Pitch Estimation of the Predominant Vocal Melody from Heterophonic Music Audio Recordings

Vignesh Ishwar

MASTER THESIS UPF 2014

Master in Sound and Music Computing

Master thesis supervisor:

Xavier Serra

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

Abstract

Music being an industry with a vast digital presence, today, we have access to a large number of audio music recordings online as well as stored locally on computers, cell phones, ipads, to name a few devices. Many non-western music cultures have also made a large digital presence over the last two decades. This opens up many windows for applications with state of the art archving, automatic tagging, lyrics to audio alignment and automatic indexing of music using a vast number of cues from the user inputs. It also opens up many avenues for meaningful musical analysis of various music traditions computationally. Melody being one of the most basic entities, predominant pitch is one of the fundamental representations used in all these tasks. In this work we deal with the pitch estimation of the predominant vocal melody from heterophonic music audio recordings. We provide a novel approach for pitch estimation using a combination of the present state of the art and timbral characteristics of the various melodic sources in the audio music recording.

We perform a detailed review of the state of the art pertaining to computational analysis of music and the usage of predominant melody pitch as a basic representation in a number of tasks. We also review the state of the art with respect to predominant melody estimation and singing voice detection them being highly relevant for this task since we aim at characterizing the singing voice.

The proposed approach is a classification based approach for pitch estimation of vocal melodies. Indian art music is subjected to the approach and for this reason, the musical and cultural aspects of the music have been considered in the approach. We first extract candidate pitch contours using a state of the art predominant melody extraction algorithm. Post this timbral features are extracted corresponding to the source of the pitch contour from the audio signal. These features, instead of being derived from the spectrum of the audio are derived from a representation of the extracted harmonics of the candidate pitch contours. A classification of the candidate pitch contours into vocal and non-vocal classes is then performed. The music specific information is incorporated in a contour selection methodology which uses the tonic pitch which is a fundamental aspect of Indian art music, the test case for this approach. A detailed explanation of the entire approach with implementation details are provided in this thesis.

The approach is evaluated on a database of Karṇāṭik music for which ground truth is manually curated. A novel evaluation methodology based on adaptive thresholding to incorporate

the properties of the music at hand is proposed. The evaluation results surpass the state of the art for predominant melody. This reinforces the hypothesis that the combination of salience based methods and timbral properties of the singing voice aids estimation of pitch of singing voice. A detailed analysis of the results obtained with plausible reasons is performed. The thesis is concluded with the summary of the work, the main conclusions and the contributions made in the course of this work.

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. Xavier Serra for his guidance and unwavering support. I am extremely grateful to him for having given me this brilliant opportunity to join the Music Technology Group at UPF and for considering me to contribute to the CompMusic project. I thank him for supporting my research throughout the year. His guidance and thought process has had an immensely valuable influence not only on the work of this thesis but also on my attitude towards research and life in general. His enthusiasm for research is something that I can only aspire to emulate.

Next, I would like to thank Emilia Gómez and Agustin Martorell for their suggestions, advice and the valuable time for discussions. The music technology courses at UPF which have provided me with an abundance of knowledge played an important role in the shaping up of the work of this thesis. I would like to thank all my professors for their knowledge and insights at every step, it has been a wonderful experience.

I am immensely grateful to the members of 55.302, for their support and encouragement. I would like to thank Ajay Srinivāsamūrthy, Sankalp Gulāti and Gōpālakrishna Kōduri in particular, for their invaluable guidance, support and encouragement. They have been a constant source of motivation in all aspects. The various debates, discussions and brainstorming sessions with them has not only shaped my outlook on research, but on life as well. I would also like to thank Rafael Caro for the valuable time spent in walking me through the basics of Bejing Opera and Western Music. This year has been made memorable by the entire batch of SMC and I thank them from the bottom of my heart.

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

I would like to thank Prof. Hema A. Murthy for her invaluable guidance and support. I am grateful to her for letting me join the IIT madras team of the CompMusic project and introducing me to research in Music Information Retrieval. Her guidance, support and constant encouragement have played an important role in instilling confidence in me as a researcher. I would also like to thank Shrey, Raghav, Srikanth and Anusha and other members of the Don Lab for their support. I would like to thank Ashwin Bellur in particular for being a constant source of motivation.

This work would not have been complete without the help of Suresh Gopalan and Charubala Natarajan of Charsur. It was very kind of them to go out of the way to provide audio for evaluation of this task.

I am highly obliged to Prof. Preeti Rao of IIT Bombay for her valuable insights and discussions for this work. I would also like to thank, Amruta Vidvans and Kaustuv Kanti Ganguly for the valuable time they invested in extracting pitch for ground truth generation.

I would like to thank my guru (teacher) T. M. Krishna for his unwavering support and encouragement throughout this journey. He has been an immense source of inspiration to me and has shaped my outlook not only towards music but also towards life. I am grateful to him for the invaluable discussions, suggestions and ideas for this research. I can only aspire to emulate his thought process, unabated energy and enthusiasm towards life. I would also like to thank Sangeetha Sivakumar for her care and support.

Lastly, I would like to thank my parents for their unreserved support and encouragement. Knowing that they are always there for me has made my life much easier and given me the courage and confidence to pursue any path I wish to choose.

Contents

List of Figures

List of Tables

Chapter 1

Inroduction

Music Information Retrieval (MIR) has grown tremendously over the past few decades. MIR deals with the study of music repertoires using computational methodologies. Owing to the fact that a vast amount of music is in digital form and on the internet, MIR has gained vast importance in the past decade. There has been a major paradigm shift in the way music is retrieved, archived, discovered and created in the past few years. Computational methods for analysing, archiving and retrieving music and information related to it have taken a step forward into main stream commercial music. Methods based on a combination of musicological studies along with signal processing and machine learning techniques are some of the sought after solutions to the challenges faced in MIR.

Most of the methodologies developed in MIR are vastly applied on western popular music and western classical music. The vast consumer base and avalibility of large database of high quality music is a possible justification to this bias. The methodologies developed are highly dependant on the music repertoires analysed. Thus non-western music cannot be directly subjected to these methodologies without considering the cultural specificities of the music. A sufficient understanding of the characteristics of the music which is being analysed is necessary for this purpose (Serra, 2011).

This work addressess the fundamental problem of vocal melody extraction from heterophonic audio music recordings with a single vocal melody source. Accurate estimation of predominant melody is crucial for many melody analysis tasks such as motific analysis, intonation analysis and audio score alignment. A large body of the

literature on melodic analysis of music use predominant melody as the basic representation. Thus estimation of predominant melody is a viable and rampantly developing problem in the field of MIR. Unlike predominant melody estimation, the scope of this work is the estimation of the pitch of the vocal melody. This is motivated from the fact that the predominant source of melody in a large number of musical traditions is the singing voice.

1.1 Motivation and Goals

Poliner et al. (2007), define melody as *"The single(monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison"*. The source of the predominant melody however, may vary across different music cultures. In many music traditions like Indian art music, singing voice is the source of the main melody. The presence of secondary melodic instruments (violin, harmonium, flute, veena etc.) does not mar the status of singing voice being the predominant melody since they are accompanying instruments with the singing voice.

The evidence of certain genres of music being vocal centric can be seen in the fact that even during instrumental solo performances, the artist tries to emulate the characteristics of the voice. This style of playing is known as *gāyakī* (Viswanathan & Allen, 2004) in Indian art music. In the case of performances by vocal duos, triplets or group renditions of music, the concept of a single voice being the main source of melody no longer holds true. The concept of melody, however, agrees with the definition given by Poliner et al. (2007) stated above. Examples of this phenomenon are the Turkish makkam and Arab Anadalusian music genres where in there are multiple vocalists singing the same melody in different registers (Şentürk, Gulati, & Serra, 2013). Melody in such music aptly fits the definition given by Paiva, Mendes, and Cardoso (2006) which is, "The dominant individual pitch line in a musical ensemble".

A large body of work on melodic analysis of music use predominant pitch as the basic representation of melody. Various problems such as melodic motif discovery (Ishwar, Dutta, Bellur, and Murthy (2013), P. Rao et al. (2014)), intonation analysis (Koduri, Serrà, & Serra, 2012), tonic identification (Gulati, 2012) use predominant melody as the basic representation. In addition to the errors that occur in the respective tasks

mentioned above, errors in pitch estimation also propogate into each task with different grades of consequences thus decreasing the overall accuracy of the analysis at hand. This necessitates the accurate estimation of predominant melody's pitch. Eventhough singing voice is the predominant melody in many music traditions, the direct application of the state of the art predominant melody estimation algorithms leads to various errors in estimating the vocal melody's pitch. These errors are discussed as a case study in section 3.1.

This work focuses on the estimation of pitch of the vocal lead in heterophonic audio music recordings. We present our work on Indian art music, a brief introduction to which is given in section 1.3. The goals of this work can be summarized as follows:

- To perform an extensive review of the present state of the art in predominant melody estimations
- To study the various types of errors in estimation of the pitch of the vocal melody using the present state of the art
- To incorporate the musical characteristics of the musical tradition at hand in the estimation of the pitch of the vocal melody
- To propose a meaningful evaluation methodology that entails the limitations and the descrepencies pertaining to the characteristics of the musical tradition being analysed

1.2 Some Definitions and Terminology

This section deals with the certain terms such as melody, pitch, heterophonic music, pitch extraction/estimation to name a few. It is important that we have a clear understanding of these terms and the way they are used throughout this entire thesis.

- **Melody**: The primary topic of interest with respect to this work is that of melody - extracting melody, characterizing it, and using it for other applications. Going back to the beginning of the chapter, melody is defined as *"The single(monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of*

that music when heard in comparison" by Poliner et al. (2007). This is one of the various definitions of melody. Some definitions like the one by Paiva et al. (2006) define melody to be, *"the dominant individual pitch line in a musical ensemble"* which assumes a certain melodic source to always be dominant in a musical ensemble. In this work we work with music cultures which have a dominant vocal melody source throuhg out the performance.

- **Pitch and fundamental frequency**: Poliner et al. (2007) defines melody as a single monophonic pitch sequence and Paiva et al. (2006) defines it as the dominant individual pitch line in a musical ensemble, which makes it necessary for us to understand the meaning of pitch in the first place. A definition for pitch is provided by Klapuri and Davy (2006) which is: *"pitch is a perceptual attribute which allows the ordering of sounds on a frequency-related scale extending from high to low"*. It is thus evident that pitch although related with frequency is a perceptual phenomenon. The closest scientific quantity that we can relate this perceptual quantity called pitch thus making it meausrable is to *fundamental frequency* ($f0$) (Salamon, 2013). For periodic or nearly periodic sounds, $f0$ is defined as the inverse of the period of the waveform (Klapuri & Davy, 2006). Unlike pitch which is a perceptual property, $f0$ is a physical quantity that can be measured and quantified.

- **Heterophony**: There are three main musical textures in Western music: *1) monophony, 2) homophony, 3) polyphony*. One of the less frequent textures that occur in non-western music traditions is heterophony. Heterophony in music is the musical texture when two or more sources simultaneously perform variations of the same melody line (Salamon, 2013). For example, in many non-western musical traditions like Indian art music, there are two melodic sources, one being the accompaniment which follows closely the lead melodic source at the same time playing variations of the same melody line embellishing the melody performed by the lead melodic source. This is important to understand for this work since we will be dealing with Indian art music in detail as a test case.

- **Melody Pitch extraction/estimation**: *Fundamental frequency estimation of a simgle predominant pitched source from polyphonic music signals with a lead voice or*

instrument. Note that pitch estimation and pitch extraction have been used interchangeably throughout this work depending on context.

1.3 Indian Art Music

Indian art music in this work, refers to the two major art music traditions of the Indian sub-continent, Hindustani music and Karṇāṭik music. Hindustani music is vastly popular in the northern regions of the sub-continent including, Pakistan, Afghanistan, Bangladesh and Nepal (Gulati, 2012). Karṇāṭik music is also known as South Indian music is popular in the southern regions of the Indian peninsular (Viswanathan & Allen, 2004). This section entails a few of the fundamental concepts and details pertaining to Indian art music.

Tonic is one of the most fundamental concepts in both the traditions of Indian art music. The tonic is the pitch chosen by the performer which serves as a reference throughout the performance. Melodies in both Karṇāṭik and Hindustani music are defined relative to the tonic. The tonic is one of the major factors affecting the identity of a melody in Indian art music (Krishna & Ishwar, 2012). The seven basic notes in Indian art music called svaras are Ṣaḍja, Ṛṣabha, Gāndhāra, Madhyama, Pañcama, Dhaivata and Niṣāda. These are represented in short form as *Sa, Ri/Re, Ga, Ma, Pa, Dha, and Ni* respectively. The Ṣaḍja or the svara Sa is the note which is the tonic pitch chosen by the performer. The frequency of the svara *Sa* determines the position of the other svaras or notes on the frequency scale. Thus a change in the tonic pitch will lead to a change in the frequencies of the other notes/svaras thus changing the identity of the melody.

Melody in Indian art music is based on the concept of rāga. Krishna and Ishwar (2012) define raga in the following manner, "A rāga is a collective melodic expression that consists of phraseology which is part of the identifiable macro-melodic movement. These phrases are collections of expressive svaras." . Thus a rāga is a concept which is an expression of the svaras (notes) with inflections which in turn form the phrases defining its phraseology and movements. The inflections on the svaras (notes) are called gamakās (*gamak* in Hindustani music). Due to the gamakās in Indian art music, the notes or svaras are a range of frequencies rather than a single point on the frequency scale. The characteristics of the pitch regions of the svaras are decided by

the characteristics of the rāga and its phraseology (Viswanathan & Allen, 2004). Serra (2011) study the characteristics of the svaras and tuning of Indian art music in comparison with Western classical music by means of pitch histograms. It is suggested that in Indian art music, there are more divisions in an octave than the 12 semitones found in Western classical music with spread out pitch distrubutions for each interval. It is also suggested that Indian art music follows a tuning system close to the just intonation system of tuning. Rhythm in Indian art music is based on the concept of tāḷa (Clayton (2000), Sen (2008)).

Though the fundamental concepts of melody and rhythm are similar in both Karṇāṭik music and Hindustani music, the music in both the traditions are significantly different from each other. Each tradition imbibes a specific cultural background and approach to music respectively.

Indian art music, over the centuries, has been an oral tradition that has been passed on from teacher to student hierarchically following the principle of school of music or gharānā[1] as referred to in Hindustani music (Saraf (2011), Mehta (2008)). Each gharānā has a unique ideology and thought process towards music defining its style of music performance.

Indian art music is heterophonic in nature with the main melody being sung or played by a lead artist (Bagchee, 1998). In a music performance there is generally a secondary melodic accompaniment closely following the melody of the lead artist (Viswanathan & Allen, 2004). A typical concert of Indian art music entails a lead artist (vocal or instrumental, occasionally a duo), a melodic accompaniment generally provided by a Violin in Karṇāṭik music and Harmonium or Sarangi in Hindustani music. The rhythmic accompaniment is generally provided by the mṛdaṅgaṁ and Tabla in Karṇāṭik and Hindustani music respectively. In both the traditions of Indian art music, a drone is constantly sounded in the background provided by an instrument called the *Tambura* which is a stringed instrument(Gulati, 2012). The *Tambura* provides a compound sound consisting of the tonic pitch and its lower octave (Sa) along with the lower octave fifth (Pa). This forms the reference of tuning for all the instruments on the concert platform. The only common instrument between Hindustani and Karṇāṭik music is the *Tambura* providing the reference drone. This is perhaps the only harmonic element inthe performance of Indian art music (Bagchee, 1998).

[1] http://en.wikipedia.org/wiki/Gharana

Chapter 2

Scientific Background and Related Work

This chapter reviews the large body of work done with respect to the main audio features, techniques and works relevant to the task of vocal melody pitch estimation. There is a large amount of literature on the extraction of pitch from digital audio. We review the scientific work on the different aspects such as signal representation, predominant melody pitch estimation and singing voice characterization which are fundamental to this work. We evaluate our approach on Indian art music. In order to emphasize the importance of the pitch of the vocal melody as a representation for the analysis of Indian art music, we also summarize the literature on the various computational tasks performed on Indian art music.

2.1 Computational analysis of Indian Art Music

Indian art music, as mentioned in Section 1.3 has two traditions, Hindustani Music and Karṇāṭik music. Recently, there has been significant work on the computational analysis of Indian art music. The melodic framework or rāga and the rythmic framework or tāḷa have a very specific unique identity. The similarities between the different rāgas and tāḷas can be exploited to obtain meaningful automatic content based information from Indian art music. With this motivation, there have been many works handling different tasks such as tonic identification, intonation analysis, melodic motivic anal-

ysis, rāga identification and rhythmic analysis of Indian art music.

The work by Gulati (2012) deals with tonic identification for Indian art music in detail. This work uses a multipitch approach for tonic identification. Various approaches for automatic tonic identification in Indian art music are reviewed in Gulati et al. (2014). All the approaches in this work use predominant F0 as the basic representation based on which various features emphasizing the presence of the tonic note are developed.

Indian music being highly melody centric, predominant F0 trajectory is used extensively as a representation to capture the melodic characteristics of the music. Indian art music, being replete with ornamentations (gamakās) (Viswanathan & Allen, 2004), makes intonation an important distinguishing factor between notes (svaras) and melodies (rāgas). Koduri, Serrà, and Serra (2012) perform a detailed intonation analysis of rāgas in Karṇāṭik music. They first perform a qualitative analysis of intonation by analysing varṇaṁs, a specific type of composition in Carnatic music. A quantitative analysis is later performed on a larger dataset by using predominant F0 characteristics and obtaining pitch distributions for the different svaras (notes) of the rāgas. Intonation analysis aids in the computation of melodic similarity which in turn leads to other similarities such as between artists, schools (gharānās) to name a few.

Ishwar et al. (2013), and P. Rao et al. (2014) propose methods for motivic analysis of Indian art music. Ishwar et. al. use a two pass dynamic programming algorithm through the Rough Common Longest Subsequence (RLCS) algorithm to search for the queried motif. The database consisted of a large number of audio music recordings of ālāpanas[1] annotated by a professional musician. A sparse representation of the F0 trajectory by quantising it at the saddle points[2] were used in this approach. P. Rao et al. (2014) propose a method for motif classification by learning the global constraints for Dynamic Time Warping(DTW). Ross and Rao (2012) propose a method for spotting the title phrases of songs using DTW. They use the rhythmic cues given by the rhythmic accompaniment for segmentation of the phrases and then determine similarity using DTW. Koduri, Gulati, Rao, and Serra (2012) propose a method for rāga recognition in Indian art music using pitch histograms and pitch dyads.

It is evident from all the works described above that accurate predominant F0 estimation is fundamental for melodic analysis of Indian music. Errors in predominant

[1] ālāpana is an improvisational aspect of Indian art music sans rhythmic accompaniment.
[2] http://en.wikipedia.org/wiki/Saddle_point

F0 estimation cumulatively increase the errors in the other predominant F0 dependant tasks.

2.2 Predominant Melody Pitch Extraction

In this section, the literature on predominant melody extraction is briefly reviewed. For a detailed review of the state of the art in predominant F0 estimation and multipitch analysis, the reader is referred to the work by Gulati (2012). Predominant melody pitch extraction is the estimation of the pitch of the predominant melody line from an audio recording of polyphonic music. Here, the predominant melody is considered to be the time varying F0 trajectory of the lead artist or the dominant melody source. Salamon, Gómez, Ellis, and Richard (2014) review in detail the various predominant melody pitch extraction algorithms in the last decade. The work performs a comparitive analysis of the various algorithms based on the results of MIREX (an international campaign for evaluation of various tasks in MIR)[3], also discussing the various challenges involved in the design of these algorithms. One of the largest group of methods for predominant melody pitch extraction have been described as "salience based" in Salamon and Gomez (Aug. 2012.). This approach involves four steps viz. 1) *Preprocessing*, 2) *Spectral Processing*, 3) *Salience Function* and 4) *Melody Selection.* A detailed illustration of all the steps is given in (Salamon et al., 2014) which also covers a detailed review of the vast state of the art in melody extraction using the salience based approach.

Salamon and Gomez (Aug. 2012.) propose a method of predominant melody pitch extraction using the salience approach. After computing a salience function based on harmonic summation, a peak picking is performed on the salience function, a time salience representation. These peaks are streamed and grouped into pitch contours which are the candidates for melody selection. Melody selection in this work involves the characterization the pitch contours using various features and setting thresholds learnt from the distribution of the features for the two classes, predominant melody and non-predominant melody. Figure 2.1 gives an overview of the approach taken by them.

The concept of a secondary melodic instrument is prevalent in many music tra-

[3]http://www.music-ir.org/mirex/wiki/MIREX_HOME

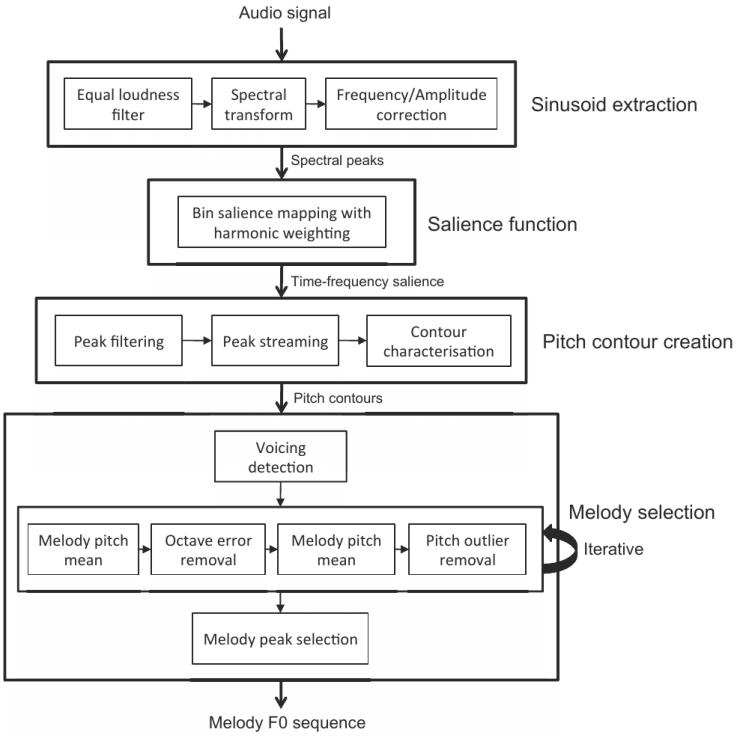

Figure 2.1: Overview of the Algorithm proposed by Salamon et.al.

ditions. The voice of the lead vocalist is considered the source for the predominant melody in such genres (Viswanathan & Allen, 2004). The work of V. Rao and Rao (2010) describes a method of extracting predominant melody pitch in the presence of pitched accompaniments from Indian art music. The assumption in this work is that the source of the predominant melody is the singing voice. This work is another salience based method where the salience function is computed using the two way mismatch error. Instead of tracking all the fundamental frequencies, from all the sources simultaneously using multi-F0 estimation methods, this work tracks the F0 of two sources at a time simultaneously. The justification suggested for this decision is that there cannot be more than one source more salient than the voice at a given time instant in a vocal music audio recordings. A feature called Sinusoidal Track Harmonic Energy (STHE) described in V. Rao, Ramakrishnan, and Rao (2009) is used to narrow down the F0 corresponding to the predominant melody.

Another set of approaches to extract predominant melody is by separating the

source of the desired melody from the rest of the audio using source separation methodologies. For instance, Durrieu, Richard, David, and Fevotte (2010) use the source filter model of the voice for unsupervised main melody pitch extraction from polyphonic audio signals. The model parameters are estimated using the expectation maximization framework. A Viterbi search is carried out for a smooth trajectory through the model parameters (which contains the f0 of the source) to get the final F0 trajectory of the voice. Tachibana, Ono, Ono, and Sagayama (2010) exploit the temporal variation of melody in contrast with the chord notes that are more sustained. The authors use harmonic percussive source separation (HPSS) to obtain an enhanced signal with accompaniment reduction. The final predominant melody pitch trajectory is obtained from the spectogram of the enhanced signal using dynamic programming, finding the path which maximizes the MAP(maximum a posteriori) estimation of the frequency sequence. For an exhaustive review of state of the art in predominant melody extraction, the reader is referred to Salamon (2013).

2.3 Singing Voice Detection

The human voice has many distinguishing qualities which allows its unique characterization. In this section we look at the literature on singing voice detection and characterization in order to study the various features used to distinguish singing voice from other accompanying instruments in polyphonic audio music reordings.

Singing Voice Detection (SVD) involves the classification of a signal into vocal and non-vocal segments. The classification may be of two kinds:

1. Singing Voice Detection in audio music recordings where isolated vocal and non-vocal segments follow one another seamlessly and are non-overlapping.

2. Singing Voice Detection in audio which is polyphonic with both vocal and accompanying instruments play simultaneously.

Feng, Nielsen, and Hansen (2008), in their work use MFCCs with a Multivariate Auto-Regression mechanism for feature selection to classify one second audio segments as vocal or non-vocal. The database consisted of 147 songs cut into segments of one second each. In Tzanetakis (2004), a semi-automatic method for classification of vocal

and non-vocal audio segments is proposed. The hypothesis is that the characteristics of singing voice and instruments change from song to song due to differing styles of music. Hence, song specific training data is collected by presenting the user with random excerpts from the song for annotation. A classifier is then built from this data and the rest of the song is classified into vocal and non-vocal segments. A detailed review of the state of the art in singing voice detection is given in (Mesaros, 2012).

The main difficulty in detecting singing voice from polyphonic music audio is the presence of accompanying instruments along with the vocals. Sometimes, the accompanying instruments have larger amplitude than the singing voice in certain regions of the audio recording. We now discuss state of the art with respect to singing voice detection in the presence of accompaniments with a special stress on works which deal with music traditions that entail pitched accompaniments. V. Rao et al. (2009), in their work, extract harmonic components from the spectrum of the polyphonic audio using predominant F0, for singing voice detection. They use energy based features derived from the harmonic components based on the knowledge from F0 trajectory of the source. They account for loud instrument accompaniment by exploiting the relative instability in singing voice pitch with respect to that of other instruments. Their method to isolate the dominant source spectrum enhances the features towards being more representative of the dominant source.

Shenoy, Wu, and Wang (2005), in their work, pass the spectrum through a series of inverse comb filters. This attenuates the harmonic spectral content of pitched instruments. The basic hypothesis here is that voice, having an unstable pitch, its harmonic content will be partially attenuated in relation to other pitched instruments with stable pitch. They attribute the instability in the "singers F0" to vibrato and intonation. This instability in the "singers F0" is exaggerated in Indian art music due to the presence of gamakās (see Section 1.3). In addition to regions with ornamentations and intonation, the singing voice at stable pitch regions will still have jitters and flutters (V. Rao et al., 2009) [4]. An adaptive threshold applied to the energy of the residual signal after the inverse comb filtering is used in Shenoy et al. (2005) for making frame level vocal/non-vocal decisions. The delay used in the design of the inverse comb filters is driven by key estimation.

In Fujihara, Goto, Kitahara, and Okuno (2010), a spectral processing technique

[4] Jitter is a terminology used in speech processing

called accompaniment sound reduction using predominant F0 for robust singing voice modeling in polyphony is presented. Two techniques, namely, accompaniment reduction and reliable frame selection are suggested in this work. The harmonic components of the spectrum are first extracted using the knowledge of predominant F0 at that instant and then synthesized using sinusoidal modelling. The reliability (influence of accompaniments on the harmonic spectral content for each frame) is estimated using separate GMMs for vocal and non-vocal frames. Finally each song is represented by its GMM consisting of the reliable frames.

V. Rao, Gupta, and Rao (2011) propose a method similar to the work of Fujihara et al. (2010) by performing dominant source spectrum isolation. A combination of static and dynamic features is used for timbre description of voice and instruments. This method is applied across five genres of music. The authors claim significant increase in the performance of the static timbral features for polyphonic music over the baseline features like MFCC's. The use of features representing the timbral dynamics and F0 and harmonic frequency dynamics provide information accounting for different signal conditions related to singing styles and accompanying instruments across genres. Bapat, Rao, and Rao (2007), use the apriori information about the instruments accompanying the singing voice to build GMMs for vocal and non-vocal segments using frame level feature vectors from a representative dataset. Due to the similarity in harmonic content of the voice and the accompanying instruments, spectral subtraction is used to supress the harmonic spectral content of the latter.

2.4 Features for Singing Voice Description

Bapat et al. (2007) use features distinguishing between the drone, percussion and voice in North Indian classical (Hindustani) music. The range of the frequency content of percussion(*tabla*) in Hindustani music is about 0-1500Hz as compared to singing voice which has a range from 0-5kHz. Thus features are chosen such that the difference in energies of the voice and *tabla* is accentuated above 1500Hz. Based on apriori knowledge about the acoustic properties of voice and accompanying instruments in Hindustani music spectral roll off was chosen as one of the features in Bapat et al. (2007). Spectral roll off is the frequency below which X% of the signal energy is concentrated (Peeters, n.d.). In the work by Bapat et al. (2007), 70% was found to provide good separation be-

tween tonal tabla strokes, drone and the voice. The harmonic energy is defined as the sum of the strength of individial harmonics. With apriori knowledge of the extracted predominant F0, this feature was computed using the algorithm described in V. Rao, S, and Rao (2008). This feature has a high value during voiced regions and lower values in instrumental regions. Sub-band energy ratios between; 1) frequency bands 5kHz to 8kHz and 0 to 1.5kHz and 2) frequency bands 2.5kHz to 5kHz and 0 to 1.5kHz are also computed. This feature is expected to have low values in voiced regions. Another feature in Bapat et al. (2007) is the sub-band spectral flux which is expected to have higher values in sung regions of the audio.

In the work Thibault and Depalle (2004) a number of voice timbre descriptors are computed for adaptive processing of the voice timbre. Harmonic and stochastic modelling of the signal is used in order to separate the harmonic and stochastic content of the signal. Instantaneous and global descriptors computed in this work are, harmonic deviation, noise spectral centroid and harmonic to noise ratio each of which is defined in Equations 2.1, 2.2 and 2.3 respectively where a_i, I, f_s and N represent respectively the amplitude of the i^{th} harmonic, the total number of harmonics present in the current frame, the sampling frequency and the length of the spectrum, for the m^{th} frame.

$$HD_m = \sum_{i=1}^{I} |fi - (iF0)| \frac{a_i}{\sum_{i=1}^{I} a_i} \tag{2.1}$$

$$NoiSC_m = \sum_{k=1}^{N} \frac{k}{N} f_s . \frac{|E[k]|}{\sum_{k=1}^{N} |E[k]|} \tag{2.2}$$

$$HNR_m = \frac{\sum_{i=1}^{I} a_i}{\sum_{k=1}^{N} |E[k]|} \tag{2.3}$$

$$Fojitter = \frac{\sum_{m=2}^{M-1} F_0[m] - \frac{F_0[m-1] + F_0[m] + F_0[m+1]}{3}}{M-2} \tag{2.4}$$

F0 jitter (Eq 2.4) which is the variation of F0 for a voiced segment is also computed as a global feature across a number of frames. This feature helps in distinguishing

between voice and accompanying instruments in steady F0 regions.

Ricard Maxer in his thesis Marxer (2012) estimates the timbre for a set of F0 candidates using a variant of the MFCCs. A harmonic spectral envelope is estimated from the harmonics of the F0 contour. The first 13 coefficients of the DCT of the harmonic spectral envelope are used as a timbral feature. This feature is computed per candidate F0 contour and the target instrument from which it is generated is estimated. In addition to the timbral feature the work also computes the static features described in the work by Salamon and Gomez (Aug. 2012.). A support vector machine classifier with radial basis function kernel is used for classification. The output of the timbral classification is then combined with the F0 likelyhoods to perform instrument pitch tracking (Marxer, 2012).

Rocamora and Harrera in Martin and Perfecto (2007) explore various descriptors to perform SVD and compare the performance of a statistical classifier using each descriptor. The work suggests Mel Frequency Cepstral Coefficients (MFCC) to be the most appropriate for the task of SVD. MFCC's, Perceptually derived LPC (PLPC), Log Frequency Power Coefficients (LFPC) and Harmonic Coefficient (HC) were the spectral features implemented. In addition, a general purpose musical instruments classification feature set was built entailing Spectral Centroid, Roll-off, Flux, Skewness, Kurtosis and Flatness. Pitch was also included, being the only non-spectral feature reported. Harrera et. al. in their book chapter Herrera, Klapuri, and Davy (2006) implement a set of various features describing various instruments in their task of building an automatic instrument classifier. The set of features are given in the Table 2.1 taken from Herrera et al. (2006).

A combination of static and dynamic features were experimented with for SVD by V. Rao et al. (2011). Dominant source isolation on the same lines as Fujihara et al. (2010) was performed post which a harmonic spectral envelope is estimated given the predominant F0. The harmonic spectral envelope is said to be representative of the resonances or formants of the instrument or voice respectively. The sub-band spectral centroid and the sub-band energy is computed as given by equations 2.5 and 2.6, where f(k) and |X(k)| are the frequency and magnitude spectrum values of the k_{th} frequency bin, and k_{low} and k_{high} are the nearest frequency bins to the lower and upper frequency limits on the sub-band respectively. The band for Spectral Centroid ranges from 1.2-4.5 kHz and that for Spectral Energy ranges from 300 to 900 Hz.

Table 2.1: MPEG-7 Descriptors, organized according to category. Each of these can be used to describe an audio segment with a summary value or with a series of sampled values. Timbral spectral descriptors are computed after extracting the relevant harmonic peaks from the spectrum of the signal. Spectral basis descriptors are a spectral representation of reduced dimensionality.

Category	Descriptors
signal parameters	fundamental frequency, harmonicity
basic	instantaneous waveform, power values
basic spectral	log-frequency power spectrum envelopes, spectral centroid, spectral spread, spectral flatness
timbral spectral	harmonic spectral centroid, harmonic spectral deviation, harmonic spectral spread, harmonic spectral variation
timbral temporal	log attack time, temporal centroid
spectral basis	spectrum basis, spectrum projection

$$SC = \frac{\sum_{k=k_{low}}^{k_{high}} f(k)|X(k)|}{\sum_{k=k_{low}}^{k_{high}} |X(k)|} \tag{2.5}$$

$$SE = \sum_{k=k_{low}}^{k_{high}} |X(k)|^2 \tag{2.6}$$

The explicit modelling of temporal dynamics of spectral features are used by Lagrange, Raspaud, Badeau, and Richard (2010) for characterizing timbre. On similar lines, V. Rao et al. (2011) use features linked to the temporal evolution of spectral envelope designed to capture specific attributes of the instrument sound. They compute the standard deviation in the spectral centroid over 0.5s, 1s, and 2s intervals of audio in order to extract meaningful temporal variations. The modulation energy ratio (MER) is extracted by computing the Discrete Fourier Transform of the feature trajectory over a 0.5s, 1s or 2s window called a texture window and then computing the ratio of the energy in the 1- 6 Hz region in this modulation spectrum to that in the 1- 20 Hz region as shown in equation 3.10.

$$MER = \frac{\sum_{k=k_{1Hz}}^{k_{6Hz}} |Z(k)|^2}{\sum_{k=k_{1Hz}}^{k_{20Hz}} |Z(k)|^2} \tag{2.7}$$

Singing differs from several musical instruments in its expressivity, which is shown in its pitch instability. In the work of V. Rao et al. (2011), some statistical descriptors such as standard deviation and mean of general pitch instability features over a texture window of note duration (approx 200ms) are computed. These features are the first order difference in the predominant F0 and the subsequent harmonic frequency tracks. The harmonic frequency tracks are first normalized by harmonic index and then converted to logarithmic cents scale so as to maintain the same range of variation across harmonics and singers' pitch ranges.

The review of the works above gives us an insight into the tremendous scope for improvement using features and methods that incorporate the intricacies and cultural backgrounds of the various genres of music. It is clear from the work of Salamon et al. (2014) that a generalized algorithm for predominant melody pitch extraction cannot be subjected to all genres of music.

Chapter 3

Vocal Melody Pitch Extraction

This chapter covers a detailed description and illustration of the approach taken for vocal melody pitch extraction from audio music recordings of heterophonic music. We propose a method for vocal melody pitch extraction incorporating the timbral characteristics of the singing voice and the musical concepts specific to the genre of music at hand. This work has been tested on Karṇāṭik music and hence certain aspects of the music have been considered as important cues for selecting the correct pitch contours corresponding to the vocal melody. We propose that the combination of timbral features and features characterizing the pitch contour aid in the selection of appropriate vocal melody pitch contours. Before we delve into the details of the approach, we perform a case study of one of the recent state of the art algorithms in predominant melody pitch extraction, in section 3.1. A brief overview of the entire system is described in the section 3.2. We describe the features used in the work and the approach taken for feature extraction in section 3.4. Sections 3.5 and 3.6 detail the classification metodology used and the post classification processing applied respectively.

3.1 Case Study

In this section, a case study is presented where in we use the state of the art algorithm proposed by Salamon and Gomez (Aug. 2012.) to extract predominant melody from different excerpts of Karṇāṭik music (see 1.3). We restrict ourselves to the algorithm proposed by Salamon and Gomez (Aug. 2012.) for conciseness. Figure 3.1 shows the

output of the pitch extraction algorithm by Salamon and Gomez (Aug. 2012.) for two excerpts of Karṇāṭik music.

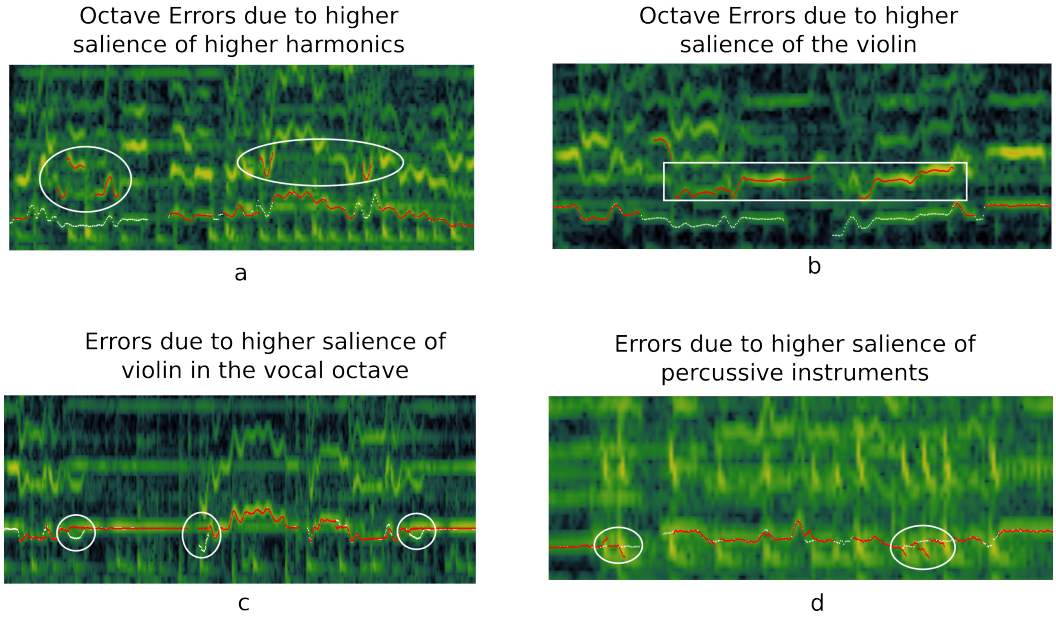

Figure 3.1: Various errors encountered in Predominant Melody Pitch Extraction, White - Ground Truth, Red - Pitch from Salamon et.al. 2012

In Karṇāṭik music, octave errors are caused due to two types of confusion one of them being the confusion between instruments and the other being that between harmonics. The first one is caused by one of the accompanying melodic instruments being more salient (loud) than the singing voice. This case is shown in Figure 3.1b & 3.1c. The white pitch contour is the ground truth where as the red pitch contour is that which is extracted using the algorithm proposed by Salamon and Gomez (Aug. 2012.). In Figure 3.1b, the errors are due to the higher salience of the melodic accompaniment (violin) playing in the higher register. This looks like an octave error due to higher salience of the second harmonic of the singing voice but on listening we observed that it is the violin which is more salient in this region and is very closely following the vocalist giving rise to the error. In Figure 3.1c the violin is more salient than the singing voice in the same register as that of the singing voice thus giving rise to the error indicated. These observations are a result of careful scrutiny of the pitch contours and intent listening sessions with a professional musician.

Figure 3.1a, illustrates octave errors caused because of the higher salience of the higher harmonics. In heterophonic music with vocal lead, the vocalist produces a highly resonant sound so that he is heard clearly over the accompaniments. As pointed out in (Salamon et al., 2014), at lower frequencies, this phenomenon causes the second harmonic(2f0) to be more salient and have higher amplitude. In male singers, this style of singing brings about an extra formant called the "singers formant " (see Sundberg (1977)), which makes the algorithm tend towards estimating the pitch of the second harmonic rather than that of the predominant vocal melody.

Plot 3.1 d illustrates the errors occurring due to the percussion instrument being more salient than the singing voice. As seen around the circled regions, the pitch corresponding to the onset of the percussion is estimated which is due to the greater salience of the percussion at that time in the audio. These onsets are those of the base side of the mṛdaṅgaṁ(the main percussion instrument in Karṇāṭik music). This excerpt is from a different audio recording of a different artist.

The case study described above clarifies that there are a number of errors that occur due to timbral confusions when there are two pitched melody sources. We can also conclude that some errors occur due to recording conditions which accentuate the salience of certain instruments and diminish that of the others. These factors cannot be accounted for by using salience based methods alone and thus algorithms incorporating timbral characteristics of the melody sources are required to be developed. For example, Figure 3.1d is from the recording of a different concert by a different artist where the percussion is a little louder than the vocal which causes the algorithm to identify the pitch corresponding to percussive onsets as against that of the melody. Thus, in order to try and alleviate these errors we propose a method including the timbral characteristics of the voice along with a salience based approach which uses pitch contour characteristics. The section that follows gives an overview of the entire approach, post which, each step in the approach is described and illustrated in detail.

3.2 Overview of the System

This work proposes an approach for vocal melody pitch extraction from heterophonic music audio recordings with a single vocal melody source using timbral features, pitch contour characteristics and a combination of both. The human voice has unique char-

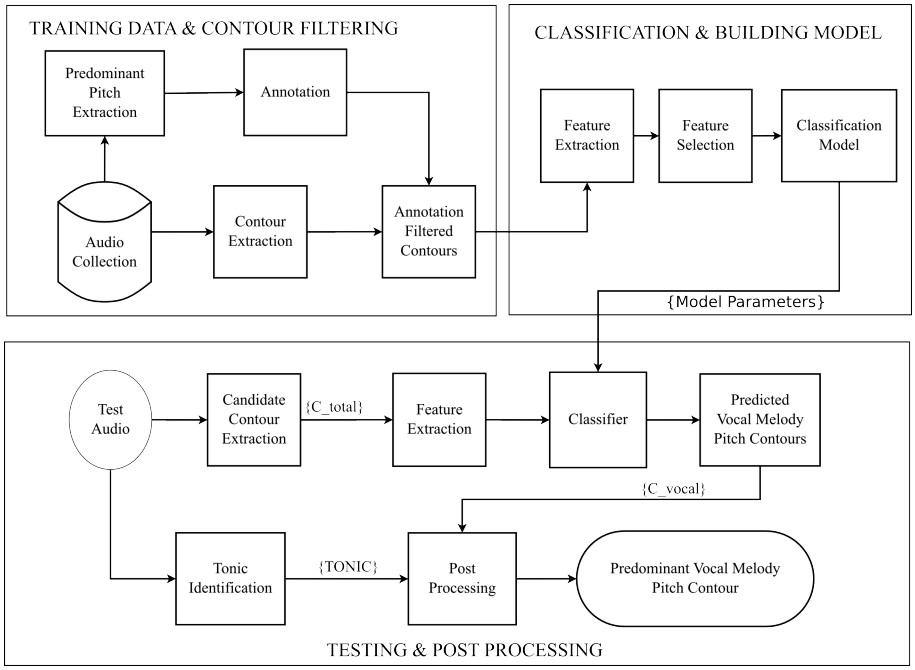

Figure 3.2: Overview of the system

acteristics which can be captured using various timbral features computed from the spectogram of the audio signal. Also from the case study performed in section 3.1 it is clear that many false positives which arise due to timbral confusions are not clearly distinguishable using just salience derived features. Hence in this work, we try to characterize the voice using various timbral features in addition to the pitch contour characteristics to achieve our final goal of extracting the pitch corresponding to the main vocal melody. The Figure 3.2 illustrates the approach in brief.

This approach is tailored to work at the contour level which means that the basic unit is not a frame but is a pitch contour which we obtain as a candidate using the algorithm proposed by Salamon and Gomez (Aug. 2012.). This candidate contour is basically classified as being vocal or non-vocal under the assumption that there is only one vocal melody source in the audio music recording and also under the assumption that it is the predominant melody source.

The candidate contours obtained are characterized by extracting features based on the static and dynamic characteristics of the pitch and pitch salience function corresponding to that contour. Timbral features from the audio corresponding to the pitch

contours are also extracted. The pitch contour features are the mean, median and standard deviation of the pitch and of the salience function peaks corresponding to the pitch contour. The total salience of a pitch contour from the salience function is also computed. The features are as described in Salamon and Gomez (Aug. 2012.) and V. Rao et al. (2011).

In addition to these features, we also compute certain timbral features for every contour and obtain a timbral feature vector C_t for each candidate pith contour. For this we capatalize on the information we have from the candidate contours and compute the harmonic amplitudes and bin frequencies for the first 30 harmonics for every frame from the audio corresponding to that candidate contour. We compute the harmonics using harmonic modelling of the signal. With this representation of harmonics in time (see Figure 3.5) we compute several timbral features for the source of the candidate pitch contour. A detailed description of the features and its implementation are given in section 3.4.

With these features at hand, we perform a feature selection for the timbral features alone. The features characterizing the pitch contour dynamics are not included in the feature selection process since they are already hand picked features from (Salamon & Gomez, Aug. 2012.). Thus, we form three feature sets which are timbral features, pitch contour features and a combination of pitch contour features and timbral features. Support vector machine classifier with an RBF kernel is used to build models for this classification task. Contours predicted as vocal contours are considered for further processing.

The vocal contours obtained after classification are further processed incorporating the properties of the music and some heuristics in order to obtain a single predominant vocal melody pitch contour for the audio music recording. The classified contours may contain multiple contours at the same time instant for which we perform two steps of contour selection described below.

- The longest contour at a time is chosen under the assumption that melody contours are longer than non-melody contours (Salamon & Gomez, Aug. 2012.).

- We exploit the knowledge of the range of the vocalists voice in Karṇāṭik music. This range is defined by the tonic pitch chosen by the lead vocal artist (see Section 1.3). In this work, we have excluded contours that are outside the frequency

range of 0.5τ to 3τ, τ being the tonic pitch. Tonic for each recording is computed automatically using the work by Gulati (2012).

It is possible that at a given time more than one candidate contour be classified as vocal. In this case a choice is made based on length of the contour. At a given time, the longest contour is chosen as the correct contour. This is based on the assumption in the work of Salamon et. al. (Salamon & Gomez, Aug. 2012.) wherein it is suggested that the contours belonging to the harmonics and other suprious contours are of lesser length than that of the predominant melody pitch contours. The contour we obtain after these contour selection steps is the final pitch contour of the vocal melody for that audio music recording. This contour is evaluated against the ground truth obtained for that file the details of which are given in chapter 4.3. The following sections illustrate in detail each step of the approach.

3.3 Data Pre-processing

The dataset consists of a carefully chosen set of songs which are a subset of a diverse representative dataset of Karṇāṭik music compiled by the CompMusic Project[1]. The contents of the dataset are explained in detail in section 4.1. This section entails the pre-processing on the audio and pitch data which helps facilitate the usage of the data for further analysis.

The audio data in the CompMusic collection is in the form of stereo mp3 recordings at 160 KBPS. The representative set of songs chosen from the collection is first converted into mono wave files which are sampled at 44.1 kHz. The audio is converted to mono by by summing up the two stereo channels using the audio editing tool audacity[2]. Predominant melody pitch contours are then extracted for each of the files using the algorithm proposed by Salamon et.al. in Salamon and Gomez (Aug. 2012.). The implementation in the essentia library is used for this purpose [3]. The audio is then annotated manually for time instants in the audio where the pitch contour for the vocal melody has been tracked accurately. The annotations are carried out using the audio

[1]http://compmusic.upf.edu/corpora
[2]http://audacity.sourceforge.net/
[3]http://essentia.upf.edu/

analysis and visualization tool Sonic Visualiser[4]. The annotations were performed by a professional musician. The analysis and feature extraction is done only in the annotated segments in order to create a curated training dataset. This data is further used for the classification experiments and model building.

3.4 Feature Extraction

In this section we detail the extraction process of various features used in the approach proposed in this work. We start with detailing the process of obtaining the candidate pitch contours in Section 3.4.1 which is the basis of every step that follows in this approach. In Section 3.4.2 we detail the process of extracting harmonics and the representation we use for extraction of timbral features which is explained in Section 3.4.3.

3.4.1 Candidate Pitch Contour Extraction

In this work, features are extracted for every candidate contour obtained using the algorithm proposed by Salamon and Gomez (Aug. 2012.). We obtain these candidate pitch contours by intervening the algorithm at the contour creation step as shown in the Figure 3.3 (See Figure 2.1 for the complete block diagram). A brief explanation of the approach by Salamon and Gomez (Aug. 2012.) is given in this section.

Given a music audio recording it is first passed through an equal loudness filter. This is perceptually motivated in that the frequencies to which the human ear is sensitive are accentuated. The implementation details are similar to that which is given in Salamon and Gomez (Aug. 2012.). The frequency representation of the filtered signal is then obtained by taking its Short Time Fourier Transform (STFT). This work uses a window size of 46ms and a window hop of 2.9ms for the STFT computation. This is followed by performing frequency and amplitude correction so that the error in spectral peak estimation at low frequencies is minimized (Salamon & Gomez, Aug. 2012.). An instantaneous frequency approach proposed by Keiler and Marchand (2002) is used for this purpose by Salamon and Gomez (Aug. 2012.). In this approach, the phase spectrum, $\phi(k)$ is used to calculate the peaks instantaneous frequency (IF) and amplitude which provide a more accurate estimate of the peak's frequency and amplitude.

[4]http://www.sonicvisualiser.org/

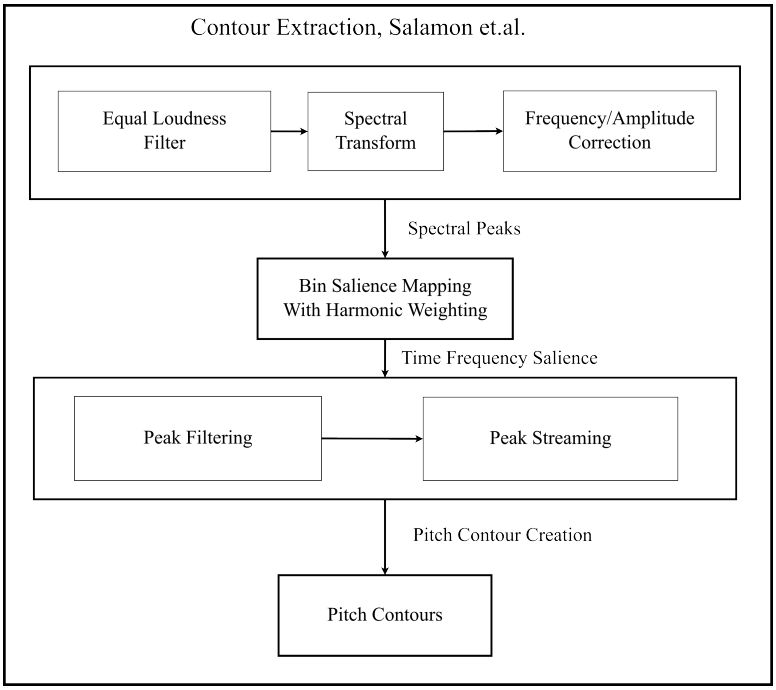

Figure 3.3: Candidate Contour Extraction, Salamon et.al.

The IF $\widehat{f_i}$ of a peak p_i found at bin k_i is computed from the phase difference $\Delta(k)$ of successive phase spectra using the phase vocoder method as follows:

$$\widehat{f_i} = (k_i + \kappa(k_i))\frac{f_s}{N} \tag{3.1}$$

where the bin offset $\kappa(k_i)$ is calculated as:

$$\kappa(k_i) = \frac{N}{2\pi H}\Psi\left(\phi(k_i) - \phi_{l-1}(k_i) - \frac{2\pi H}{N}\right) \tag{3.2}$$

where Ψ is the principal argument function which maps the phase to the $\pm\pi$ range. The instantaneous magnitude $\widehat{a_i}$ is calculated using the peak's spectral magnitude $|X_l(k_i)|$ and the bin offset $\kappa(k_i)$ as follows:

$$\widehat{a_i} = \frac{1}{2}\frac{|X_l(k_i)|}{W_{Hann}\left(\frac{M}{N}\kappa(k_i)\right)} \tag{3.3}$$

where W_{Hann} is the Hann window kernel.

The extracted spectral peaks are used to formulate a salience function which is the representation of the pitch salience over time. The salience function computation is based on harmonic summation similar to Klapuri (2006), where the salience of a given pitch frequency is computed as the sum of the weighted energies of the harmonics of that frequency. Only the spectral peaks are used in the approach proposed by Salamon and Gomez (Aug. 2012.). The salience function designed by Salamon and Gomez (Aug. 2012.) covers a pitch range of nearly 5 octaves from 55Hz to 1.76kHz, quantized into b = 1...600 bins on a cent scale (10 cents per bin). Given a frequency $\hat{f}$ in Hz, its corresponding bin $B(\hat{f})$ is calculated as:

$$B(\hat{f}) = \left[\frac{1200 \cdot log_2\left(\frac{\hat{f}}{55}\right)}{10} + 1 \right] \tag{3.4}$$

At each frame the salience function $S(b)$ is constructed using the spectral peaks p_i (with frequencies $\hat{f}_i$ and linear magnitudes $\hat{a}_i$) returned by the sinusoid extraction step (i = 1...I, where I is the number of peaks found). The salience function is defined as:

$$S(b) = \sum_{h=1}^{N_h} \sum_{i=1}^{I} e(\hat{a}_i) \cdot g(b, h, \hat{f}_i) \cdot (\hat{a}_i)^{\beta} \tag{3.5}$$

where is a magnitude compression parameter, $e(\hat{a}_i)$ is a magnitude threshold function and $g(b, h, \hat{f}_i)$ is the function that defines the weighting scheme for the peaks. The magnitude threshold function is defined as:

$$e(\hat{a}_i) = \begin{cases} 1, & \text{if } 20log_{10}(\hat{a}_M/\hat{a}_i) < \gamma \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

where $\hat{a}_M$ is the magnitude of the highest spectral peak in the frame and γ is the maximum allowed difference (in dB) between the i_{th} amplitude and $\hat{a}_M$. The weighting function $g(b, h, \hat{f}_i)$ defines the weight given to peak p_i when it is considered the h^{th}

harmonic of the bin b:

$$g(b, h, \hat{f}_i) = \begin{cases} cos^2(\delta \cdot \frac{\pi}{2}) \cdot \alpha^{h-1}, & \text{if } |\delta| \leq 1 \\ 0, & \text{if } |\delta| > 1 \end{cases} \tag{3.7}$$

where $\delta = |B(\hat{f}_i/h)|/10$ is the distance in semitones between the harmonic freqency and the centre frequency of the bin b and α is the harmonic weighting parameter. In this work we extract 30 harmonics making $N_h = 30$ and the parameter is set to 1. The non zero values for $|\delta| < 1$ suggests that each sinusoid contributes not only to one single bin but also to the neighboring bins with a cos^2 weighting. This avoids the potential problems that may arise due to quantization of the salience function into bins and inharmonicities present in the audio. Figure 3.4 shows the salience function for an excerpt of Karṇāṭik music. Once the salience function representation is

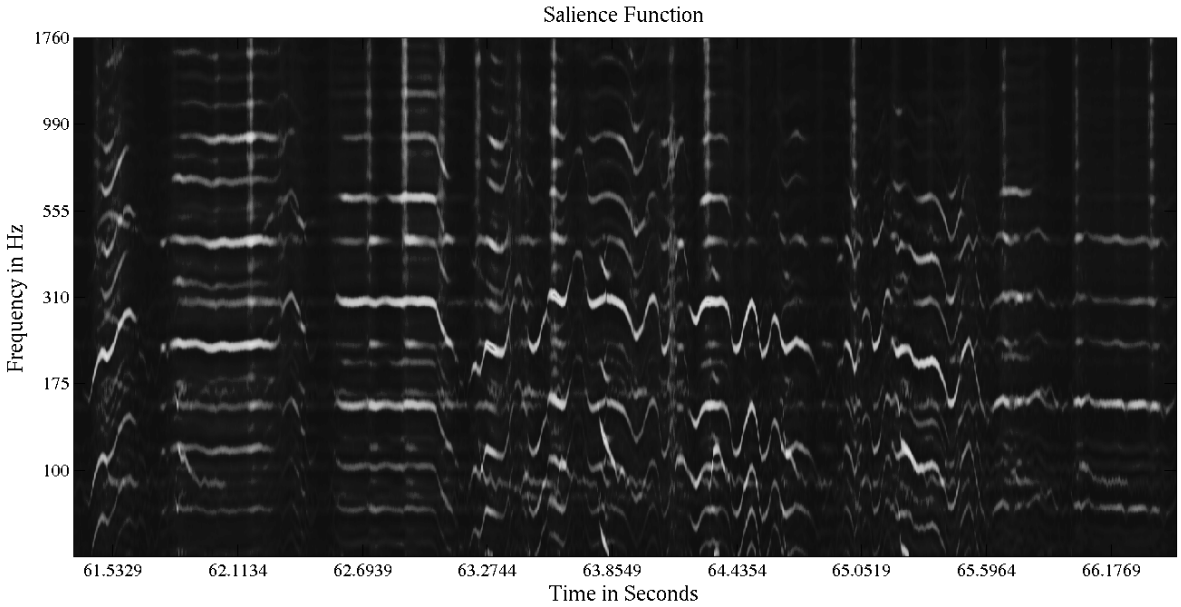

Figure 3.4: Salience Function Representation

obtained, the peaks of the salience function at every frame are considered as potential F0 candidates. This gives a multi-pitch representation of pitch over time. Unlike other sate of the art that attempt to track the melody directly from these peaks, Salamon and Gomez (Aug. 2012.) group these F0 candidates into groups of smaller pitch con-

tours. The length of these contours can range from that of a note to that of a small phrase. Salamon and Gomez (Aug. 2012.) further perform melody selection using pitch contour characteristics alone to select the pitch contours corresponding to that of the predominant melody. This work stops at the contour creation step and uses the candidate pitch contours for further analysis and classification. Since the salience function gives a multi-pitch representation, the contours obtained from it by grouping the peaks of the salience function represent the contours from the different sources in the audio. In this work, the implementation of this algorithm from the essentia library was used. We further classify these contours into vocal and non-vocal contours and finally obtain the vocal melody's pitch contour.

3.4.2 Estimation of Harmonics

In this work, an attempt is made to compute features that represent the timbre of the source of the candidate pitch contour obtained as described in the previous sub section. Taking advantage of the pitch information at hand, for each candidate contour, a representation of the harmonics over time is extracted using harmonic sinusoidal modelling. Given a candidate pitch contour, the harmonics of the pitch contour are extracted from the spectrum of the audio corresponding to the pitch contour. We estimate 30 harmonics for every frequency $\hat{f}_i$ constituting the pitch contour ($i = 1...I$, where I is the number of samples in a candidate pitch contour). We estimate 30 harmonics in order to exploit the instability of the vocal pitch at higher harmonics as compared to other instruments. The estimation of 30 harmonics also enables to distinguish voice from other instruments due to the spectral brightness of the voice at higher frequencies lacking in other instruments.

The process of estimating harmonics is illustrated in Figure 3.5. The frequency representation of the audio signal is obtained by performing Short time Fourier analysis. Parabolic interpolation is used for locating the peaks corresponding to the fundamantal $\hat{f}_i$ and its harmonics $h_n\hat{f}_i$, where h_n is the harmonic index and $n = 1...30$. This gives a more accurate estimate of the peak locations and amplitudes. Once the harmonics amplitudes and the corresponding bin frequencies are extracted for every $\hat{f}_i$ in the candidate pitch contour, a representation of harmonic amplitudes over time is obtained by placing the harmonic amplitudes corresponding to every frame adjecent

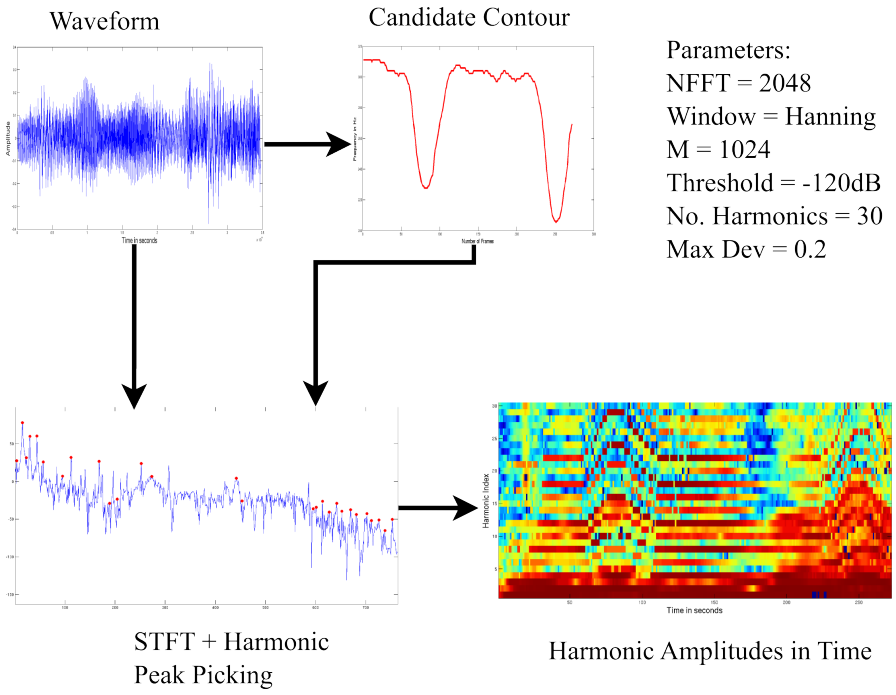

Figure 3.5: Representation of Harmonic Amplitude in Time

to each other without taking into consideration the locations. This is the timbral representation we obtain pertaining to the candidate pitch contour. The implementation of harmonic sinusoidal modelling from SMS tools[5] was used in this work. It is from this representation of the harmonics that all the features characterizing the timbre of the source of the candidate pitch contour are extracted.

3.4.3 Timbral Feature Extraction

One of the major bottle necks in the effective use of timbral dynamics for the classification of instruments in polyphonic or heterophonic music was found to be the isolation of individual instrument spectra and the inability to pay selective attention to them. V. Rao et al. (2011) propose a method using isolated dominant source spectral representation to extract features corresponding to the isolated spectra of the lead instrument. In this work we obtain a representation of the estimated harmonic amplitudes of the candidate pitch contour over time as a timbral representation of the source of that

[5]https://github.com/MTG/sms-tools

candidate pitch contour from the corresponding audio signal.

Singing differs from several musical instruments in its expressivity, which is shown in its pitch instability. This work, tries to capitalize on the differing timbral dynamics in time between voice and the other accompanying instruments. Features representing the various static and dynamic characteristics of the timbre of the source of a candidate pitch contour are extracted in this work. We extract static features such as Spectral Centroid(SC) given by Equation 3.9 and Spectral Energy(SE) given by Equation 3.8. These features are extracted not from the spectrum corresponding to the audio but from the harmonic representation obtained as explained in Section **??**.

$$SC = \frac{\sum_{k} f(k)|X(k)|}{\Sigma|X(k)|} \tag{3.8}$$

$$SE = \sum |X(k)|^2 \tag{3.9}$$

where $f(k)$ and $|X(k)|$ are the frequency and magnitude spectral value of the k^{th} bin.

The work by V. Rao et al. (2011) suggests that the characteristics of the timbre of the instrument is also captured by studying the temporal variations of the harmonic amplitudes and that of the static timbral features. Thus, in this work, we compute the mean and standard deviation (StD) over time, of the harmonic amplitudes (h_{mag}), their first order difference(Δh_{mag}), the SC and its first order difference (ΔSC). We also compute the mean and StD of the harmonic bin frequencies ($h_n \hat{f}_i$) over time. The harmonic frequencies are first normalized by harmonic indices (h_n) so as to maintain the same range of variation across harmonics and singers' pitch ranges. We take the mean and StD of the harmonic frequencies in the 0-2kHz and 2-5kHz sub bands and also of the first five harmonics and the first ten harmonics. These features capture the pitch instability of the voice which is expected to vary to a greater extent than other instruments at higher frequencies.

We also compute the Modulation energy ratio (MER) of the spectral centroid. For computing the MER of the SC, first the DFT of the SC over the duration of the candidate pitch contour is computed. The MER is the ratio of the energy in the 1-6 Hz region in this modulation spectrum to the energy in the 1-20Hz region. The MER is given by Equation 3.10 where $|Z(k)|$ is the fourier transform of the feature vector $z(k)$ (Here,

$z(k)$ is the spectral centroid). This feature is expected to have a higher value for voice than for other instruments.

$$MER = \frac{\sum_{k=k_{1Hz}}^{k_{6Hz}} |Z(k)|^2}{\sum_{k=k_{1Hz}}^{k_{20Hz}} |Z(k)|^2} \tag{3.10}$$

In addition to timbral features, we also extract the features characterizing the pitch contour which are the features computed in the work by Salamon and Gomez (Aug. 2012.). In this work the feature characterizing the vibrato in singing voice from Salamon and Gomez (Aug. 2012.) is excluded from the featureset since in Karṇāṭik music vibrato is not the same as that in western music cultures. The extent of the vibrato in Karṇāṭik music is variable and in addition to this there are many other ornamentations called gamakās (see section 1.3). These ornamentations are an intrinsic part of the music and not only a characteristic of the singing voice. Thus the feature does not provide relevant distinguishing information between vocal and non-vocal candidate pitch contours. The features that are used in this work, as described in Salamon and Gomez (Aug. 2012.) are as follows:

- **Pitch mean** : Mean of Pitch of the contour denoted by C_{mp}
- **Pitch standard deviation** : Standard deviation of the contour trajectory denoted by C_{stdp}
- **Pitch Length** : Length of the pitch contour denoted by C_{len}
- **Contour mean salience** : Mean salience of all peaks comprising the contour denoted by C_{mssal}
- **Contour total salience** : Sum of salience of all peaks comprising the contour denoted by C_{ts}
- **Contour salience standard deviation** : Standard Deviation of the salience of all peaks comprising the contour over time denoted as C_{stdsal}

To account for contours which are octave errors that occur due to higher salience at the harmonic frequencies, we introduce a feature which computes the ratio of the

harmonic energy of candidate pitch contour to that of its subharmonic pitch contour (i.e. $\hat{f}_i/2$ (the harmonic peaks of which are determined from the same spectrum). This energy ratio (HeR) is high for contours which are actual fundamental frequency contours and is low for contours which are octave errors. HeR is formulated in equation 3.11. The features extracted are summarized in Table 3.1.

$$HeR = \frac{\xi(\hat{f}^*)}{\xi(\hat{f}^*/2)} \tag{3.11}$$

where ξ is the function computing the Spectral Energy (SE, see equation 3.8) from the harmonic representation obtained in section 3.4.2 and $\hat{f}^*$ and $\hat{f}^*/2$ are the pitch contour and its subharmonic pitch contour.

Table 3.1: Timbral Features Computed

Timbral Features	Dynamic Frequency Features	Pitch Contour Features
Spectral Centroid (Mean and StD)	Mean and StD of Delta Harmonic Bin Frequencies from 0-2 kHz	Length of Contour
Delta Spectral Centroid(Mean and StD)	Mean and StD of Delta Harmonic Bin Frequencies from 2-5 kHz	Mean Pitch of Contour
MER in (1to6Hz /1to 20 Hz) Bands	Mean and StD of Delta of Harmonics 1 to 5	Mean Salience of Contour
Mean and StD of First 30 Harmonic Amplitudes	Mean and StD of Delta of Harmonics 6 to 10	Total Salience of Contour
Mean and StD of Delta of First 30 Harmonic Amplitudes	Mean and StD of Delta of Harmonics 1 to 10	StD of Pitch Contour
Harmonic Energy Ratio of contour to its subharmonic		StD of Salience of Contour

3.5 Classification - Vocal Contours v/s Non-Vocal Contours

This section describes the methodology and the classifiers adpoted for classification of the candidate pitch contours contours into vocal and non-vocal classes. This work chooses a classification based approach in order to learn the timbral characteristics of the singing voice vis-a-vis those of the other instruments. We reiterate the fact that this work tries to isolate the source to which a candidate pitch contour belongs by estimating the harmonics from the spectrum of the audio (see Section 3.4.3). We first perform a feature selection to choose relevant distinguishing features for the task post which a classification is performed. The details of this classification are described in this section.

Feature Selection: In order to narrow down on relevant timbral features that distinguish vocal and non-vocal characteristics, we perform feature selection on the timbral features extracted. The feature selection procedure is run only on the timbral features and features characterizing the static and dynamic properties of the pitch contour and its harmonic frequencies. The features extracted as given by Salamon and Gomez (Aug. 2012.) are not subject to the feature selection procedure. The machine learning tool Weka[6] was used for this purpose. We used the CfsSubsetEval attribute selection algorithm implemented in Weka for this purpose. The algorithm as described in Weka states the following, *"Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred."* For further details on correlation based feature subset selection see Hall (1998). A 20 fold cross validation was performed using the CfsSubsetEval feature selection algorithm and the features selected in all the 20 folds were chosen for further analysis.

Classification: This work uses five classifiers for classification the details of which are given in Section 4.2. The features are split into three sets and the three classifiers were built, one for each feature set using each of the five classifiers. The three feature

[6]http://www.cs.waikato.ac.nz/ml/weka/

sets are *1) Timbral Feature set 2) Pitch Contour Feature set 3) Salamon feature set.* The details of each feature set are given in Section 4.2. Leave one out validation methodology was followed in this work. The machine learning tool called sci-kit learn[7] was used for building the classifiers. The details of the cross validation results with the training dataset are given in section 4.2. Models for classifying vocal and non-vocal candidate pitch contours are built using the classifiers.

3.6 Post Classification - Contour Selection

In this section, the various steps taken for contour selection after the classification of candidate pitch contours into vocal and non-vocal classes are discussed. Given an audio music recording, the candidate contours are extracted post which feature extraction is done as described in 3.4 and the candidate contours are classified based on the model built in the classification step of the approach. It has to be noted that even after the classification process, it is highly possible that at the same time instant, two different pitch contours are classified as vocal (see Figure 3.6). Thus it is not necessary that we obtain a unique pitch contour after the classification process.

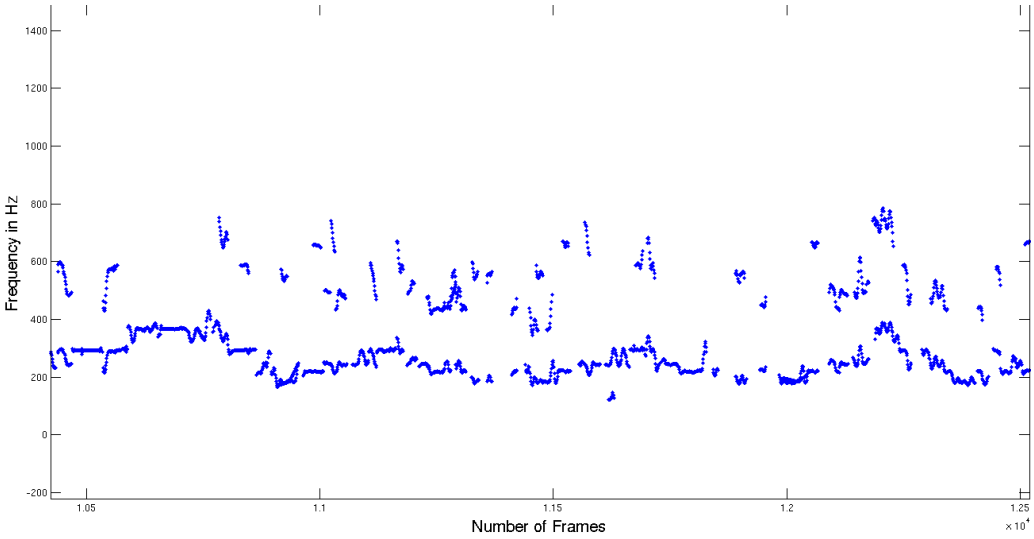

Figure 3.6: Output of the model with vocal predicted contours

[7]http://scikit-learn.org/

We follow a two step contour selection approach for resolving this issue. First the tonic pitch of the audio music recording at hand is estimated using the work by Gulati (2012). The candidate contours are then subjected to a range based on the tonic pitch in such a way that the contours falling outside this range are discarded. This frequency range is defined by the vocal range of the musician in music. We have seen that the tonic is chosen by the lead vocalist according to his/her vocal comfort (see Section 1.3. The tonic pitch is generally chosen such that the vocalist has a vocal reach of one octave below the tonic pitch to two octave above the tonic pitch (Viswanathan & Allen, 2004). Thus it is evident from this fact that the candidate pitch contours that surpass this range of frequencies cannot be vocal pitch contours and hence are rejected. In order to account for vocal yoodling and other high frequencies we impose a range of one octave below the tonic pitch to three octaves above the tonic pitch, $0.5 \cdot \tau - 3 \cdot \tau$, where τ is the tonic pitch. This setp discards the contours outside the range and reduces the number of contours simultaneously occurring at the same time instant.

After the limiting of the contours by tonic pitch range there may still be cases where more than contours occur simultaneously at the same time period. As a solution to this, if two contours occur at the same time instant, the contour which is larger in length is chosen as that of the vocal melody. This assumption is motivated from the work of Salamon and Gomez (Aug. 2012.), where it is suggested that the pitch contours corresponding to melody will be longer than that of the other non melody contours. In this case the distinction is not made between melody and non-melody contours but between vocal melody and non-vocal melody pitch contours. The assumption holds true for our case since the contours of non-vocal melody sources like the violin are smaller in length than that of the vocal pitch contours when vocal melody is the predominant melody. Thus after these two steps of contour selection, we obtain a single vocal melody pitch contour for the audio music recording at hand. The pitch contour obtained after contour selection is shown in Figure 3.7.

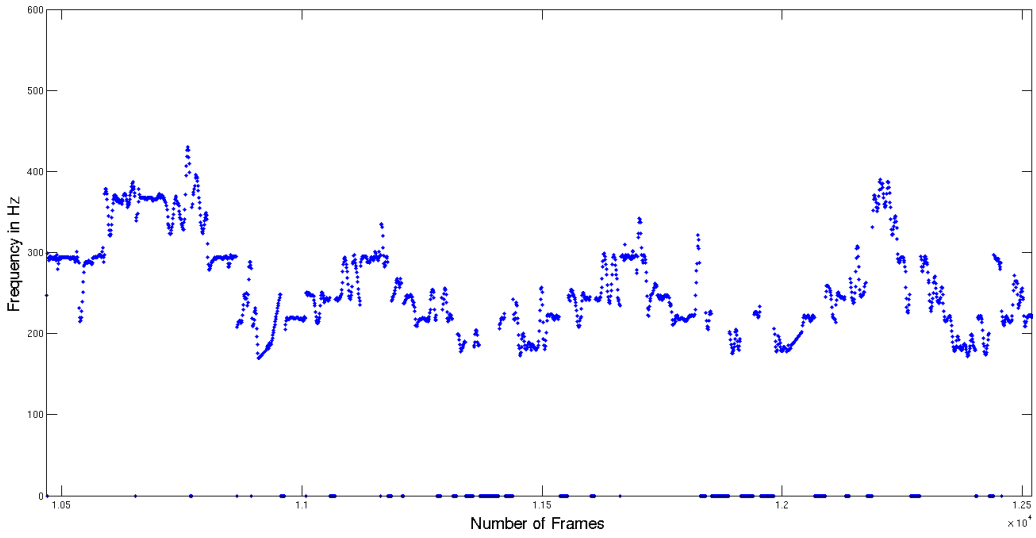

Figure 3.7: Final Pitch contour after the Contour Selection process

Chapter 4

Experiments, Evaluation and Results

In this chapter, we describe the experiments that have been conducted in the course of this work and the results obtained. We perform an evaluation of the system and also propose a musically meaningful way for evaluation of this task, especially when working with musics which are replete with ornamentations. The following section describes the datasets used for training and testing in detail.

4.1 Database - Train and Test

Train Data: The audio data used in is work consist of 4 hours of audio music recordings comprising ten male artists, ten female artists and three recordings of violin solos. This data is a subset of a carefully compiled diverse representative collection of Karṇāṭik music for the CompMusic project[1]. The recordings are chosen from different concert albums to create a representative dataset. The vocal timbre in the recordings range from a rich barritone voice to a high pitched voice to cover different singing voice timbres. The audio recordings are 160 kbps stereo MP3 recordings converted to mono wav audio files at 32 bit with a 44.1 kHz sampling rate. The violin recordings are included in the data to represent the violin accompaniment in the vocal concert recordings. For each audio music recording, the number of candidate pitch contours are high in number. The vocal concert audio recordings contain violin accompaniment due to which there are some candidate contours in each vocal concert audio recording belonging to

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

violin. We also observed that choosing a large number of violin recordings leads to overfitting the system towards violin contours. Hence in order to ensure a balanced number of instances for the vocal and non-vocal classes, we choose only three violin solo recordings as against 20 vocal concert recordings. This data will be referred to as *TrainDataMix* in the following sections. All the recordings in the dataset *TrainDataMix* have been annotated for regions where the F0 contour for the vocal melody has been tracked accurately as mentioned in section 3.3. Not all regions where the vocal melody pitch contour was tracked accurately were annotated but a representative number of segments were annotated. Figure 4.1 gives details about the annotation process and the details about the dataset. The annotations were performed by a professional musician after keen observation and listening to each and every recording using the audio visualising and analysis tool Sonic Visualiser[2]. Figure 4.2 shows the interface along with some annotations for better understanding. The candidate pitch contours in these annotated regions are used for training and building models.

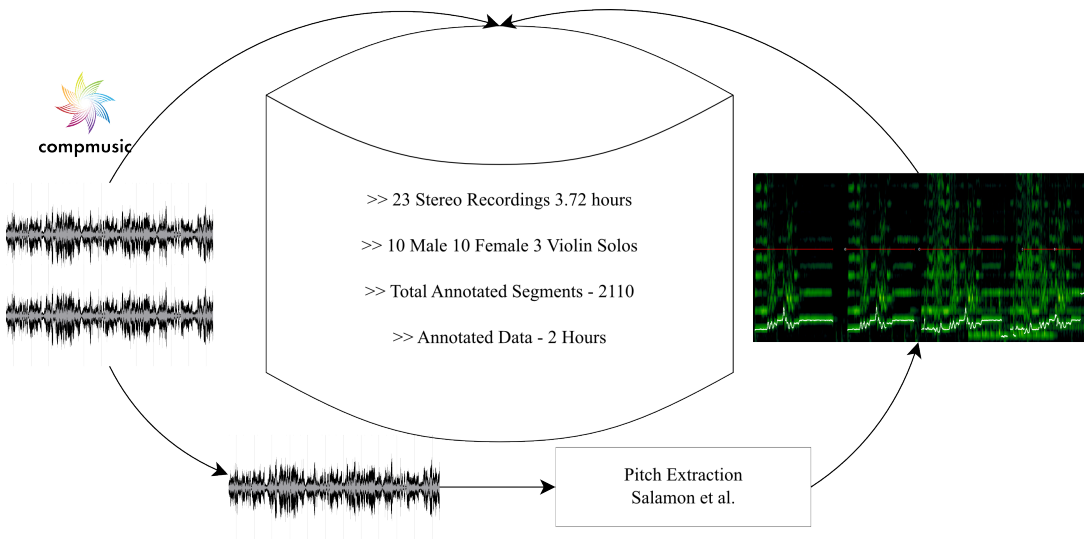

Figure 4.1: Training Data

Test Data: The test data for this work comprises a representative set of seven audio music recordings belonging to the same collection from which the *TrainDataMix* data

[2]www.sonicvisualiser.org

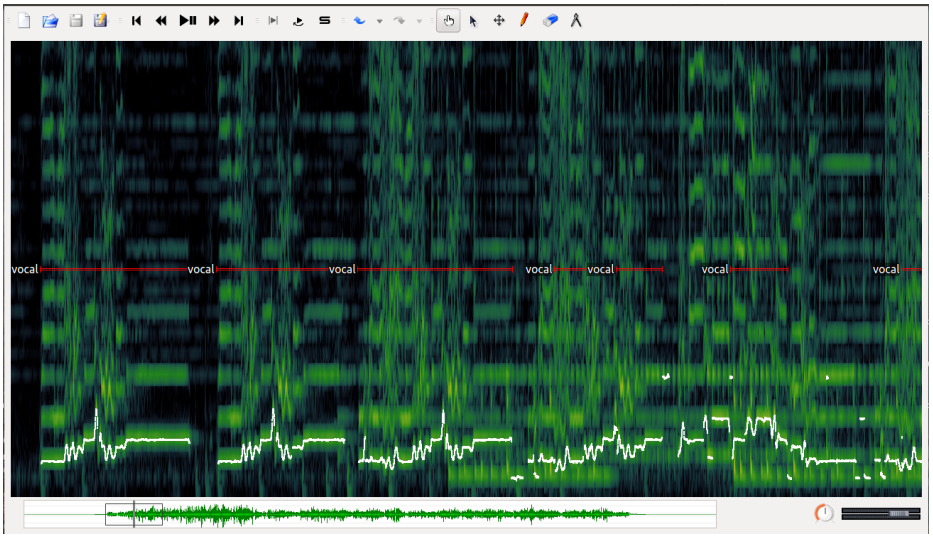

Figure 4.2: Sonic Visualiser interface with Annotations

was chosen. These seven recordings were restricted to 10 minutes each for uniformity of time and pitch contour length and also to facilitate the ground truth creation process. The dataset were in the form of 160 kbps stereo MP3 and were converted to mono wav audio files at 32 bit with a 44.1 kHz sampling rate. For each of the recordings features were extracted as described in section 3.4. This dataset which will be called *TestDataMix* in the following sections comprises four male artists and three female artists.

Ground Truth: For the extracting the ground truth pitch contours of the vocal melody for the *TestDataMix* we acquired the monophonic tracks of the vocals from multitrack recordings of the final commercial mix from the record label *Charsur*[3] under the premise that the content cannot be shared openly. Since these recordings are a part of different live concert recordings the vocal mono tracks have a mild leakage from the other channels belonging to the accompaniments. Ignoring this leakage, pitch was extracted for these mono track recordings using the semi-supervised algorithm proposed by V. Rao and Rao (2010) called PolyPDA. This algorithm is a salience based pitch detection algorithm which uses the Two Way Mis-Match error for designing the salience function. The algorithm demands the wav files to be sampled at a rate of 16kHz and

[3]www.charsur.com

hence the audio music recordings of *TestDataMix* were downsampled to 16 kHz using the audacity audio analysis tool[4]. A 30ms window was used for analysis of audio recordings of female artists and a 40ms window was used for analysis of audio recordings of male artists with a common hop size of 10ms. There were unvoiced regions in the recording for which the algorithm had erroneously estimated a finite pitch. Also due to the leakage from other channels, there was pitch estimated in some of the instrumental regions where the leakage was loud enough to have high salience. Thus the pitch contours obtained from the output of this algorithm were then manually corrected for these errors by a professional musician. Details of the PolyPDA algorithm are dealt with in (V. Rao & Rao, 2010).

4.2 Classification Experiments and Results

This section details the Classification experiments and the results obtained. As mentioned in Section 3.5 feature selection was performed on the computed timbral features in order to choose relevant distinguishing features highlighting the differences between the timbral characteristics of vocal and non-vocal candidate pitch contours. The CfsSubSetEval algorithm in the machine learning tool Weka[5] was used with a 20 fold cross validation (See Section 3.5). Features selected in all the 20 folds of the experiment were chosen for further analysis. The selected features are given in Table 4.1.

The selected features were then used for classification of the candidate pitch contours. We used five classifiers for this experiment. They are listed below:

- Support vector machine (SVM) classifier with a radial basis function kernel.
- K- Nearest Neighbor (KNN) Classifier with 5 nearest neighbors.
- L2 Norm Regularized Logistic Regression (LogReg) Classifier.
- Decision Tree Classifier.
- Random base-line classifier.

[4]http://audacity.sourceforge.net/
[5]http://www.cs.waikato.ac.nz/ml/weka/

Table 4.1: Features Selected, HP - Harmonic Power, StD - Standard Deviation, SC - Spectral Centroid, MER - Modulation Energy Ratio

Timbral Features	Pitch Contour Features (Salamon et al.)
Mean Delta HP1, StD HP2, StD HP3, Std of Delta of HP14, HP16, HP22, HP24, Mean SC, MER-SC, Mean, Median, StD of Harmonics from 2kHz - 5kHz and the 6th to 10th harmonics, Harmonic Energy Ratio.	Length of Pitch, Mean Pitch, Total Salience, Mean Salience, StD of Salience, StD of Pitch.

For this experiment, we combine the features selected into three feature sets, The Timbral Feature Set, Salamon et al. Feature Set and the Combined Feature Set. The Combined Feature Set is a combination of the features in the first two feature sets. The classification problem is a two class problem of the candidate pitch contour either being vocal or non vocal. The vocal contours were labelled with numeric class 1 and the non-vocal contours were labelled with numeric class 0. A leave one out method of cross validation was adopted in order to ensure generalization of the approach. In this method of crossvalidation we make sure that in every fold, the candidate pitch contours pertaining to a particular audio music recording are exculded from the training feature set and are used for testing. The results of the classification for each of the classifiers used are detailed in Table 4.2. From the Table 4.2 we can see that the SVM classifier performs best for all the three FeatureSets. Hence we choose SVM for building the classifier models. Since the number of instances of both the classes were not equal, we build 10 models in this work by randomly sampling the training set 10 times to balance the number of instances of each class in every iteration. Given an audio recording, all its instances were tested against each of the 10 models. A majority voting was performed on the outcomes of the 10 models for each instance and its outcome was decided to be the predicted class of that instance.

Table 4.2: Results of the classification experiment with five classifiers across three datasets.

FeatureSet	KNN	Tree	LogReg	SVM	Random
Timbral Features (T)	88.70	87.10	83.60	90.40	49.90
Salamon et al. Features (P)	86.60	84.20	82.40	87.00	50.30
Combination (T + P)	88.70	88.70	84.50	89.70	50.00

4.3 Evaluation

Post the classification using the models built, the candidate contours predicted as vocal contours are subjected to the vocal contour selection process as described in Section 3.6. The final vocal melody pitch contour obtained is subject to evaluation against the ground truth pitch contours obtained as mentioned in Section 4.1. The evaluation measures used are described in the subsection that follows.

4.3.1 Evaluation Measures

In this subsection, we discuss the evaluation measures used for evaluation of the system proposed in this work. The evaluation measures used are the same as those ditcated by MIREX, an evaluation campaign in the field of Music Information Retrieval. These measures were first used in MIREX 2005 (Poliner et al., 2007) and have since then become the de facto set of measures for evaluating melody extraction algorithms. These evaluation measures are based on frame based comparison of the estimated pitch contour with the ground truth.

Let the algorithm's estimated melody pitch contour be represented by a unidimensional vector **f**, and the ground truth pitch contour sequence by f^*. Let the voicing indicator vector be defined as **v** where the τ^{th} element takes the value $v_\tau = 1$ when the algorithm estimates the τ^{th} frame to be voiced (i.e. it estimates the melody to be present in this frame), with corresponding ground truth v^*. We also define an "unvoicing indicator" $\overline{v_\tau}$ = 1 - v_τ. In this work, the frame is considered to be unvoiced if the $f_\tau = 0$ for the τ^{th} frame. The measures are defined as follows:

- **Voicing Recall Rate**: The proportion of frames labelled as voiced in the ground

truth that are estimated as voiced by the algorithm:

$$Rec_{vx} = \frac{\Sigma_\tau v_\tau v_\tau^*}{\Sigma_\tau v_\tau^*} \tag{4.1}$$

- **Voicing False Alarm Rate**: The proportion of frames labelled as unvoiced in the ground truth that are mistakenly estimated as voiced by the algorithm:

$$FA_{vx} = \frac{\Sigma_\tau v_\tau \overline{v_\tau^*}}{\Sigma_\tau \overline{v_\tau^*}} \tag{4.2}$$

- **Raw Pitch Accuracy**: The proportion of voiced frames in the ground truth for which f_τ is considered correct (i.e. within the threshold of f_τ^*):

$$Acc_{pitch} = \frac{\Sigma_\tau v_\tau^* \mathcal{T}\left[C(f_\tau - C(f_\tau^*)\right]}{\Sigma_\tau *_\tau} \tag{4.3}$$

where τ is a threshold function defined by:

$$\mathcal{T}[a] = \begin{cases} 1, & \text{if } |a| < \lambda \\ 0, & \text{if } |a| \geq \lambda \end{cases} \tag{4.4}$$

where λ is the deviation of the estimated frequency value from the ground truth in cents. C is a function that maps a frequency value in Hz to a perceptually motivated axis where every semitone is divided into 100 $cents$, thus expressing the frequency value as a real valued number of cents above a reference frequency f_{ref}.

$$C(f) = 1200 \cdot log_2 \left(\frac{f}{f_{ref}} \right) \tag{4.5}$$

- **Overall Accuracy**: This measure combines the performance of the pitch estimation and voicing detection tasks to give an overall performance score for the system. It is defined as the proportion of all frames correctly estimated by the algorithm, where for unvoiced frames this means the algorithm labelled them as unvoiced and for voiced frames the algorithm both labelled them as voiced and

provided a correct f_0 estimate for the vocal melody.

$$Acc_{overall} = \frac{1}{L}\sum_{\tau} v_\tau^* \mathcal{T}\left[C(f_\tau - C(f_\tau^*)\right] + \overline{v}_\tau^* \overline{v}_\tau \tag{4.6}$$

where L is the total number of frames.

Note that all the measures have a worst case of 0 and the best case of 1 except for voicing false alarm rage FA_{vx} where 0 represents the best case and 1 the worst case. The performance of the algorithm on the entire music collection for a given measure is obtained by averaging the per-excerpt scores for that measure over all excerpts in the collection.

4.3.2 Glass Ceiling Analysis

The best performance that we can obtain using this system is as good as the best performance of the State of the Art algorithm which is used to obtain the candidate pitch contours. In this glass ceiling analysis we obtain the best possible pitch contour from the candidate pitch contours by directly comparing them with the ground truth for each song in *TestDataMix* dataset (see Section 4.1). The results of the Glass Ceiling are given in Table 4.4 in the first row. One can observe that eventhough the system's performance is behind the Glass Ceiling performance by a large extent, the Glass Ceiling performance however is not 100%. This is due to the fact that the performance of the algorithm proposed by Salamon and Gomez (Aug. 2012.) is not a 100%. The glass ceiling results of the overall accuracy reported by Salamon and Gomez (Aug. 2012.) are 84%, 85% and 83% on the ADC2004, MIREX05 and MIREX09 datasets respectively. The glass ceiling results reported in this work using the same methodology is 91.51% overall accuracy on the *TestDataMix* dataset described in Section 4.1. This is possible given the argument that the music dealt with in the work by Salamon and Gomez (Aug. 2012.) is polyphonic and has a larger number of instruments as compared to Karṇāṭik music, which is heterophonic and has lesser number of instruments. An improvement in the algorithm for estimation of the candidate pitch contour will push the Glass Ceiling performance along with that of the system proposed in this work.

4.3.3 Evaluation Results and Adaptive Threshold

Evaluation of the final pitch contour corresponding to the vocal melody extracted using the system proposed in this work is performed against the ground truth pitch contours corresponding to the vocal melody curated as described in Section 4.1. The results of the Evaluation are given in Table 4.3.

Table 4.3: Evaluation Results in Percentage

Type of Contour	**Raw Pitch Accuracy**	**Voicing Accuracy**	**Voicing False Alarm**	**Overall Accuracy**
Glass Ceiling	88.64	90.89	0.17	91.00
Contour-Timbralfeat	79.48	84.57	12.83	81.18
Contours-Salamonfeat	77.40	83.48	13.30	79.38
Contour-Combinedfeat	79.67	84.83	13.59	81.13
State of the art (Salamon et al.)	74.82	81.11	25.22	73.94

The results given in Table 4.3 are misleading to an extent. The condition suggested by the MIREX standards for evaluation of pitch contours is that the estimated pitch at a given time is correct if it is within one quartertone of the ground truth pitch at the same time instant, one quartertone being $50cents$. Thus if the estimated frequency $\hat{f_{est}}(t_i)$ is within $\pm 50cents$ of the ground truth frequency $\hat{f_{gt}}(t_i)$ it is considered to be accurate. The MIREX campaign deals majorly with western music genres and this threshold is set considering the paradigm of western music genres. However for non-western music, this threshold may not hold true. Especially for Indian art music, which is replete with ornamentations involving a large number of inflexions, this threshold for evaluation fails.

An experiment was conducted to justify this claim. As detailed earlier (see Section 4.1), the ground truth pitch for evaluation is prepared first by extracting pitch for the monophonic vocal audio music track using the PolyPDA algorithm (V. Rao & Rao, 2010) which was then manually corrected by a professional musician. Thus the ground

truth pitch is subject to discrepencies that occur due to change in parameters while extracting pitch using the PolyPDA algorithm. Thus these discrepencies affect the accuracies in evaluation to a very small extent thus providing misleading results.

Monophonic recordings of two small snippets by a professinal Karṇāṭik musician are used for this experiment. The pitch for these recordings are extracted using the PolyPDA algorithm with window sizes of 30 ms and 46 ms for analysis. Figure 4.3 shows the extracted pitch using PolyPDA, the green pitch contour extracted using a 30 ms analysis window and the red pitch contour extracted using a 46 ms analysis window in both the snippets. The figure also shows the absolute difference between the two pitch contours extracted using different analysis window lengths. One can observe that in the regions with large ornamentations or inflexions ranging more than 50 Hz, the difference between the pitch contours extracted using two different analysis windows is more than 80 cents. The value of the difference crosses 100 cents in some regions. Thus if these pitch contours were subject to the same evaluation conditions and thresholds as suggested by MIREX the algorithm would fail in regions with highly varying ornamentations. One can infer from this that when the difference in values of pitches occurring between the pitch contours obtained from the same algorithm with different analyses window sizes is not fitting the conditions proposed by MIREX for evaluation, those belonging to some other algorithms when compared with the ground truth extracted from a third algorithm and curated manually, will not fit the condition, thus providing misleading results. Thus in order to accommodate for this anomalie, this work proposes an adaptive threshold based on the difference between the frequency values of consequitive pitch values. This adaptive threshold takes into account the extent of variation in pitch between consequitive frames thus considering the extent of variation in the ornamentation. The formula for the adaptive threshold is given in Equation 4.7 where d_{f0} and d_{max} are the first order difference in pitch and the maximum difference respectively.

$$\lambda = 50 + 50 * \left(\frac{d_{f0}}{d_{max}} \right) \tag{4.7}$$

The parameter d_{max} is set according to the musical characteristics of Indian art music and also by taking an engineering decision after analysing the maximum differ-

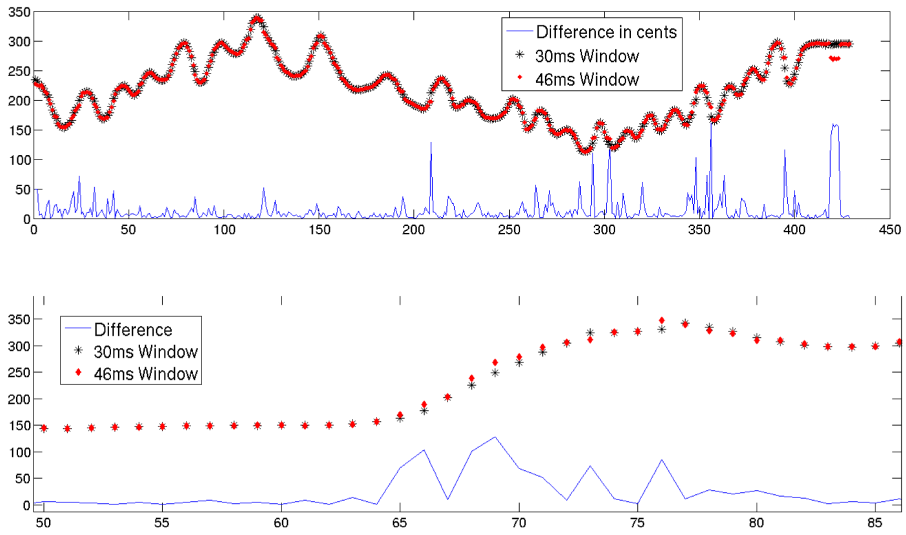

Figure 4.3: Pitch contour for two monophonic recordings of ornamented phrases for 30ms and 46ms window.

ences in frequencies that occur between consequitive frames of the seven ground truth pitch contours obtained before (see Section 4.1). For this work, the paramter d_{max} is set at 400 cents which implies that if the difference between the present frame and the previous frame crosses 400 cents, then the threshold for evaluation will be set to 100 cents which is one semitone. The threshold, until then will be weighted by the ratio of the d_{f0} to that of the parameter d_{max}. This parameter will be different for differnet music traditions. The evaluation results after adaptive thresholding have been given in Table 4.4. We see an improvement in the accuracies eventhough the difference in the previous results and the ones after adaptive thresholding is small (see Tables 4.3 & 4.4).

4.4 Discussions

In this section we analyze the results obtained on the test data *TestDataMix* using the proposed approach. The evaluations are carried out with the standard MIREX evaluation measures as discussed in Section 4.3. The results with and without adaptive thresh-

Table 4.4: Evaluation Results in Percentage with Adaptive Thresholding

Type of Contour	**Raw Pitch Accuracy**	**Voicing Accuracy**	**Voicing False Alarm**	**Overall Accuracy**
Glass Ceiling	89.20	90.89	0.17	91.51
Contour-Timbralfeat	80.00	84.57	12.83	81.57
Contour-Salamonfeat	77.89	83.48	13.30	79.76
Contour-Combinedfeat	80.18	84.83	13.59	81.53
State of the art (Salamon et al.)	75.18	81.11	25.22	74.22

olding as suggested in Section 4.3 are provided in Tables 4.4 and 4.3 respectively. One can observe that the difference in values of the various evaluation measures before and after adaptive thresholding is in the range of 0.5% to 1%. The reason for the difference in accuracies being very small is that the pitch contour being evaluated is subjected to an adaptive threshold only in regions where the condition that the difference between frequencies of consequetive frames is more than 400 cents in this work. This condition is satisfied in regions that have heavy ornamentations. Such regions are interspersed with the non-ornamented regions in an almost equal proportion (Krishna & Ishwar, 2012). Thus the number of such ornamented regions is not as significantly large that it would make a very large and significant difference to the accurcies. Nevertheless, it provides a more musically meaningful evaluation methodology.

The evaluation is carried out using three feature sets, the first being only timbral features (Contour - Timbralfeat), the second being the same features as extracted in the work by Salamon and Gomez (Aug. 2012.) and the third being a combination of the first two. The final row in the Tables 4.3 and 4.4 correspond to the results obtained by subjecting the audio music recordings in *TestDataMix* dataset to the state of the art predominant melody extraction algorithm proposed by Salamon and Gomez (Aug. 2012.). We can see that for all feature sets, the proposed algorithm in this work performs better than the state of the art for this task. We can see that the usage of classification based technique for the task of vocal melody pitch extraction with the same features as extracted by Salamon and Gomez (Aug. 2012.) itself gives an improvement

(80 %, 77.89 %, 80.18 % raw pitch accuracy with the three featuresets, Table 4.4)over the state of the art (75.18 % raw pitch accuracy, Table 4.4). We can see that the raw pith accuracy of the proposed approach using the features extracted by Salamon and Gomez (Aug. 2012.) is practically the same as that obtained using the state of the art whereas that using the other two featuresets has improved by almost 5 %.

One can also observe that the overall accuracies using the proposed approach with all the three featuresets is giving almost a 6 % improvement than the state of the art. This is majorly due to the large improvement in the voicing false alarm that is obtained using the proposed approach. One can observe that the false alarm rate using the approach proposed in this work (12.83 %, 13.30 %, 13.59 %) is almost 50 % lesser than that of the state of the art (25.22 %). This is due to the fact that for the proposed approach the instrument regions are taken to be unvoiced in the ground truth since the ground truth is extracted only for the vocal melody. The state of the art is an algorithm which estimates the predominant melody at a given instant of time which does not take into account the presense or absence of an instrument. Due to the presense of more than one melodic sources in heterophonic music the algorithm estimates pitch also in the instrumental regions where the conditions for it being a predominant melody are satisfied irrespective of the timbre of the instrument. This increases the voicing false alarm rate given by the state of the art (Salamon & Gomez, Aug. 2012.) with respect to the task of vocal melody pitch extraction.

One can also observe from Table 4.4 that the combining of the timbral featureset and the Salamon et al. featureset has not brought about a significant improvement in results. This shows that the timbral characteristics dominate the classification of candidate pitch contours into vocal and non-vocal classes. The assumption was that a combination of timbral features along with features characterizing the pitch contour will aid the extraction of pitch for the singing voice along with it being characterized as predominant melody since the features extracted by Salamon et al. were designed for characterizing pitch contours as melody and non-melody contours. But the timbral features are highly sensitive since they are extracted from the harmonic amplitudes and bin frequencies and hence dominate the combination of features giving almost similar results for the proposed approach using both just timbral features and the combination of timbral features and Salamon et al. features.

Chapter 5

Concusion and Future Work

In this final chapter, we summarize the work of this thesis. We draw conclusions and inferences from the results obtained. We also discuss issues which are to be resolved and present some suggestions for future work.

5.1 Summary of the work

We started with a brief introduction to how the culture specific aspects and a thorough understanding of the music being studied through computational methods is important and relevant for Music Information Retrieval. We then introduce the task of this thesis which is extraction of pitch of the vocal melodies from heterophonic music audio recordings. Later we illustrate the motivation for doing this task and define a definite set of goals pertaining to the research work. We also provide definitions and explanations for certain terminology, a basic understanding of which is required for this work. Since this work uses Indian Art music as a test case, we give an overview of the music tradition and go on to illustrate the key aspects required for a basic comprehension of the music.

We have reviewed the main audio features, techniques and works relevant to the computational analysis of the melodic aspect of music and timbral analysis of music signals. The state of the art pertaining to computational analysis of Indian art music is first reviewed in order to obtain an idea about the complexity involved and reiterate the importance of predominant vocal melody in the various tasks taken up for research.

We then studied the state of the art with pertaining to prefominant melody extraction. For knowledge about characterization of the singing voice and to have a background knowledge about the features used for the characterization of singing voice, we review the state of the art pertaining to singing voice detection. We review in detail the various features used for the characterization of singing voice which are relevant to the task we have taken up for research.

We proposed and implemented a classification based approach for extraction of the pitch of vocal melody in this work. We have given a detailed explanation of the approach, the motivation behind the approach and all the implementation details. We have evaluated the approach on a decently sizeable database and have obtained good accuracies following this approach. We have also proposed a novel evaluation methodology which is musically meaningful in that it takes into account the properties of the music at hand. We finally present a discussion of the various results obtained and have tried to give plausible justifications for the results obtained following the proposed approach.

5.2 Conclusions

The state of the art predominant melody extraction algorithm have been designed majorly for western music genres. From the state of the art review done in this work, we learn that subjecting non-western music to the state of the art algorithm as if it were a general melody extraction algorithm yields results that are erroneous. It is also seen that predominant melody is a representation that is widely used for many applications. Thus it is important that the musical and cultural characteristics be included in the design of melody extraction algorithms for better results.

In this work, we have first carried out a case study in which we have extracted predominant melody for two excerpts of Karṇāṭik music and analysed the various errors that were made by the algorithm. A detailed analysis of the errors is performed. The analysis enlightens the various errors possible in the extraction of the pitch of vocal melodies using the current state of the art. Post this study, we propose an approach for extraction of pitch of the vocal melody. After a brief overview of the method we delve into the details of the approach at hand. We first extract candidate pitch contours using the salience based state of the art proposed by Salamon and Gomez (Aug. 2012.).

On obtaining the candidate pitch contours, a representation of the isolated harmonic spectrum pertaining to the source of the pitch contour is obtained. From this representation, timbral features characterizing the source of the candidate pitch contour are extracted. We also describe the feature extraction process for the pitch contour characterization proposed by Salamon and Gomez (Aug. 2012.). We then describe the classification of these candidate pitch contours into vocal and non-vocal classes based on models built after learning the distinguishing characteristics of the singing voice from the heterophonic mix using the timbral and pitch contour features extracted. The classification results reveal that Support Vector Machines or similar classifiers yield good results in classification of timbral features. A large amount of representative data is required so that the approach does not bias itself towards the recording conditions. Most importantly we then illustrate the contour selection steps in which the music specific aspects are used for filtering out false positives and obtaining the final pitch contour corresponding to the vocal melody. We use the tonic pitch (the fundamental aspect of Indian art music) as a quantity to design a filtering method which discards the false positive candidate pitch contours. The approach was evaluated on seven unseen audio music recordings and a novel evaluation methodology was proposed which takes into account the property of the music being analysed. An adaptive threshold is designed in order to incorporate the high variation in pitch at transcient regions of heavy ornamentations. Eventhough the results show only a minor improvement in accuracies, this gives a more musically meaningful evaluation. This shows us that for music traditions with heavy ornamentations and inflexions it is important to take into account the discrepencies that occur in the process of formation of the ground truth for evaluation.

The work persented in this thesis is primarily from a computational point of view. We plan to include more knowledge about the timbre of singing voice from the point of view of spectral shapes and also from the point of view of timbre perception and cognition of a human listner since melody is a perceptive property.

5.3 Open Issues for Future Development

While we have achieved the goals set for this work, there are a number of issues which are unresolved and require further investigation. This sections lists out some of the

interesting issues which could yield more robust results if resolved.

- **Source Separation**: Eventhough we surpass the state of the art with respect to evaluation measures, there still are false positive candidate pitch contours the timbral properties of which are confused with that of the voice (example: certain violin phrases in the lower register). One of the possible solution for this issue is subjecting the heterophonic mix to source separation in order to separate out the singing voice from the mix.

- **Spectral Shape**: The spectral shape of the human voice is very unique and differs from that of other instruments. It has also been proven that the singing voice has characteristics specific to it that distinguish it from other instruments. One of the possible ways to exploit this property of the voice is to extract the spectral shape of the singing voice from the heterophonic mix. Due to the change in the shape of the vocal tract in the process of singing, the spectral shape is expected to change in time where as that of other instruments are expected to remain constant through out. This dynamic property of the spectral shapes of the voice vis-a-vis that of the instruments is a viable feature for this task.

- **Perceptual Studies**: Studies pertaining to human perception and how in the cognitive domain the human brain distinguishes between different timbres must be performed for this task. This would aid in the separation of the singing voice and also provides a baseline for the amount of audio data required to match the human performance.

- **Musical Influences**: It is very important to consider the type of vocalization involved in different types of music for this task. This would help in designing features specific to the usage of singing voice in a particular type of music which would help in better distinguishing the singing voice from other instruments. This also aids in resolving issues with respect to instrument-vocal timbre confusions by paying attention to the properties of the voice under certain music specific circumstances.

5.4 Contributions

This section summarizes the relevant contributions associated with the work done during the course of this thesis.

- Review of the relevant works and techniques pertaining to the extraction of pitch of the vocal melody in heterophonic audio music recordings.
- A novel classification based approach for vocal melody pitch extraction by combining the salience based pitch extraction methods with timbral properties of the singing voice.
- Compiling of a comprehensive and representative training data set of about 4 hours with annotations for the task of vocal melody pitch extraction [1].
- A manually curated ground truth of the pitch of the vocal melody for 70 minutes of audio music recordings of Karṇāṭik music[2].
- A novel evaluation approach using adaptive thresholding for musically meaningful evaluation of the task at hand.

[1] link to dataset page
[2] link to ground truth pitch data

References

Bagchee, S. (1998). *Understanding raga music.* Business Publications Inc.

Bapat, A., Rao, V., & Rao, P. (2007). Melodic contour extraction for indian classical vocal music. In *Proc. international workshop on artificial intelligence and music (ijcai-07).*

Clayton, M. R. L. (2000). *Time in Indian music: rhythm, metre, and form in North Indian rag performance.* Oxford University Press.

Durrieu, J. L., Richard, G., David, B., & Fevotte, C. (2010, March). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(3), 564-575.

Feng, L., Nielsen, A. B., & Hansen, L. K. (2008). Vocal segment classification in popular music. In *Ismir* (p. 121-126).

Fujihara, H., Goto, M., Kitahara, T., & Okuno, H. (2010, March). A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(3), 638-648. doi: 10.1109/TASL.2010.2041386

Gulati, S. (2012). *A tonic identification approach for indian art music* (Unpublished master's thesis). Universitat Pompeu Fabra, Barcelona.

Gulati, S., Bellur, A., Salamon, J., H. G., R., Ishwar, V., Murthy, H. A., & Serra, X. (2014). Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation. *Journal of New Music Research*, *43*(01), 55–73. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/09298215.2013.875042` doi: 10.1080/09298215.2013.875042

Hall, M. A. (1998). *Correlation-based feature subset selection for machine learning* (Unpublished doctoral dissertation). University of Waikato, Hamilton, New Zealand.

Herrera, P., Klapuri, A., & Davy, M. (2006). *Automatic classification of pitched musical instrument sounds* (A. Klapuri & M. Davy, Eds.). Springer US.

Ishwar, V., Dutta, S., Bellur, A., & Murthy, H. (2013, November 4-8). Motif spotting in an alapana in carnatic music. In *Proceedings of the 14th international society for music information retrieval conference.*

Keiler, F., & Marchand, S. (2002). Survey on extraction of sinusoids in stationary sounds. In *Proceedings of the 6th international conference on digital audio effects (dafx-03)* (pp. 51–58).

Klapuri, A. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Ismir* (pp. 216–221).

Klapuri, A., & Davy, M. (Eds.). (2006). New York: Springer.

Koduri, G. K., Gulati, S., Rao, P., & Serra, X. (2012). Raga Recognition based on Pitch Distribution Methods. *Journal of New Music Research*, *41*(4), 337–350. doi: 10.1080/09298215.2012.735246

Koduri, G. K., Serrà, J., & Serra, X. (2012). Characterization of Intonation in Karnataka Music by Parametrizing Context-Based Svara Distributions. In *2nd compmusic workshop* (pp. 1–5). Istanbul. Retrieved from `http://mtg.upf.edu/node/2623`

Krishna, T. M., & Ishwar, V. (2012, July). Svaras, gamaka, motif and raga identity. In *Workshop on computer music.* Instanbul, Turkey.

Lagrange, M., Raspaud, M., Badeau, R., & Richard, G. (2010, September). Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity. *Pattern Recogn. Lett.*, *31*(12), 1498–1506. Retrieved from `http://dx.doi.org/10.1016/j.patrec.2009.09.008` doi: 10.1016/j.patrec.2009.09.008

Martin, R., & Perfecto, H. (2007, September). Comparing audio descriptors for singing voice detection in music audio files. In *Brazilian symposium on computer music, 11th. san pablo, brazil.* Retrieved from `http://iie.fing.edu.uy/publicaciones/2007/RH07`

Marxer, R. (2012). *Audio source separation for music in low-latency and high-latency scenarios* (Unpublished doctoral dissertation). Universitat Pompeu Fabra, Barcelona, Spain.

Mehta, R. (2008). *Indian Classical Music and Gharana Tradition* (First ed.). Readworthy Publications Pvt. Ltd.

Mesaros, A. (2012). *Singing voice recognition for music information retrieval* (Unpublished doctoral dissertation). Tampereen teknillinen yliopisto - Tampere University of Technology, Tampere, Finland.

Paiva, R. P., Mendes, T., & Cardoso, A. (2006). Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Computer Music Journal*, *30*, 80-98. Retrieved from http://dblp.uni-trier.de/db/journals/comj/comj30.html#PaivaMC06

Peeters, G. (n.d.). *A large set of audio features for sound description (similarity and classification) in the cuidado project* (Tech. Rep.). IRCAM.

Poliner, G. E., Ellis, D., Ehmann, A., Gomez, E., Streich, S., & Ong, B. (2007, May). Melody transcription from music audio: Approaches and evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, *15*(4), 1247-1256. doi: 10.1109/TASL.2006.889797

Rao, P., Ross, J. C., Ganguli, K. K., Pandit, V., Ishwar, V., Bellur, A., & Murthy, H. A. (2014). Classification of melodic motifs in raga music with time-series matching. *Journal of New Music Research*, *43*(1), 115-131. Retrieved from http://dx.doi.org/10.1080/09298215.2013.873470 doi: 10.1080/09298215.2013.873470

Rao, V., Gupta, C., & Rao, P. (2011). Context-aware features for singing voice detection in polyphonic music. In *Adaptive multimedia retrieval* (p. 43-57).

Rao, V., Ramakrishnan, S., & Rao, P. (2009). Singing voice detection in polyphonic music using predominant pitch. In *Interspeech* (p. 1131-1134). ISCA. Retrieved from http://www.isca-speech.org/archive/interspeech_2009/i09_1131.html

Rao, V., & Rao, P. (2010, Nov). Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *Audio, Speech, and Language Processing, IEEE Transactions on*, *18*(8), 2145-2154.

Rao, V., S, R., & Rao, P. (2008). Singing voice detection in north indian classical music. In *Proc. of the national conference on communications (ncc) 2008.* Hyderabad, India.

Ross, J. C., & Rao, P. (2012, October). Detecting melodic motifs from audio for hindustani classic music. In *Ismir.* Portugal.

Salamon, J. (2013). *Melody extraction from polyphonic music signals* (Doctoral dissertation, Universitat Pompeu Fabra, Barcelona). Retrieved from http://mtg.upf.edu/node/2827

Salamon, J., & Gomez, E. (Aug. 2012.). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, 20(6):1759-1770.

Salamon, J., Gómez, E., Ellis, D. P. W., & Richard, G. (2014, 02/2014). Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine*, *31*, 118-134. Retrieved from `http://mtg.upf.edu/node/2773` doi: 0.1109/MSP.2013.2271648

Saraf, R. (2011). *Development of Hindustani Classical Music (19th & 20th centuries)* (First ed.). Vidyanidhi Prakashan.

Sen, A. K. (2008). *Indian Concept of Rhythm* (Second ed.). New Delhi: Kanishka Publishers, Distributors.

Şentürk, S., Gulati, S., & Serra, X. (2013). Score informed tonic identification for makam music of turkey. In *14th international society for music information retrieval conference (ismir).*

Serra, X. (2011). A Multicultural Approach to Music Information Research. In *Proc. 12th International Conference on Music Information Retrieval (ISMIR).*

Shenoy, A., Wu, Y., & Wang, Y. (2005). Singing voice detection for karaoke application. In *Proc. visual communications and image processing (vcip).* Beijing, China.

Sundberg, J. (1977). *The acoustics of the singing voice.* W.H. Freeman.

Tachibana, H., Ono, T., Ono, N., & Sagayama, S. (2010, March). Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *Acoustics speech and signal processing (icassp), 2010 ieee international conference on* (p. 425-428). doi: 10.1109/ICASSP.2010.5495764

Thibault, F., & Depalle, P. (2004, May). Adaptive processing of singing voice timbre. In *Electrical and computer engineering, 2004. canadian conference on* (Vol. 2, p. 871-874 Vol.2). doi: 10.1109/CCECE.2004.1345253

Tzanetakis, G. (2004, June). Song-specific bootstrapping of singing voice structure. In *Multimedia and expo, 2004. icme '04. 2004 ieee international conference on* (Vol. 3, p. 2027-2030 Vol.3). doi: 10.1109/ICME.2004.1394662

Viswanathan, T., & Allen, M. H. (2004). *Music in south india.* Oxford University Press.

Karṇāṭik Music

ālāpana: An unmetered melodic improvisation
Karṇāṭik: is an art music tradition of India
gamakā: Ornamantation in Indian art music
gāyakī: A style of playing emulating the vocal qualities
mṛdaṅgaṁ: The primary percussion accompaniment
rāga: The melodic framework of Carnatic music
Ṣaḍja: Musical note Carnatic music
tāḷa: The rhythmic framework of Carnatic music
varṇaṁ: A musical form in Carnatic music