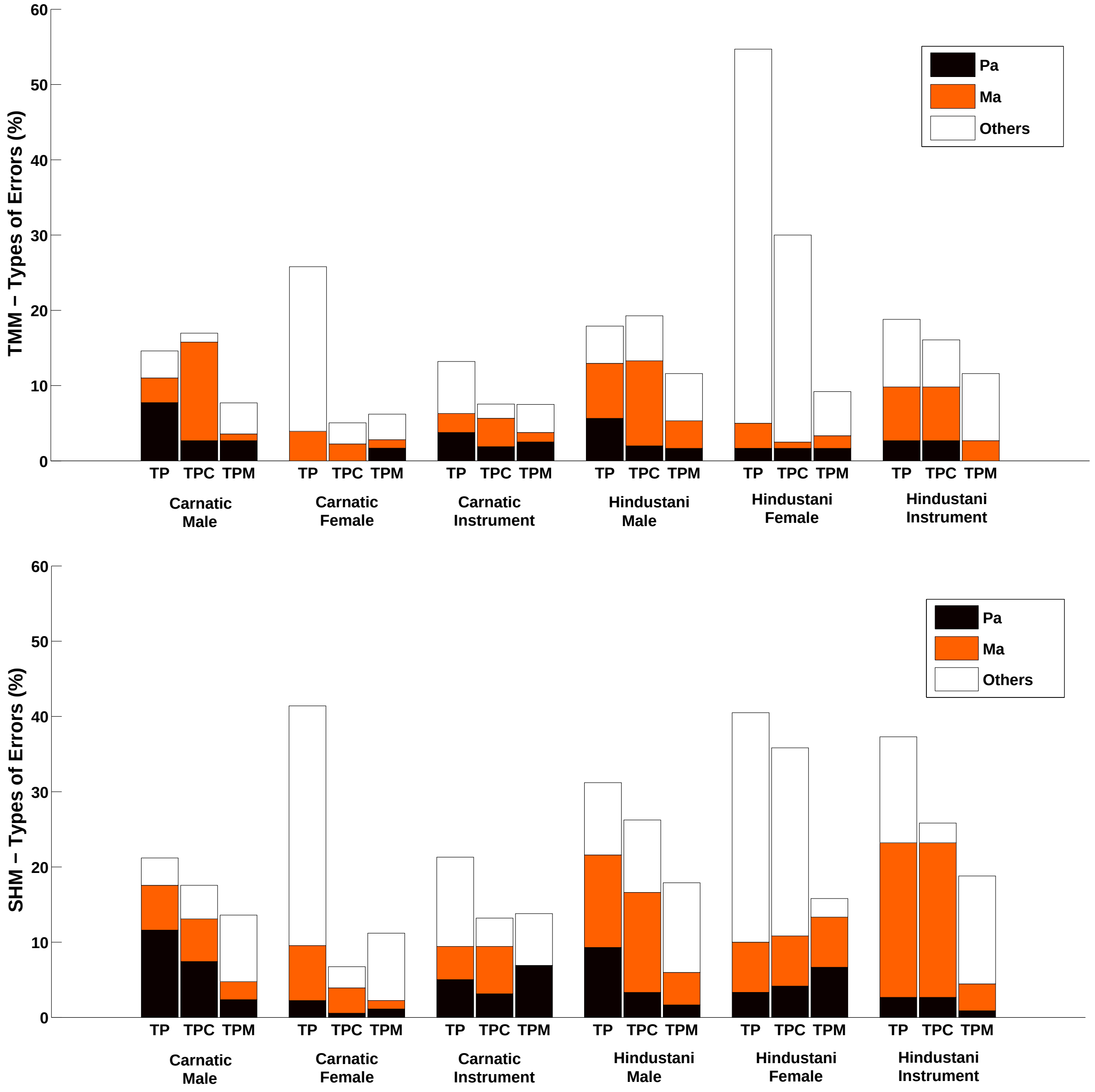

Figure 3.16: The performance of TMM and SHM using GD histograms on database DB4. TP – Tonic Pitch, TPM – Tonic Pitch with Metadata, TPC – Tonic Pitch class

CHAPTER 4

Processing the drone for tonic identification

4.1 Motivation and context

In Indian classical music as is practiced today, the drone is a definite component of a musical performance. There is some amount of debate regarding the period during which the drone first became a fixed component of a performance. Though there are claims that drone featured prominently even in ancient music, the understanding nowadays is that its usage as a definite component in a performance dates back to the 17th century (Deva, 1952). The drone in Indian classical music as mentioned in Chapter 2, serves to establish the reference tonic pitch in a performance. The drone is tuned by the performer and played in the background continuously throughout the concert. The drone is used as reference pitch by both the performer and the accompanying artists.

A variety of instruments such as the *tambura/tānpura*, electronic *śruti* box, *svarmanḍal* or the sympathetic strings of an instrument like the *vīṅa* or the *sitār* are used to produce the reference pitch in Indian classical music. The *tambura* or the electronic *śruti* box features in almost all vocal and most of instrumental performances. The electronic *śruti* box replicates the sound of the *tambura*. The term *drone* used in this chapter refers to the sound produced by the *tambura* or the electronic *śruti* box.

The *tambura* is a multi string instrument and the *tambura* used by majority of the musicians consists of four strings. Though the primary purpose of the instrument is to establish the reference pitch, the drone produced by the *tambura* or the electronic *śruti* box is not just a single tone of that of the tonic. Rather the drone is a richer tonal sound, providing a harmonic base to the music. In order to achieve the rich harmonic base, the strings of the *tambura* are tuned in a specific manner. Generally, facing the *tambura*, from left to right, the first string is tuned to the lower octave of the reference note (*mandra Sa*). The next two strings are tuned to the tonic pitch that is middle octave

Sa or the *madhya Sa*. They are tuned in such a way that no beats are heard when plucked one after the other. They also have to be tuned to exactly one octave higher than the first string. The fourth string may be tuned either to the *Pa* ($1.5\times$*Sa*), the fourth which is the *Ma* ($\approx 1.33\times$ *Sa*) or the seventh which is the *Ni* ($\approx 1.88\times$*Sa*). This tuning depends on the musician and the *Rāga* performed. Thus, the sound of the *tambura* as mentioned earlier, does not consist of the reference note *Sa* alone. It is a homogeneous sound consisting of the *Sa* and other notes, the dominant note corresponding to that of the *Sa*.

It was Salamon *et al.* (2012), that first proposed processing the drone in the background to identify tonic for both Carnatic and Hindustani music. Given the ubiquitous nature of the tonic due to presence of the drone, Salamon *et al.* (2012); Gulati *et al.* (2012) employed a multi pitch based approach to establish tonic. It was shown that by analyzing histograms constructed on the multi pitch extracted, the pitch of the drone in the background can be detected to establish tonic. Due to the non trivial nature of identifying tonic from the histogram constructed, machine learning techniques are used to learn the relationship between the peaks of the histogram.

In this work, we show that by analyzing the relation between the drone and the music performed, a simple cepstrum based pitch extraction technique suffices to identify the pitch of the tonic. It is shown that instead of extracting pitch contours (mono pitch or multi pitch), selecting and processing a few optimal frames lead to not only highly accurate but also almost instantaneous tonic identification. As no machine learning techniques are employed to identify tonic, the performance is not affected by the mismatch between train and test data.

4.2 Drone and its characteristics in Indian classical music

In order to determine the tuning of the drone in an item of Indian classical music it is first necessary to understand the characteristics of the signal corresponding to that of music vis-a-vis the drone. Figure 4.1 shows the spectrogram of the drone produced by a *tambura* played in isolation and Figure 4.2a shows the plot of the pitch contour extracted from the recording. The pitch was extracted using Yin (Cheveigne and Kawahara, 2002.). Figure 4.2b is the histogram of the pitch extracted with a bin width of $1Hz$. Three clear peaks can be seen in the histogram constructed. As denoted in Figure 4.2b the first and the third peak indicate the lower and middle octave *Sa* respectively. The second peak corresponds to the *svara Pa* $(1.5 \times tonic)$ in the middle octave. The pitch of the second peak can vary depending on the tuning of the drone which in turn depends on the *Rāga* being performed.

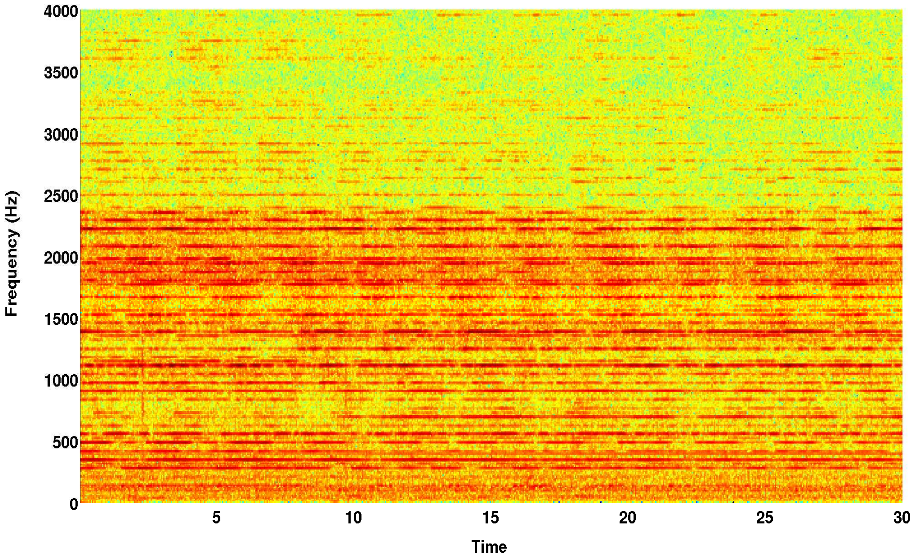

Figure 4.1: Spectogram of the *tambura* in isolation

In an actual performance, the drone though omnipresent, is in the background making it difficult to determine the tuning of the drone. Figure 4.3 shows the spectrogram of an excerpt of Carnatic music. Harmonics belonging to the lead vocal and the drone can be clearly seen in Figure 4.3 with the harmonics of the drone present in all the frames.

Given this ubiquitous nature of drone, we hypothesize that by further investigating the properties of the drone and its relationship with the predominant melody, simple signal processing techniques might suffice for identifying the tonic. In order to understand the interaction of the drone with the other sources, a large number of excerpts across Hindustani and Carnatic music were therefore analyzed. The predominant pitch contour across musicians and instruments was extracted using Yin and the following features were observed.

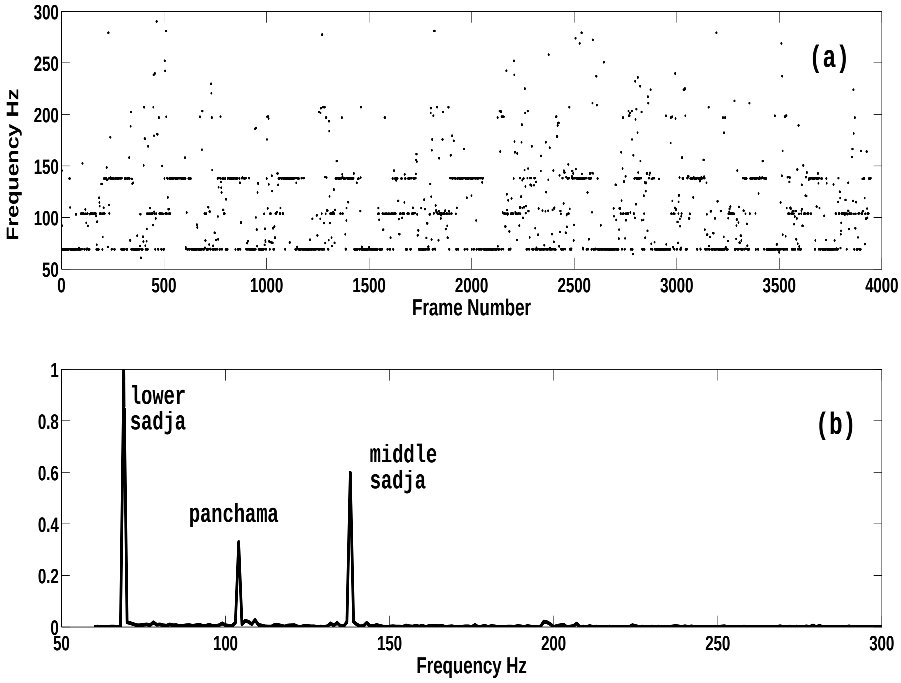

Figure 4.2: Figure a is the pitch extracted from a *tambura* excerpt and Figure b is the corresponding histogram

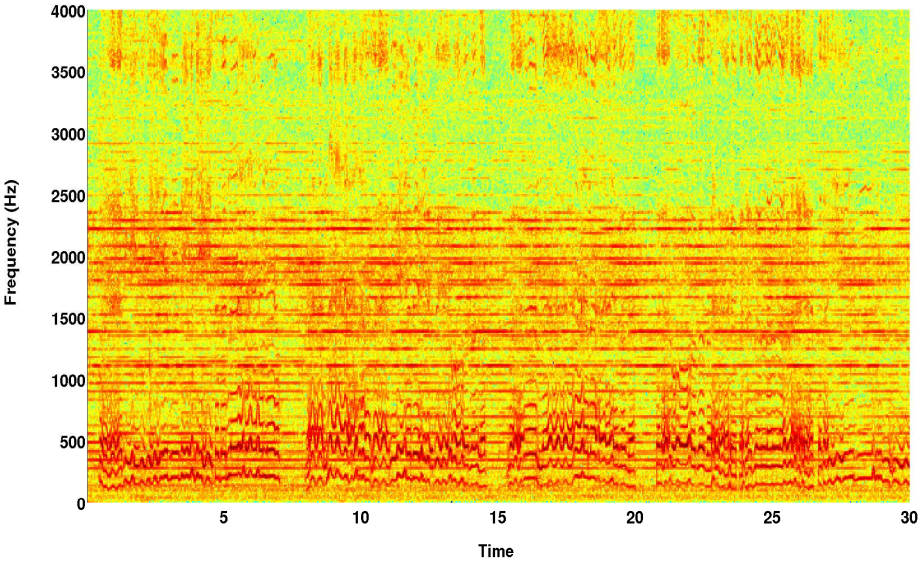

Figure 4.3: Spectogram of an excerpt of Carnatic music

- There is a limited range of frequencies within which performers chose their reference *svara*. To reiterate, for the database collected, it was seen that the tonic pitch i.e middle octave *ṣaḍja* ranged from
  - Male vocal 110 to 180 Hz
  - Female vocal 140 to 250 Hz

    - Instruments 120 to 220 Hz

- Though the drone is omnipresent in the background, pitch of the drone is seldom registered in the prominent mono pitch extracted.
- In the silence regions where the drone served as the prominent source, it was observed that it frequently registered pitch values corresponding to that of the lower octave *Sa*. This characteristic of the drone was also illustrated in Figure 4.2.
- The lead vocal or the accompanying instrument's range of melodic frequencies were predominantly found to occupy the middle and upper octaves.
- The drone in general was found to sustain at a particular frequency for a longer duration when compared with the pitch of the predominant melody.

In the following section we propose different ways in which a cepstrum based pitch extraction technique can be optimized to exploit these observations, in order to identify the tuning of the drone and the tonic pitch accurately.

4.3 Cepstrum based pitch extraction for tonic identification

Cepstrum is defined as the real part of the inverse Fourier transform of the log power spectrum. Given that it has a strong peak corresponding to the pitch period for a voiced frame (Noll, 1967), it has been used for the purpose of pitch extraction of voiced speech. It was shown in Noll (1967) that even in the case of missing fundamental, the fine structures in the spectrum due to the harmonics give rise to a cepstral peak indicating the fundamental pitch. Figure 4.4 shows the block diagram for estimating pitch using cepstrum. If $f0_i$ is the pitch of source i in a given frame, the cepstrum would show a peak at $sr/f0_i$, where sr is the sampling rate.

In the case of Indian classical music, there are multiple sources that can be active at a given instant. A general ensemble in a performance consists of the lead vocal/instrument, accompanying instruments like the violin in Carnatic music or the *sārangi* in Hindustani music, and the percussionists. This implies identifying the tuning of the drone, present in the background, using the cepstral pitch method is non-trivial. To further illustrate the workings of the cepstrum pitch method, *tambura* and male vocal music were first recorded in isolation. Figures 4.5a, 4.5d and 4.5g show the cepstrum in the singing range for *tambura* in isolation. Figures 4.5b, 4.5e and 4.5h show the cepstrum for the male vocal in isolation. Figures 4.5c, 4.5f and 4.5i show the cepstrum for the mixed signal.

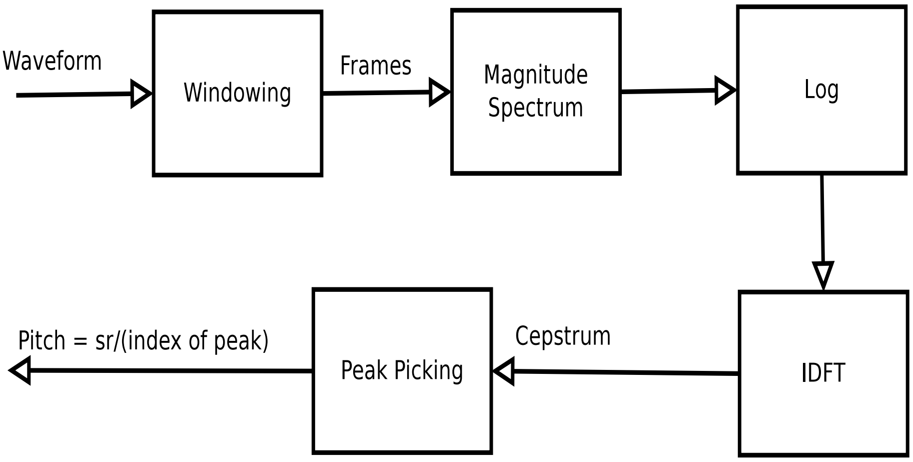

Figure 4.4: Procedure for cepstrum based pitch extraction

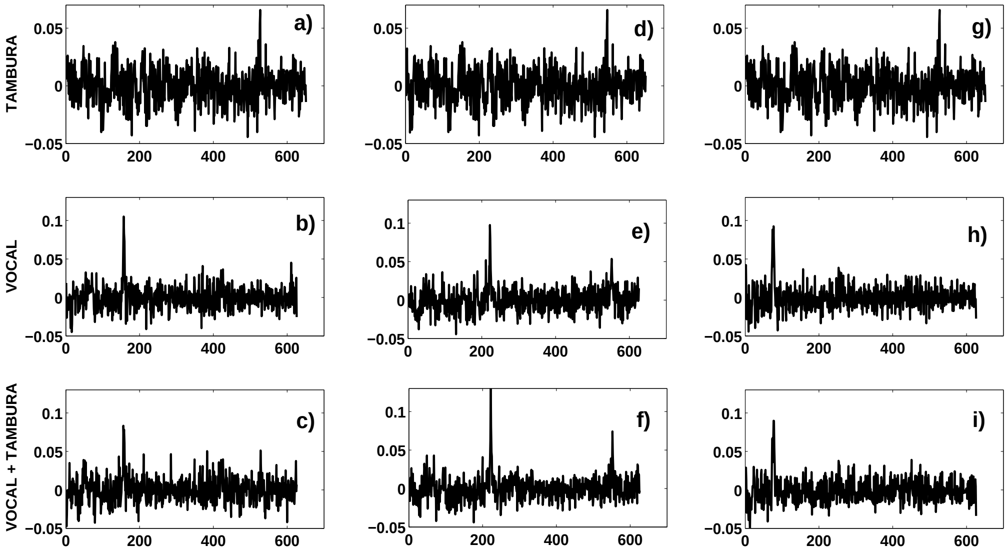

Figure 4.5: Cepstrum for three frames of data. Plots a, d and g are the cepstrum plots for the *tambura* in isolation. Plots b, e and h are cepstrum plots for male vocal in isolation. Plots c, f and i are cepstrum plots for the mixed signal

It can be seen that for the mixed signal, the peak due to the drone gets masked by the peak due to that of the lead performer. Given that the drone is of lower energy, it is seldom the most prominent peak. Also, cepstral peaks decrease in amplitude with increasing quefrency (L. R. Rabiner and R. W. Schafer, 1978). This is important as the drone often registers middle or lower octave *Sa* with the melody being rendered at a much higher pitch. All these factors result in the peak corresponding to the melody dominating the peak corresponding to the drone. This implies that the cepstral pitch method will not help in determining the tuning of the drone if the whole melodic range is considered. Having said that, observations made in Section 4.2 can be used to modify the existing process. The following customization of the cepstral pitch method is suggested

- It is evident that the simple peak picking across the entire pitch range will pick the peak of the most prominent sound and not that of the pitch of the drone. In this work, we therefore explore the possibility of reducing the range for peak picking, given the observations in Section 4.2. By just restricting the peak picking range to the lower octave regions in the cepstral domain, one can attempt to identify the pitch produced by the drone. Given that the performers seldom perform in the lower octave range, peak picking range in the cepstral pitch method is restricted to 50 to 90 Hz for male vocal and 70 to 125 hz for female vocal. For instrumental lead, the range is restricted to 60 to 110 Hz. These are the respective lower octave *Sa* ranges.

- Another important factor is the length of the window applied to the signal to compute the cepstrum. The length of the analysis window does have an effect on the height of the cepstral peak (L. R. Rabiner and R. W. Schafer, 1978), with a minimum of two pitch period required to see a strong peak indicating the period in the cepstrum. It was observed in Section 4.2 and as can be seen in Figure 4.2, the drone sustains at a particular frequency for a longer period relative to the performer. This implies that larger window sizes can be used for analysis. Sustained nature of the drone will result in a strong peak indicating the drone. Variation in the melody rendered, within the analysis window, will also dampen the cepstral peak corresponding to the melody.

Using the customized cepstral pitch method for tonic identification, in the following section we propose several methods for identifying the tonic pitch. We also evaluate the usefulness of the proposed changes by evaluating the proposed methods on the database described in Chapter 2.

4.3.1 Tonic identification using cepstral pitch method and pitch histograms

In this method, given an excerpt, pitch is extracted for all the frames of the excerpt using the cepstral pitch method. After experimentation with various window sizes for analysis, window size of 0.050 seconds was found to be optimal for determining the tuning of the drone. This analysis length is considerably larger than the usual analysis length adopted for quasi stationary analysis. Then depending on the type of excerpt i.e male, female or instrumental, pitch is estimated by picking the cepstral peak with the respective prescribed ranges mentioned in the previous Section. A histogram is then computed using the extracted pitch with bin width of $1Hz$. The bin value of the tallest peak in the histogram is deemed to be the lower octave *Sa*. The desired tonic pitch is then twice the lower octave *Sa*. Figure 4.6 shows the flowchart of the process proposed for identifying tonic.

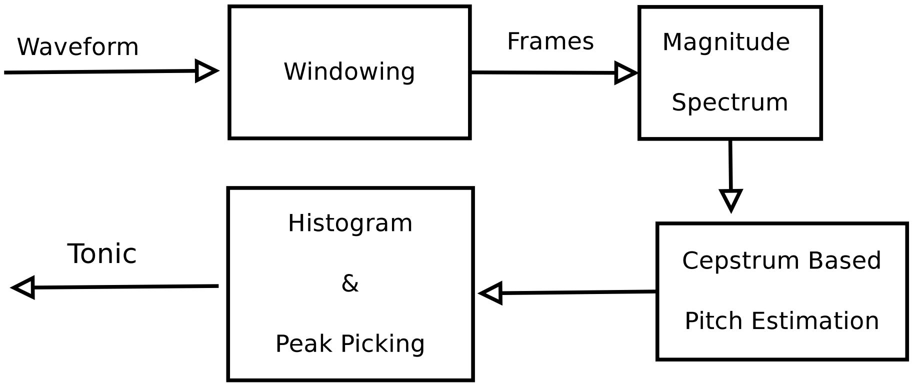

Figure 4.6: Block diagram for method 1

4.3.2 Audio descriptors to select frames

Even though an effort is made in the previous method to extract pitch corresponding to that of the drone given an excerpt of music, an error is foreseeable when even the prominent melody performed in the excerpt is in the lower octave regions. In order to negate this error, we propose using low level audio descriptors to select an optimal set of frames from a given excerpt which can then be analyzed for tonic identification. These desired optimal set of frames are those frames that are indicated of having the drone as the prominent source by the audio descriptors. In order to identify the descriptor that best indicates the prominence of the drone in a frame, a host of temporal and spectral features were analyzed. Table 4.1 lists the descriptors used and their relative values for the frames with a prominent drone when compared to the other frames. Relative values of the descriptors can also be seen in Figure 4.7, where the features were extracted from an $30s$ excerpt of a Hindustani male vocal item. The regions of the waveform marked in red are regions where drone is clearly audible in isolation. The descriptors experimented with in this work are part of a slew of audio descriptors that have been used for instrumentation classification (Deng *et al.*, 2008). As indicated in Benetos *et al.* (2006) and Fu *et al.* (2011) these features are some of the most popular descriptors used for describing the characteristics of instruments.

Table 4.1: Low level audio descriptors

Spectral entropy	High
Short term energy	Low
Spectral Centroid	High
Zero crossing rate	High
Spectral roll-off	High
Spectral Flux	High

Given these relative values for frames with prominent drone, these features were then used to select optimal frames from the given audio for analysis. The procedure adopted for using each of these features is illustrated next using the one of the features – short term energy.

Figure 4.9 shows the waveform and the short term energy contour (hop size of $10ms$) for a two minute excerpt of a Carnatic music piece rendered by a male vocalist. The low energy regions are the silence regions where the drone present in the background can be assumed to be the prominent sound. We propose that instead of estimating pitch for all the frames of a given excerpt, one can detect pitch of the drone better by processing only the low energy frames. For a given excerpt, short term energy is first computed. The short term energy values are sorted in descending order. The bottom 300 frames ($3s$ of data), signifying frames with low energy are selected. Pitch using cepstrum is computed only on these frames. As detailed earlier, a histogram is constructed on the pitch extracted and the bin value of the tallest peak is deemed to be the tonic. Figure 4.8 gives the flow chart for computing the tonic using the short term

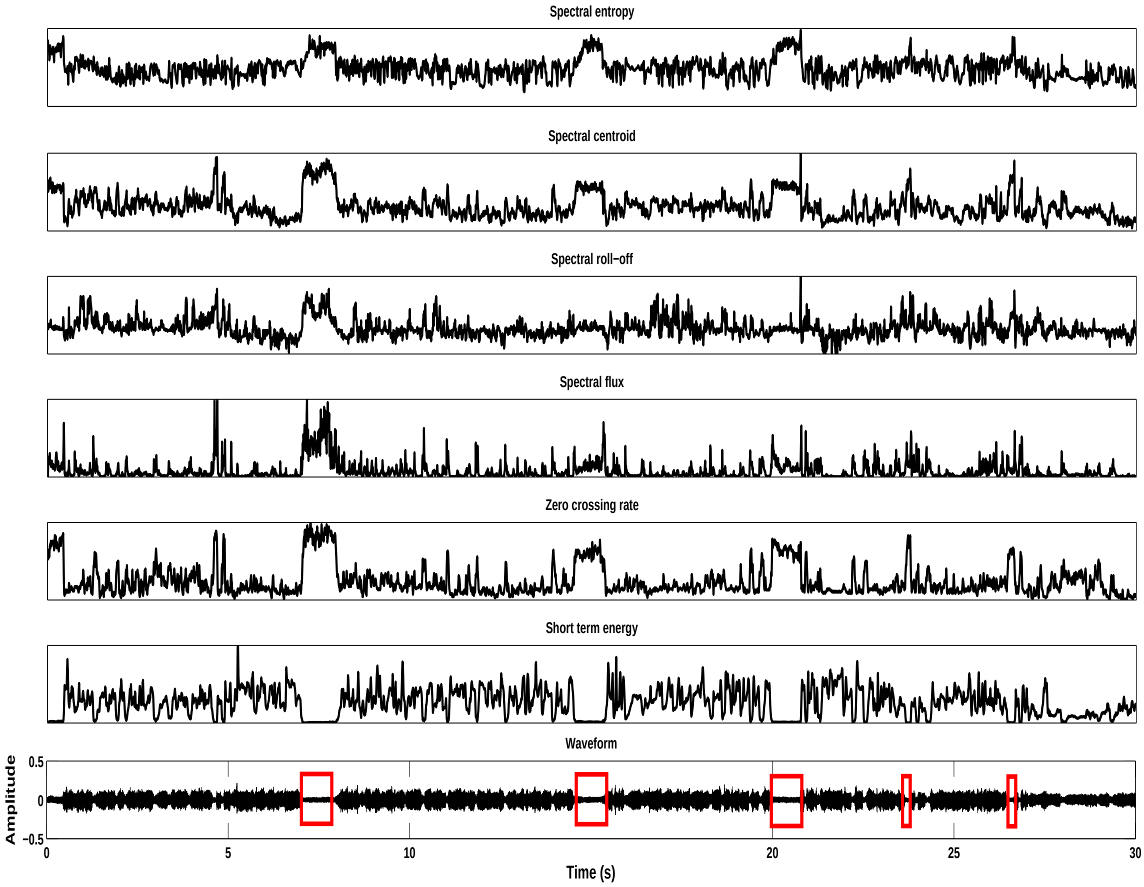

Figure 4.7: Low level audio descriptors and their relative values for prominent drone frames

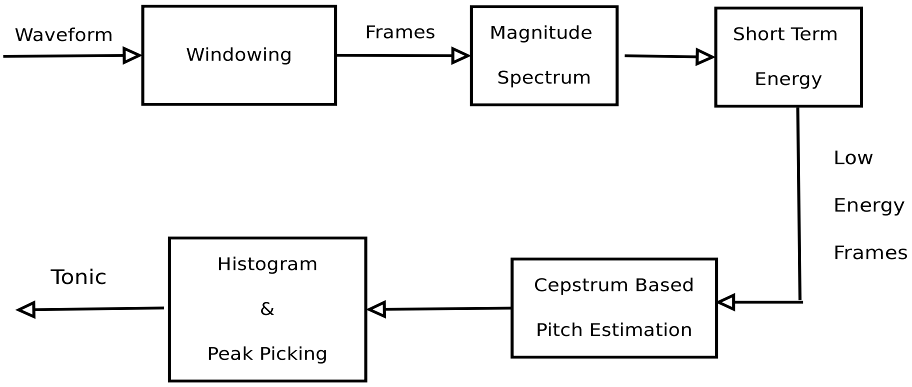

Figure 4.8: Block diagram for method 2

energy based approach. Figure 4.10 illustrates a typical example on using short term energy thresholds. In Figure 4.10a, the histogram is computed on the pitch extracted for the entire excerpt in the lower octave region. Though the *svara Sa* in the lower octave does register a prominent peak, it is dominated by the peak corresponding to that of the melody being performed in the lower octave in the excerpt, leading to erroneous identification of tonic. Figure 4.10b shows the histogram of the estimated pitch on frames with least energy in the given excerpt. It can be seen that the lower octave *Sa* of the drone is the tallest peak leading to accurate tonic identification.

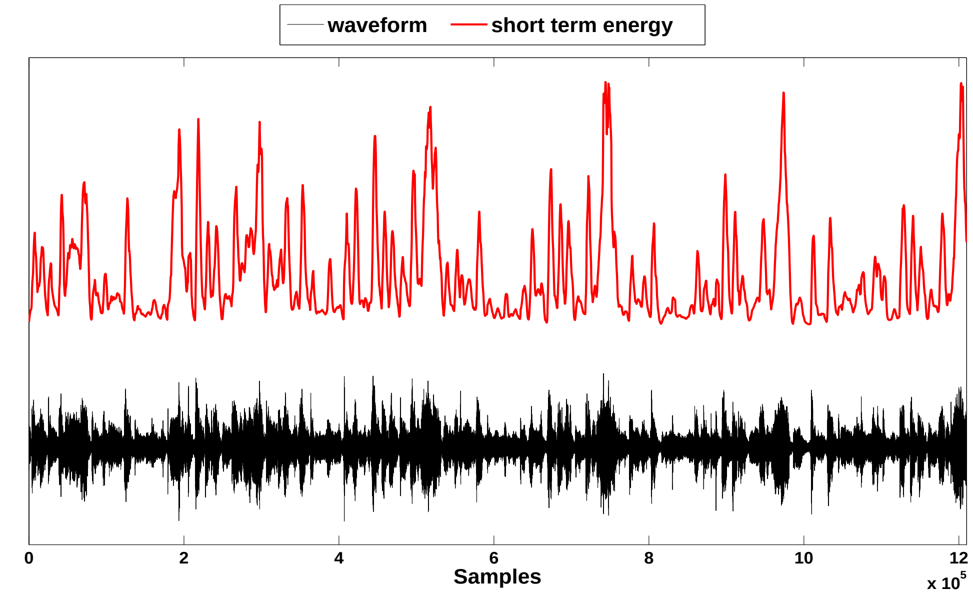

Figure 4.9: Waveform of 60s duration and short term energy

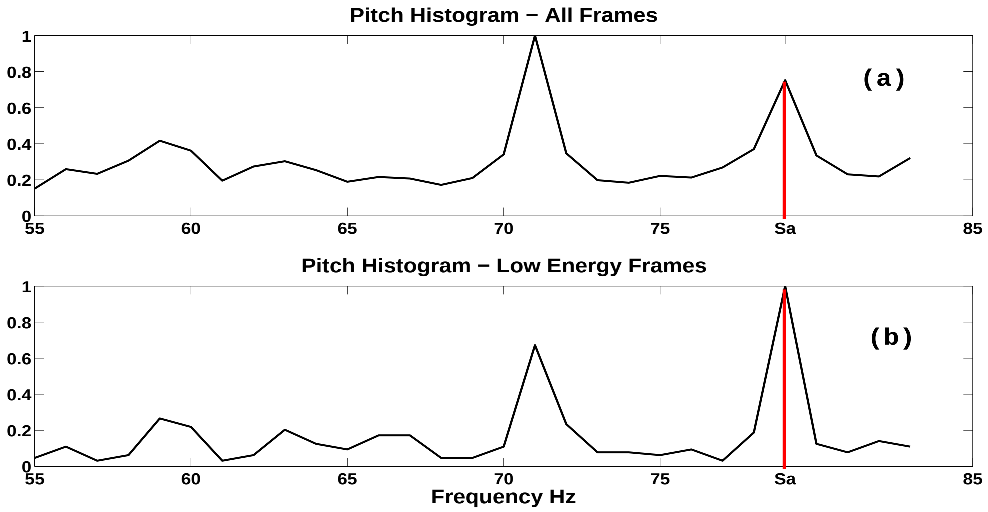

Figure 4.10: Plot (a) is the pitch histogram computed from pitch extracted from all the frames. Plot (b) is the pitch histogram computed using pitch from short energy frames

Similar procedure was followed using the each of the audio descriptors in Table 4.1. While frames with least energy were chosen in the previous case, for rest of the descriptors, frames that register the highest values of the respective descriptors were considered for analysis. The performance of these features and the results of method 2 will be demonstrated and inferred in Section 4.4

As an alternate to constructing histograms on the pitch extracted for a bag of frames, in the following Section 4.3.3 we describe a Non-negative Matrix Factorization (NMF) based technique to process the same bag of frames for tonic identification. Use of NMF to identify tonic further enhances the speed of tonic identification and also serves as an elegant illustration of the ubiquitous nature of the drone in Indian classical music.

4.3.3 Non-negative matrix factorization based tonic identification

Non-negative matrix factorization (NMF) is a technique using which a non-negative matrix V is factorized into two matrices W and H as shown in Equation 4.1. The matrix factors W and H derived using NMF are also non-negative. The factorized matrices are non unique and there are various different constraints that can be used within the NMF framework to arrive upon the matrix factors. While there are several methods for factorizing a given matrix like principle component analysis and independent component analysis, owing to the non-negative nature of the factor matrices, NMF has found various application in speech and audio processing.

$$V \approx WH \tag{4.1}$$

In Smaragdis and Brown (2003), it was shown that NMF technique can be used for factorizing a spectrogram into its major components. In Smaragdis and Brown (2003), the magnitude spectrum of an excerpt of music was taken as the matrix V of dimension $m \times n$. Using NMF, the factor matrices, W of dimension $m \times k$ and H of dimension $k \times n$ were obtained. By selecting $k << n$, it was shown that any spectrogram can be factorized into matrices W and H, such that the columns of W are spectral basis vectors representing the main elements of the original spectrogram and H represents the activations of these basis spectral vectors over time. Based on such factorization of the spectra, NMF based techniques have found various other applications. These include tasks like source separation (Grindlay and Ellis, 2009), instrument classification

(Benetos *et al.*, 2006) and music transcription (Smaragdis and Brown, 2003; Ananthapadmanabhan *et al.*, 2013) amongst many others.

Even within the NMF frame work there are many ways by which the factor matrices can be derived. The methods differ on the metric used to quantify the approximation in Equation 4.1. In this work we use the popular euclidean measure and multiplicative update rules proposed in Lee and Seung (2001) as shown in Equation 4.2 to iteratively estimate W an H. The multiplicative update rules given in Equation 4.2 ensure that $||V - WH||_F$ is minimized. $||.||_f$ is the Frobenius norm.

$$\begin{aligned} W &\leftarrow W.\frac{VH^T}{WHH^T} \\ H &\leftarrow H.\frac{W^TV}{W^TWH} \end{aligned} \tag{4.2}$$

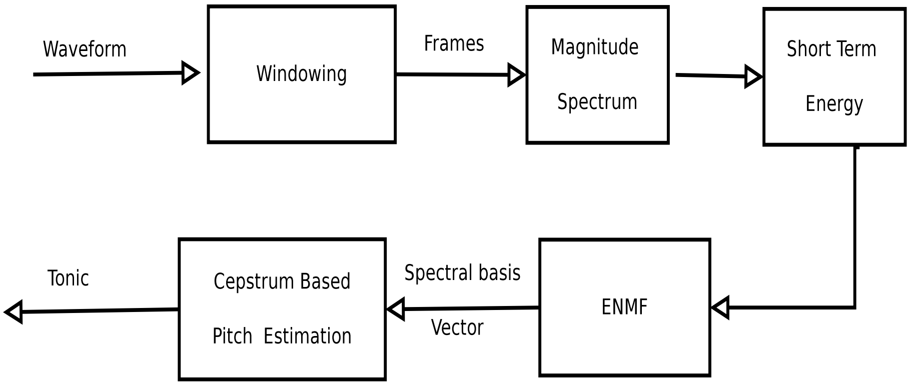

Figure 4.11: Block diagram for method 3

Estimating tuning of the drone using NMF

For the task of tonic identification, an attempt is made to employ NMF to exploit the fact that the drone is present through out the excerpt. Given a excerpt of music, short term Fourier transform is first performed. Then, a certain number frames that indicate a strong presence of a drone are collected. Any of the features mentioned in Section 4.3.2 can be used to select the bag of frames. In this work, short term energy is the preferred choice. The magnitude spectrum of these bag of frames is the matrix V of dimension $m \times n$. m in this case would be the order of the discrete Fourier transform and n the number of frames. NMF is then performed to determine W and H of dimensions $m \times k$ and $k \times n$ respectively. The drone being present in all the frames of the excerpt and being relatively more prominent in the low energy frames, forms a major element in V. This implies that the k spectral basis vectors which best describe V will contain the drone information. We proposed that it should then suffice to just analyze only the k spectral vectors to determine the tonic. We propose that by choosing $k = 1$ and estimating pitch on the decomposed single spectral vector of W one can identify tonic. Pitch is then estimated for the given spectral vector, by picking peak within a narrow range in the cepstral domain, as described in the previous sections. The estimated pitch is the pitch corresponding to that of the tonic. Figure 4.11 shows the flowchart of the NMF based approach.

Non-negative matrix factorization also allows fixing once of the factor matrices and estimating the unknown matrix factor using the same multiplication rules given in Equation 4.2. This facility can be used to further emphasize the omnipresent nature of the drone. We propose that the activation matrix H can be fixed as a single row matrix with a constant value for all frames. This implies that through NMF, we are now seeking that element of V which is prominent and activated with constant energy. The single column of W derived can now be processed in the same manner as described above to identify the tonic. Hence in this work, two flavors of NMF were attempted.

1. W and H are randomly initialized and matrix factors are iteratively estimated
2. W is randomly initialized and H is kept initialized as a constant row vector and kept constant throughout the iterative process. Only W is re-estimated.

For both the variations, it was found that a very small number of iterations, around 10 iterations, are sufficient to get a good estimate of the spectral basis vector i.e the matrix W . This implies that using NMF, given an audio excerpt, tonic can be identified almost instantaneously, as only short term energy and few iterations of NMF is sufficient. The fact that this method is not based on the pitch histogram also gives it an advantage over the previous methods. The pitch histograms based methods are very much dependant on the melody being performed by the artist in frames being analyzed and tonic pitch identified might turn out be erroneous in the case of the histogram indicating the melody being performed rather than the tuning of the drone. Whereas the

NMF based method is less likely to be effected by the melody as the emphasis is on that source which is omnipresent in all the frames.

4.4 Experiments and Results

The proposed methods in this Chapter were tested on the database, the details of which are provided in Chapter 2. In each of the methods proposed, the metadata regarding the performer i.e male vocal, female vocal or instrumental was used to restrict the peak picking ranges in the cepstrum.

4.4.1 Method 1 - pitch histogram with cepstrum based pitch extraction

In this method, given an audio excerpt, pitch is extracted using cepstral based pitch extraction. The objective of this method is to illustrate that by restricting the pitch picking range to the lower octave ranges, one can identify the tonic pitch fairly accurately. Figure 4.12 shows the accuracy of identifying the tonic pitch for databases DB2, DB3 and DB4 (Chapter 2, Table 2.1). Figure 4.12 also illustrates the performance of the method with varying amounts of data, ranging from 3 seconds randomly extracted from an item to using the whole item. For extracting the pitch, a window size of 0.050 seconds was used with a window shift of $0.01ms$. A pitch histogram was then constructed and the bin value of the tallest peak is the tonic pitch. Figure 4.12 shows the performance of the method on using different amounts of data, randomly extracted from an item provided for tonic identification. As can be seen in Figure 4.12 there is major improvement on using the lower octave ranges when compared with the use of middle octave ranges in the cepstral pitch method.

It can also be seen that the performance improves as the duration of the excerpts increase. This is expected, given that excerpts are picked randomly. Errors might occur due to masking of the drone pitch by other sources within this excerpt extracted at random. When small excerpts of music are chosen at random from an entire item, the prominent of the sources of ensemble might even be performing in the lower octave

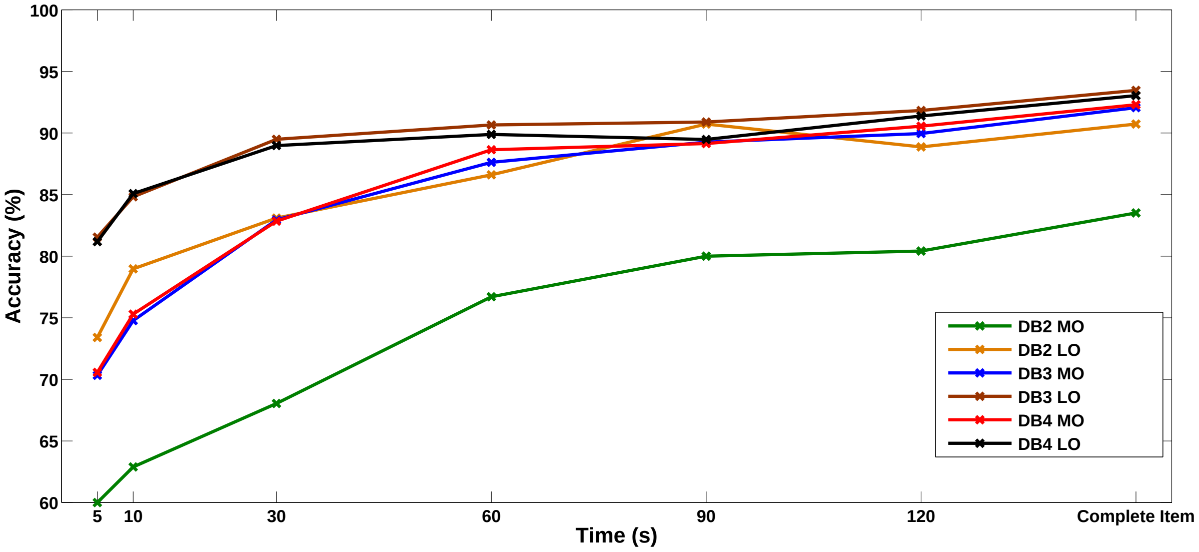

Figure 4.12: The performance on determining pitch in the lower and middle octave performance range, using varying quantities of data. LO - Lower Octave, MO- Middle Octave

regions. Naturally as length of the excerpts increase, performance improves as majority of the frames will register the pitch of the drone at least in the lower octave. It can also be seen that the performance for DB3 and DB4 is better than that for DB2. This naturally stems from the fact that the DB2 contains amateur recordings where the drone might not feature in the background.

Having seen that identifying the lower octave *Sa* is preferable, the performance of Method 1 with respect to various sub-classes was then analyzed. Figure 4.13 shows the results of Method 1 on Carnatic and Hindustani music, as well as male, female and instrumental music. The performance can be seen to be consistently high for male artists, but not as good for female vocal and instrumental music. One of the reasons for the poor performance can be the fact that the accompanying instruments, were observed to be playing the same melody in lower octave regions especially in the case of female vocal artists. As will be seen in Section 4.4.2, on carefully choosing the drone frames, this problem for female vocal and instrument excerpts is alleviated.

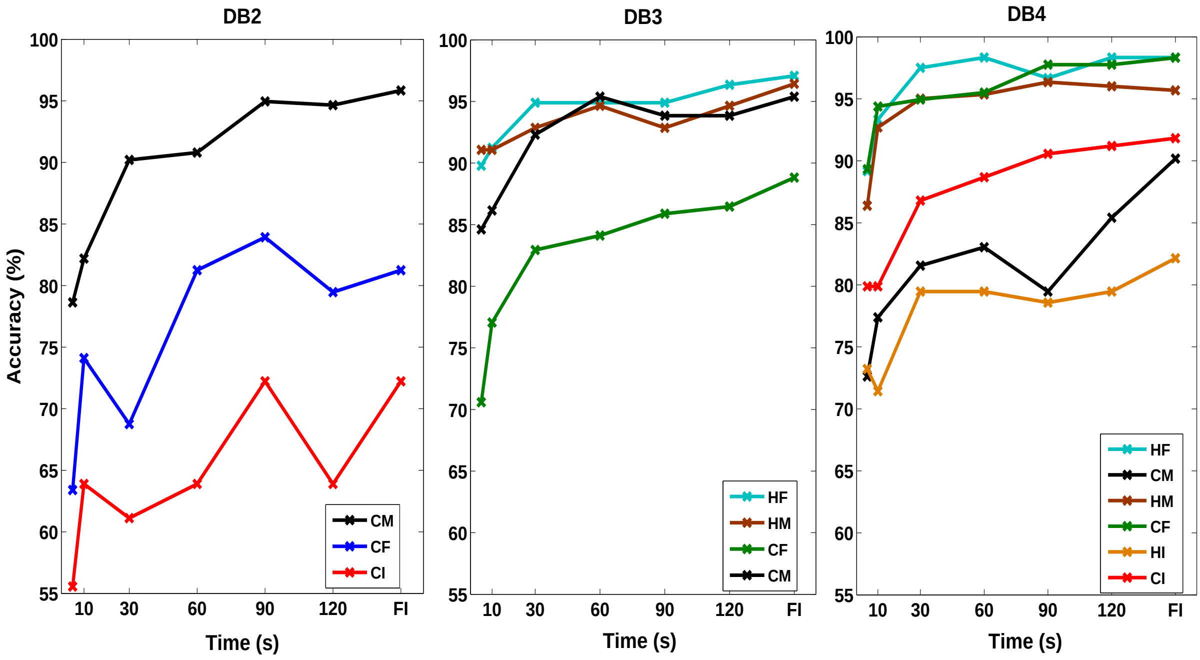

Figure 4.13: The performance on determining pitch in the lower octave range, using varying quantities of data. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument, FI - Full Item

4.4.2 Method 2 - audio descriptors

In this method, the same pitch histogram approach of method 1 is employed, but the histogram is computed on only those frames where the drone is deemed to be prominent. In order to analyze which of the audio descriptors mentioned in Section 4.3.2 are best suited for choosing the optimal set of frames, tonic pitch was first identified for database DB4 using all the descriptors. Given an item from DB4, $90s$ of data was randomly extracted from the item, from which frames that have drone as a prominent source, as indicated by the audio descriptor were retained. The total number of frames retained for pitch extraction amounted to $3s$ of data for each excerpt. A histogram was then constructed using pitch extracted from the retained frames, the bin value of the highest peak being the tonic pitch. The performance of different audio descriptors can be seen in Figure 4.14. Database DB4 was chosen for analysis considering that it has larger number of instances of Carnatic and Hindustani music, as well as male, female and instrumental music.

As can be seen in Figure 4.14, short term energy seems to work best in almost all cases when evaluated in terms of tonic pitch accuracy. Spectral flux also seems to

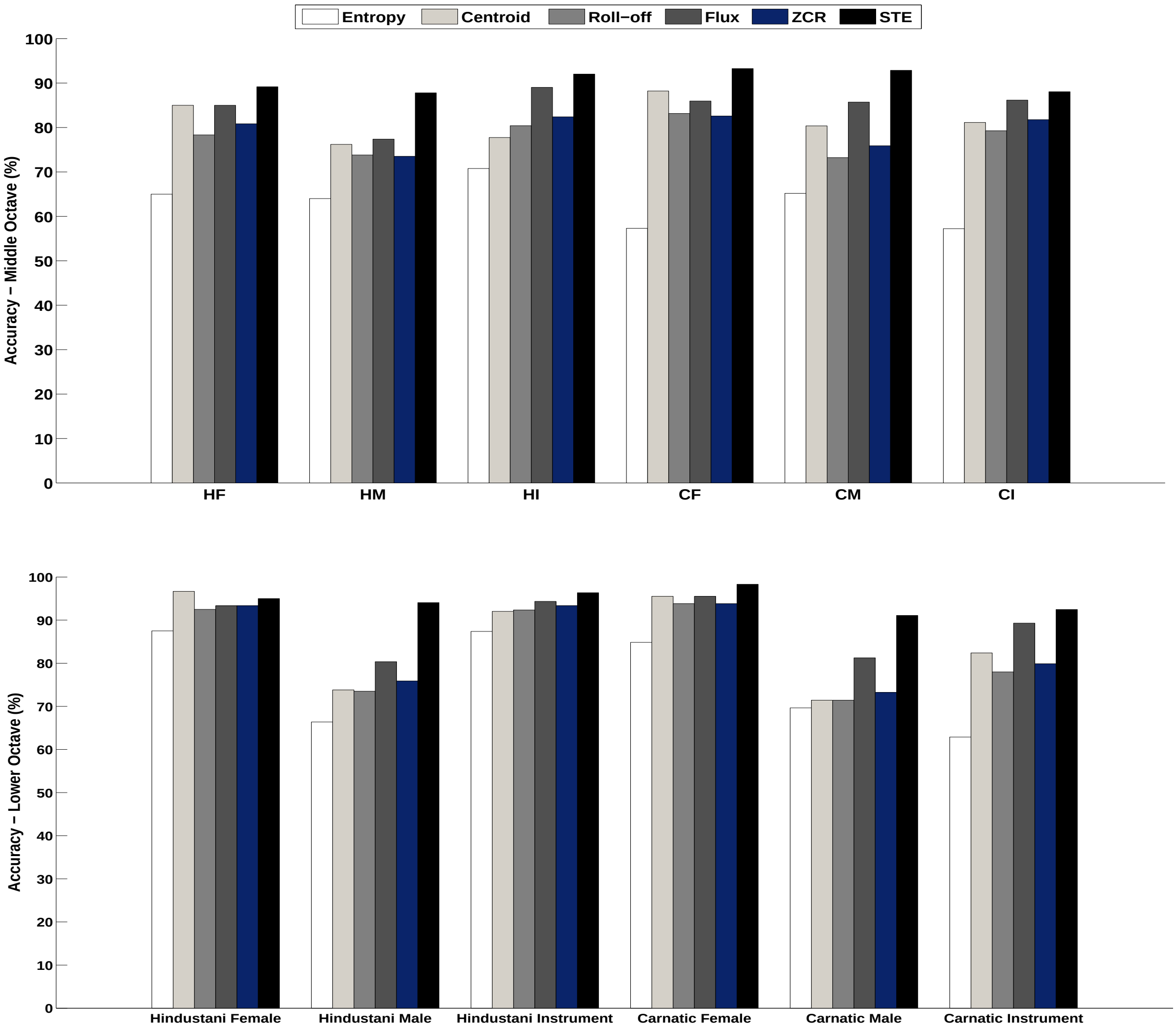

Figure 4.14: The performance on using various audio descriptors. ZCR - Zero crossing rate, STE - Short term energy. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument

perform well with spectral entropy yielding poor results.

Having observed that the short term energy performs best amongst the audio descriptors, it was then used as the descriptor to select frames for the purpose of pitch extraction and tonic identification for all the remaining databases. Figure 4.15 shows the result of such a process under various conditions.

Presented an item, first a part of the item was randomly extracted, from which frames with least short term energy were retained for pitch extraction. Figure 4.15 shows the performance on analyzing just the low energy frames. The X axis denotes the duration of excerpt extracted at random from an item. Figure 4.15 also shows the performance on retaining varying amounts of low energy frames. The number of frames retained from the excerpt extracted can amount to 3, 5 or 10 seconds as indicated in Figure 4.15. The performance on estimating the lower octave *Sa* and the middle octave *Sa* is also illustrated in Figure 4.15.

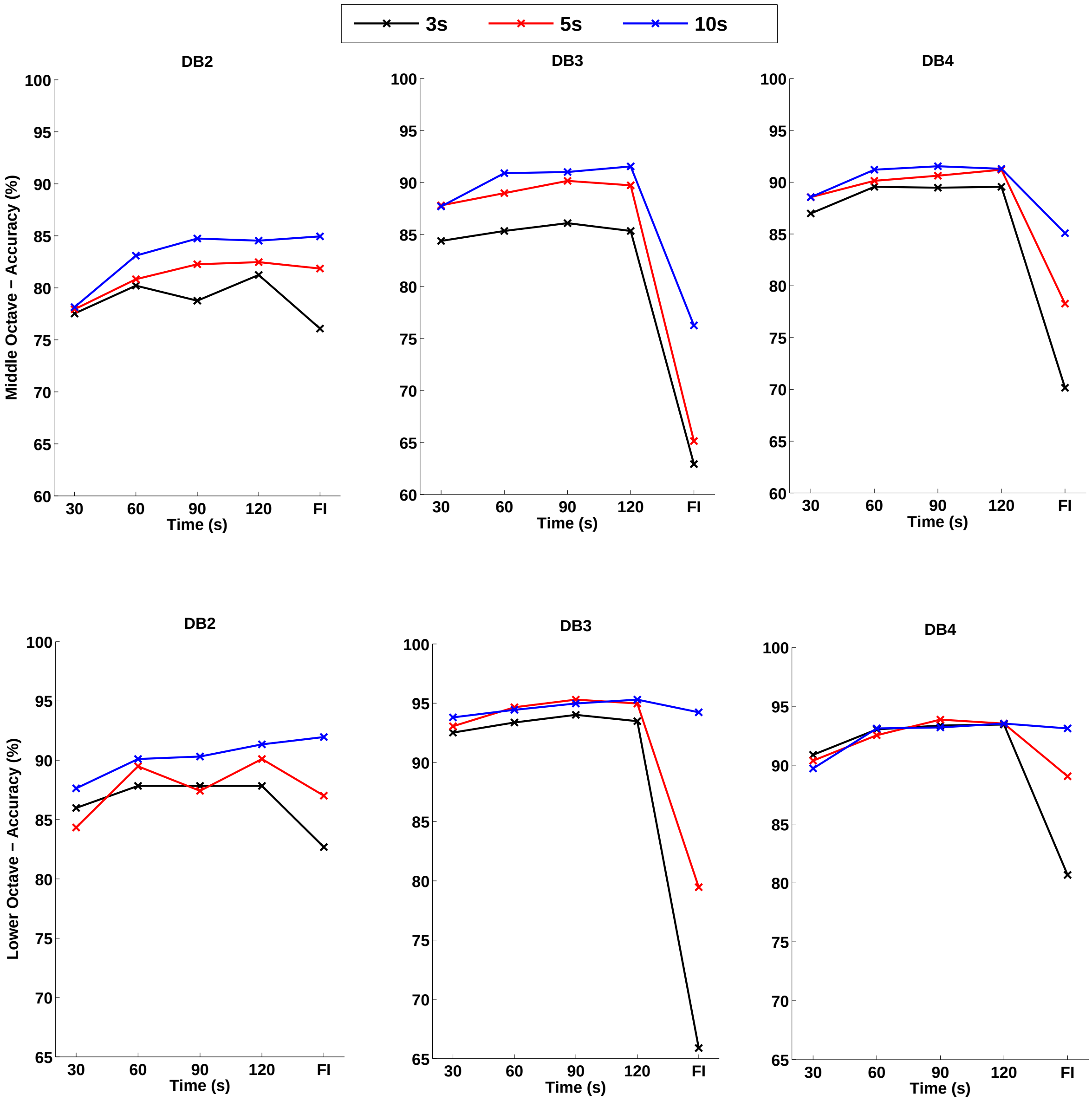

Figure 4.15: The performance on selecting drone prominent frames using short term energy. FI – Full Item

It can be seen from Figure 4.15 that very high accuracies of tonic identification are obtained on using the short energy frames to retain frames with prominent drone. There is considerable improvement in the performance when compared with using all the frames (Figure 4.12). Similar to the previous method, performance on estimating the lower octave *Sa* is superior.

It should also be noted that unexpectedly the performance of the methods seem to reduce on using the full item, especially for database DB3 and DB4, which consist of audio CD recordings. On analyzing the errors, it was the seen that the errors can be attributed to the nature of studio recordings. In many recordings, it was observed that there are silence regions with even the drone absent. These regions were mainly found at the starting of an item. Thus on using the complete item, these regions get selected for pitch extraction leading erroneous tonic identification. As can be seen in Figure 4.15, this problem is alleviated by extracting smaller excerpts from an item (ignoring the few seconds at the beginning of an item) for tonic identification or retaining a larger number of frames ($10s$).

It is interesting to note that the number of frames selected using short term energy does not seem to matter in the case of database DB3 and DB4, especially when estimating the lower octave *Sa*. While for the other cases, on database DB2 and when estimating the middle octave ṣaḍja, larger the number of frames selected, better is the performance. This is mainly due to the fact that the cepstral pitch method does not consistently estimate the pitch corresponding to that of the drone in these cases. With the drone many a time absent in database DB2 and the melody also being performed in the peak picking range, the pitch extracted might also correspond to that of the melody instead of the drone. In which case, as seen in Figure 4.12 and 4.13, the performance improves with larger amounts of data.

4.4.3 Method 3 - non-negative matrix factorization

This method explores the use the NMF based technique for tonic identification. As was discussed in Section 4.3.3, in order to estimate the single basis vector for analysis, 2 variants of the activation matrix can be used. 1) A randomly initialized iteratively re-estimated matrix H. 2) A matrix initialized as a constant row vector and re-estimated. Figure 4.16 shows the performance of the NMF based technique using these two variations on database DB4. In this analysis, for each item in database DB4, an excerpt was first randomly extracted. Spectrogram of the excerpt was then estimated, using which a single basis vector was estimated using the NMF technique. Lower octave *Sa* was then identified by extracting pitch using the single basis vector. X axis denotes the duration of the excerpt randomly extracted.

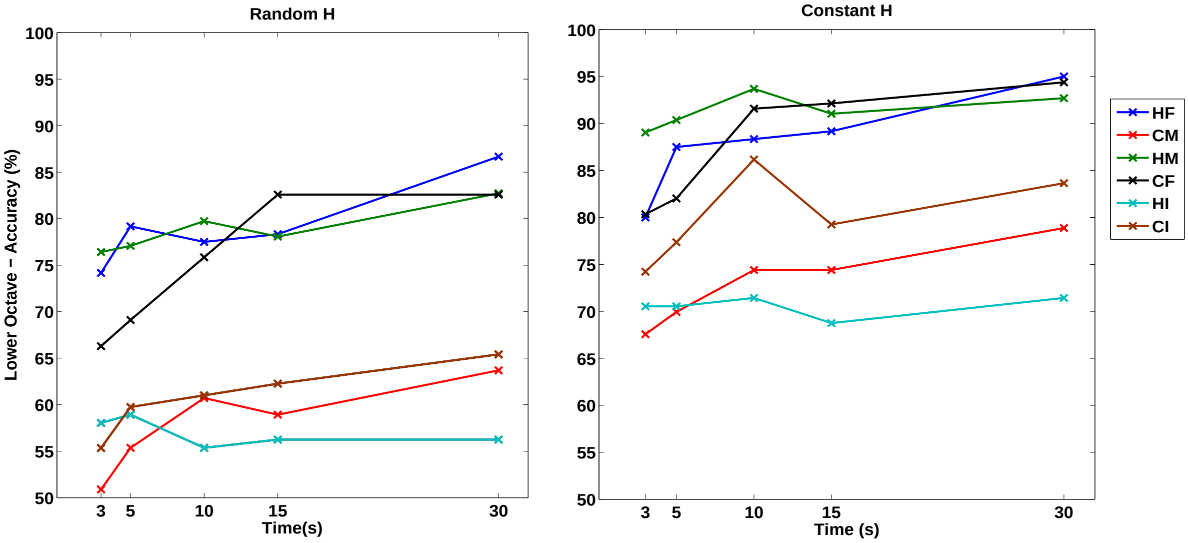

Figure 4.16: The performance on using Random and Constant activation matrix on DB4. FI – Full Item

As can be seen in Figure 4.16 , the constant activation matrix does lead to considerable better performance. There is an improvement in the performance for each of the sub-classes of database DB4 on using a constant activation matrix.

Having seen that using a constant activation does aid in better identifying the tonic, tonic was then identified for all the databases using a constant H. Before estimating the spectral basis vector, frames with lower energy were first selected. Figure 4.17, shows the performance of the NMF based technique on all the databases. The black, red and blue plots indicate the performance on retaining frames between $3s, 5s$ and $10s$ respectively.

As can be seen, very high accuracies are obtained using the NMF based technique. Comparing with performance with the pitch histogram based technique (Figure 4.15), it can be seen that the NMF based techniques performs well even when identifying the middle octave *Sa*. This can be attributed to the fact that by forcing a constant activation, the NMF technique is better enabled to track the pitch of the drone when compared to YIN in the pitch histogram based techniques, where the pitch of the prominent source

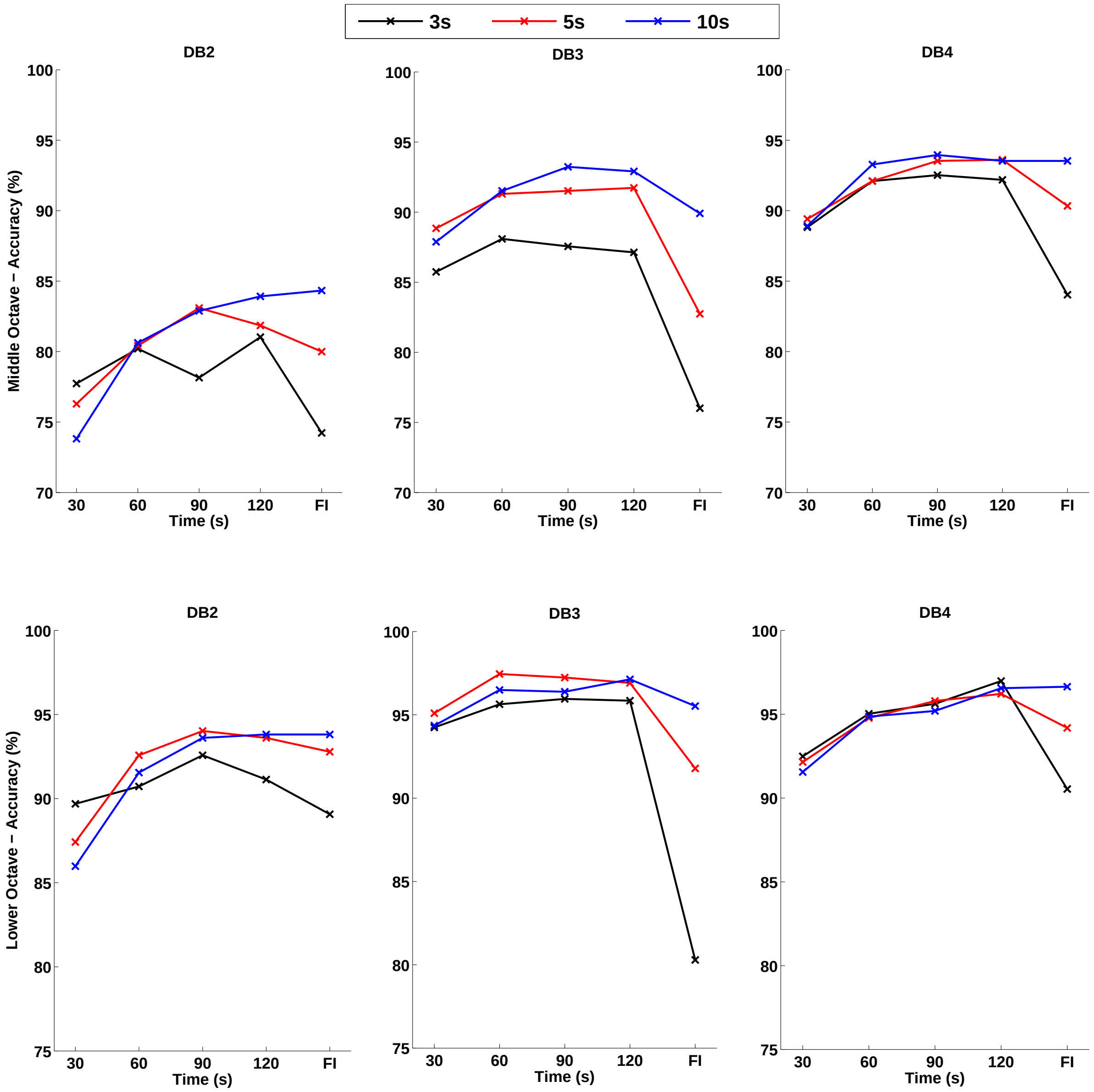

Figure 4.17: The performance of NMF based method in tonic identification. FI – Full Item

is tracked. Accuracies upwards of 95% is obtained for databases DB3 and DB4 on estimating the lower octave *Sa* using just 90 seconds of data. Similar to method 2, even in this approach the performance reduces on using the complete item due the presence of silence regions at the beginning of an item in Audio CD recordings.

4.4.4 Error Analysis

In this section we analyze the nature of errors for both the pitch histogram and the NMF based methods. In order to perform the error analysis, given an item, 90 seconds of data was randomly extracted and the data was divided into frames using window length of 0.050 seconds and hop size of 0.01 seconds. Frames with low energy amounting to 5 seconds of data were then retained for further analysis. Figure 4.18 shows the nature of errors for the pitch histogram based method and Figure 4.19 for the NMF based method. Similar to the analysis performed in Chapter 3, the errors are categorized into *Pa*, *Ma* and 'others'.

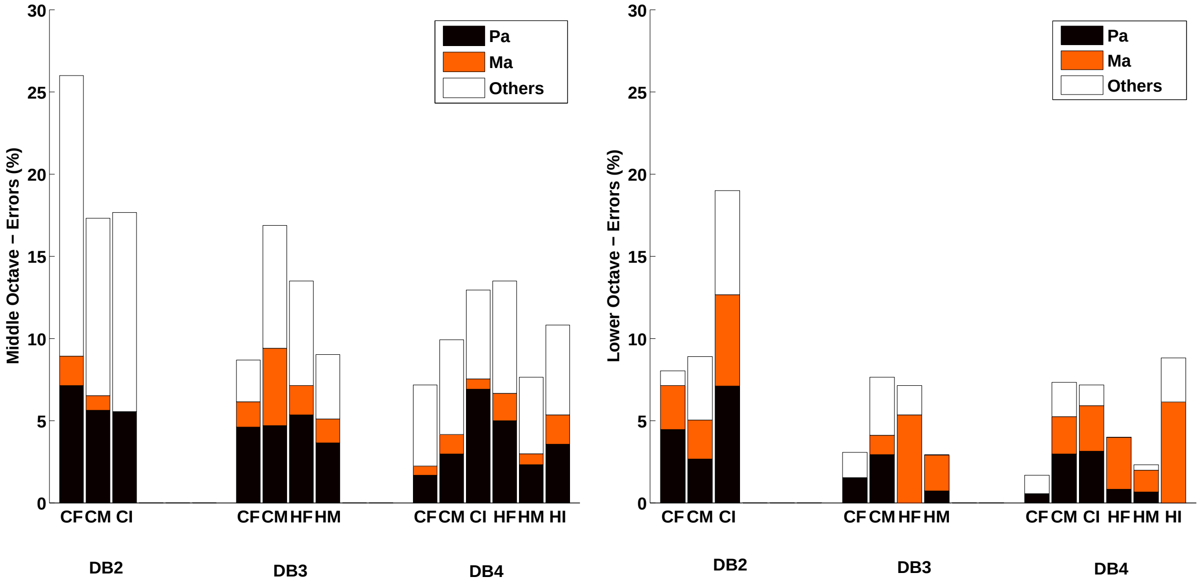

Figure 4.18: Error analysis - pitch histogram on low energy frames based method . H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument

It can be seen from both Figures 4.18 and 4.19 that a large number of errors fall into the 'others' category on attempting to identify the middle octave *Sa*. This is primarily because the histograms or the NMF end up capturing the melody being performed in a given excerpt rather than the *Sa*. However, on attempting to identify the lower octave *Sa*, the number of errors that fall into the 'others' category reduces considerably.

It can also be seen that for both cases, the relative number of *Pa* and *Ma* errors increases in most cases on attempting to identify the lower octave *Sa*. This can be primarily attributed to the fact that the drone not only produces the pitch of the *Sa* but also that of *Pa* and *Ma* depending on the melody being performed. In some of the cases,

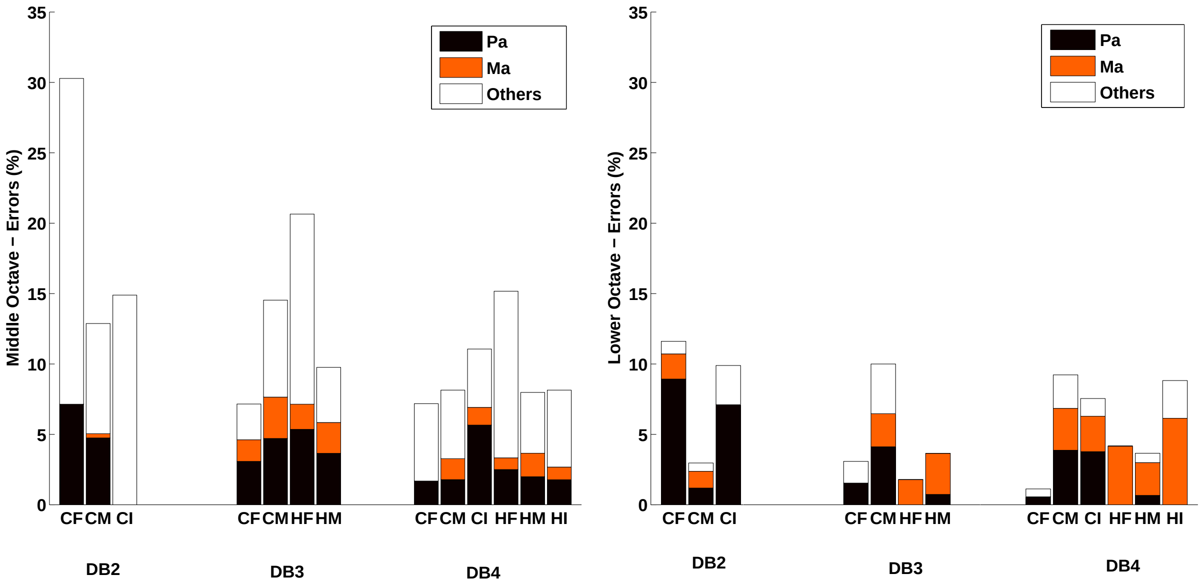

Figure 4.19: Error analysis - NMF based method. H – Hindustani, C – Carnatic, F – Female, M – Male, I – Instrument

histogram or the NMF basis vectors registers this *svara* produced by the drone more prominently than that of the *Sa* leading to erroneous identification of the tonic. While for Carnatic music, *Pa* errors seem to be more prevalent, *Ma* errors are relatively higher for Hindustani music, in both the pitch histogram and the NMF based method.

4.5 Summary

The drone which is used to establish the tonic pitch was processed to identify the tonic. First, the melodic aspects of the drone in relation to the main melody being performed was studied. It was seen that the drone often renders pitch in the lower octave ranges of a performance. This range being infrequently visited by the artists, a cepstrum pitch method was developed to extract pitch in the prescribed ranges. Using the pitch extracted, several methods were then developed to identify the tonic. The pitch histogram based method was shown to perform reasonably well. In order to further enhance the performance, it was proposed that spectral and temporal features can be used to select those frames which feature the drone prominently. Short term energy was found to perform best when tested with a set of audio descriptors popular in the literature. Pitch histogram based method on the pitch from these select frames was shown to boost the performance considerably. As an alternate to pitch histograms, a non-negative matrix factorization based method was also proposed to exploit the ubiquitous nature of the drone. The NMF based method was shown to be considerably fast and was found to perform best in identifying tonic pitch.

CHAPTER 5

Summary and Conclusion

Automatic tonic pitch identification in Indian classical music is a prerequisite to apply computational methodologies to study melodic concepts. As each artist can choose his own reference pitch i.e the tonic, by identifying the tonic automatically, pitch which is the basic representation of melody can be normalized with respect to the tonic. Such a normalization process would then enable study of melodic concepts across artists with large amounts of data.

The primary objective of this thesis was to develop techniques to automatically identify the tonic pitch when presented with an audio excerpt of Indian classical music. Some of the salient points presented in this thesis are as follows:

5.1 Salient Points

- Indian classical music repertoire presents the listener various cues regarding the tonic pitch. In this work, it was shown that the cues emanating from the melodic characteristics and the drone in the background can be used gainfully to identify the tonic pitch.

- It was shown using pitch histograms that the *svara* in Indian classical music is more a pitch region than a specific pitch. The less inflected properties of the *svaras Sa* and *Pa* and them manifesting as narrower regions in pitch histograms were then illustrated, indicating that this characteristic of the music can be used to identify the tonic pitch.

- To further emphasize the less inflected nature of the *svaras Sa* and *Pa*, a novel application of the group delay function was proposed. Pitch histograms are modeled as a set of resonators in parallel and interesting properties of the group delay function for such a setup were illustrated. It was shown that the group delay functions due to their additive nature, along with resolving peaks also better retain the pole radius information. This aspect of the group delay response was used to further accentuate the peaks corresponding to the *svaras Sa* and *Pa*.

- Several strategies were developed to use the group delay processed histogram to identify the tonic pitch. The strategies relied on the fixed ratio between the *svara Sa* and *Pa* and the ubiquitous nature of *svara Sa* to determine the identity of the tonic.

- Next, an effort was made to process the drone information present in the audio to determine the tonic. Analyzing the melody of the drone and the performers, it was shown that the drone registers pitch in the lower octave regions, which is seldom registered by the main artists. It was shown that by using a cepstral pitch method with the pitch range restricted to lower octave regions, tuning of the drone can be determined accurately.
- To further enhance the performance, it was proposed to select only those frames where drone features prominently. Amongst a host of low level features experimented with, it was shown that short term energy performs best. High accuracies, upward of 90% was obtained on all the databases.
- An NMF based technique was also developed for identifying the tonic pitch. It was shown that the omnipresent nature of the drone can be indicated using a constant activation matrix. By determining a single spectral vector representing a set of frames, it was shown that accuracies upward of 95% can be obtained on all the databases.

5.2 Criticism of the work

In this Section we discuss of some of the drawbacks of the techniques proposed for tonic identification.

- The techniques that employ melodic cues, require larger amount of data to determine the tonic. Given that the melodic traits manifest over an item, a small excerpt might not suffice to determine the tonic. Pitch extraction for items of longer duration is computationally expensive and time consuming.
- The use of pitch histograms to characterize melodic traits have a few drawbacks. The height of the peaks plays a cardinal role in determining the identity of the tonic. In the case of an artist extensively improvising around a particular *svara* that is not *Sa* or *Pa*, the proposed methods can go wrong, even after group delay processing.
- Strategies developed to use both sets of cues rely on the meta data to identify the tonic pitch. This requires manual tagging of pieces as male, female or instrumental music.

5.3 Future work

Given some of the shortcomings of the methods, as observed in the previous section, there is scope for further improving the techniques developed in this work. At the same time, in this thesis, quite a few characteristics of the music and the manner in which they manifest on signal processing have been illustrated. These observations and the processing techniques developed can be employed for further MIR studies in Indian classical music.

- For identifying the tonic pitch using melodic cues, the *rāga* information can be exploited. Having seen that group delay processing does provide better information regarding the *svaras* of the melody, appropriate templates indicating all the notes of the melody can be used instead of just the *Sa*-*Pa*-*Sa* template. Preliminary work on these lines was attempted and reported in Bellur *et al.* (2012) but was not pursued further given the lack of availability of metadata for the complete database.

- Instead of using the metadata information for the drone based methods, machine learning may be used to build decision trees as in Salamon *et al.* (2012); Gulati *et al.* (2012), to determine which of the peaks of the histogram correspond to that of the tonic. This can especially prove useful for methods which attempt to determine the tuning of the drone, where a lot of the confusion is just between the *svaras Sa*, *Pa* and *Ma*.

- Pitch histograms and variants of it have been used to study intonation and as features for *rāga* recognition systems. Given some of the favorable properties of group delay processed histograms observed in this work, use of GD histograms as a feature can be explored for studying intonation or build *rāga* recognition systems for Indian music.

REFERENCES

1. **Ananthapadmanabhan, A.**, **A. Bellur**, and **H. A. Murthy** (2013). Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorisation. *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 181–185.

2. **Bellur, A.**, **V. Ishwar**, **X. Serra**, and **H. A. Murthy** (2012). A knowledge based signal processing approach to tonic identication in indian classical music. *2nd Compmusic Workshop*, 113 – 116.

3. **Benetos, E.**, **M. Kotti**, and **C. Kotropoulos** (2006). Musical instrument classification using non-negative matrix factorization algorithms. *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 221–224.

4. **Bozkurt, B.**, **L. Couvreur**, and **T. Dutoit** (2007). Chirp group delay analysis of speech signals. *Speech Communication*, **49**(3), 159–176.

5. **Cheveigne, A. D.** and **H. Kawahara** (2002.). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930.

6. **Chordia, P.**, **J. Jayaprakash**, and **A. Rae** (2009). Automatic carnatic raag classification. *Journal of the Sangeet Research Academy (Ninaad).*

7. **Chordia, P.** and **A. Rae** (2007). Raag recognition using pitch class and pitch class dyad distributions. *In Proc. of ISMIR*, 431–436.

8. **Deng, J. D.**, **C. Simmermacher**, and **S. Cranefield** (2008). A study on feature analysis for musical instrument classification. *Cybernetics, IEEE Transactions*, 429–438.

9. **Deva, B. C.** (1952). The emergence of the drone in Indian music - a psychological approach. *The Journal of the Music Academy of Madras*, 126–152.

10. **Deva, B. C.** (1965). The problem of continuity in music and sruti. *The Journal of the Music Academy of Madras*, 56–66.

11. **Fu, Z.**, **G. Lu**, **K. M. Ting**, and **D. Zhang** (2011). A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions*, 303–319.

12. **Grindlay, G.** and **D. P. W. Ellis** (2009). Multi-voice polyphonic music transcription using eigeninstruments. *in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA*, 53–56.

13. **Gulati, S.**, **J. Salamon**, and **X. Serra** (2012). A two stage approach for tonic identification in indian art music. *In 2nd Compmusic woorkshop*, 119–127.

14. **Hegde, R. M.**, **H. A. Murthy**, and **V. R. R. Gadde** (2007). Significance of the modified group delay features in speech recognition. *IEEE International Transactions on Audio,Speech and Language Processing*, **15**, 190–202.

15. **Ishwar, V.**, **S. Dutta**, **A. Bellur**, and **H. A. Murthy** (2013). Motif spotting in an alapana in carnatic music. *In Proc. of ISMIR*.

16. **Koduri, G. K.**, **S. Gulati**, **P. Rao**, and **X. Serra** (2012). Raga recognition based on pitch distribution methods. *Journal of New Music Research*, 337–350.

17. **Krishnaswamy, A.** (2003). Application of pitch tracking to south indian classical music. *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 557–560.

18. **Krishnaswamy, A.** (2004). Inflexions and microtonality in south indian classical music. *Frontiers of Research on Speech and Music*.

19. **L. R. Rabiner and R. W. Schafer**, *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

20. **Lee, D. D.** and **H. S. Seung** (2001). Algorithms for nonnegative matrix factorization. *In Neural Inf. Process. Syst*, 556–562.

21. **Levy, M.**, *Intonation in North Indian music: a select comparison of theories with contemporary practice*. Biblia Impex, 1982.

22. **Meer, W. V. D.**, *Hindustani music in the 20th century*. Allied publishers, 1980.

23. **Murthy, H. A.** and **B. Yegnanarayana** (1991). Formant extraction from minimum phase group delay function. *Speech Communication*, **10**, 209–221.

24. **Nagarajan, T.**, **H. A. Murthy**, and **R. M. Hegde** (2003*a*). Segmentation of speech into syllable-like units. *In Proc. of EUROSPEECH*, 2893–2896.

25. **Nagarajan, T.**, **V. K. Prasad**, and **H. A. Murthy** (2003*b*). Minimum phase signal derived from the root cepstrum. *IEE Electronics Letters*, **39**, 941–942.

26. **Noll, A. M.**, Cepstrum pitch determination. *In J. Acoust. Soc. Amer.*. 1967.

27. **Oppenheim, A. V.** and **R. W. Schafer**, *Discrete Time Signal Processing*. Prentice Hall, Inc, New Jersey, 1990.

28. **Padmanabhan, R.** and **H. A. Murthy** (2009). Dynamic selection of magnitude and phase based acoustic feature streams for speaker verification. *In Proc. of European Conference on Signal Processing*, 1244–1248.

29. **Pandey, G.**, **C. Mishra**, and **P. Ipe** (2003). Tansen: A system for automatic raga identification. *Indian International Conference on Artificial Intelligence*, 1350–1363.

30. **Rajan, R.** and **H. A. Murthy** (2013). Group delay based melody monopitch extraction from music. *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 186–190.

31. **Ranade, G. H.**, *Eternal Pradox in Indian music: The Shrutis*. Division, Ministry of Information and Boradcasting, 1957.

32. **Ranjani, H. G.**, **S. Arthi**, and **T. V. Sreenivas** (2011). Shadja, swara identification and raga verification in alapana using stochastic models. *In 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 29–32.

33. **Rao, M. N.**, **S. Thomas**, **T. Nagarajan**, and **H. A. Murthy** (2005). Text-to-speech synthesis using syllable-like units. *National Conference on Communication (NCC)*, 227–230.

34. **Rao, S.** and **W. V. D. Meer**, *The construction, reconstruction and deconstruction of shruti*. Manohar, 2010.

35. **Ross, J. C.**, **T. P. Vinutha**, and **P. Rao** (2012). Detecting melodic motifs from audio for hindustani classical music. *In Proc. of ISMIR*, 193 – 198.

36. **Salamon, J.**, **S. Gulati**, and **X. Serra** (2012). A multipitch approach to tonic tdentification in indian classical music. *In Proc. of ISMIR*, 157–162.

37. **Sengupta, R.**, **N. Dey**, **D. Datta**, and **A. K. Mukerjee** (2005). Automatic tonic (sa) detection algorithm in indian classical vocal music. *In National Symposium on Acoustics*, 1–5.

38. **Serra, J.**, **G. K. Koduri**, **M. Miron**, and **X. Serra** (2011). Tuning of sung indian classical music. *In Proc. of ISMIR*, 157–162.

39. **Serra, X.** (2011). A multicultural approach to music information research. *In Proc. of ISMIR*, 151–156.

40. **Smaragdis, P.** and **J. C. Brown** (2003). Non-negative matrix factorization of polyphonic music transcription. *in Proc. IEEE Workshop on Applicaion of Signal Processing to Audio and Acoustics*, 177–180.

41. **Subramanian, M.** (2007). Carnatic ragam thodi - pitch analysis of notes and gamakams. *in Journal of the Sangeet Natak Akademi*, 3–28.

42. **Yegnanarayana, B.** (1979). Formant extraction from linear prediction phase spectra. *Acoustical Society of America*, **63**, 1638–1640.

43. **Yegnanarayana, B.** and **H. A. Murthy** (1992). Significance of group delay functions in spectrum estimation. *IEEE Trans. Signal Processing*, **40**(9), 2281–2289.

44. **Yegnanarayana, B.**, **D. K. Saikia**, and **T. R. Krishan** (1984). Significance of group delay functions in signal reconstruction from spectral magnitude or phase. *IEEE Trans. Acoustics Speech and Signal Processing*, **ASSP-32**(3), 610–623.

LIST OF PAPERS BASED ON THESIS

1. Ashwin Bellur, Vignesh Ishwar, Xavier Serra and Hema A Murthy, A knowledge based signal processing approach to tonic identication in Indian classical music, *2nd CompMusic Workshop*, 113-116 (2013)

2. Ashwin Bellur and Hema A Murthy, A cepstrum based approach for identifying tonic pitch in Indian classical music, *National Conference on Communications*, 1-5 (2013).

3. Ashwin Bellur and Hema A Murthy, A novel application of group delay function for identifying tonic in Carnatic music, *EUSIPCO*, (2013)