

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/263725309>

Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation

Article in *Journal of New Music Research* · March 2014

DOI: 10.1080/09298215.2013.875042

CITATIONS

55

READS

6,648

7 authors, including:



Sankalp Gulati

Pompeu Fabra University

20 PUBLICATIONS 918 CITATIONS

[SEE PROFILE](#)



Justin Salamon

Adobe Research

87 PUBLICATIONS 9,639 CITATIONS

[SEE PROFILE](#)



Ranjani H.G.

Indian Institute of Science Bangalore

8 PUBLICATIONS 136 CITATIONS

[SEE PROFILE](#)



Vignesh Ishwar

Pompeu Fabra University

9 PUBLICATIONS 307 CITATIONS

[SEE PROFILE](#)

Automatic Tonic Identification in Indian Art Music: Approaches and Evaluation

Sankalp Gulati, Ashwin Bellur, Justin Salamon, Ranjani H.G, Vignesh Ishwar,
Hema A Murthy and Xavier Serra *

[This is an Author's Original Manuscript of an Article whose final and definitive form, the Version of Record, has been published in the Journal of New Music Research, Volume 43, Issue 1, 31 Mar 2014, available online at: <http://www.tandfonline.com/doi/full/10.1080/09298215.2013.875042>]

Abstract

The tonic is a fundamental concept in Indian art music. It is the base pitch, which an artist chooses in order to construct the melodies during a rāg(a) rendition, and all accompanying instruments are tuned using the tonic pitch. Consequently, tonic identification is a fundamental task for most computational analyses of Indian art music, such as intonation analysis, melodic motif analysis and rāg recognition. In this paper we review existing approaches for tonic identification in Indian art music and evaluate them on six diverse datasets for a thorough comparison and analysis. We study the performance of each method in different contexts such as the presence/absence of additional metadata, the quality of audio data, the duration of audio data, music tradition (Hindustani/Carnatic) and the gender of the singer (male/female). We show that the approaches that combine multi-pitch analysis with machine learning provide the best performance in most cases (90% identification accuracy on an average), and are robust across the aforementioned contexts compared to the approaches based on expert knowledge. In addition, we also show that the performance of the latter can be improved when additional metadata is available to further constrain the problem. Finally, we present a detailed error analysis of each method, providing further insights into the advantages and limitations of the methods.

Keywords: Tonic, Drone, Indian art music, Hindustani, Carnatic, Tānpūrā, Śadja, Indian classical music

*This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). S. Gulati, J. Salamon and X. Serra are affiliated with Universitat Pompeu Fabra, Barcelona, Spain. A. Bellur, V. Ishwar and H. A. Murthy are affiliated with the Indian Institute of Technology Madras, Chennai, India. Ranjani H. G. is affiliated with the Indian Institute of Science, Bangalore, India. e-mail: sankalp.gulati@upf.edu, ashwinbellur@gmail.com, justin.salamon@upf.edu, ranjanigh@ece.iisc.ernet.in, vigneshishwar@gmail.com, hema@cse.iitm.ac.in, xavier.serra@upf.edu

1 Introduction

The tonic is the foundation of melodies in both Hindustani and Carnatic music (Viswanathan & Allen, 2004). It is the base pitch of a performer, carefully chosen to explore the pitch range effectively in a rāg rendition (Danielou, 2010) (the term rāg is used in the Hindustani music tradition, whilst in the Carnatic music tradition the term rāga is used; for consistency, in this article we shall always use Hindustani terminology). The tonic serves as a reference and the foundation for melodic integration throughout the performance (Deva, 1980). That is, all the tones in the musical progression are always in reference and related to the tonic pitch. All the accompanying instruments such as tablā, violin and tānpūrā are tuned using the tonic of the lead performer. It should be noted that tonic in Indian art music refers to a particular pitch value and not to a pitch-class. The frequency range in which the tonic pitch may reside when considering both male and female vocalists spans more than one octave, typically between 100-260 Hz (Sengupta, Dey, Nag, Datta, & Mukerjee, 2005).

Indian art music encapsulates two music traditions of the Indian subcontinent: *Hindustani* music (also known as North Indian music), prominent in the northern regions of India, Pakistan, Nepal, Afghanistan and Bangladesh (Bor, Delvoye, Harvey, & Nijenhuis, 2010; Danielou, 2010); and *Carnatic* music, widespread in the southern regions of peninsular India and Sri Lanka (Singh, 1995; Viswanathan & Allen, 2004). In both Hindustani and Carnatic music, the *rāg* is the fundamental melodic framework upon which the music is built (Bagchee, 1998; Danielou, 2010; Viswanathan & Allen, 2004), and the *tāl* (tāla in the Carnatic music tradition) provides the rhythmic framework (Clayton, 2000; Sen, 2008). Though Hindustani and Carnatic music traditions share fundamental music concepts of rāg and tāl, the music is significantly different in each tradition (cf. (Narmada, 2001) for a comparative study of rāgs).

Indian art music is basically heterophonic, with the main melody sung or played by the lead artist (Bagchee, 1998). Often, an instrument provides a melody accompaniment by closely following the melody rendered by the lead artist (Viswanathan & Allen, 2004). A typical arrangement in a performance of Indian art music consists of a lead performer (occasionally a duo), a melody accompaniment provided by harmonium or *sārangī* in Hindustani music and by violin in Carnatic music, a rhythm accompaniment usually provided by *tablā* in Hindustani music and *mṛdangam* in Carnatic music and a constantly sounding drone in the background. The drone sound, which is typically produced by the *tānpūrā*, is the only component that adds a harmonic element to the performance (Bagchee, 1998).

The seven solfège symbols (Sa, Re, Ga, Ma, Pa, Dha and Ni in short-form) used in Indian art music are called *svars* (svaras in the Carnatic music tradition) (Danielou, 2010; Bagchee, 1998). With the exception of Sa (also referred to as *Ṣaḍja*) and Pa (also referred to as *Pancham*, fifth with respect to Sa), every other svar has two or three variations, where each variation is either a komal (flat), śudh (unmodified, literally means pure) or tivṛ (sharp) of the basic svar and has a specific function in a rāg rendition (Viswanathan & Allen, 2004).

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

The problem of tonic identification and the related problem of key identification have received considerable attention in the context of Western music (Krumhansl & Kessler, 1982; Peeters, 2006; Chew, 2002; Gómez & Herrera, 2004). However, the tonic as understood in the context of Indian art music is considerably different, consequently requiring the development of new and context-specific algorithms for automatic tonic identification. Whilst we shall focus on Indian art music in this paper, it is worth mentioning that context-specific algorithms have also been proposed for other music traditions, for example the work by Bozkurt, Yarman, Karaosmanoğlu, and Akkoç (2009) and Şentürk, Gulati, and Serra (2013) on tonic identification in the Makam music of turkey.

In this paper we review the existing approaches for automatic tonic identification of the lead performer in Indian art music. Our main focus is to consolidate existing work on tonic identification in Indian art music and to evaluate these approaches on representative datasets. We begin by presenting a general block diagram that shows the methodology adopted by different approaches. Further discussion is organized block-wise, where we present a brief overview of each approach and highlight the differences between them in every block. We evaluate seven methods on six diverse datasets and discuss the results. We also analyze the advantages and shortcomings of each of these methods in terms of the music material being analyzed, such as Hindustani versus Carnatic, male versus female lead singer, vocal versus instrumental music, and the amount and type of data used by the methods (both in terms of audio length and complementary metadata).

In Section 1.1 we present a brief overview of the use of drone in Indian art music along with a short introduction to tānpūrā and its tonal characteristics. Subsequently, in Section 1.2 we discuss the main musical cues used by a human listener to identify the tonic in a performance. In Section 2 we describe the existing methods for tonic identification. In Section 3 we describe the evaluation methodology, annotation procedure and the different datasets used for evaluation in this study. The results are presented and discussed in Section 4, followed by conclusions and directions for future work in Section 5.

1.1 Drone and tānpūrā in Indian art music

Indian art music is a performance centric music tradition where both the performer and the audience need to hear the tonic pitch of the lead artist throughout the concert. Hence, every performance of Indian art music has a drone sound

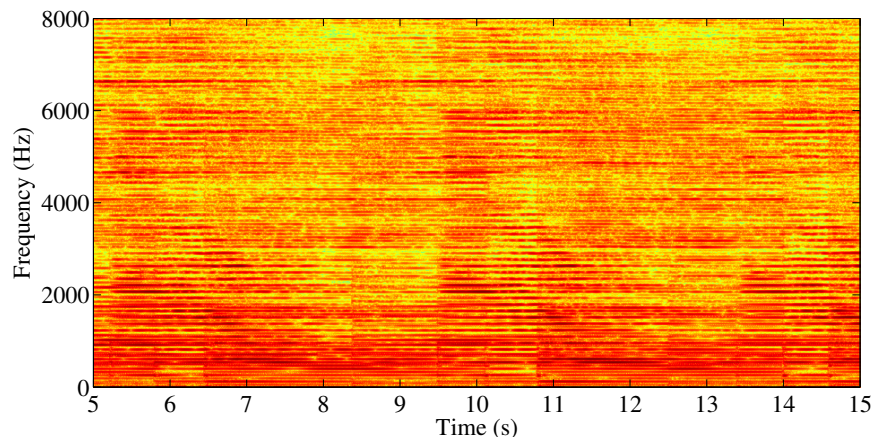


Figure 1: Spectrogram of a solo tānpūrā sound excerpt.

in the background that reinforces the tonic. Along with the tonic, the drone also emphasizes other svars such as the fifth, fourth and sometimes the seventh with respect to the tonic, depending on the choice of the rāg. Essentially, the drone provides the reference pitch that establishes all the harmonic and melodic relationships between the pitches used during a performance. Typically the drone is produced by either a tānpūrā, an electronic tānpūrā or a śruti box for vocal music and by the sympathetic strings of the instruments such as sitār, sārangī or viṇā for instrumental performances.

The emergence of a drone sound in Indian art music dates back to 1600 AD (Bagchee, 1998). The drone emphasizes facets such as intonation and consonance. As described by Deva (1980), without a drone, the intonation and the tonality of the music is governed by tonal memory (a matter of retrospect and post relation of tones). But with the employment of a drone, a musician is forced to constantly refer to this tonal background both for intonation and consonance resolution. The tonal structure of a drone instrument is thus a very important aspect of this music tradition. We briefly describe the tonal structure of the tānpūrā, which is the main drone instrument used to accompany a lead performer.

Tānpūrā is a long-necked plucked string instrument, which comes in three different sizes that correspond to the different pitch ranges it can produce (Figure 2). The largest tānpūrā, which has the lowest pitch range, is used to accompany male singers. A smaller size is used for female singers and the smallest tānpūrā is used to accompany instrumentalists. Typically a tānpūrā has 4 strings, where the two middle strings are tuned to the tonic of the lead artist (Sa), the fourth string to an octave below the tonic pitch (Sa) and the first string to either Pa, Ma or Nī between the tonic and the octave below. The tānpūrā sound is bright and has a dense spectrum, as illustrated in the spectrogram shown in Figure 1 and in the spectrum shown in Figure 3. The higher overtones in the sound add energy to various pitch-classes. Deva (1980) presents an in-depth analysis of the spectral characteristics of

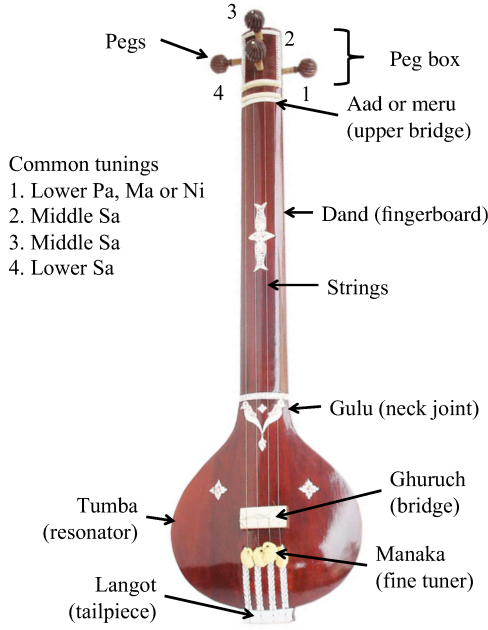


Figure 2: The tãnpūrā drone instrument with labeled parts.

the tãnpūrā sound. He also provides an interesting historical perspective on the emergence of the tãnpūrā and its significance in Indian art music.

The tãnpūrā is played by repeatedly rolling the fingers gently over the strings to create a constant slow rhythmic pattern (which bears no relation to the speed of the song). The playing style differs slightly in Hindustani and Carnatic music. The unique resonant sound of the tãnpūrā is mainly due to the special characteristics of the bridge and the javārī (thread) that is inserted between the bridge and the strings. Raman (1921) studied this phenomenon in depth, describing how the javārī makes the vibration modes of the tãnpūrā strings violate the Helmholtz law. Bagchee (1998) provides a detailed explanation of the experiments conducted to observe this phenomenon.

1.2 Prominent musical cues for tonic identification

A listener uses various musical cues to identify the tonic pitch in a performance of Indian art music, and some of these cues are exploited by the automatic tonic identification methods described later. After interacting with musicians and expert listeners we have compiled a non-comprehensive list of these musical cues:

1. Melodic characteristics: In Carnatic music, *gamakas* are an inseparable component of the music. Gamakas can be classified into multiple categories: one class of gamakas is broadly described as oscillatory movement around a svar (Koduri, Gulati, et al., 2012). Another class of gamakas includes a glide from one svar to another. The use of gamakas on the Sa and Pa svars is minimal compared to the other svars used in the rāg rendition. Ranjani,

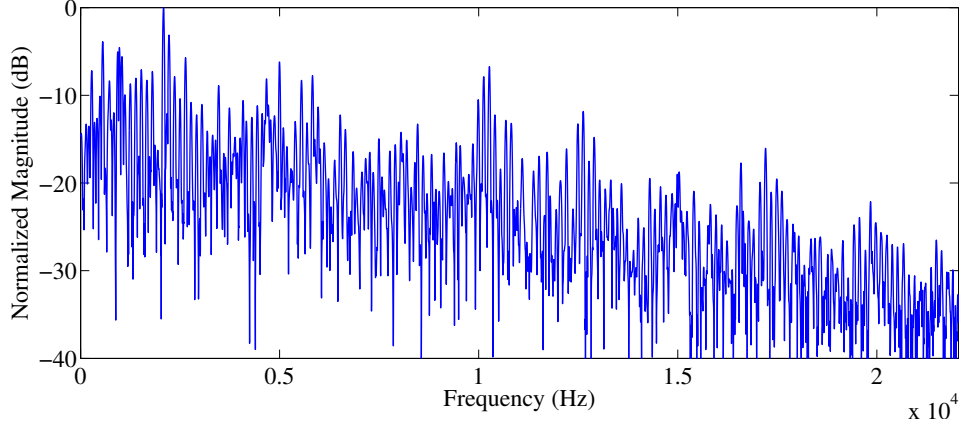


Figure 3: Spectrum of a tānpūrā recording highlighting its richness and brightness.

Arthi, and Sreenivas (2011) and Bellur, Ishwar, Serra, and Murthy (2012) utilize this melodic characteristic of Carnatic music for the identification of the tonic pitch.

2. Presence of a drone: A characteristic feature of Indian art music is the presence of a drone in the background of the performance, which primarily reinforces the tonic pitch. Salamon, Gulati, and Serra (2012) and Gulati, Salamon, and Serra (2012) use a multi-pitch analysis of the audio signal to exploit this specific property in order to identify the tonic pitch.
3. Rāg knowledge: A rāg is typically characterized by a set of svars along with their relative salience and a set of characteristic melodic phrases (*pakad*). If the rāg of a performance is identified, one can then backtrack the tonic of the performer, as the melodic phrases and dominant svars have a known relationship with the tonic pitch. Bellur et al. (2012) utilize the information regarding the two most salient svars of a rāg (Vadi and Samvadi) to identify the tonic pitch.

2 Methods

There have been various efforts to automatically identify the tonic pitch of the lead artist in a performance of Indian art music (Sengupta et al., 2005; Ranjani et al., 2011; Salamon et al., 2012; Bellur et al., 2012; Gulati et al., 2012). These approaches mainly differ in terms of the musical cues that they utilize to identify the tonic, the amount of input audio data used to perform this task and the type of music material they are devised for (Hindustani or Carnatic, vocal or instrumental, etc.). Despite the differences, all these approaches can be divided into three main processing blocks, as shown in Figure 4. The only exception to this schema is the approach proposed by Sengupta et al. (2005).

In all the aforementioned approaches, the three main processing blocks are the following: feature extraction,

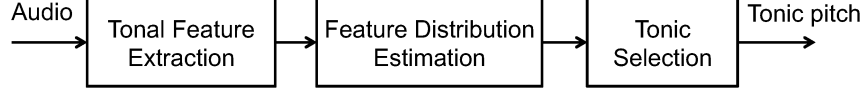


Figure 4: Block diagram of the processing steps used by tonic identification approaches.

Method	Features	Feature Distribution	Tonic Selection
RS (Sengupta et al., 2005)	Pitch (Datta, 1996)	N/A	Error minimization
RH1/2 (Ranjani et al., 2011)	Pitch (Boersma & Weenink, 2001)	Parzen-window-based PDE ¹	GMM fitting
JS (Salamon et al., 2012)	Multi-pitch salience (Salamon, Gómez, & Bonada, 2011)	Multi-pitch histogram	Decision tree
SG (Gulati et al., 2012)	Multi-pitch salience (Salamon et al., 2011)	Multi-pitch histogram	Decision tree
	Predominant melody (Salamon & Gómez, 2012)	Pitch histogram	Decision tree
AB1 (Bellur et al., 2012)	Pitch (De Cheveigné & Kawahara, 2002)	GD ² histogram	Highest peak
AB2 (Bellur et al., 2012)	Pitch (De Cheveigné & Kawahara, 2002)	GD histogram	Template matching
AB3 (Bellur et al., 2012)	Pitch (De Cheveigné & Kawahara, 2002)	GD histogram	Highest peak

Table 1: Summary of existing tonic identification approaches.

feature distribution estimation and tonic selection. Since the task of tonic identification involves an analysis of the tonal content of the audio signal, the features extracted in the first block are always pitch related. In the second block, an estimate of the distribution of these features is obtained using either Parzen window based density estimation or by constructing a histogram. The feature distribution is then used in the third block to identify the tonic. The peaks of the distribution correspond to the most salient pitch values used in the performance (usually the *svars* of the *rāg*), one of which corresponds to the tonic pitch. As the most salient peak in the distribution is not guaranteed to be the tonic, various techniques are applied to select the peak that corresponds to the tonic. In subsequent sections we describe the steps applied by each approach in every processing block (cf. Figure 4). In Table 1 we provide a summary of the methods reviewed in this paper, where the main differences between them become evident.

2.1 Feature extraction

In the tonal feature extraction block, pitch-related features are extracted from the audio signal for further processing. With the exception of the approaches by Salamon et al. (2012) and Gulati et al. (2012), all other approaches use a single feature, the pitch of the lead artist, which is represented by its fundamental frequency (f_0). Note that whilst pitch and f_0 are not the same (the former being a perceptual phenomenon and the latter a physical quantity), for the purpose of tonic identification the f_0 is considered as a reliable representation of pitch. Salamon et al. (2012) use a multi-pitch salience feature in order to exploit the tonal information provided by the drone instrument. Finally, Gulati et al. (2012) use both the multi-pitch salience feature and the f_0 of the lead artist. The two features (f_0 and

¹Pitch density estimation

²Group delay

multi-pitch salience) and the different algorithms used to extract them are described in the following sections.

2.1.1 Fundamental frequency

As mentioned before, Indian art music is fundamentally heterophonic, where the essence of the music is in the main melody which delineates the rāg in a performance. Hence, the melody is a valuable source of information for tonic identification. A simple melody representation that uses the f_0 contour of the predominant source (lead artist) is shown to be a promising feature for tonic identification (Sengupta et al., 2005; Ranjani et al., 2011; Bellur et al., 2012). However, the method used for extracting the f_0 contour plays a crucial role in determining the performance of the tonic identification approach. We next discuss the f_0 estimation methods used by the different tonic identification approaches.

F0 estimation for monophonic music signals:

- **Autocorrelation based method:** Ranjani et al. (2011) use the f_0 contours obtained from the Praat software (Version 5.3) (Boersma & Weenink, 2001). The software implements the algorithm proposed by Boersma (1993). In his work, Boersma proposes to estimate the autocorrelation of the original signal as the ratio between the autocorrelation of the windowed signal and the autocorrelation of the window function. Additionally, cost functions are introduced to detect voiced/unvoiced transitions and octave jumps. This aids in finding the best possible path across frames from the set of candidate peaks obtained from the estimated autocorrelation function (searched within a specified range). Ranjani et al. (2011) down-mix the audio to a mono channel prior to obtaining the f_0 contours from Praat. Further, in the f_0 computation a fixed time step of 10 ms and a frequency range of 80-800 Hz is used.
- **Average magnitude difference function (AMDF) based method:** Bellur et al. (2012) use YIN (De Cheveigné & Kawahara, 2002), an AMDF based f_0 extraction algorithm developed for speech and music sounds. In YIN, the authors propose a set of modifications to the standard autocorrelation based methods in order to reduce the estimation errors. Bellur et al. (2012) apply a low-pass filter with a cut off frequency of 1200 Hz as a pre-processing step before the f_0 extraction using YIN. The authors use a window size of 93 ms, a hop size of 10 ms, and consider a frequency range of 40-800 Hz for the f_0 extraction.
- **Phase-space based method:** Sengupta et al. (2005) use a method based on Phase-Space Analysis (PSA) (Datta, 1996) for the f_0 extraction. In a phase-space diagram, for periodic signals, the trajectory of two points which are separated by a phase of 2π is a straight line with a slope of $\pi/4$. For a quasi-periodic signal such points would lie in a close, highly flattened loop around the same line. As the phase difference increases (wrapped between 0 and 2π) the loop widens, successively increasing the deviation of points from the straight line of

slope $\pi/4$. The deviation is found to be minimal when the phase difference is 2π . This is the underlying logic applied to estimate the fundamental period of the signal.

Sengupta et al. (2005) report three types of estimation errors frequently observed in the obtained f_0 contours: half or doubling of f_0 value, a f_0 value greater than the defined frequency range for a valid f_0 (70-700 Hz in this case) and spikes in the f_0 sequence. A post processing step is applied to correct these three types of errors. Subsequently, a steady state detection is performed on the f_0 contours in order to consider only the steady note regions for the analysis. Only the segments in the f_0 contour with a minimum steady-state duration of 60 ms are used. Note that this method was evaluated on solo vocal performances (monophonic audio), which were carefully recorded in a studio without any accompaniment.

Predominant f_0 estimation for polyphonic music signals:

One of the possible caveats in the aforementioned f_0 estimation methods (or *pitch trackers*) is that they are all designed for monophonic signals containing a single sound source. This means that the number of estimation errors could increase as we add more instruments into the mixture. Due to the heterophonic nature of Indian art music, monophonic pitch trackers to an extent detect the f_0 of the lead melodic source, even in the presence of accompaniment instruments. One way of overcoming this problem is by using a predominant melody (f_0) extraction algorithm. Gulati et al. (2012) use the method proposed by Salamon and Gómez (2012) for estimating the f_0 sequence of the predominant melody from the audio signal. This method was shown to obtain state of the art results in an international evaluation campaign for a variety of musical genres including Indian art music³. Gulati et al. (2012) exploit the pitch information of the predominant melody in the second stage of their approach to identify the specific octave of the tonic (the tonic pitch-class is identified during the first stage of the method).

- **Salience based predominant pitch estimation method:** The melody extraction approach proposed by Salamon and Gómez (2012) consists of four blocks: sinusoid extraction, salience function, contour creation and melody selection. In the first block spectral peaks (sinusoids) are extracted from the audio signal. First, a time-domain equal loudness filter (Vickers, 2001) is applied to attenuate spectral components belonging primarily to non-melody sources (Salamon et al., 2011). Next, the short-time Fourier transform (STFT) is computed with a 46 ms Hann window and 2.9 ms hop size. Finally the frequency and amplitude estimates for the selected peaks are refined by calculating each peak's instantaneous frequency (IF) using the phase vocoder method (Flanagan & Golden, 1966). In the second block, the spectral peaks are used to compute a multi-pitch time-frequency representation of pitch salience over time, a *salience function*. The salience function is based on harmonic summation with magnitude weighting, and spans a range of five octaves from 55Hz to 1760Hz. The peaks

³http://nema.lis.illinois.edu/nema_out/mirex2011/results/ame/indian08/summary.html

of the salience function at each frame represent the most salient pitches in the music recording. In the third block, the peaks of the salience function are grouped over time using heuristics based on auditory streaming cues (Bregman, 1990). This results in a set of pitch contours, out of which the contours belonging to the melody need to be selected. The contours are automatically analyzed and a set of contour characteristics is computed. In the final block of the algorithm, the contour characteristics and their distributions are used to filter out non-melody contours. Then, the melody f_0 at each frame is selected out of the remaining pitch contours based on their salience. Salamon and Gómez (2012) describe the method in detail.

2.1.2 Pitch salience function

As noted earlier, some recently proposed methods for tonic identification (Salamon et al., 2012; Gulati et al., 2012) use a multi-pitch approach. Instead of extracting the predominant melodic component from the audio signal, the methods compute a multi-pitch time-frequency representation of pitch salience over time (salience function) (Salamon et al., 2011). The salience function used in these methods is taken from the first block of the melody extraction algorithm proposed by Salamon and Gómez (2012) (cf. Section 2.1.1). The motivation for using multi-pitch analysis is twofold: first, as noted earlier, the music material under investigation is non-monophonic (includes many instruments playing simultaneously). Second, the tonic is continuously reinforced by the drone instrument, and this important cue cannot be exploited if we only extract a single pitch value for each frame of the audio recording. To illustrate this point, in Figure 5 we show the spectrogram of a short audio excerpt of Hindustani music. Two types of harmonic series are clearly visible in the plot: the first consists of nearly straight lines and corresponds to the drone instrument (playing Sa and Pa). The second harmonic series (which start approximately at time 1 s) corresponds to the voice of the lead performer. If we only consider the pitch of the lead performer (which is the most dominant component in this recording) in our analysis, we lose the information provided by the drone instrument, which in this case is an important indicator of the tonic pitch. Salamon et al. (2012) and Gulati (2012) provide a detailed description of the method and required implementation details. An implementation of the method proposed by Salamon et al. (2012) can be found in *Essentia*⁴ (Bogdanov et al., 2013), an open-source C++ library for audio analysis and content-based music information retrieval.

To further illustrate this point, in Figure 6 we plot the peaks of the salience function computed from the signal whose spectrogram was presented in Figure 5. We see that the tonic pitch (Sa) and the fifth (Pa) played by the *tānpūrā* are clearly visible along with the peaks corresponding to the voice. Since the drone instrument is constantly present in the signal, a histogram of the peaks of the salience function will have prominent peaks at the pitches of the drone instrument, and this is exploited by Salamon et al. (2012) and Gulati et al. (2012) for identifying the tonic. The main

⁴<http://essentia.upf.edu/>

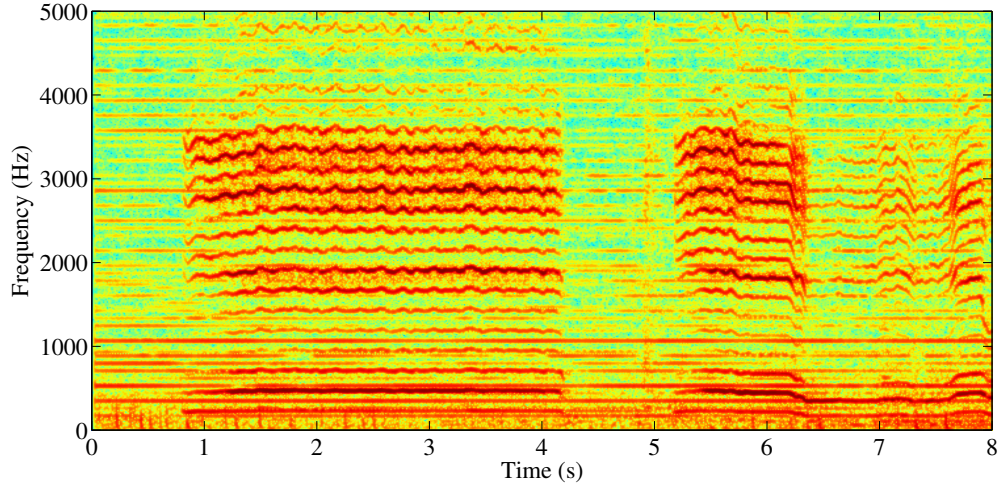


Figure 5: Spectrogram of an excerpt of Hindustani music with two clearly visible types of harmonic series, one belonging to the drone and the other to the lead voice.

difference between the two approaches is that whilst Salamon et al. (2012) directly identifies the tonic pitch from the histogram, Gulati et al. (2012) divides the task into two stages: first, the tonic pitch-class is identified using an extension of the method proposed by Salamon et al. (2012), and then the correct tonic octave is identified using the predominant melody information (cf. Gulati (2012)).

2.2 Pitch distribution functions

The audio features extracted by the different tonic identification approaches are subsequently analyzed in a cumulative manner (cf. block two in Figure 4). The pitch values from all frames (whether a single value is computed per frame or multiple values) are aggregated into a pitch distribution function, which reflects the rate of occurrence (possibly weighted) of different pitch values in the entire audio excerpt. The peaks of the pitch distribution function represent the most frequent (or salient if weighting is used) pitches in the recording, one of which is the tonic. The only exception to this is the approach proposed by Sengupta et al. (2005), which instead of analyzing the distribution of the features, computes an aggregate error function in order to select the tonic. The methods used by the different tonic identification approaches for estimating the pitch distribution function are described below.

2.2.1 Pitch histograms

In the approaches proposed by Salamon et al. (2012) and Gulati et al. (2012), the pitch values of the peaks of the salience function (cf. Section 2.1.2) in every frame are aggregated into a histogram. The top 10 peaks in every frame are used, ensuring that in addition to the lead instrument/voice, the pitch content of other accompanying instruments

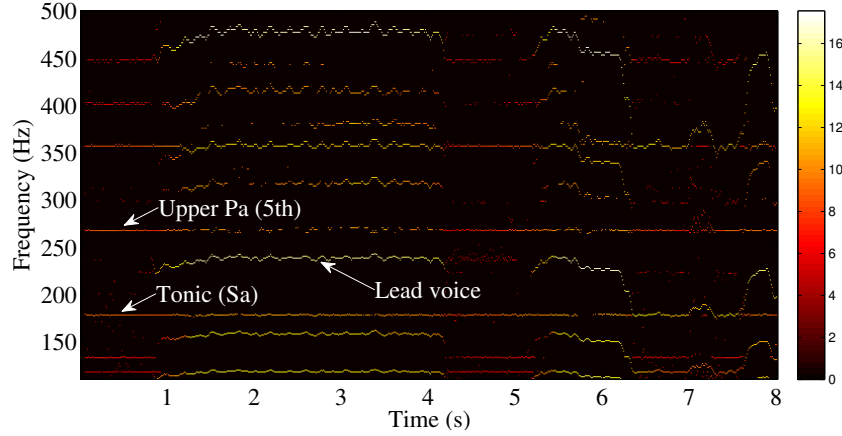


Figure 6: Peaks of the salience function computed for the excerpt from Figure 5. The top 10 peaks of the salience function are shown for each frame, where the magnitude of the peaks is plotted using a logarithmic scale (dB).

is also captured, most importantly the svars played by the drone instrument. The frequency range considered for selecting the peaks of the salience function for constructing the histogram is restricted to 100-370 Hz. Note that the typical frequency range for the tonic pitch is 100-260 Hz. The reason for computing the histogram beyond 260 Hz even though the tonic rarely goes above this frequency, is that in some cases the aforementioned methods can exploit the presence of a peak corresponding to the fifth/fourth (Pa/Ma) above the tonic in order identify the tonic pitch.

Since in many cases the lead voice/instrument is considerably louder than the drone sound (cf. Figure 6), the weights of the peaks in the salience function are ignored in the computation of the pitch histogram, meaning only the rate of occurrence is taken into account. As noted earlier, the result is that the pitches produced by the drone instrument (the tonic and Pa, Ma or Ni) manifest in the form of high peaks in the histogram, since the drone sounds continually in the recording. The resulting pitch distribution thus depends heavily on the svars produced by the drone instrument. This would not be the case if we only considered the predominant melody for computing the histogram, in which case the pitch distribution would depend on the rāg, thus increasing the complexity of the task. In Figure 7 we show two pitch histograms, computed using (a) the pitch of the predominant melody and (b) the peaks of a multi-pitch salience function. Both histograms are computed from the same three-minute audio excerpt. We see that in the histogram computed using the predominant melody (see Figure 7 (a)), the prominent peaks correspond to svars Sa, Ga and Re (the prominent svars of the rāg *Sindh Bhairavī*), whereas in the multi-pitch histogram (see Figure 7 (b)), the top three peaks correspond to Sa (in two octaves) and Pa, which are the prominent svars produced by the drone instrument.

Bellur et al. (2012) construct a pitch histogram using a frequency range of 40-800 Hz with a 1 Hz resolution, which is later post processed using a group delay function. The authors show that by assuming that the constructed

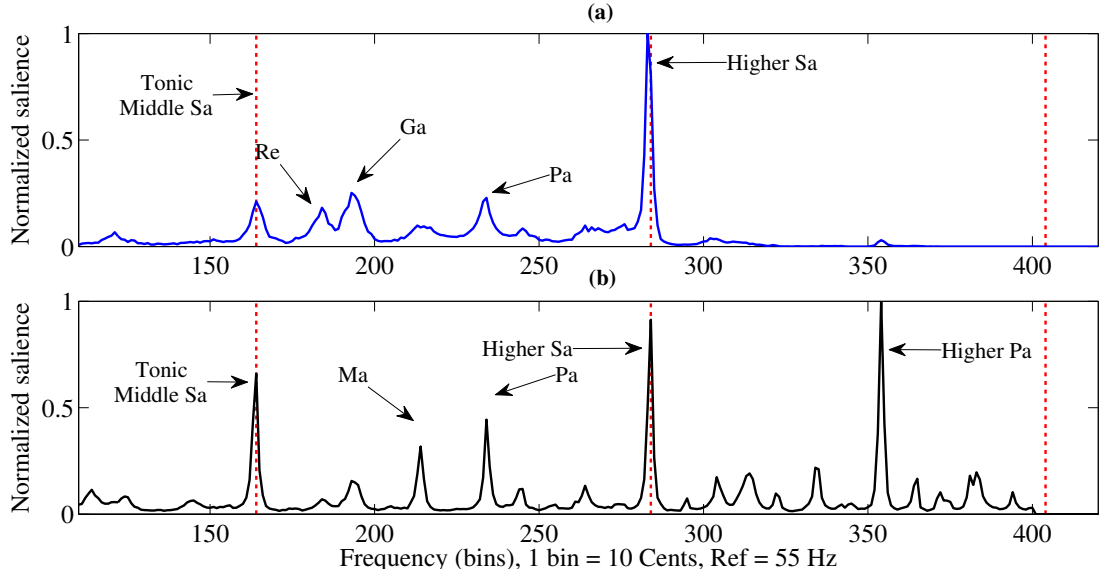


Figure 7: Pitch histograms for the same excerpt constructed using (a) predominant melody (in blue) and (b) peaks of a multi-pitch salience function (in black). The tonic pitch-class locations are indicated with red dotted lines.

pitch histogram is the squared magnitude of resonators in parallel, group delay functions can be applied to obtain a better resolution for the peaks in the resulting histogram. It is also shown that a group delay function accentuates peaks with lesser bandwidths. Given that the *şadja* (Sa, the tonic pitch-class) and *panchama* (Pa, fifth with respect to the tonic) in all octaves are relatively less inflected, this characteristic of the group delay function is shown to be beneficial for improving the accuracy of tonic identification. The processed histograms are referred to as group delay (GD) histograms.

Bellur et al. (2012) also propose the concept of segmented histograms. In order to exploit the continuous presence of the *şadja*, the authors propose to segment the f_0 contour of a music excerpt into smaller units and compute a GD histogram for each unit. Later, the individual histograms computed for each unit are combined by taking a bin-wise product. Given that the *şadja* is present in all the units, the peak corresponding to the *şadja* is enhanced in the combined GD histogram. This also helps in reducing the salience of the non-*şadja* peaks which might not be present in all the segments. Tonic selection is then performed on the combined histogram, referred to as the segmented GD histogram.

2.2.2 Pitch density function

Instead of using a histogram, Ranjani et al. (2011) use a Parzen window estimator to compute a pitch density function. Parzen window estimators (or kernel density estimators) are non-parametric density estimators. The choice of kernel

function can control the smoothness of the estimated density. They are widely used as an alternative to histograms to alleviate the problem of discontinuity at the boundaries of the bins of the histogram, and aid in a smoother peak picking process. In addition, they do not require partitioning of the data into distinct bins. Given n samples of pitch data x_i ($i = 1 \dots n$) in Hz, the Parzen window pitch density estimate for any (unobserved) pitch value k is given by equation 1 (Duda, Hart, & Stork, 2000).

$$\hat{p}_n(k) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n} \phi\left(\frac{k - x_i}{h_n}\right) \quad (1)$$

where, ϕ denotes the kernel function for the estimation. Kernel density estimators are sensitive to the choice of variance (Bishop, 2006; Duda et al., 2000). Ranjani et al. (2011) use Parzen window estimators with Gaussian kernels for estimating the density of the extracted pitch frequencies. The smoothing parameter h_n is kept fixed and was set after careful experimentation to 0.7.

2.3 Tonic selection

In the previous section we discussed different ways to compute the pitch distribution function. This section presents the last processing block shown in Figure 4, where the pitch distribution function is used to identify the tonic pitch. The peaks of the pitch distribution function correspond to the most frequent (or salient) pitches present in the audio signal. Depending on how the pitch distribution is computed, the peaks will either coincide with the svars of the rāg or with the svars produced by the drone instrument. The problem of tonic identification is thus reduced to selecting the peak of the distribution function that corresponds to the tonic of the lead artist. As noted earlier, the peak corresponding to the tonic pitch is not always the highest peak in the distribution. For this reason, various strategies are proposed for analyzing the pitch distribution and selecting the peak that corresponds to the tonic. The complexity of the approaches varies from simply selecting the highest peak of the histogram to the application of machine learning algorithms in order to automatically learn the best set of rules for selecting the tonic peak. We briefly describe the different tonic selection strategies used in the aforementioned approaches.

2.3.1 Semi-continuous GMM fitting

Ranjani et al. (2011) model the pitch distribution using semi-continuous Gaussian mixtures, motivated by the following two musical cues in Indian art music: first, the relative positions of the svars with respect to the tonic hover around a mean ratio (Krishnaswamy, 2003b) and second, the śadja (Sa, tonic pitch-class) and panchama (Pa, fifth with respect to the tonic pitch-class) are the prakrthi (natural) svars which means that they are sung or played without any inflections (Manikandan, 2004; Krishnaswamy, 2003a).

From the obtained pitch density function (using the Parzen window technique), J peaks are chosen within a

suitable pitch range (P_{min}, P_{max}). The frequencies corresponding to these peaks constitute possible tonic candidates, $S_0(j)$; $j \in 1 : J$. As noted above, one of the key characteristics of şadja and panchama is that they do not vary (in pitch) throughout the performance. The variance can be inferred by modeling each tonic candidate (i.e. peak) with a Gaussian distribution. Motivated by this, Ranjani et al. (2011) use a semi-continuous (SC) GMM (Huang, Acero, & Hon, 2001) fit for each of the J candidates. The means μ_i of the SC-GMM are fixed to the 12 possible svar ratios across three octaves (i.e. $i \in [1 : 36]$). The weights α_i and variances σ_i of the mixture model are inferred using the EM algorithm (Dempster, Laird, & Rubin, 1977), with the means kept fixed during the maximization step. The likelihood of the fit is not used as the criterion for determining the şadja, but the inferred parameters are used in the decision process. The authors study five different tonic estimators for using the SC-GMM parameters to identify the tonic, two of which are included in the comparative evaluation conducted in this study:

$$\theta_1 = \arg \min_{S_0(j)} \left\{ \frac{\sigma_{S_0}}{\alpha_{S_0}} \mid S_0(j) \right\} ; j \in [1 : J] \quad (2)$$

$$\theta_2 = \arg \min_{S_0(j)} \left\{ \frac{\sigma_{S_0} + \sigma_{P_0} + \sigma_{S_+}}{\alpha_{S_0} + \alpha_{P_0} + \alpha_{S_+}} \mid S_0(j) \right\} ; j \in [1 : J] \quad (3)$$

Here, S_0 , P_0 and S_+ denote the madhya şadja (middle Sa), panchama (fifth) and tara şadja (higher Sa). The performance of these two estimators are reported under the labels RH1 (equation 2) and RH2 (equation 3). Ranjani et al. (2011) provide further details of this method. For the present evaluation, J is set to 10. For evaluating performance without the availability of song metadata, P_{min} and P_{max} are set to 100 Hz and 250 Hz respectively. When the information regarding the gender of the vocalist is added to the algorithm, P_{min} and P_{max} are set to 100 Hz and 195 Hz for the excerpts corresponding to male vocalists and to 135 Hz and 250 Hz for the female vocalists.

2.3.2 Classification based approach

Salamon et al. (2012) and Gulati et al. (2012) use a classification based approach to identify the peak of the multi-pitch histogram which corresponds to the tonic pitch (Salamon et al., 2012) or tonic pitch-class (Gulati et al., 2012). Since all the pitches in a performance are in relation to the tonic, the relationships between the peaks of the histogram (height and distance) are used to compute a set of features, which are then used to train a classifier for identifying which peak corresponds to the tonic. In this way, rather than having to manually define a template for selecting the tonic, an optimal set of rules can be learned automatically using machine learning.

Given a pitch histogram, the authors select the top 10 peaks as the candidates for the tonic pitch (or pitch-class). Subsequently, they compute the distance between every tonic candidate p_i and the highest candidate in the histogram p_1 . This gives a set of pitch interval features d_i ($i = 1 \dots 10$), where d_i is the distance in semitones between p_i and p_1 . Another set of amplitude features a_i ($i = 1 \dots 10$) is computed, where a_i is the amplitude ratio between p_i and p_1 . For training the classifier, every audio excerpt is annotated with a class label, “first” if the highest peak of the

pitch histogram is the tonic, “second” if the second-highest peak is the tonic, and so on. The goal of the classifier is thus to identify the rank of the peak in the histogram that corresponds to the tonic. To reduce the amount of features necessary for classification and increase the generalizability of the approach, the authors apply attribute selection using the *CfsSubsetEval* attribute evaluator and *BestFirst* search method (Hall, 1999) in a 10-fold cross validation framework, only keeping features that were used in at least 80% of the folds. After feature selection, the number of features is reduced from 20 to just 3: d_2 , d_3 and d_5 .

For classification, the Weka data-mining software is used (Hall et al., 2009). Salamon et al. (2012) and Gulati et al. (2012) experiment with many classification algorithms, including the C4.5 decision tree (Quinlan, 1993), support vector machines (SMO) and an instance based classifier (k^*) (Witten, Frank, & Hall, 2011). The authors show that for the tonic identification task the decision tree classifier yields the highest classification accuracy. For the comparative evaluation in this paper, a C4.5 decision tree classifier is used with the same parameter settings reported by Salamon et al. (2012) and Gulati et al. (2012).

Since Gulati et al. (2012) use the classifier to identify the tonic pitch-class (and not the pitch), in this approach each excerpt is labeled with the rank of the highest peak in the histogram that corresponds to the tonic pitch-class (since the frequency range considered for the histogram computation spans more than one octave, there could be multiple peaks in different octaves representing the tonic pitch-class (Gulati, 2012)). The second stage of the approach proposed by Gulati et al. (2012) is also classification based, only now the goal is to identify the correct octave of the tonic, as the pitch-class is already identified in the previous step. To do this, the authors use the pitch histogram computed from the f_0 sequence of the predominant melody. For every candidate pitch (candidates have the same pitch-class but are in different octaves) a set of 25 features is computed h_i ($i = 1 \dots 25$). The features are the values of the melody histogram at 25 equally spaced locations spanning two octaves centered around the tonic pitch candidate. An example is provided in Figure 8 for a tonic pitch candidate at bin 166 (143.5 Hz). The 25 melody histogram values used as features are marked by the stars.

In this case, the classification task is a two-class problem: either the pitch candidate is in the correct octave, or it is not. For training, a class label is assigned to every pitch candidate: ‘TonicOctave’ if the tonic candidate is in the correct octave, or ‘NonTonicOctave’ otherwise. As before, a C4.5 decision tree is trained using the Weka data-mining software with attribute selection. Gulati (2012) provides a detailed description of the method.

2.3.3 Error minimization

Sengupta et al. (2005) use an error minimization technique to identify the tonic. This is a brute force approach in which a large number of pitch values within a pre-defined frequency range are considered as candidates for the tonic pitch. A cumulative deviation is computed between the steady state regions of the pitch contour (described

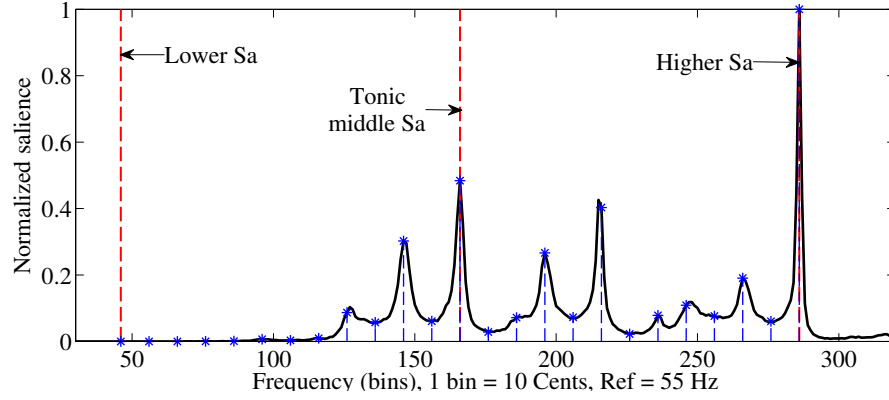


Figure 8: An example of the predominant melody histogram extracted from an audio excerpt. The red lines mark the tonic pitch-class locations

in Section 2.1.1) and the pitch values of the closest svars to these regions, which are obtained using three different tuning schemas given a tonic candidate. The tonic candidate which results in the minimum deviation is selected as the tonic of the musical excerpt.

2.3.4 Highest peak

Bellur et al. (2012) propose a simple approach of selecting the highest peak of the pitch distribution as the tonic. In methods AB1 and AB3, the bin value of the highest peak of the segmented GD pitch histogram is selected as the tonic pitch. The frequency range of the histogram is restricted to 100-250 Hz. When the information regarding the gender of the vocalist is available, this range is further restricted.

2.3.5 Template matching

In addition to the simple highest peak selection approach mentioned above, Bellur et al. (2012) also propose a template matching approach to identify the tonic pitch (AB2). This approach also exploits the smaller degree of pitch variation around şadja and panchama svars, like in the approach by Ranjani et al. (2011). The procedure is as follows: a GD pitch histogram is computed for a given piece of music. Peaks of the GD pitch histogram within a certain range are selected and the frequency values of the bins serve as candidates for the tonic. Let G represent a vector with the magnitude of the candidate peaks at corresponding frequency values and zero in all other bins. For a tonic candidate with frequency i , the following template summation is computed:

$$T(i) = \sum_{k=-\Delta}^{\Delta} G(i/2 + k) + G(3i/4 + k) + G(i) + G(3i/2 + k) + G(2i + k) \quad (4)$$

where $\Delta = 3$. The frequency value for which $T(i)$ is highest is selected as the tonic pitch value.

3 Evaluation Methodology

In this paper we evaluate seven of the eight reviewed tonic identification methods, denoted JS (Salamon et al., 2012), SG (Gulati et al., 2012), RH1 and RH2 (Ranjani et al., 2011), AB1, AB2 and AB3 (Bellur et al., 2012) (cf. Table 1). RS (Sengupta et al., 2005) was not available for evaluation. Each approach is evaluated on six different datasets, denoted CM1, CM2, CM3, IITM1, IITM2 and IISCB1 (cf. Section 3.1). AB1 requires several excerpts from the same concert in order to compute the segmented GD histogram, and this kind of data (and metadata) is only available for the IITM1 dataset. Hence, AB1 is only evaluated for IITM1 dataset.

For vocal performances we evaluate the accuracy of correctly identifying the tonic pitch, whereas for instrumental music we evaluate the accuracy of estimating the tonic pitch-class only (i.e. the identified tonic pitch is allowed to be in any octave). This is because whilst for vocal music the idea of the tonic pitch being in a specific octave is clearly defined (because it is restricted by the pitch range of the singer), this notion is not as clear for Hindustani instrumental music. For vocal performances, the tonic identified by a method is considered correct if it is within 50 cents of the ground truth annotation. For instrumental music, a method’s estimate is considered correct if it is within 50 cents of the correct tonic pitch-class. Classification based approaches, which require training (JS and SG) are evaluated by performing 10-fold cross-validation on every dataset, repeating every experiment 10 times and reporting the mean accuracy over the 10 repetitions. All parameters are kept fixed for all methods across all datasets.

3.1 Datasets

The datasets used for evaluation in this study are subsets of three different music collections, which are described below. A summary of all the datasets, including relevant statistics, is provided in Table 2.

3.1.1 CompMusic Music Collection

This audio music collection has been compiled as part of the CompMusic project (Serra, 2011). The audio recordings are ripped from commercial quality audio CD releases and stored in 160 kbps mp3 format (stereo). The metadata corresponding to every recording is stored in Musicbrainz⁵. Currently the audio collection contains approximately 400 CDs comprising 2400 recordings spanning roughly 520 hours of audio data, including both Hindustani and Carnatic music. A small randomly selected subset of this large collection was manually annotated and divided into three datasets (CM1, CM2 and CM3) to be used for evaluation. The datasets CM1 and CM2 contain 3-minute-long

⁵<http://musicbrainz.org/>

Dataset	Avg. length (min)	#Excerpts	Hi.(%)	Ca.(%)	Voc. (M/F)(%)	Inst. (%)	#U song	#Uartists
CM1	3	271	41	59	0	100	169	33
CM2	3	935	45	55	100 (68 / 32)	0	547	81
CM3	14.8	428	45	55	100 (72 / 28)	0	428	71
IITM1	144.6	38	0	100	89 (79 / 21)	11	N/A	22
IITM2	12.3	472	0	100	92 (77 / 23)	8	472	22
IISCB1	7.4	55	0	100	100 (80 / 20)	0	55	5

Table 2: Dataset summary, including average excerpt length (Avg. length), number of excerpts (#Excerpts), percentage of Hindustani music (Hi), Carnatic music (Ca), vocal excerpts (Voc.), instrumental excerpts (Inst.), number of unique songs (#U song) and number of unique artists (#Uartists) in each dataset. For vocal excerpts we also provide the breakdown into male (M) and female (F) singers. Percentage (%) values are rounded to the nearest integer.

excerpts extracted from full length songs. When the full song was longer than 12 minutes, 3 excerpts were extracted from the beginning, middle and end of the recording. When the song was shorter, only one excerpt was extracted from the beginning of the recording. By taking excerpts from different sections of a song we ensure that the datasets are representative, since the musical characteristics can change significantly between different parts of a recording. CM1 contains exclusively instrumental performances, and does not overlap with CM2 and CM3. The latter two contain exclusively vocal performances, where CM3 contains full performances and CM2 contains excerpts taken from the performances in CM3.

3.1.2 IITM Music Collection

This collection was compiled by selecting 40 concerts from a private collection of hundreds of live concert recordings. The 40 concerts consist of 472 pieces. In order to study the robustness of tonic identification methods, the concerts that were selected range from artists from the 1960's to present day artists. The quality of the recordings vary from poor to good, usually depending on the period in which they were made. IITM1 is comprised of 38 concerts. IITM2 consists of pieces extracted from the 40 selected concert recordings. The performances are of varying duration, ranging from 46 seconds to 85 minutes.

3.1.3 IISCB Music Collection

The audio material in this dataset is obtained from an online Carnatic music archive⁶ compiled by Carnatic musician and enthusiast Dr. Shivkumar Kalyanaraman, for the benefit of music amateurs and hobbyists as an online educa-

⁶<http://www.shivkumar.org/music/index.html>

tional resource. The archive includes various forms of Carnatic music. The IISCB1 dataset is comprised of 55 songs in the *alapana* form, recorded by 5 singers across 7 rāgs. The total duration of the dataset is 6.75 hours. It includes recordings from the last 50 years, many of which were recorded live on analog audio tapes. The overall quality of the recordings is not very high.

3.2 Annotations

The tonic pitch for vocal performances and tonic pitch-class for instrumental performances was manually annotated for each excerpt in the CM1, CM2 and CM3 datasets by Gulati (2012). The annotations were later verified by a professional Carnatic musician and the number of discrepancies was very small. To assist the annotation process, the author used the candidate generation part of the approach proposed by Salamon et al. (2012). For every excerpt the top 10 tonic candidates were synthesized together with the original audio file to help identify and label the correct candidate. Note that the correct tonic pitch was always present amongst the top 10 candidates. A detailed description of this procedure is provided by Gulati (2012).

The tonic pitch for the IITM1 and IITM2 datasets was manually annotated by a professional musician, and for IISCB1 was manually annotated by two professional musicians, S. Raman and S. Vijayalakshmi.

4 Results and Discussion

In this section we present the results obtained by the different tonic identification methods and discuss various types of errors made by them. The section is divided into three parts: in Section 4.1 we present the results obtained when only the audio data is used and no additional metadata is provided to the methods. Subsequently, we report the performance accuracy obtained when information regarding the gender of the singer (male or female) and performance type (instrumental or vocal) is provided to the methods in addition to the audio data (Section 4.2). Finally in Section 4.3 we present an analysis of the most common errors made by the methods and make some general observations regarding their performances.

4.1 Results obtained using only audio data

4.1.1 Overall results

In Table 3 we summarize the identification accuracies (in percentage) for tonic pitch (TP) and tonic pitch-class (TPC) obtained by seven methods on six datasets, using only audio data.

We see that most of the methods perform well on all datasets, and the accuracy of the best performing method on each dataset ranges from 84-97%. We note that the identification accuracy obtained for instrumental music (CM1) by

Methods	CM1		CM2		CM3		IISCB1		IITM1		IITM2	
	TP	TPC	TP	TPC	TP	TPC	TP	TPC	TP	TPC	TP	TPC
JS	-	88.9	87.4	90.1	88.4	91	75.6	77.5	89.5	97.4	90.8	94.1
SG	-	92.2	87.8	90.9	87.7	90.5	79.8	85.3	97.4	97.4	93.6	93.6
RH1	-	81.4	69.6	84.9	73.2	90.8	81.8	83.6	92.1	97.4	80.2	86.9
RH2	-	63.2	65.7	78.2	68.5	83.5	83.6	83.6	94.7	97.4	83.8	88.8
AB1	-	-	-	-	-	-	-	-	89.5	89.5	-	-
AB2	-	88.9	74.5	82.9	78.5	83.4	72.7	76.4	92.1	92.1	86.6	89.1
AB3	-	86	61.1	80.5	67.8	79.9	72.7	72.7	94.7	94.7	85	86.6

Table 3: Accuracies for tonic pitch (TP %) and tonic pitch-class (TPC %) identification by seven methods on six different datasets using only audio data. The best accuracy obtained for each dataset is highlighted using bold text. The dashed horizontal line divides the methods based on supervised learning (JS and SG) and those based on expert knowledge (RH1, RH2, AB1, AB2 and AB3). TP column for CM1 is marked as ‘-’, because it consists of only instrumental excerpts for which we not evaluate tonic pitch accuracy.

each method is comparable to the accuracy obtained for vocal music, meaning the approaches are equally suitable for vocal and instrumental music. The approaches based on multi-pitch analysis and classification (JS and SG) are more consistent and generally perform better across different datasets compared to the approaches based only on the predominant pitch (with the exception of IISCB1, most likely due its poor recording quality). Within the multi-pitch based approaches, SG obtains slightly better identification accuracy than JS. This is most likely due to the additional predominant melody information used in SG, and indeed the difference between the two approaches is mainly in the TP measure and less so in the TPC measure (i.e. correctly identifying the octave of the tonic pitch). As could be expected, the simple maximum peak selection approach employed by AB1 and AB3 is too simplistic and the template matching approach employed in AB2 yields better results in most cases.

SG obtains the best results for the instrumental dataset CM1, with AB2 and JS reporting comparable accuracies. For the CM2 and CM3 datasets, we see that the multi-pitch based approaches (SG and JS) obtain the best performance, whilst the predominant pitch based methods exhibit a considerable difference between the TP and TPC accuracies. This means that in many cases these approaches are able to identify the tonic pitch-class correctly but fail to identify the correct octave of the tonic pitch. In the case of RH1, RH2, AB2 and AB3, this can be attributed primarily to the tonic selection procedure employed by these approaches. The group-delay processing used in AB2 and AB3, and the estimators used in RH1 and RH2, accentuate the peaks corresponding to all svars which have a low degree of pitch variance. This includes both the lower and higher octave *ṣadja* and *panchama* in addition to the middle octave

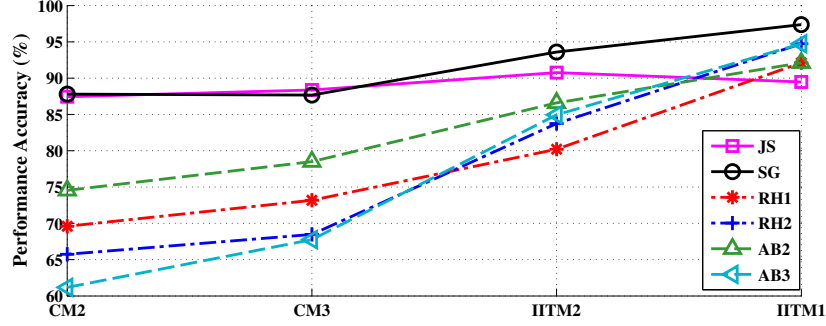


Figure 9: Accuracy (%) of different methods on four datasets arranged by increasing order of mean duration.

şadja (the tonic pitch). Furthermore, the magnitude of peaks corresponding to şadja in higher and lower octave is sometimes further accentuated by pitch halving and doubling errors produced by the pitch extraction algorithm. This makes identification of the correct tonic octave more difficult and as seen in Table 3, results in a higher degree of octave errors.

When considering the results for the IISCB1 dataset, we note that the performance drops for all methods. The main reason for this is the poor audio quality of the excerpts in this collection. The recordings are relatively old and noisy, and contain a humming sound in the background. This makes pitch tracking very difficult. Furthermore, the drone sound in the recordings is very weak compared to the lead artist, which explains the drop in performance for the multi-pitch based approaches.

If we consider performance for IITM1 on the other hand, we see that all methods perform very well. This is because each excerpt in this dataset is a full concert, which includes many performances in different rāgs. Usually different set of svars are used in different performances, but with the same tonic pitch throughout the concert. As a result, the melody histogram contains a very high peak corresponding to the Sa svar, making it considerably easier to identify the tonic pitch.

4.1.2 Accuracy as a function of excerpt duration

As shown in Table 2, different datasets contain audio excerpts of different lengths. In order to investigate a possible correlation between the accuracy of a method and the length of an audio excerpt, in Figure 9 we plot the identification accuracies of the different methods for four of the six datasets ordered by the mean duration of the excerpts: CM2 (3 min), CM3 (full song), IITM2 (full song) and IITM1 (full concert). CM1 and IISCB1 are excluded because the characteristics of these datasets are very different compared to the rest of the datasets (CM1 contains only instrumental performances and IISCB1 has poor quality audio). As could be expected, we note that practically for all methods there is an improvement in the performance as we increase the duration of the excerpts. Interestingly, the

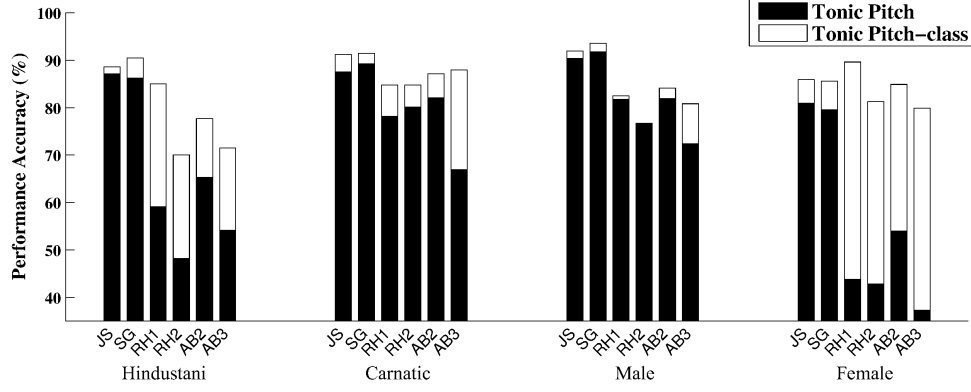


Figure 10: Accuracy (%) as a function of different attributes (Hindustani, Carnatic, male, female).

improvement is very significant for the predominant pitch based methods (RH1, RH2, AB2 and AB3) compared to the multi-pitch based methods (JS and SG). This implies that the latter approaches, which exploit the pitch information of the drone instrument, are more robust to the duration of audio data.

4.1.3 Accuracy as a function of excerpt characteristics

In addition to analyzing the performance accuracy for the whole dataset, we also examine the results as a function of different attributes of the audio excerpts, namely music tradition (Hindustani or Carnatic) and the gender of the lead singer (male or female). For this analysis we use the CM2 dataset, as it has the most balanced representation of excerpts from the different categories. In Figure 10 we show the accuracies obtained by the different methods as a function of the different attributes. We see that the performance of the multi-pitch based approaches (JS and SG) is relatively independent of the music tradition (Hindustani or Carnatic). On the other hand, for the predominant pitch based approaches there is a significant difference in performance for Hindustani and Carnatic music (they obtain considerably better results on Carnatic music). The most notable difference for these approach is the increased amount of octave errors made for Hindustani music compared to Carnatic music. A possible reason for this is that in the Hindustani recordings the tānpūrā is generally more salient compared to the Carnatic recordings. This results in the monophonic pitch estimators tracking the tānpūrā in some frames, in particular when the lead artist is not singing. As a result the pitch histogram includes high peaks at octave multiples or sub-multiples of the correct tonic pitch. In the case of AB2, AB3, RH1 and RH2, most octave errors were found to be sub-multiples of the tonic pitch, caused by the stable and salient lower Sa played by the drone instrument.

Now we turn to examine the performance as a function of the gender of the lead artist (male or female). We see that in general, all approaches perform better on performances by male singers compared to those by female singers. As in the case of Hindustani versus Carnatic music, the difference is once again considerably more significant for the

Methods	CM1	CM2	CM3	IISCB1	IITM1	IITM2
	TPC	TP	TP	TP	TP	TP
JS	88.9	93.6	92.4	80.9	97.4	92.3
SG	92.2	90.9	90.5	85.3	97.4	93.6
RH1	87.7	83.5	88.9	87.3	97.4	91.7
RH2	79.55	76.3	82	85.5	97.4	91.5
AB1	-	-	-	-	97.4	-
AB2	92.3	91.5	94.2	81.8	97.4	91.1
AB3	87.5	86.7	90.9	81.8	94.7	89.9

Table 4: Accuracies (tonic pitch-class (%)) when using additional information regarding the gender of the lead singer (male/female) and performance type (vocal/instrumental). The dashed horizontal line divides the methods based on supervised learning (JS and SG) and those based on expert knowledge (RH1, RH2, AB1, AB2 and AB3).

predominant pitch based methods, which make a lot of octave errors for performances by female singers. As noted earlier, in methods RH1, RH2, AB2 and AB3 a range of 100-250 Hz is considered for finding the tonic pitch when no additional metadata about the artists is available. In the case of female singers, the tonic usually resides in the higher end of this range. However, the presence of the drone, the tonal sounds produced by percussive instruments and the octave errors produced by the pitch tracker, all contribute to the appearance of a high peak one octave below the tonic of the female singer. This is especially the case for 3-minute excerpts where a limited amount of vocal pitch information is available. In the case of the approaches based on multi-pitch analysis and classification (JS and SG), a probable reason for obtaining better performance for male singers is the larger amount of excerpts with male singers in the database. As a result, it is possible that the rules learned by the classifier are slightly biased towards the performances of male singers.

4.2 Results obtained using metadata together with the audio

In order to reduce the amount of octave errors caused by the different tonic pitch ranges used by male and female singers, one can use additional information regarding the gender of the singer (when available) to guide the method, usually by adjusting the frequency range considered for the tonic pitch. In this section we analyze the effect of including information regarding the gender of the singer and the performance type (vocal or instrumental) on the identification accuracy obtained by the different methods.

In Table 4 we present the identification accuracies obtained when gender information (male/female) and perfor-

mance type (vocal/instrumental) is available to the methods in addition to the audio data. Note for this evaluation we only report the tonic pitch accuracy for vocal excerpts (and not pitch-class accuracy) since when this metadata is available the pitch range of the tonic is known and limited to a single octave, meaning the TP and TPC accuracies will be the same.

Comparing the results with those in Table 3 we see that the identification accuracies for all methods are higher when gender and performance metadata is available. With the additional information the performance of the predominant pitch based approaches (AB2, AB3 and RH1) becomes closer to that of the multi-pitch based approaches (JS and SG). Whilst the performance of all methods is improved, the increase in accuracy is more considerable for the predominant pitch based approaches which use template matching (in particular AB2 and AB3) compared to classification-based approaches (JS and SG). This can be attributed to the fact that the rules learned automatically using machine learning are more complete compared to the relatively simple Sa-Pa templates, meaning that the classification based approaches can correctly identify the octave of the tonic even without using gender metadata. That is, since both male and female excerpts are used during training, the influence of the gender of the singer on the pitch features is implicitly learned by the classifier, thus producing rules that can handle both male and female performances, even without explicit metadata about the gender of the singer. On the other hand, manually defined template-based approaches require this extra information to fine-tune the frequency range considered for the tonic, after which they obtain comparable performance to that of the classification-based methods.

A potential advantage of the template-based approaches is that they do not require training. This, in theory, could make them more generalizable compared to the classification-based methods. To assess this, we ran an experiment in which the classification-based approaches were trained on one dataset and tested on a different dataset (CM2 and IITM2). We found that the results only went down by approximately 2% compared to the results obtained using 10-fold cross validation on a single dataset. Furthermore, the datasets used for this experiment contained relatively different music material (percentage of Carnatic music excerpts and length of the audio files). This suggests that for tonic identification the rules learned by the classification-based approaches are generalizable and can be used to obtain high identification accuracies for the different types of excerpts evaluated in this study.

4.3 Error Analysis

We now turn to analyze the different types of errors made by the methods, both with and without using additional metadata for each dataset. Three common types of errors were identified: Pa errors, where the fifth (Pa) is selected instead of the tonic, Ma errors, where the fourth (Ma) is selected instead of the tonic, and the previously mentioned octave errors, where the correct pitch is identified but in the wrong octave (usually one octave above or below the tonic). Since octave errors are already discussed at length in the previous section, here we focus on all other types

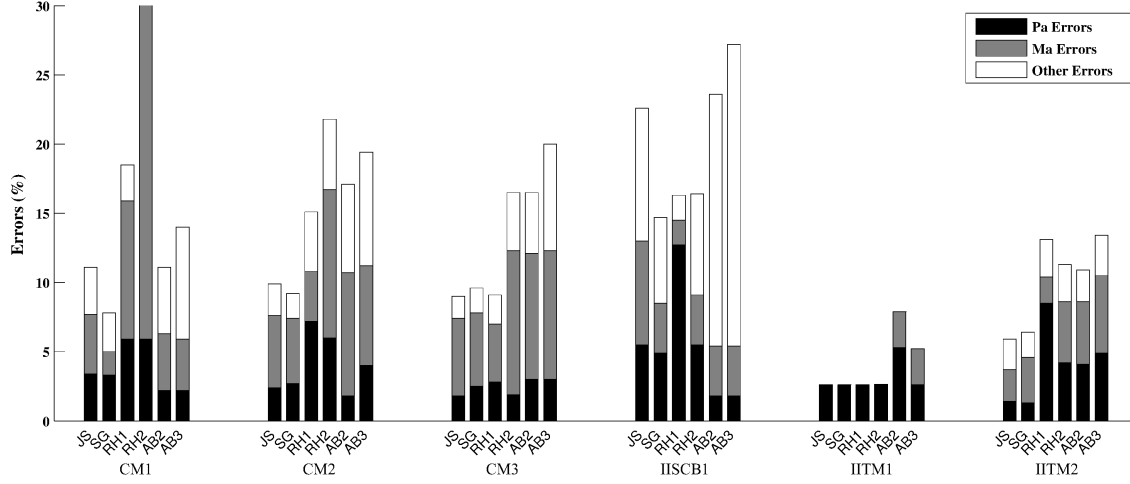


Figure 11: Percentage of excerpts containing each of the three different categories of errors (excluding octave errors): Pa, Ma and Other, when no additional metadata is used.

of errors, which we divide into three categories: Pa (for Pa errors), Ma (for Ma errors) and “Other”, which includes all errors that are neither Pa, Ma nor octave errors (e.g. selecting the seventh (Ni) instead of the tonic Sa).

4.3.1 Errors when only audio data is used

In Figure 11 for each dataset we present the percentage of excerpts containing each of the three categories of errors for every method (when no additional metadata is used). We see that for most datasets Pa and Ma errors constitute a large proportion of the total amount of errors made by each method. These confusions make sense from a musical perspective, since in every Indian art music performance one of these two svars (Pa or Ma) is always present in the melody in addition to Sa (the tonic pitch-class). Furthermore, the pitch distance between Sa and Pa (fifth) is the same as the distance between Ma and higher Sa, and the pitch distance between Sa and Ma (one fourth) is same as the distance between between Pa and higher Sa. Since most approaches are based on templates or rules that consider the pitch distance between the peaks of the feature histogram, these equivalences can cause four types of confusions: considering a Sa-Pa pair to be Ma-Sa leading to a Pa error, considering Ma-Sa to be Sa-Pa leading to a Ma error, considering Sa-Ma to be Pa-Sa leading to a Ma error and considering Pa-Sa to be Sa-Ma leading to a Pa error.

For the approaches based on multi-pitch analysis (JS and SG) we observe that the only case where we get more ‘Other’ errors compared Pa and Ma errors is for the IISCB1 dataset. Since the drone sound is very weak in the excerpts of this dataset, there are cases in which the prominent peaks of the multi-pitch histogram correspond to svars other than Sa, Ma and Pa (which will depend on the choice of the rāg). Since these approaches assume that the multi-pitch histogram represents the svars of the drone instrument, the peaks of the histogram are mistakenly

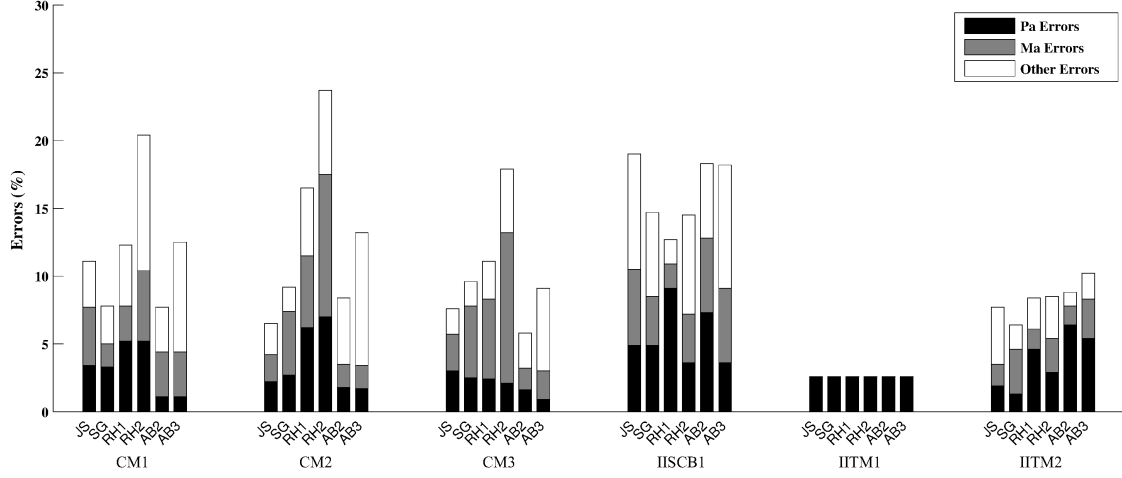


Figure 12: Percentage of different type of errors (Pa, Ma and Others) by different methods on all the datasets using information regarding the gender of the singer and performance type

identified as Sa and Pa or Sa and Ma, leading to an error in identification. For these specific type of excerpts the RH1 method produces slightly better results, as the şadja is not inflected (i.e. there is little pitch variation) regardless of the rāg.

In many cases we observe that the percentage of Ma errors is greater than the percentage of Pa errors. For the classification based approaches, this can be attributed to the fact that in most excerpts the drone instrument is tuned to *Pa tuning* (lower Sa, middle Sa, middle Sa, lower Pa). This creates a bias in the training and the rules learned by the classifier work better for Pa tuning. Ma errors are also common in RH2, as the estimator looks for a Sa-Pa-higher Sa pitch relation, which would also fit a Ma-tuned performance. RH1 on the other hand does not search for a Sa-Pa-Sa template, resulting in a low proportion of Ma errors compared to the other methods. Finally we note that most methods do not make any Ma errors on the IITM1 dataset. This is because the items in this dataset are full concerts, each concert consisting of several pieces. Whilst Ma may be included in the melody of some of the pieces, Pa and Sa are always present. As a result, the pitch histogram for the complete concert does not contain a prominent Ma peak, meaning that it is highly unlikely for it to be selected as the tonic.

4.3.2 Errors when metadata is provided

We examine how the errors are affected once we allow methods to use gender and performance metadata, shown in Figure 12. If we compare the results to those of Figure 11, we see that Ma and Pa errors are reduced more than “Other” errors. By restricting the tonic frequency range to a single octave we prevent the appearance of a high Sa peak, thus avoiding the possible confusion between fourths and fifths explained earlier and reducing the amount of

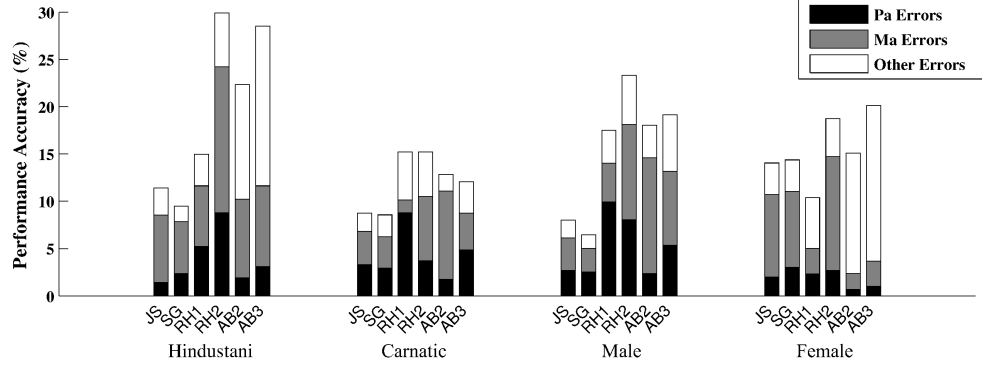


Figure 13: Percentage of excerpts with the different categories of errors (Pa, Ma and Others) for every method as a function of different excerpt attributes (Hindustani, Carnatic, male, female).

Pa and Ma errors.

For RH1 and RH2 the percentage of Ma errors actually increases slightly after including male/female information. A large proportion of these errors were observed in excerpts with female singers. For these excerpts, the range for *ṣadja* candidates is limited to 130-250 Hz. For this range, candidates fitting a lower Ma-middle Sa-middle Ma template would also satisfy the minimization criteria used in RH2. In the case of RH1, the reduced frequency range results in relatively weak peaks also being considered, and their small pitch variance can result in the wrong candidate being selected during the minimization process.

4.3.3 Errors as a function of excerpt attributes

Finally, we analyze the errors as a function of the different attributes of the excerpts (Hindustani versus Carnatic, male versus female). As in Section 4.1, we use the CM2 dataset for the analysis because it is the most balanced, and methods are not provided with any metadata in addition to the audio signal. The percentage of excerpts containing each of the three categories of errors (Pa, Ma and Other) for every approach as a function of the different excerpt attributes is shown in Figure 13. We see that for the classification based methods, the proportion of Ma errors is much higher in performances by female singers compared to performances by male singers. The pitch range of the tonic for female singers is such that the lower Ma resides in the frequency range where the tonic of most male singers lies. Thus, the lower Ma-middle Sa (fifth) relationship for female singers is often confused with middle Sa-Pa relationship for male singers, resulting in a high number of Ma errors.

For further details and insights regarding the types of error made by the different methods and their underlying causes we refer the reader to the publications where these methods are discussed in depth (Salamon et al., 2012; Gulati, 2012; Ranjani et al., 2011; Bellur et al., 2012).

5 Conclusions and future work

In this paper we presented a comparative study of the existing approaches for tonic identification in Indian art music. We introduced the main concepts related to tonic of the lead performer and highlighted the importance of automatic tonic identification for the computational analysis of Indian art music. We presented a generic block diagram that summarizes the main processing blocks of all the approaches and described briefly each processing step in the context of different methods. The main differences between the approaches were highlighted at each step.

We evaluated seven published methods on six datasets and reported the results. The best accuracy for most of the datasets is approximately 90%. The only exception is the IISCB1 dataset, which contains audio files of low quality. The approaches based on multi-pitch analysis were found to be more successful and consistent across datasets. However, when using information regarding the gender of the singer and performance type, the predominant pitch based approaches perform on par with multi-pitch based approaches. We also observed that the predominant pitch based approaches are sensitive to the length of the audio files and their performance increases with the duration of the audio. We found that a big percentage of the total errors made by most methods are Ma and Pa errors. These errors arise due to confusion between pitch distances that can be interpreted as Sa-Pa or Ma-Sa (fifths) and distances that can be interpreted as Sa-Ma or Pa-higher Sa (fourths).

Future work and efforts could be directed towards using rāg information to reduce the amount of Ma and Pa errors. Another aspect which can be explored for tonic identification is that of melodic motifs. Specifically, characteristic phrases (Pakads) could be useful for tonic identification, as they have a fixed relationship with the tonic pitch for a given rāg. Finally, the rhythm pattern of the tānpūrā could also be studied as a cue for the tonic, as the middle octave Sa svar is played twice in every cycle whilst the Pa, Ma or Nī svar are only played once.

References

- Bagchee, S. (1998). *Nad: understanding raga music*. Business Publications Inc.
- Bellur, A., Ishwar, V., Serra, X., & Murthy, H. (2012). A knowledge based signal processing approach to tonic identification in Indian classical music. In *2nd CompMusic Workshop* (pp. 113–118).
- Bishop, C. M. (2006). *Pattern recognition and machine learning (Information science and statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings 17* (pp. 97–110).
- Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glott Int.*, 5(9/10), 341–345.

- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... Serra, X. (2013). Essentia: an audio analysis library for music information retrieval. In *Proc. of int. society for music information retrieval conf. (ismir)* (pp. 493–498).
- Bor, J., Delvoye, F. N., Harvey, J., & Nijenhuis, E. T. (Eds.). (2010). *Hindustani music: Thirteenth to twentieth centuries* (First ed.). New Delhi: Manohar Publishers and Distributors.
- Bozkurt, B., Yarman, O., Karaosmanoğlu, M. K., & Akkoç, C. (2009). Weighing diverse theoretical models on Turkish Maqam music against pitch measurements: A comparison of peaks automatically derived from frequency histograms with proposed scale tones. *Journal of New Music Research*, 38(1), 45–70.
- Bregman, A. (1990). *Auditory scene analysis*. Cambridge, Massachusetts: MIT Press.
- Chew, E. (2002). The spiral array: An algorithm for determining key boundaries. In *Music and artificial intelligence* (pp. 18–31). Springer.
- Chordia, P., & Rae, A. (2007). Raag recognition using pitch-class and pitch-class dyad distributions. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 431–436).
- Clayton, M. R. L. (2000). *Time in Indian music: rhythm, metre, and form in North Indian rag performance*. Oxford University Press.
- Danielou, A. (2010). *The ragas of Northern Indian music*. New Delhi: Munshiram Manoharlal Publishers.
- Datta, A. K. (1996). Generation of musical notations from song using state-phase for pitch detection algorithm. *Journal of Acoustical Society of India*, 24.
- De Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917–1930.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, series B*, 39(1), 1–38.
- Deva, B. C. (1980). *The music of India: A scientific study*. Delhi: Munshiram Manoharlal Publishers.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd edition ed.). Wiley, New York.
- Flanagan, J. L., & Golden, R. M. (1966). Phase vocoder. *Bell Systems Technical Journal*, 45, 1493–1509.
- Gómez, E., & Herrera, P. (2004). Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*.
- Gulati, S. (2012). *A tonic identification approach for Indian art music*. Master's thesis, Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.
- Gulati, S., Salamon, J., & Serra, X. (2012). A two-stage approach for tonic identification in Indian art music. In *2nd CompMusic Workshop* (pp. 119–127).
- Hall, M. (1999). *Correlation-based feature selection for machine learning*. Phd thesis, University of Waikato.

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.
- Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Prentice Hall PTR.
- Koduri, G. K., Gulati, S., Rao, P., & Serra, X. (2012). Rāga recognition based on pitch distribution methods. *Journal of New Music Research*, 41(4), 337–350.
- Koduri, G. K., Serrà, J., & Serra, X. (2012). Characterization of intonation in Carnatic music by parametrizing pitch histograms. In *Proc. of Int. Society for Music Information Retrieval Conf. (ISMIR)* (pp. 199–204).
- Krishnaswamy, A. (2003a). Application of pitch tracking to South Indian classical music. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 5, pp. 557–560).
- Krishnaswamy, A. (2003b). On the twelve basic intervals in South Indian classical music. In *Audio Engineering Society Convention 115*.
- Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, 89(4), 334.
- Manikandan, T. V. (2004). *Lakshana and laksya of Carnatic music: A quest*. New Delhi, India: Kanishka Publishers and Distributors.
- Narmada, M. (2001). *Indian music and sancharas in raagas*. Delhi: Somnath Dhall, Sanjay Prakashan.
- Peeters, G. (2006). Chroma-based estimation of musical key from audio-signal analysis. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 155–120).
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Raman, C. V. (1921). On some Indian stringed instruments. In *Indian Association for the Cultivation of Science* (Vol. 33, pp. 29–33).
- Ranjani, H. G., Arthi, S., & Sreenivas, T. V. (2011). Carnatic music analysis: Shadja, swara identification and raga verification in Alapana using stochastic models. *Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE Workshop*, 29–32.
- Ross, J. C., Vinutha, T. P., & Rao, P. (2012). Detecting melodic motifs from audio for Hindustani classical music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 193–198).
- Salamon, J., & Gómez, E. (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., & Bonada, J. (2011). Sinusoid extraction and salience function design for predominant melody estimation. In *Proc. of Int. Conf. on Digital Audio Effects (DAFx)* (pp. 73–80).

- Salamon, J., Gulati, S., & Serra, X. (2012). A multipitch approach to tonic identification in Indian classical music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 499–504).
- Sen, A. K. (2008). *Indian concept of rhythm* (Second ed.). New Delhi: Kanishka Publishers, Distributors.
- Sengupta, R., Dey, N., Nag, D., Datta, A. K., & Mukerjee, A. (2005). Automatic tonic (SA) detection algorithm in Indian classical vocal music. In *National Symposium on Acoustics* (pp. 1–5).
- Şentürk, S., Gulati, S., & Serra, X. (2013). Score informed tonic identification for Makam music of Turkey. In *Proc. of Int. Society for Music Information Retrieval Conf. (ISMIR)*.
- Serrà, J., Koduri, G. K., Miron, M., & Serra, X. (2011). Assessing the tuning of sung indian classical music. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 157–162).
- Serra, X. (2011). A multicultural approach to music information research. In *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)* (pp. 151–156).
- Singh, J. (1995). *Indian music* (First ed.). Munshiram Manoharlal Publishers Pvt Ltd.
- Vickers, E. (2001). Automatic long-term loudness and dynamics matching. In *Audio Engineering Society Convention 111*.
- Viswanathan, T., & Allen, M. H. (2004). *Music in South India*. Oxford University Press.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining : practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.