Intonation analysis of rāgas in Carnatic music

Gopala Krishna Koduri[a], Vignesh Ishwar[b], Joan Serrà[c], Xavier Serra[a], Hema Murthy[b]

[a]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.*
[b]*Department of Computer Science and Engineering, Indian Institute of Technology - Madras, India*
[c]*Instituto de Investigación en Inteligencia Artificial, Consejo Superior de Investigaciones Científicas, Bellaterra, Spain.*

Abstract

Intonation is a fundamental music concept that has a special relevance in Indian art music. It is characteristic of a rāga and key to the musical expression of the artist. Describing intonation is of importance to several music information retrieval tasks such as developing similarity measures based on rāgas and artists. In this paper, we first assess rāga intonation qualitatively by analyzing varṇaṁs, a particular form of Carnatic music compositions. We then approach the task of automatically obtaining a compact representation of the intonation of a recording from its pitch track. We propose two approaches based on the parametrization of pitch-value distributions: performance pitch histograms, and context-based svara distributions obtained by categorizing pitch contours based on the melodic context. We evaluate both approaches on a large Carnatic music collection and discuss their merits and limitations. We finally go through different kinds of contextual information that can be obtained to further improve the two approaches.

Keywords: Music Information Research, Carnatic Music, Histogram parametrization, Pitch analysis

1. Introduction

1.1. Carnatic music and basic melodic concepts

The Indian subcontinent has two prominent art music traditions: Carnatic music in south India, and Hindustani music in north India, Pakistan and Bangladesh. Rāga is the melodic framework on which both art music traditions thrive (Narmada, 2001). The basic structures which make up a rāga are svaras, gamakas and phrases. Figure 1 shows melodic phrases obtained from a Carnatic performance with different types of gamakas labeled. Notice that the span of each of the gamakas extends from a few cents to several semitones.

There are seven svaras per octave: Sa, Ri, Ga, Ma, Pa, Da, Ni. Each svara has two/three variants, except for the tonic and the fifth: Sa, Ri_1, $\mathrm{Ri}_2/\mathrm{Ga}_1$, $\mathrm{Ri}_3/\mathrm{Ga}_2$, Ga_3, Ma_1, Ma_2, Pa, Da_1, $\mathrm{Da}_2/\mathrm{Ni}_1$, $\mathrm{Da}_3/\mathrm{Ni}_2$, Ni_3. These are termed svarastānas and are twelve in number. It is common to use the terms svara and svarastāna interchangeably when their distinction is not necessary. Svaras in a rāga are organized into ascending and descending progressions and have a specific function. Depending on their properties, and the distance from the neighboring svaras, a subset of them are sung with pitch modulations, viz. gamakas (for an elaborate discussion of svaras and gamakas, see Krishna & Ishwar, 2012). A rāga's identity is embodied in a set of phrases which encapsulate these properties.

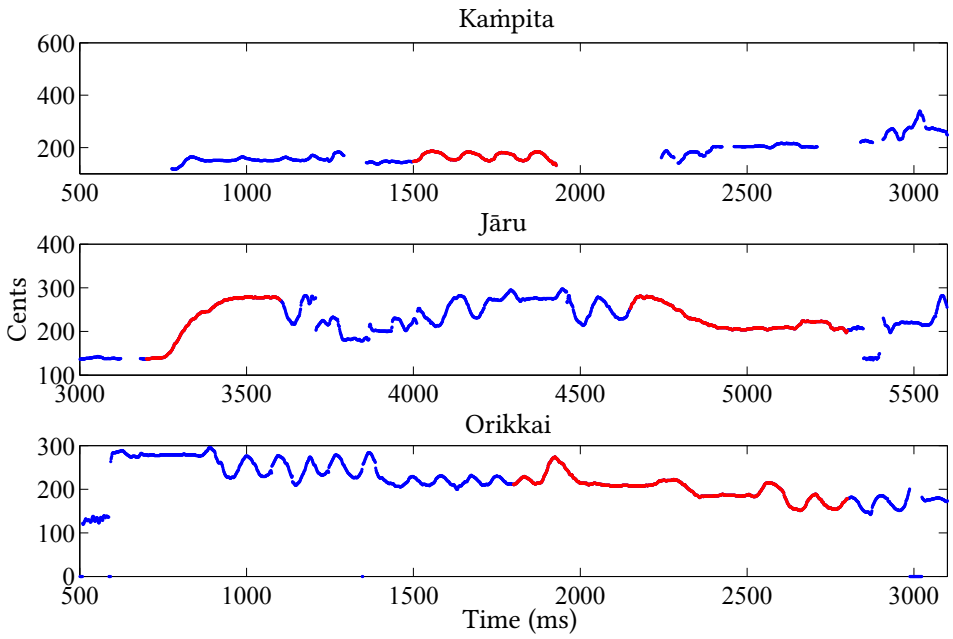


Figure 1: Melodic phrases from a Carnatic performance with three different types of gamakas labeled: Kampita, Jaaru and Orikkai.

A given svara can be sung steady in one rāga or with heavy melodic modulations in another rāga. Thus, because of the gamakas and differing roles of svaras, two rāgas can be different while having the exact same set of svaras. Given a melodic context, the way the pitch of a svara is interpreted in a performance is referred to as its intonation (Levy, 1982; Krishnaswamy, 2004). Therefore, it is evident that, to computationally understand and model Indian art music, intonation analysis is a fundamental step (Krishnaswamy, 2004). In this paper, we propose to parametrize pitch distributions in order to describe intonation in Carnatic music. To this extent, we present two approaches and evaluate them on a large Carnatic music collection.

1.2. Intonation

For computational purposes, we define intonation as the pitches and pitch modulations used by an artist in a given musical piece. From this definition, our approach will consider a performance of a piece as our unit of study. In Carnatic music practice, it is known that the intonation of a given svara varies significantly depending on the style of singing and the rāga (Swathi, 2009; Levy, 1982). In this paper, intonation refers to rāga intonation unless specified otherwise.

The study of svara intonation differs from that of tuning in its fundamental emphasis. Tuning refers to the discrete frequencies with which an instrument is tuned, thus it is more of a theoretical concept than intonation, in which we focus on the pitches used during a performance. The two concepts are basically the same when we study instruments that can only produce a fixed set of discrete frequencies, like the piano. On the other hand, given that in Indian art music there is basically no instrument with fixed frequencies (the harmonium is an important exception), tuning and intonation can also be considered practically the same. In the following discussion, we will maintain the terms, tuning or intonation, used by the different studies as they are intended.

Approaches to tuning analysis of real musical practice usually follow a so-called `stable region' approach, in which only stable frequency regions are considered for the analysis (cf. Serrà et al., 2011). However, it is known that most of a given performance in Carnatic music is gamaka-embellished (cf. Subramanian, 2007). Since gamakas are crucial to the identity of a rāga, the stable-region approach is not suitable as it undervalues the crucial information contained in gamaka-embellished portions of the recording. So far, tuning analysis has been employed to explain the interval positions of Carnatic music with one of the known tuning methods like just-intonation or equal-temperament (Serrà et al., 2011; Krishnaswamy, 2003). But considering that these intervals are prone to be influenced by factors like rāga, artist (Levy, 1982) and instrument (Krishnaswamy, 2003), computational analysis of svara intonation for different rāgas, artists and instruments has much more relevance to the Carnatic music tradition.

Krishnaswamy (2003) discusses various tuning studies in the context of Carnatic music, suggesting that it uses a hybrid tuning scheme based on simple frequency ratios plus various tuning systems, specially equal temperament. His work also points out the lack of empirical evidence for the same thus far. Recently, Serrà et al. (2011) have shown existence of important quantitative differences between the tuning systems in the current Carnatic and Hindustani music traditions. In particular, they show that Carnatic music follows a tuning system which is very close to just-intonation, whereas Hindustani music follows a tuning system which tends to be more equi-tempered.

Levy (1982) conducted a study with Hindustani music performances in which pitch consistency was shown to be highly dependent on the nature of gamaka usage. The svaras sung with gamakas were often found to have a greater variance within and across performances and different artists. Furthermore, the less dissonant svaras were also found to have greater variance. However, it was noted that across the performances of the same rāga by a given artist, this variance in intonation was minor. The same work concluded that the svaras used in the analyzed performances did not strictly adhere to either just-intonation or equal-tempered tuning systems. More recently, Swathi (2009) conducted a similar experiment with Carnatic music performances and draws similar conclusions about the variance in intonation. Belle et al. (2009) show the usefulness of intonation information of svaras to classify a limited set of five Hindustani rāgas. Noticeably, the approaches to intonation description employed in the studies conducted by Levy (1982) and Swathi (2009) cannot easily be scaled to a larger set of recordings due to the human involvement at several phases of the study, primarily in cleaning the data and the pitch tracks, and also in interpreting the observations.

1.3. Outline of the paper

To better understand the concept of intonation, we chose a particular form of compositions in Carnatic music, called varṇaṁs. In Section 2, we report our observations from a qualitative analysis of intonation in 7 rāgas using 28 varṇaṁs. In Sections 3 & 4, we discuss the two quantitative approaches we propose to automatically obtain an intonation description from a given audio recording, and report the results from their evaluation over a large Carnatic music collection. We conclude the paper in Section 5 with a discussion on a number of possible improvements to the two approaches.

2. Qualitative assessment of rāga intonation in varṇaṁs

Varṇaṁ[1] is a compositional form in Carnatic music. They are composed in different rāgas (melodic framework) and tālas (rhythmic framework). Though they are lyrical in nature, the fundamental emphasis

[1]This Sanskrit word literally means color, and varṇaṁs in Carnatic music are said to portray the colors of a rāga

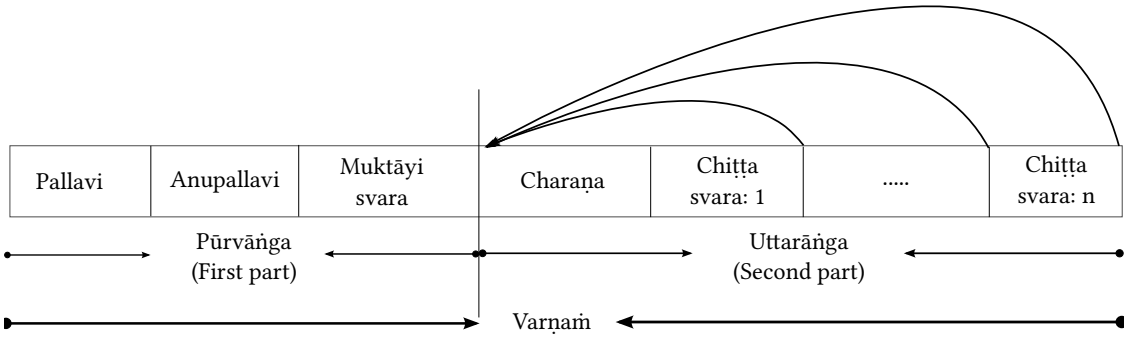


Figure 2: Structure of the varṇaṁ shown with different sections labeled. It progresses from left to right through each verse (shown in boxes). At the end of each chiṭṭa svara, charaṇa is repeated as shown by the arrows. Further, each of these verses is sung in two speeds.

lies in the complete exploration of the melodic nuances of the rāga in which it is composed. Hence, varṇaṁs are indispensable in an artist's repertoire of compositions. They are an invariable part of the Carnatic music curriculum, and help students to perceive the nuances of a rāga in its entirety. The coverage of the properties of svaras and gamakas covered in a varṇaṁ within a given rāga is exhaustive. This makes the varṇaṁs in a particular rāga a good source for many of the characteristic phrases of the rāga.

The macro structure of varṇaṁ has two parts: pūrvāṅga and uttarāṅga. The pūrvāṅga consists of the pallavi, anupallavi and muktāyi svara. The uttarāṅga consists of the charaṇa and the chiṭṭa svaras. Figure 2 shows the structure of the varṇaṁ with two parts and different sections labelled. A typical varṇaṁ performance begins with the singing of pūrvāṅga in two different speeds, followed by uttarāṅga, where in after each chiṭṭa svara, the singer comes back to charaṇa. Different variations to this macro structure give rise to various types of varṇaṁs: pada varṇaṁs, tāna varṇaṁs and dhāru varṇaṁs (Rao, 2006). Varṇaṁs are composed in a way such that the structure includes variations of all the improvisational aspects of Carnatic music (for an in-depth understanding of the relevance of varṇaṁs in Carnatic music, see Vedavalli, 2013a,b). For example, chiṭṭa svaras[2] are composed of svaras that capture all their possible combinations and structures in a given rāga. This helps singers in an improvisational form called *kalpana svaras*, where they permute and combine svaras as allowed by the rāga framework to create musically aesthetic phrases.

Due to the varṇaṁ structure, the rendition of varṇaṁs across musicians is fairly less variant than the variations seen in the renditions of other compositional forms. This is because most performances of the varṇaṁs deviate less from the given notation. Though the artists never use the notations in their actual performances, they have been maintained in the tradition as an aid to memory. This paper exploits this rigidity in structure of the varṇaṁ to align the notation with the melody and extract the pitch corresponding to the various svaras. Rāgas were chosen such that the 12 svarastānas in use in Carnatic music are covered (Serrà et al., 2011; Krishna & Ishwar, 2012). This would allow us to observe the impact of different melodic contexts (i.e., in different rāgas) on each of the svaras.

2.1. Music collection

For the aforementioned analysis, we recorded 28 varṇaṁs in 7 rāgas sung by 5 young professional singers who received training for more than 15 years. To make sure we have clean pitch contours for the analysis, all the varṇaṁs are recorded without accompanying instruments, except the drone. The structure

[2]Chiṭṭa svaras in Sanskrit literally mean the *svaras in the end.*

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

Table 1: Details of the varṇaṁ collection recorded for our analysis.

of varṇaṁ allows to attribute each part shown in Figure 2 to one/two tāla cycles depending on the speed. We take advantage of this information to semi-automate the synchronization of the notation and the pitch-contour of a given varṇaṁ. For that, we annotated all the recordings with tāla cycles. Also, in order to further minimize the manual intervention in using the annotations, all the varṇaṁs are chosen from the same tāla (adi tāla, the most popular one (Viswanathan & Allen, 2004)). Table. 1 gives the details of the varṇaṁ collection recorded for this analysis. This data is accessible online[3].

2.2. Svara synchronization and histogram computation

Our aim is to obtain all the pitch values corresponding to each svara, and analyze their distribution. The method consists of five steps: (1) The pitch contour of the recording is obtained (see Sec. 3.3). (2). Tāla cycles are manually annotated. (3) These tāla cycles are semi-automatically synchronized with the notation. (4). Pitch values corresponding to each svara are obtained from the pitch-contour. (5). A histogram from the pitch values of each svara is computed and interpreted (see Sec. 3).

We confine our analysis to a predetermined structure of the varṇaṁ in its sung form: pūrvāṇga in two speeds, followed by a verse-refrain pattern of charaṇa and chiṭṭa svaras, each in two speeds. Using Sonic Visualizer (Cannam et al., 2010), we marked the time instances which correspond to the start and end of tāla cycles which fall into this structure. A sequence of tāla cycles is generated from the notation such that they correspond to those obtained from the annotations. Hence, we now have the start and end time values for each tāla cycle (from annotations) and the svaras which are sung in that cycle (from notation).

Recall that we chose to analyze the varṇaṁs sung only in adi tāla. Each cycle in ādi tāla corresponds to 8 or 16 svaras depending on whether the cycle is sung in fast or medium speed. Each cycle obtained from annotations is split into appropriate number of equal segments to mark the time-stamps of individual svaras. The pitches for each svara are then obtained from the time locations in the pitch contour as given by these time-stamps. A histogram is then computed for each svara combining all its pitch-values.

2.3. Evaluation & results

Figure 3 shows pitch histograms for performances in two rāgas: Kalyāṇi and Śankarābharaṇaṁ. Even though they theoretically have all but one svara in common, the pitch histograms show that the peak locations and their characteristics are different. This implies that the rāgas cannot be differentiated by using just their svarastānas.

[3]URL to be included later.

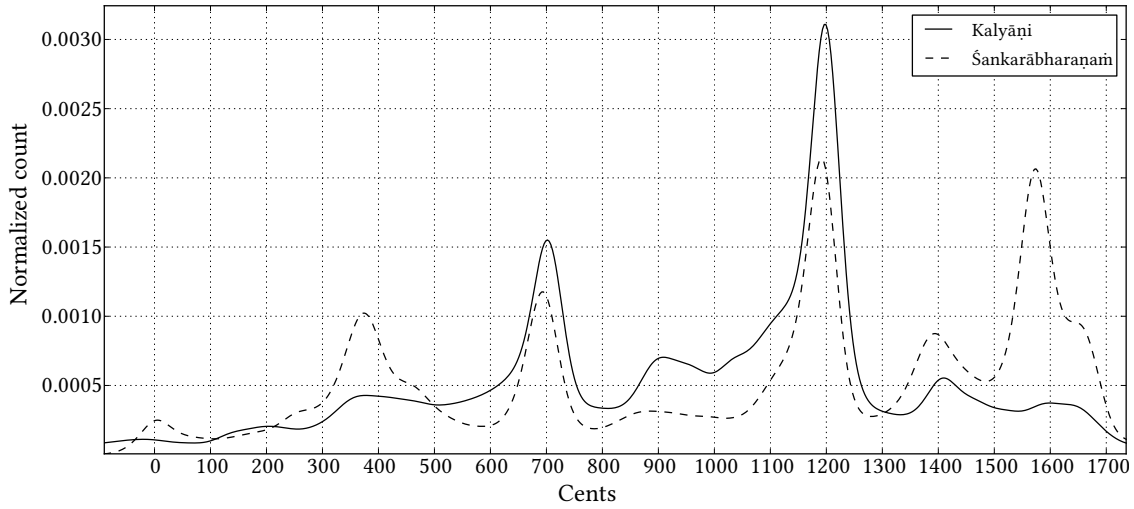


Figure 3: Histograms of pitch values obtained from recordings in two rāgas: Kalyāṇi and Śankarābharaṇaṁ. X-axis represents cent scale, normalized to tonic (Sa).

There are many such rāgas which have common svaras. However, their intonation is very different depending on the rāga's characteristics and context[4]. For instance, the svara Ga is common between the rāgas Mōhanaṁ and Bēgaḍa, but due to the context in which the Ga is sung in each of the rāgas, the intonation and the gamakas expressed on the svara change. Figure 4 shows that the svara Ga in Bēgaḍa corresponds to one sharp dominating peak at 400 cents. This concurs with the fact that the Ga in Bēgaḍa is always sung at its position with minimum gamakas. It is a steady note in the context of the rāga Bēgaḍa. On the other hand, Figure 4 shows that Ga in Mōhanaṁ corresponds to two peaks at 400 and 700 cents with a continuum from one peak to the other. The dominant peak is located at 400 cents (i.e., Ga's position). This is in line with the fact that Ga in Mōhanaṁ is rendered with an oscillation around its pitch position. The oscillation may vary depending on the context in which it is sung within the rāga. Ga in Mōhanaṁ, generally, starts at a svara higher (Ma or Pa) even though it may not be theoretically present in the rāga, and ends at its given position after oscillation between its own pitch and a higher pitch at which the movement started.

Another example of such svara is Ga in Ābhōgi and Śrī. Figure 4 shows that Ga in Ābhōgi is spread from 200 cents to 500 cents, with peaks at 200 cents and 500 cents respectively. These peak positions correspond to the svaras Ri and Ma, respectively. The inference one can make from this is that the Ga in Ābhōgi is sung as an oscillation between Ri and Ma of the rāga Ābhōgi, which is true in practice. The pitch histogram for Ga of Śrī in Figure 4 shows that the peak for Ga in Śrī is smeared with a peak at 200 cents which is the Ri in Śrī. This is consistent with the fact that Ga in Śrī is rendered very close to Ri. A comparison of the pitch histograms of the Ri in Śrī (Figure 5) and the Ga in Śrī shows that the peaks of Ga and Ri almost coincide and the distribution of the pitch is also very similar. This is because the movement of Ga in Śrī always starts at Ri, touches Ga and lands at Ri again. Ga in Śrī is always a part of any phrase that ends with RGR sequence of svaras, and in this context Ga is rendered as mentioned above.

Insights such as the ones we discussed in this Section require musical knowledge about the svaras and their presentation in the context of a rāga. To complement this, we have derived the transition matrices of svaras in each varṇaṁ from notations. The transition statistics of a given svara are observed to usually

[4]To be concise, we discuss only a few of them here.

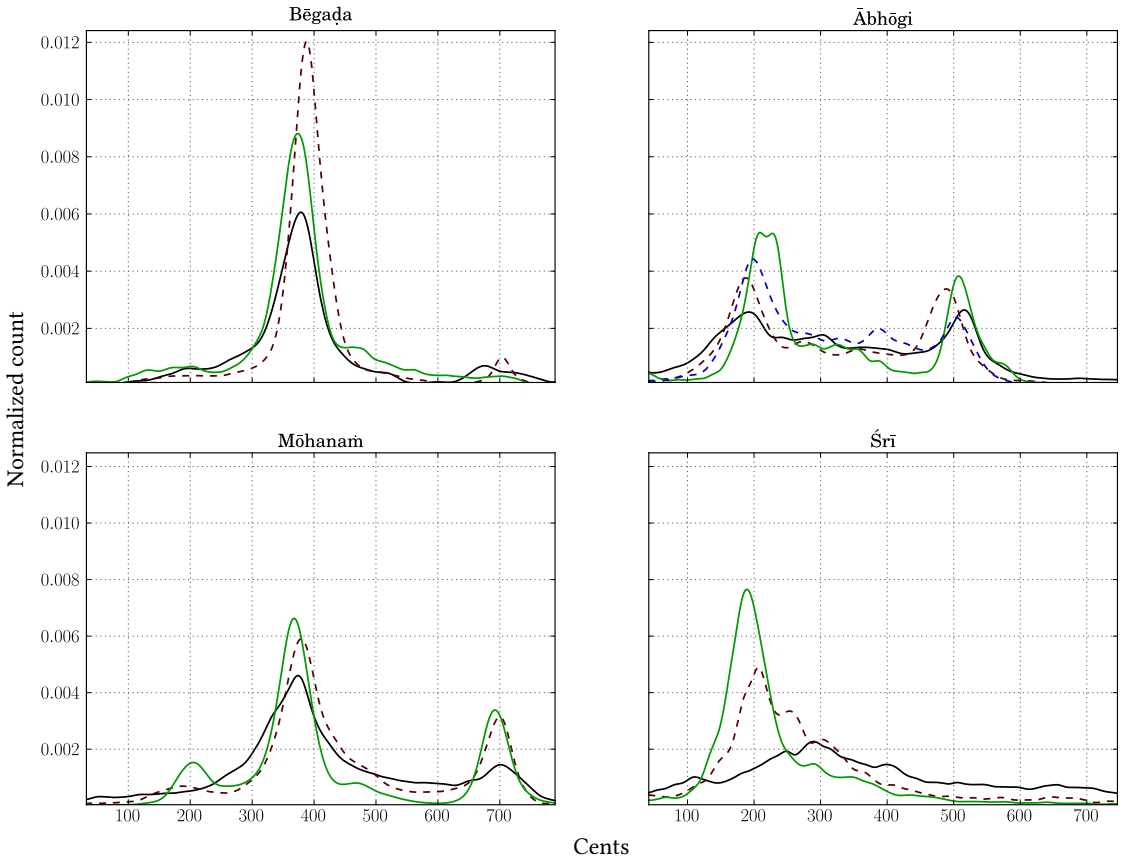


Figure 4: Pitch histograms of Ga svara in four rāgas: Bēgaḍa, Mōhanaṁ, Ābhōgi and Śrī. X-axis represents cent scale. Different lines in each plot correspond to different singers.

correspond to the pattern of peaks we see in its pitch histogram. Table. 2 lists the transitions involving Ga in Bēgaḍa, Mōhanaṁ, Ābhōgi and Ga, Ri in Śrī[5].

With the exception of Ga in Bēgaḍa, we notice that the other svaras to/from which the transitions occur are the ones which are manifest in the pitch histogram of the given svara. Combining this information with peak information in pitch histogram yields interesting observations. For instance, a svara such as Ga in Bēgaḍa rāga records a number of transitions with Ri and Ma svaras, but the pitch histogram shows a single peak. This clearly indicates that it is a svara sung steadily without many gamakas. On the other hand, in the case of svaras like Ga in Mōhanaṁ, we see that there are a number of transitions with Ri and Pa svaras, while there are also several peaks in the histogram. This is an indication that the svara is almost always sung with gamakas, and is anchored on other svara or sung as a modulation between two svaras. The transitions are also indicative of the usage of svaras in ascending/descending phrases. For instance, the transitions for Ga svara in Śrī rāga mark the limited context in which it is sung.

2.4. Conclusions

We chose varṇaṁs to analyse the differences in intonation of svaras in different rāgas. The observations clearly show that the intonation for the svara in different rāgas differs substantially depending on the

[5]Ga in Bēgaḍa and Mōhanaṁ correspond to a svarastāna which is different from the one that Ga in Ābhōgi and Śrī correspond to.

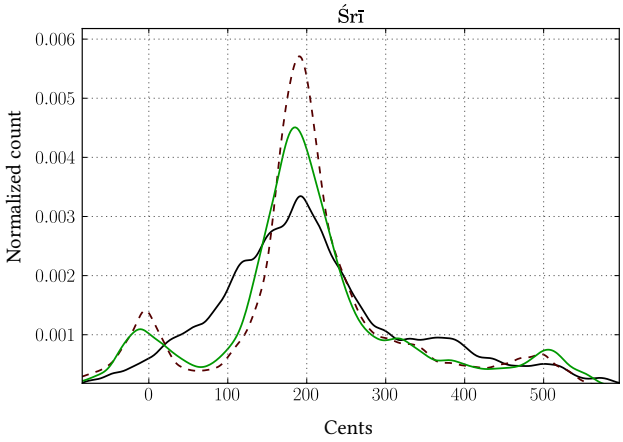


Figure 5: Pitch histogram for Ri svara in Śrī rāga. X-axis represents cent scale. Different lines in each plot correspond to different singers.

Svara (Rāga)	Sa	Ri	Ga	Ma	Pa	Da	Ni
Ga (Bēgaḍa)	0/14	74/56	-	80/64	0/18	2/0	0/4
Ga (Mōhanaṁ)	4/2	72/71	-	-	68/96	28/4	-
Ga (Ābhōgi)	24/0	44/68	-	55/58	-	2/0	-
Ga (Śrī)	0/2	88/88	-	0/0	0/0	0/0	2/0
Ri (Śrī)	106/132	-	88/88	52/46	6/6	0/0	26/6

Table 2: Transition statistics for svaras discussed in the section. Each cell gives the ratio of number of transitions made from the svara (corresponding to the row) to the number of transitions made to the svara.

melodic context established by the rāga. Therefore, it constitutes a crucial information in the identity of a rāga. In Sections. 3 and 4 we discuss the two approaches we propose to automatically obtain a description of svara intonation from an audio recording, and present the results of their evaluation using a large Carnatic music collection.

3. Histogram peak parametrization

3.1. Music collection

In the collection put together for qualitative analysis (sec. 2), the primary emphasis was on understanding intonation differences and not on assessing the intonation description thoroughly. Such a collection is insufficient to draw meaningful quantitative conclusions. Therefore, to evaluate the methods we propose in this section, we put together a music collection which is comprehensive and representative of existing commercial Carnatic music releases and live concerts. It is derived from CompMusic project's Carnatic music collection (Serra, 2012), by choosing only those rāgas for which there are at least 5 recordings. Table 3 shows the current size of the whole collection and the sub-collection we use for evaluation in this paper. Table 13 in Appendix A gives detailed statistics of the collection used for evaluation in this paper.

3.2. Segmentation

In a typical Carnatic ensemble, there is a lead vocalist who is accompanied by a violin, drone instrument(s), and percussion instruments with tonal characteristics (Raman, 1934). Based on the instruments

	Rāgas	Recordings	Duration (minutes)	Artists	Releases/Concerts
Collection used for evaluation	45	424	5617	38	62
CompMusic collection	180	1986	22805	65	180

Table 3: Statistics of the music collection used for evaluation in this paper, compared to the CompMusic collection.

being played, a given performance is usually a mix of one or more of these: vocal, violin and percussion. The drone instrument(s) is heard throughout the performance. The order and interspersion of these combinations depend on the musical forms and their organization in the performance. For different melodic and rhythmic analysis tasks, it is required to distinguish between these different types of segments. Therefore, it is necessary to have a segmentation procedure which can automatically do this.

In this study, we do not address the intonation variations due to artists. However, as we consider each recording as a unit for describing intonation, there is a need to assert the artist and rāga which characterize the intonation of the recording. For this reason, we have considered those recordings in which only one rāga is sung, which is the case for most of the recordings. Furthermore, we also distinguish between the segments where the lead artist exerts a dominant influence and the segments in which the accompanying violin is dominant. We choose the pitch values only of the former segments. In order to do this, we consider three broad classes to which the aforementioned segments belong to: vocal (all those where the vocalist is heard, irrespective of the audibility of other instruments), violin (only those where the vocalist is not heard and the violinist is heard) and percussion solo.

To train our segmentation algorithm to classify an audio excerpt into the three classes, we manually cropped 100 minutes of audio data for each class from commercially available recordings[6], taking care as to ensure diversity: different artists, male and female lead vocalists, clean, clipped and noisy data, and different recording environments (live/studio). The audio data is split into one-second fragments. There are few fragments which do not strictly fall into one of the three classes: fragments with just the drone sound, silence, etc. However, as they do not affect the intonation analysis as such, we did not consciously avoid them. This data is accessible online[7].

After manual segmentation we extract music descriptors. Mel-frequency cepstral coefficients (MFCCs) have long been used with a fair amount of success as timbral features in music classification tasks such as genre or instrument classification (Tzanetakis & Cook, 2002). Jiang et al. (2002) proposed octave based spectral contrast feature (OBSC) for music classification which is demonstrated to perform better than MFCCs in a few experiments with western popular music. Shape based spectral contrast descriptor (SBSC) proposed by Akkermans et al. (2009) is a modification of OBSC to improve accuracy and robustness by employing a different sub-band division scheme and an improved notion of contrast. We use both MFCC and SBSC descriptors, along with a few other spectral features that reflect timbral characteristics of an audio excerpt: harmonic spectral centroid, harmonic spectral deviation, harmonic spectral spread, pitch confidence, tristimulus, spectral rolloff, spectral strongpeak, spectral flux and spectral flatness (Tzanetakis & Cook, 2002).

A given audio excerpt is first split into fragments of length 1 second each. The sampling rate of all the audio recordings is 44100 Hz. Features are extracted for each fragment using a framesize of 2048 and

[6]These recordings are also derived from CompMusic collection, some of which also are part of the sub-collection we chose for evaluation.

[7]URL to be included later.

a hopsize of 1024 (double sided Hann window is used). The mean, covariance, kurtosis and skewness are computed over each 1-second fragment and stored as features. MFCC coefficients, 13 in number, are computed with a filterbank of 40 mel-spaced bands from 40 to 11000Hz (Slaney, 1998). The DC component is discarded, yielding a total of 12 coefficients. SBSC coefficients and magnitudes, 12 each in number, are computed with 6 sub-bands from 40 to 11000Hz. The boundaries of sub-bands used are 20 Hz, 324 Hz, 671 Hz, 1128 Hz, 1855 Hz, 3253 Hz and 11 kHz (see Akkermans et al., 2009). Harmonic spectral centroid (HSC), harmonic spectral spread (HSS) and harmonic spectral deviation (HSD) of the i^{th} frame are computed as described by Kim et al. (2006):

$$HSC_i = \frac{\sum_{h=1}^{N_H}(f_{h,i}A_{h,i})}{\sum_{h=1}^{N_H}A_{h,i}} \tag{1}$$

$$HSS_i = \frac{1}{HSC_i}\sqrt{\frac{\sum_{h=1}^{N_H}[(f_{h,i}-HSC_i)^2A_{h,i}^2]}{\sum_{h=1}^{N_H}A_{h,i}^2}} \tag{2}$$

$$HSD_i = \frac{\sum_{h=1}^{N_H}\left|\log_{10}A_{h,i} - \log_{10}SE_{h,i}\right|}{\sum_{h=1}^{N_H}\log_{10}A_{h,i}} \tag{3}$$

where $f_{h,i}$ and $A_{h,i}$ are the frequency and amplitude of h^{th} harmonic peak in the FFT of the i^{th} frame, and N_H is the number of harmonics taken into account, ordering them by frequency. For our purpose, the maximum number of harmonic peaks chosen was 50. $SE_{h,i}$ is the spectral envelope given by:

$$SE_{h,i} = \begin{cases} \frac{1}{2}(A_{h,i} + A_{h+1,i}) & \text{if } h = 1 \\ \frac{1}{3}(A_{h+1,i} + A_{h,i} + A_{h-1,i}) & \text{if } 2 \leq h \leq N_H - 1 \\ \frac{1}{2}(A_{h-1,i} + A_{h,i}) & \text{if } h = N_H \end{cases}$$

All the features thus obtained are normalized to the 0-1 interval. In order to observe how well each of these different descriptors perform in distinguishing the aforementioned three classes of audio segments, classification experiments are conducted with each of the four groups of features: MFCCs, SBSCs, harmonic spectral features and *other* spectral features. Furthermore, different classifiers are employed: naive Bayes, k-nearest neighbors, support vector machines, multilayer perceptron, logistic regression and random forest (Hall et al., 2009). As the smallest group has 12 features, the number of features in other groups is also limited to 12 using information gain feature selection algorithm (Hall et al., 2009). The classifiers are evaluated in a 3-fold cross validation setting in 10 runs. All three classes are balanced. Table 4 shows the average accuracies obtained.

MFCCs performed better than the other features, with the best result obtained using a k-NN classifier with 5 neighbors. The *other spectral features* and SBSCs also performed considerably well. Using paired t-test with a p-value of 0.05, none of the results obtained using harmonic spectral features were found to be statistically significant with respect to the baseline at 33% using zeroR classifier.

From among all features, we have selected 40 features through a combination of hand-picking and information-gain feature selection algorithm. These features come from 9 descriptors: MFCCs, SBSCs, harmonic spectral centroid, harmonic spectral spread, pitch confidence, spectral flatness, spectral rms, spectral strongpeak and tristimulus. The majority of these features are means and covariances of the nine descriptors. Table 4 shows results of classification experiments using all the features. In turn, k-NN classifier with 5 neighbors, performed significantly better than all the other classifiers.

	k-NN	Naive Bayes	Multilayer perceptron	Random Forest	SVM	Logistic regression
MFCCs	91.51	72.22	81.44	90.44	83.56	73.59
SBSCs	88.41	72.64	79.64	87.93	79.71	73.58
Harmonic spectral features	66.75	60.93	69.56	74.19	69.68	67.30
Other spectral features	87.45	70.22	84.56	89.15	84.0	80.79
All combined (40 features picked using feature-selection)	93.88	74.44	91.85	92.44	91.58	85.26
All combined (40 hand-picked features)	96.94	83.30	95.42	96.08	95.90	89. 42

Table 4: Accuracies obtained in classification experiments conducted with features obtained from four groups of descriptors using different classifiers.

3.3. F0 analysis

With the segmentation module in place, we minimize to a large extent the interference from accompanying instruments. However, there is a significant number of the obtained voice segments in which the violinist fills short pauses or in which the violin is present in the background, mimicking the vocalist very closely with a small time lag. This is one of the main problems we encountered when using pitch tracking algorithms like YIN (de Cheveigné & Kawahara, 2002), since the violin was also being tracked in quite a number of portions. To address this, we obtain the predominant melody using a multi-pitch analysis approach proposed by Salamon & Gomez (2012). In this approach, multiple pitch contours are obtained from the audio, which are further grouped based on auditory cues like pitch continuity and harmonicity. The contours which belong to the main melody are selected using heuristics obtained by studying features of melodic and non-melodic contours.

The frequencies are converted to cents and normalized with the tonic frequency obtained using the approach proposed by Gulati (2012). In Carnatic music, the lead artist chooses the tonic to be a frequency value which allows her/him to explore three octaves. The range of values chosen for tonic by the artist usually is confined to a narrow range and does not vary a lot. Hence, we take advantage of this fact to minimize the error in tonic estimation to a large extent, using a simple voting procedure. A histogram of the tonic values is obtained for each artist and the value which is nearest to the peak is obtained. This is considered to be the *correct* tonic value for the artist. The tonic values which are farther than 350 cents from this value are then set to the *correct* tonic value thus obtained. After all these preliminary steps are performed, we obtain the intonation description.

3.4. Method

From the observations made by Krishnaswamy (2003) and Subramanian (2007), it is apparent that steady svaras only tell us part of the story that goes with a given Carnatic music performance. However, the gamaka-embellished svaras pose a difficult challenge for automatic svara identification. Therefore, alternative means of deriving meaningful information about the intonation of svaras becomes important.

Gedik & Bozkurt (2010) present a detailed survey of histogram analysis in music information retrieval tasks, and also emphasize the usefulness of histogram analysis for tuning assessment and makam recogni-

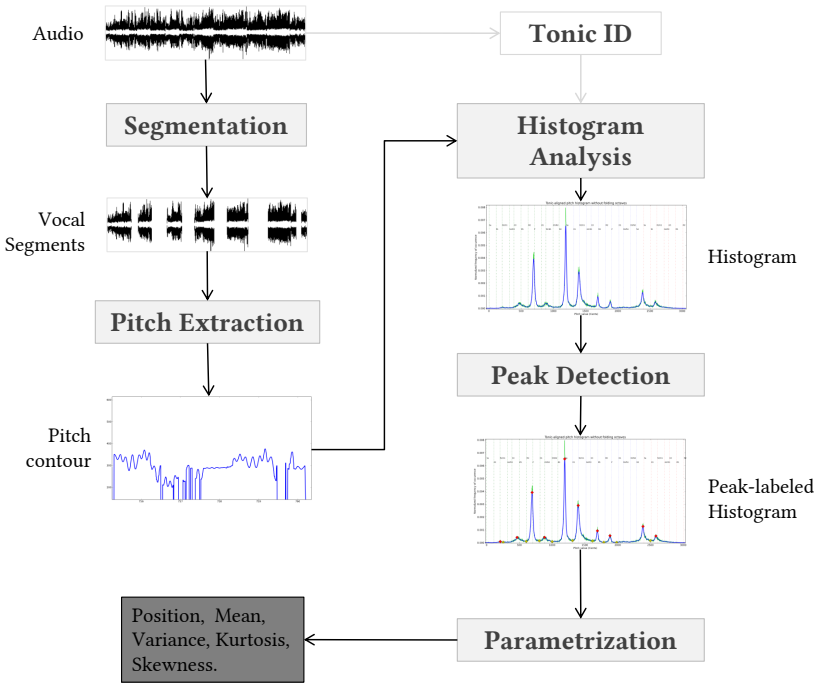


Figure 6: Block diagram showing the steps involved in Histogram peak parametrization method for intonation analysis.

tion in Makam music of Turkey. As the gamakas and the role of a svara are prone to influence the aggregate distribution of a svara in the pitch histogram of the given recording, we believe that this information can be derived by parametrizing the distribution around each svara (cf. Belle et al., 2009). Therefore, we propose an approach which is based on histogram peak parametrization that helps to describe the intonation of a given recording by characterizing the distribution of pitch values around each svara.

Our intonation description approach based on histogram peak parametrization involves five steps. In the first step, prominent vocal segments of each performance are extracted (Sec. 3.2). In the second step, the pitch corresponding to the voice is extracted using multipitch analysis (Sec. 3.3). In the third step, a pitch histogram for every performance is computed and its prominent peaks detected. In the fourth step, each peak is characterized by using the valley points and an empirical threshold. Finally, in the fifth step, the parameters that characterize each of the distributions are extracted. Figure 6 shows the steps in a block diagram. We now describe the last three steps.

As Bozkurt et al. (2009) point out, there is a trade-off in choosing the bin resolution of a pitch histogram. A high bin resolution keeps the precision high, but significantly affects the peak detection accuracy. However, unlike Turkish makam music, where the octave is divided into 53 Holdrian commas, Carnatic music uses roughly 12 svarastānas (Shankar, 1983). Hence, in this context, choosing a finer bin width is not as much a problem as it is in Turkish makam music. In addition, we employ a Gaussian kernel with a large standard deviation to smooth the histogram before peak detection. However, in order to retain the preciseness in estimating the parameters for each peak, we consider the values from the distribution of the peak before smoothing, which has the bin resolution as one cent. We compute the histogram H by placing the pitch values into their corresponding bins:

$$H_k = \sum_{n=1}^{N} q_k, \tag{4}$$

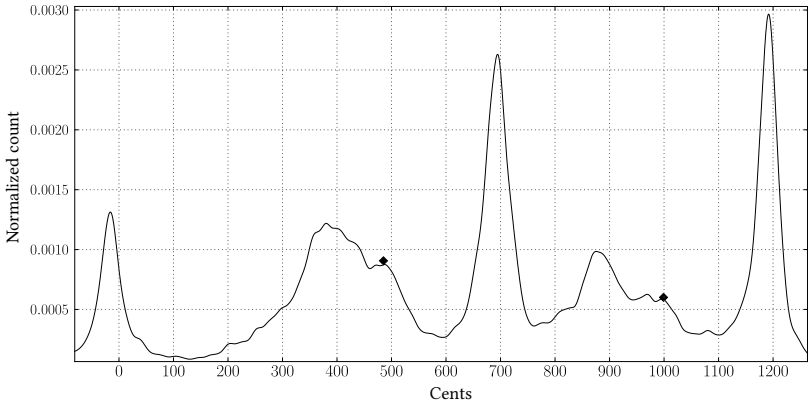


Figure 7: A sample histogram showing the peaks which are difficult to be identified using traditional peak detection algorithms. X-axis represents cent scale.

where H_k is the k-th bin count, N is the number of pitch values, $q_k = 1$ if $c_k \leq P(n) \leq c_{k+1}$ and $q_k = 0$ otherwise, P is the array of pitch values and (c_k, c_{k+1}) are the bounds on k-th bin.

Traditional peak detection algorithms can broadly be said to follow one of the three following approaches (Palshikar, 2009): (a) those which try to fit a known function to the data points, (b) those which match a known peak shape to the data points, and (c) those which find all local maximas and filter them. We choose to use the third approach owing to its simplicity.

The important step in such an approach is filtering the local maximas to retain the peaks we are interested in. Usually, they are processed using an amplitude threshold (Palshikar, 2009). However, following this approach, the peaks such as the ones marked in Figure 7 are not likely to be identified, unless we let the algorithm pick up a few spurious peaks. The cost of both spurious and undetected peaks in tasks such as intonation analysis is very high as it directly corresponds to the presence/absence of svaras.

To alleviate this issue, we propose two approaches to peak detection in pitch histograms which make use of few constraints to minimize this cost: peak amplitude threshold (A_T), valley[8] depth threshold (D_T) and intervallic constraint (I_C). Every peak should have a minimal amplitude of A_T, with a valley deeper than D_T on at least one side of it. Furthermore, only one peak is labelled per musical interval given by a predetermined window (I_C).

The first one of the peak detection approaches is based on the slope of the smoothed histogram. A given histogram is convolved with a Gaussian kernel to smooth out jitter. The length and standard deviation of the Gaussian kernel are set to 44 and 11 bins respectively. The length of the histogram is 3600 (corresponding to 3 octaves with 1 cent resolution). The local maximas and minimas are identified using slope information. The peaks are then found using D_T, and with an empirically set intervallic constraint, I_C. A local maxima is labelled as a peak only if it has valleys deeper than D_T on both sides, and it is also the maxima at least in the interval as defined by I_C.

The second one is an interval based approach, where the maximum value for every musical interval (I_C) is marked as a peak. The interval refers to one of the just-intonation or the equal temperament intervals. In the case of a just-intonation interval, the window size is determined as the range between the mean values obtained with the preceding and succeeding intervals. In the case of an equi-tempered interval, it is constant for all the intervals, which is input as a parameter. The window is positioned with the current

[8]Valley is to be understood as the deepest point between two peaks.

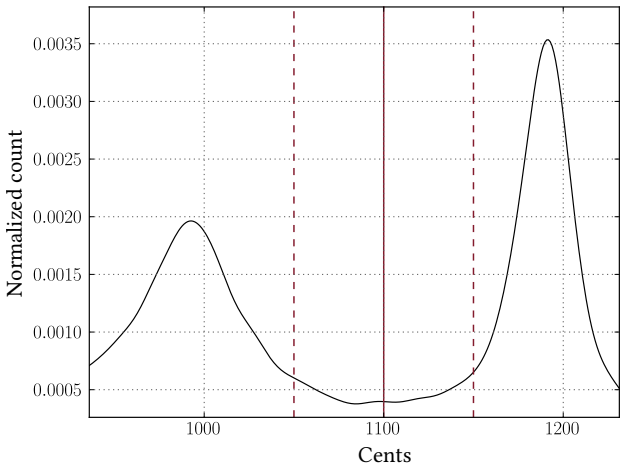


Figure 8: A semi-tone corresponding to 1100 cents is shown, which in reality does not have a peak. Yet the algorithm takes the point on either of the tails of the neighbouring peaks (at 1000 and 1200 cents) as the maxima, giving a false peak.

interval as its center. The peaks thus obtained are then subject to A_T and D_T constraints. In this approach, it is sufficient that a valley on either side of the peak is deeper than D_T.

Among all the points labelled as peaks, only a few correspond to the desired ones. Figure 8 shows three equi-tempered semi-tones at 1000, 1100 and 1200 cents. There are peaks only at 1000 and 1200 cents. However, as the algorithm picks the maximum value in a given window surrounding a semi-tone (window size is 100 cents in this case), it ends up picking a point on one of the tails of the neighbouring peaks. Therefore, we need a post-processing step to check if each peak is a genuine local maxima. This is done as follows: the window is split at the labelled peak position, and the number of points in the window that lie to both sides of it are noted. If the ratio between them is smaller than 0.15, there is a high chance that the peak lies on the tail of the window corresponding to a neighbouring interval[9]. Such peaks are discarded.

In order to evaluate the performance of each of these approaches, we have manually annotated 432 peaks in 32 histograms with pitch range limited from -1200 cents to 2400 cents. These histograms correspond to the audio recordings sampled from the dataset reported in Table A. As there are only a few parameters, we performed a limited grid search to locate the best combination of parameters for each approach using the given ground-truth. This is done using four different methods: one method from slope based approach (M_S), two methods from interval based approach corresponding to just-intonation (M_{JI}) and equi-tempered intervals (M_{ET}), and a hybrid approach (M_H) where the results of M_S and M_{JI} are combined. The intention of including M_H is to assess whether the two different approaches complement each other. The reason for selecting M_{JI} in the hybrid approach is explained later in this section.

Table 14 shows the ranges over which each parameter is varied when performing the grid search. For the search to be computationally feasible, the range of values for each parameter are limited based on the domain knowledge of the intervals and their locations, and empirical observations Shankar (1983); Serrà et al. (2011). A maximum F-measure value of 0.96 is obtained using M_H with A_T, D_T and I_C set to $5.0 \cdot 10^{-5}$, $3.0 \cdot 10^{-5}$ and 100 respectively. In order to further understand the effect of each parameter on peak detection, we vary one parameter at a time keeping the values for the other parameters as obtained in the optimum case. Figure 14 in Appendix B shows the impact of varying different parameters on different methods.

[9]This value is empirically chosen.

The kernel size for Gaussian filter was also evaluated, giving optimal results when set to 11. Higher and lower values are observed to have poor impact on the results. In the case of the window size, the larger it is, the better has been the performance of M_H and M_S. We suppose it is because the large window sizes handle deviations from the theoretical intervals with more success. Unlike equi-tempered intervals, just-intonation intervals are heterogeneous. Hence, a constant window has not been used. In M_{ET}, there does not seem to be a meaningful pattern in the impact produced by varying the window size. From Figure 14, we observe that D_T and A_T produce an optimum result when they are set to $5.0 \cdot 10^{-5}$, $3.0 \cdot 10^{-5}$ respectively. Further increasing their values results in the exclusion of many valid peaks.

As Serrà et al. (2011) have shown, Carnatic music intervals align more with just-intonation intervals than the equi-tempered ones. Therefore, it is expected that the system achieves higher accuracies when intervals and I_C are decided using just-intonation tuning. This is evident from the results in Figure 14. This is also the reason why we chose M_{JI} over M_{ET} to be part of M_H. Serrà et al. (2011) also show that there are certain intervals which are far from the corresponding just-intonation intervals. As slope-based approach does not assume any tuning method to locate the peaks, in the cases where the peak deviates from theoretical intervals (just intonation or equi-tempered), it performs better than interval-based approach. In the interval based approach, the peak positions are presumed to be around predetermined intervals. As a result, if a peak is off the given interval, it will be split between two windows with the maximums located at extreme position in each of them, and hence are discarded in the post-processing step described earlier. This is unlike the slope based approach, where the local maximas are first located using slope information, and I_C is applied later. The results from Figure 14 emphasize the advantage of a slope-based approach over an interval-based approach.

On the other hand, the interval based approach performs better when the peak has a deep valley only on one side of the peak. As a result, methods from the two approaches complement each other. Hence, M_H performs better than any other method. Therefore we choose this approach to locate peaks from pitch histograms. Most peaks are detected by both M_{JI}/M_{ET} and M_S. For such peaks, we preferred to keep the peak locations obtained using M_S.

In order to parametrize a given peak in the performance, it needs to be a bounded distribution. We observe that usually two adjacent peaks are at least 80 cents apart. The valley point between the peaks becomes a reasonable bound if the next peak is close by. But in cases where they are not, we have used a 50 cent bound to limit the distribution. The peak is then characterized by six parameters: peak location, amplitude, mean, variance, skewness and kurtosis. We extract parameters for peaks in three octaves. Each peak corresponds to a svarastāna. For those svarastānas which do not have a corresponding peak in the pitch histogram of the recording, we set the parameters to zero. Since for each octave there are 12 svarastānas, the total number of features of a given recording is 216 (3 octaves $\times$ 12 svarastānas $\times$ 6 parameters).

3.5. Evaluation & results

Intonation is a fundamental characteristic of rāga. Therefore, automatic rāga classification is a plausible way to evaluate computational descriptions of intonation. The two parameters from histogram analysis that have been used for rāga classification task in the literature are position and amplitude of the peaks (for a survey of rāga recognition approaches, see Koduri et al., 2012). We devise an evaluation strategy that tests whether the new parameters we propose are useful, and also if they are complementary and/or preferred to the ones used in the literature.

The evaluation strategy consists of two tasks: feature selection and classification. The feature selection task verifies if the new parameters are preferred to the features from position and amplitude parameters. In this task, we pool in the features from all the parameters and let the information gain measure and

	Position		Amplitude		Mean		Variance		Skewness		Kurtosis	
	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.
Information gain	0.9	0.7	1.4	0.8	1.2	0.8	0.4	0.4	0.8	0.6	0.4	0.4
SVM	1.7	0.9	0.9	0.6	1.1	0.7	0.5	0.4	0.4	0.3	0.5	0.4

Table 5: Results of feature selection on three-class combinations of all the rāgas in our music collection, using information gain and support vector machines. Ratio of total number of occurrences (abbreviated as Occ.) and ratio of number of recordings in which features from a given parameter are chosen at least once (abbreviated as Rec.), to the total number of runs are shown for each parameter. Note that there can be as many features from a parameter as there are number of svaras for a given recording. Hence, the maximum value of Occ. ratio is 5 (corresponding to 5 features selected per recording), while that of Rec. ratio is 1.

Features/Classifier	Naive Bayes	3-Nearest Neighbours	SVM	Random forest	Logistic regression	Multilayer Perceptron
Position and Amplitude	79.13	78.52	68.91	81.26	78.65	78.75
All features	78.26	78.46	71.79	81.16	78.61	78.78

Table 6: Averages of accuracies obtained using different classifiers in the two rāga classification experiments, using all the rāgas. The baseline calculated using zeroR classifier lies at 0.33 in both experiments.

support vector machine feature selection algorithms pick the top n features among them (Hall et al., 2009; Witten & Frank, 2005). We then analyze how often features from each of the parameters get picked.

The rāga classification task allows us to check if the features from the new parameters bring in complementary information compared to the features from position and amplitude. For this, we divide this task into two subtasks: classification with features obtained from the position and amplitude parameters, and classification with features obtained from all the parameters (position, amplitude and new parameters: mean, variance, skewness and kurtosis). We compare the results of the two subtasks to check if the features from the new parameters we propose carry complementary information to distinguish rāgas.

To ensure that the comparison of results in the two subtasks is fair, we use top n features in each subtask picked by information gain algorithm in feature selection task. Furthermore, six different classifiers were used: naive Bayes, k-nearest neighbours, support vector machines, logistic regression, multilayer perceptron and random forest (Hall et al., 2009; Witten & Frank, 2005)[10], and the accuracies obtained for each of them are checked if they stabilize after a few runs of the experiment.

As the number of classes is large (Table 13), it is hard to explain why the selected features are preferred over others: which classes do they distinguish and why. To address this issue, we perform numerous classification experiments each of which has 3 classes. As ${}^{45}C_3$ is a huge number, for the sake of computational feasibility, we listed all the possible combinations and picked 800 of them in a random manner. Each such combination is further sub-sampled thrice so that all the classes represented in that combination have equal number of instances, which is 5 as it is the minimum number of instances in a class in our music collection. As the total number of instances in each case is 15, we limit the number of features picked by the feature selection algorithms to 5.

Table 5 shows the statistics of outcomes of the two feature selection algorithms. For each parameter, two ratios are shown. The first one, abbreviated as Occ., is the ratio of total number of occurrences of the parameter to the total number of runs. The second one, abbreviated as Rec., is the ratio of number of

[10]The implementations provided in Weka were used with default parameters.

	Position		Amplitude		Mean		Variance		Skewness		Kurtosis	
	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.
Information gain	0.9	0.7	1.3	0.8	0.8	0.7	0.6	0.5	0.7	0.6	0.7	0.5
SVM	1.2	0.8	1.0	0.7	1.0	0.8	0.7	0.5	0.4	0.3	0.7	0.6

Table 7: Results of feature selection on sub-sampled sets of recordings in ${}^nC_2$ combinations of allied rāgas using information gain and support vector machines. Ratio of total number of occurrences (abbreviated as Occ.) and ratio of number of recordings in which the parameter is chosen at least once (abbreviated as Rec.), to the total number of runs are shown for each parameter.

Features/Classifier	Naive Bayes	3-Nearest Neighbours	SVM	Random forest	Logistic regression	Multilayer Perceptron
Position and Amplitude	86.94	88.84	86.87	85.84	82.70	86.37
All features	87.66	89.28	87.67	85.93	83.69	87.75

Table 8: Accuracies obtained using different classifiers in the two rāga classification experiments, using just the allied rāga groups. The baseline calculated using zeroR classifier lies at 0.50 in both experiments.

recordings in which the parameter is chosen at least once, to the total number of runs. The former lets us know the overall relevance of the parameter, while the latter allows to know the percentage of recordings to which the relevance scales to. Clearly, the position and amplitude of a peak are the best discriminators of rāgas given the high values for both ratios. It is also an expected result given the success of histograms in rāga classification (Koduri et al., 2012). The mean of the peak is also equally preferred to the position and amplitude, by both the feature selection algorithms.

Mean, variance, skewness and kurtosis are chosen in nearly 40-50% of the runs. Recall that each recording has 216 features, with 36 features from each of the parameters. Therefore, in 40-50% of the runs, features from the new parameters (mean, variance skewness and kurtosis) are preferred despite the availability of features from position and amplitude. This shows that the new parameters carry important information for distinguishing rāgas, than the positions and amplitudes for few svaras.

The results from the rāga classification task help us to assess the complementariness of the features from new parameters. Table 6 shows the averages of all the results obtained using each classifier over all the sub-sampled combinations for the two subtasks (classification of rāgas using features from all parameters, and those of position and amplitude). There is only a marginal difference in the results of the two subtasks, with a noticeable exception in the case of results obtained using SVM where the features from new parameters seemed to make a difference.

There is a class of rāgas which share exactly the same set of svaras, but have different characteristics, called allied rāgas. These rāgas are of special interest as there is a chance for more ambiguity in the positions of svaras. This prompted us to report separately the results of the feature selection and rāga classification tasks described earlier, on 11 sets of allied rāgas which together have 332 recordings in 32 rāgas. For those allied rāga sets which have more than two rāgas per set (say n), we do the experiments for all ${}^nC_2$ combinations of the set.

Table 7 shows the statistics over the outcomes of feature selection algorithms. One noteworthy observation is that the relevance of variance and kurtosis parameters is more pronounced in the classification of the allied rāgas, compared to the classification of all the rāgas in general (ref. table 5). This is in line with our hypothesis owing to the special property of allied rāgas.

Table 8 shows the classification results. Unlike the results from table 6, there is a small but consistent

increase in the accuracies of classification using features from all the parameters, compared to the case of using features from just position and amplitude parameters.

3.6. Conclusions

We have proposed a histogram peak parametrization approach to describe intonation in Carnatic music and evaluated it qualitatively using two tasks. The new parameters proposed were shown to be useful in discriminating rāgas, especially allied rāgas. However, as observed in the general rāga classification task, the information contained in the new parameters obtained through this approach do not seem to be very complementary of the information given by position and amplitude parameters.

There are quite a few issues in this approach. Few svaras, by the nature of the role they play, will not be manifested as peaks at all. Rather, they will appear as a slide that cannot be identified by a peak detection algorithm. The histogram peak parametrization itself is an aggregate approach which completely discards the contextual information of pitches: mainly the melodic & temporal contexts. The melodic context of a pitch instance refers to the larger melodic movement of which a given pitch is part of. The temporal context refers to the properties of the modulation: a fast intra-svara movement, a slower inter-svara movement, a striding glide from one svara to another, etc. A pitch value gets the same treatment irrespective of where it occurs in pitch contour. Consider the following two scenarios: (i) a given svara being sung steadily for some time duration, and (ii) the same svara appearing in a quick transition between two neighboring svaras. Using histogram peak parametrization, it is not possible to handle them differently. But in reality, the first occurrence should be part of the given svara's distribution, and the second occurrence should belong to either of the neighboring svaras depending on which is more emphasized. The objective of the approach we propose in the following section is to handle such cases by incorporating the local melodic and temporal context of the given pitch value.

4. Context-based svara distributions

4.1. Method

In this approach, the pitches are distributed among the 12 svarastānas based on the context estimated from the pitch contour, taking into account the modulations surrounding a given pitch instance. The pitch contour is viewed as a collection of small segments. For each segment, we consider the mean values of a few windows containing the segment. The windows are positioned in time such that, in each subsequent hop, the segment moves from the end of the first window to the beginning of the last window. The mean values of such windows provide us with useful contextual information. Figure 9 shows the positions of windows for a given segment S_k. The pitch samples of the segment are marked to belong to the svara which is nearest to the median of the mean values of all the windows that contain the segment.

More specifically, we define a shifting window with its size set to t_w milliseconds and hop size set to t_h milliseconds. For a k^{th} hop on pitch contour P, k=0,1,...$\frac{N}{t_h}$, where N is the total number of samples of the pitch contour, we define segment (S_k) as:

$$S_k = P(t_w + (k-1)t_h : t_w + kt_h) \tag{5}$$

where S_k is a subset of pitch values of P as given by Eq. 5. Notice that the width of the segment is t_h milliseconds. The mean of each window that contains the segment is computed as:

$$\mu_k = \frac{1}{t_w} \sum_{i=kt_h}^{t_w+kt_h} P(i) \tag{6}$$

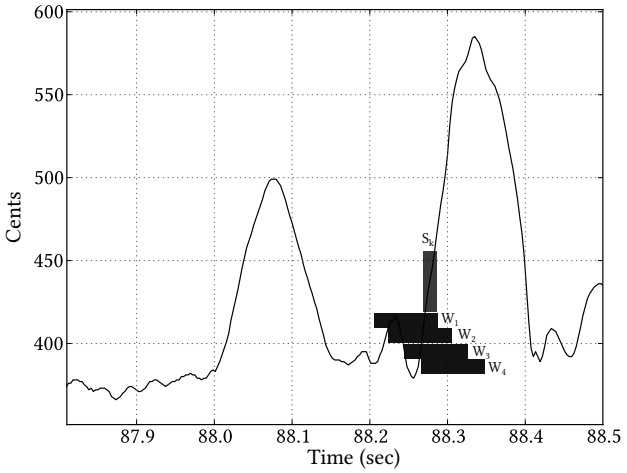


Figure 9: The positions of windows shown for a given segment S_k, which spans t_h milliseconds. In this case, the width of the window (t_w) is four times as long as the width of the segment (t_h), which is also the hop size of the window. X-axis represents time and y-axis represents cent scale.

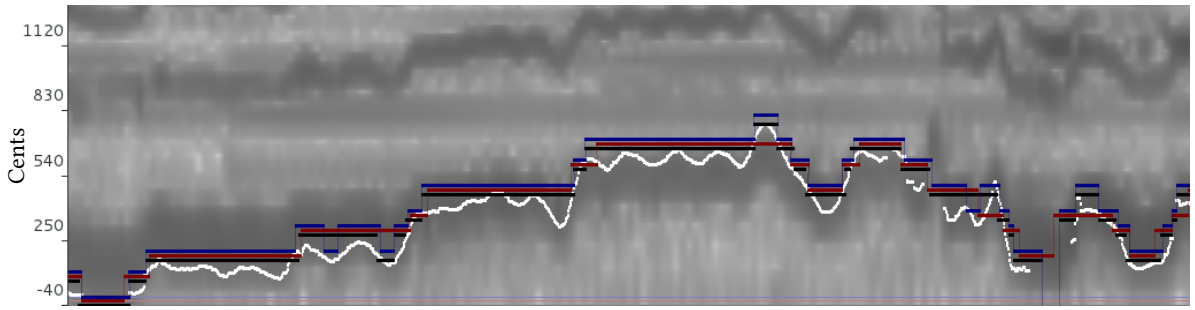


Figure 10: The pitch contour (white) is shown on top of the spectrogram of a short segment from a Carnatic vocal recording. The red (t_w = 150ms, t_h = 30ms), black (t_w = 100ms, t_h = 20ms) and blue (t_w = 90ms, t_h = 10ms) contours show the svara to which the corresponding pitches are binned to. The red and blue contours are shifted few cents up the y-axis for legibility.

The width of each window is t_w milliseconds. We now define ϵ, the total number of windows a given segment S_k can be part of, and $\bar{m}_k$, the median of the mean values of those ϵ windows as:

$$\epsilon = \frac{t_w}{t_h} \tag{7}$$

$$\bar{m}_k = \text{median}(\mu_k, \mu_{k+1}, \mu_{k+2} \ \cdots \ \mu_{k+\epsilon-1}) \tag{8}$$

Given Eqs. 5-8, a pitch-distribution $\mathbb{D}_I$ of a svara I is obtained as:

$$\mathbb{D}_I = \{S_k \ \mid \ \text{argmin}_i \ |\Gamma_i - \bar{m}_k| = I\} \tag{9}$$

where Γ is a predefined array of just-intonation intervals corresponding to four octaves. Therefore, $\mathbb{D}_I$ corresponds to the set of all those vocal pitch segments for which the median of the mean pitch in each of the windows containing that segment is closest to the predetermined just-tuned pitch (Γ_I) corresponding to svarastāna I. A histogram is computed for each $\mathbb{D}_I$, and the parameters are extracted as described in sec. 3. The key difference between the two approaches lies in the way parameters for each svara are obtained. In the earlier approach, we identify peaks corresponding to each svara from the aggregate histogram of the

recording. In this approach, we isolate the pitch values of each svara from the pitch contour and compute a histogram for each svara.

The crucial step in this approach is to determine t_w and t_h. A Carnatic music performance usually is sung in three speeds: lower, medium and higher (Viswanathan & Allen, 2004). A large part of it is in the middle speed. Also, singing in higher speed is more common than in the lower speed. From our analysis of varṇaṁs in Carnatic music, we observed the average duration each svara is sung in the middle speed to be around 200-250ms, while in the higher speed it is observed to be around 90-130ms.

Therefore, based on the choice of the window size (t_w), two different contexts arise. In the cases where the window size is less than 100ms (thus a context of 200ms for each segment), the span of the context more or less will be confined to one svara. Whereas in the other cases, the context spans more than one svara. In this paper, we explore the first case and defer the other to future work.

Hop size (t_h) decides the number of windows (ϵ) which a given segment in the pitch contour is part of. A higher value for ϵ is preferred as it provides more fine-grained contextual information about the segment S_k (See Eqs. 5 and 7). This helps to take a better decision in determining the svara distribution to which it belongs to. However, if ϵ is too high, it might be that either t_w is too high, or t_h is too low, both of which are not desired: a very high value for t_w will span multiple svaras which our method does not handle, and a very low value for t_h is not preferred as it implies more computations. Keeping this in mind, we empirically set t_w and t_h to 100ms and 20ms respectively. Figure 10 shows the results for $t_w = 150$ ms, $t_h = 30$ ms, $t_w = 100$ ms, $t_h = 20$ ms and $t_w = 90$ ms, $t_h = 10$ ms. In the figure, the intra-svara movements tend to be associated with the corresponding svara whereas the inter-svara movements are segmented and distributed appropriately.

Using this approach, we intend that the pitch segments be attributed to the appropriate svarastānas. However, the context might not be sufficient to do so. Hence we do not claim that the final distributions are representative of the actual intonation of svaras as intended by the artists. Yet, as we obtain the context for segments in every recording using the same principle, we believe there will be more intra-class correspondences than the inter-class ones.

Figure 11 shows the overview of the steps involved in this approach in a block diagram. Notice that this method alleviates the need for peak detection and finding the distribution bounds as we obtain each svara distribution independently (compare with Figure 6). These two steps which are part of histogram peak parametrization approach have their own limitations. The peak detection algorithm is prone to pick erroneous peaks and/or leave out few relevant ones. On the other hand, in order to estimate the parameters it is necessary to determine the bandwidth of peaks from the histogram. In the cases where the valley points of a peak are not so evident and the peak distribution overlapped with that of a neighboring svara, we chose a hard bound of 50 cents on either side of the peak. This affects the parameters computed for the distribution. Such issues do not arise with this approach as it does not require these two steps.

4.2. Evaluation & results

We run the same set of tasks as for histogram peak parametrization, but with the parameters obtained using context-based svara distributions. We will regard the results from the histogram peak parametrization as the baseline and compare with them. Tables 9 & 10 show the statistics over the outcome of feature selection on all rāgas and allied rāga groups respectively. Unlike the statistics from tables 5 and 7, the position parameter assumes a relatively lesser role in rāga discrimination, while amplitude still is the most discriminating parameter. With an exception of kurtosis, all the newly introduced parameters (mean, variance and skewness) also are chosen by the feature selection algorithms more frequently than before. This marks the relevance of melodic and temporal context of svaras for their intonation description, and is an indicator that the approach has been successful to, at least partially, obtain such context.

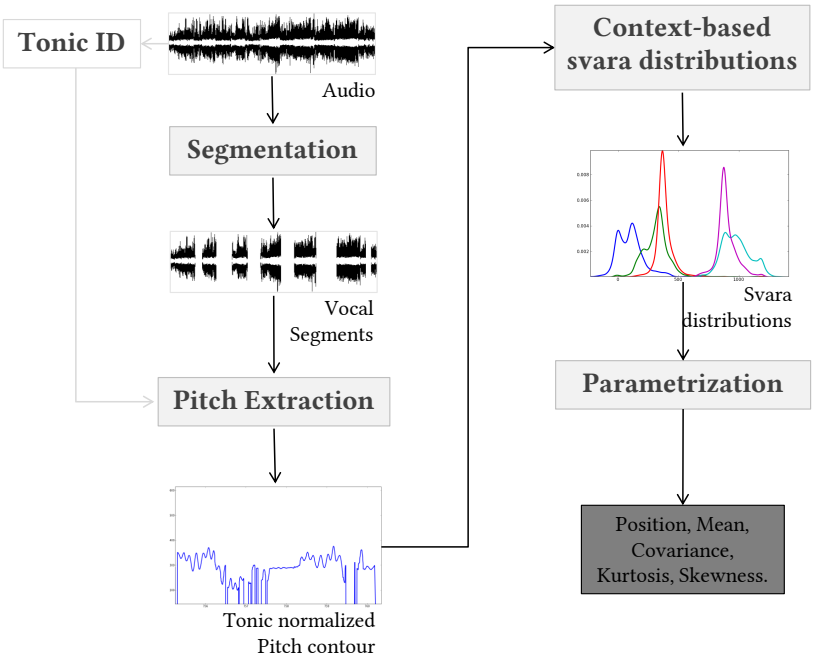


Figure 11: Block diagram showing the steps involved to derive context-based svara distributions for intonation analysis.

	Position		Amplitude		Mean		Variance		Skewness		Kurtosis	
	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.
Information gain	0.8	0.6	1.5	0.8	1.0	0.7	0.8	0.6	0.6	0.5	0.3	0.3
SVM	0.7	0.6	1.7	0.9	0.9	0.6	0.7	0.5	0.5	0.4	0.4	0.4

Table 9: Results of feature selection on sub-sampled sets of recordings in ${}^{n}C_{3}$ combinations of all rāgas using information gain and support vector machines. Ratio of total number of occurrences (abbreviated as Occ.) and ratio of number of recordings in which the parameter is chosen at least once (abbreviated as Rec.), to the total number of runs are shown for each parameter.

	Position		Amplitude		Mean		Variance		Skewness		Kurtosis	
	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.	Occ.	Rec.
Information gain	0.7	0.6	1.3	0.8	0.7	0.6	0.9	0.6	0.8	0.6	0.6	0.5
SVM	0.9	0.6	1.4	0.8	0.9	0.6	0.6	0.5	0.7	0.5	0.5	0.4

Table 10: Results of feature selection on sub-sampled sets of recordings in ${}^{n}C_{2}$ combinations of just the allied rāgas using information gain and support vector machines. Ratio of total number of occurrences (abbreviated as Occ.) and ratio of number of recordings in which the parameter is chosen at least once (abbreviated as Rec.), to the total number of runs are shown for each parameter.

Method/Classifier	Naive Bayes	3-Nearest Neighbours	SVM	Random forest	Logistic regression	Multilayer Perceptron
Histogram peak parametrization	78.26	78.46	71.79	81.16	78.61	78.78
Context-based svara distributions	82.63	82.83	79.90	82.69	81.11	82.17

Table 11: Accuracies obtained using different classifiers in the rāga classification experiment with all the rāgas using histogram peak parametrization, and context-based svara distributions. The baseline calculated using zeroR classifier lies at 0.33 in both experiments.

Method/Classifier	Naive Bayes	3-Nearest Neighbours	SVM	Random forest	Logistic regression	Multilayer Perceptron
Histogram peak parametrization	87.66	89.28	87.67	85.93	83.69	87.75
Context-based svara distributions	88.88	89.38	85.35	87.94	83.55	86.06

Table 12: Accuracies obtained using different classifiers in the rāga classification experiment with the allied rāga groups using histogram peak parametrization, and context-based svara distributions.. The baseline calculated using zeroR classifier lies at 0.50 in both experiments.

In order to assess if the context-based svara distributions bring in complementary information in the new parameters, we conducted the same set of rāga classification experiments as we have done for histogram peak parametrization. Tables 11 and 12 show the averages over all the results for classification experiments conducted over all the rāgas in our music collection, and the allied rāga groups respectively. The accuracies of each classifier stabilize after a few runs of the experiment. There is a notable improvement in the classification accuracies for the former, while the differences are marginal for the later. Recall that the svarasthānas are common among a group of allied rāgas. As a result, there is even more emphasis on differentiating between them using svara properties, which will shape the characteristics of a given peak. Hence, the difference between the performances of the two methods in the case of allied rāgas is not notable. Whereas, a notable improvement in the classification of all the rāgas using context-based svara distributions indicate that this approach has provided us with a better intonation description.

Figure 12 shows the pitch histograms of Ga svara in Ābhōgi, Bēgaḍa and Mōhanaṁ rāgas, obtained using the context-based svara distributions. We compare these with the corresponding pitch histograms in Figure 4. In the case of Ābhōgi rāga, our method is partially successful in showing two peaks (i.e., at 200 and 500 cents), resembling the peaks in the corresponding plot in Figure 4. However, this is not the case with Mōhanaṁ rāga where the pitch histograms obtained from our method failed to show peaks at 200 and 700 cents, while we still see a slight bump around 450 cents. For Bēgaḍa rāga, we observe a single peak in the histograms shown in both figures.

4.3. Conclusions

We have presented an approach to parametrize context-based svara distributions to obtain intonation description from Carnatic music recordings. In this approach, we attempted to address the major drawbacks of the histogram peak parametrization method (Sec 3): lack of melodic and temporal context, and

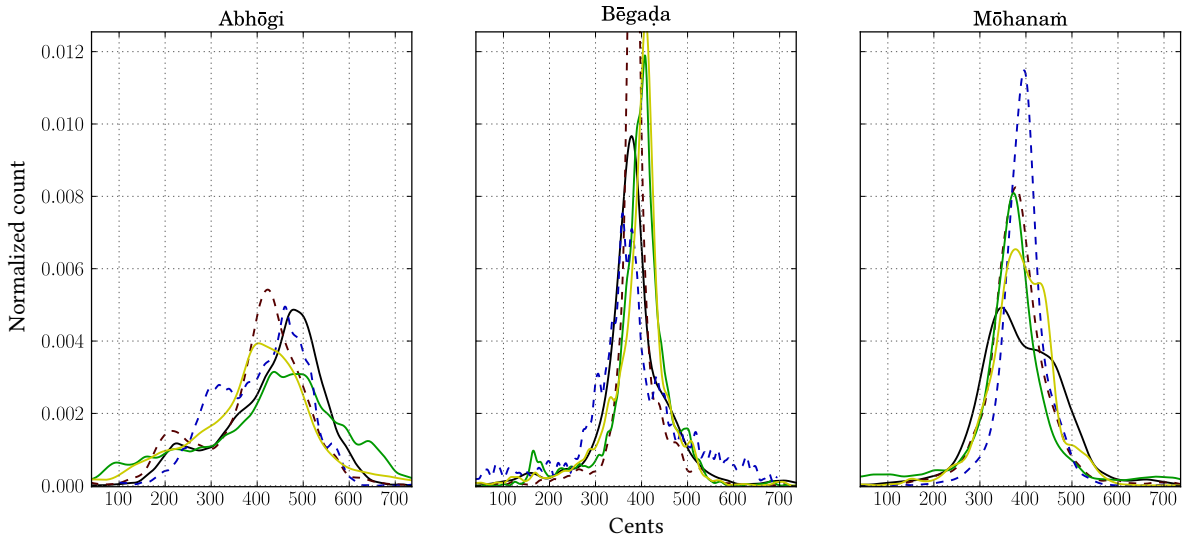


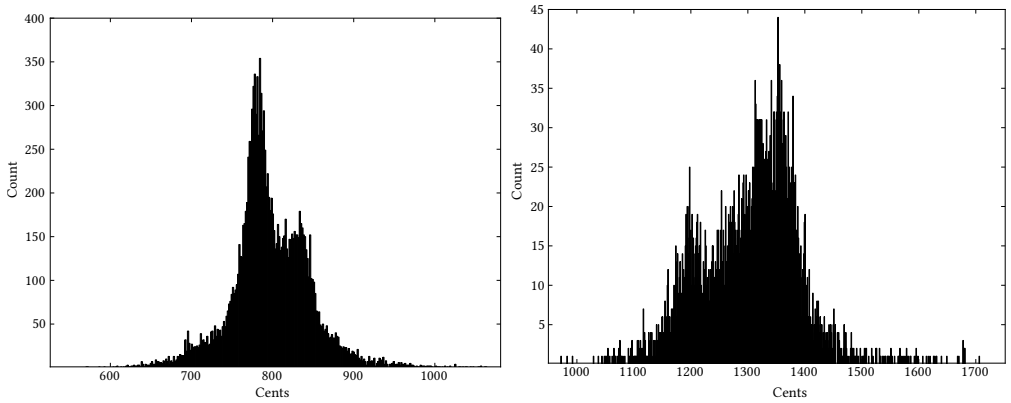
Figure 12: Pitch histograms of Ga svara in three rāgas: Ābhōgi, Bēgaḍa and Mōhanaṁ, obtained using context-based svara distributions. Different lines in each plot correspond to different singers.

finding the bandwidth of the peaks. Unlike the former approach where pitches for each svara are derived from a histogram, in this method, the pitches corresponding to each svara are obtained directly from the pitch contour by considering it's melodic and temporal context. For each svara, a histogram is then computed from which the parameters are obtained. Thus, it alleviates the need for estimating the location and bandwidth of peaks. As a result, this approach requires lesser number of parameters. The results from both the evaluation tasks show that this approach performs better compared to the earlier one, indicating that this approach provides a better intonation description.

5. Summary & Future work

Intonation is a very relevant musical concept that is fundamental for melodic analysis of Carnatic music. It is characteristic of rāgas, and also varies with artists. We have shown the qualitative differences between intonation of svaras across rāgas in a special compositional form called varṇaṁs. We have then proposed two approaches to automatically obtain intonation description from a given Carnatic recording. The results from both the approaches show that the new parameters contain useful information for discriminating rāgas. The results from the second approach also show that such information is complementary to the information contained by position and amplitude parameters.

There is a lot of scope to further improve parametrization of context-based svara distributions to obtain melodic and temporal context. At the moment, parameters of the overall distribution are the only source of information which we have taken into account. Consider the svara distributions shown in Figure 13. Though each of them has one dominant peak, they are also characterized by one or more other minor peaks. These peaks as such may or may not correspond to another svara, but they do signify a melodic context in which the svara occurs: a frequently co-occurring svara, a repeating melodic movement over this svara which involves another svara or vice-versa, etc. Though this information is partly contained by a few of the existing parameters such as skewness and kurtosis, it might greatly help to obtain such information more explicitly.



(a) Da svara in first octave (b) Ri svara in second octave

Figure 13: Sample svara distributions from a Carnatic vocal recording. X-axis is in cents, and Y-axis shows corresponding number of occurrences. Besides the dominant peak, such distributions are also characterized by the presence of other minor peaks, which can be important in characterizing the intonation of the svara.

In addition, the context of pitch values in the melodic contour is constrained by t_w and t_h, which are set to predefined values. However, as the performance includes multiple speeds, constant values to t_w and t_h are not an ideal choice. Furthermore, the speed variations that come into play between different artists turns the choice of t_w and t_h a very complex. Another possible method that would alleviate such constraints is to derive the context based on the characteristics of melodic movement which the given pitch instance is located in. For instance, the locations of nearest peak and valley on either side of the pitch instance and the corresponding slope are a window to understand the characteristics of the melodic movement. Such information over a range of pitches can be used to group them into a meaningful unit (such as a specific melodic shape/gamaka) based on their patterns. For instance, a large kampita[11] will result in a pattern of slopes, pitch differences which is different from that of a single glide surrounded by other modulations.

The histogram peak parametrization does not take into account the relative role/importance of svaras in a given rāga. The current parametrization seeks to model the svara without any reference to the other svaras. The relevance of such information is clearly observable from the results of feature selection algorithms (ref. tables 5, 7, 9 and 10), where the amplitude parameter dominates the rest. It is the relative information that the amplitudes carry (recall that the histograms are normalized), which makes it the most discriminating parameter. Therefore, tuning the methodologies to imbibe such information in other parameters helps in improving their span of context.

Both the approaches can further take advantage of more information which either is characteristic to the rāga or to the artist. An example for the first kind is the ascending and descending context of a svara. This context is determined by the ascending and descending progressions (usually referred to as ārōhaṇa and avarōhaṇa) of the given rāga, which more or less define the possible melodic movements, or rather prevent a few movements deeming them inappropriate for the melodic context of that particular rāga (for more details with an example, see Koduri et al., 2012). This is bound to impact the way svaras are sung: with or without gamakas, the extent of gamakas, svaras sung in a given gamaka and so on. Therefore,

[11]Kampita is one of the gamakas which means oscillation. It can mean anything between a vibrato on a svara and larger modulations involving different svaras.

obtaining separate distributions for each svara in its ascending and descending contexts might provide much more insight.

An example for the second kind is amplitude variations in the melody. This is one among probably a spectrum of other criteria which determine the style (usually referred to as bāṇi) of the performance. It also helps in understanding the perceptual importance of a given pitch instance. One direct way of incorporating such information in the parameters we compute would be a simple weighting of pitches with their amplitude.

Acknowledgments

We are very thankful to Ajay Srinivasamurthy for his help in annotating varṇaṁs. This research was partly funded by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). J.S. acknowledges 2009-SGR-1434 from Generalitat de Catalunya, ICT-2011-8-318770 from the European Commission, JAEDOC069/2010 from CSIC, and European Social Funds.

References

Akkermans, V., Serrà, J., & Herrera, P. (2009). Shape-based spectral contrast descriptor. In *Sound and Music Computing* July (pp. 23--25).

Belle, S., Joshi, R., & Rao, P. (2009). Raga Identification by using Swara Intonation. *Journal of ITC Sangeet Research Academy*, *Vol. 23*.

Bozkurt, B., Yarman, O., Karaosmanoğlu, M. K., & Akkoc, 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*, 45--70.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia 2010 International Conference* (pp. 1467--1468). Firenze, Italy.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*, 1917--1930.

Gedik, A., & Bozkurt, B. (2010). Pitch-frequency histogram-based music information retrieval for Turkish music. *Signal Processing*, *90*, 1049--1063.

Gulati, S. (2012). *A Tonic Identification Approach for Indian Art Music*. Masters' thesis Universitat Pompeu Fabra.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, *11*, 10--18.

Jiang, D.-N., Lu, L., Zhang, H.-j., Tao, J.-h., & Cai, L.-h. (2002). Music type classification by spectral contrast feature. In *IEEE International Conference on Multimedia and Expo* (pp. 113--116).

Kim, H., Moreau, N., & Sikora, T. (2006). *MPEG-7 audio and beyond: Audio content indexing and retrieval*. John Wiley & Sons.

Koduri, G. K., Gulati, S., Rao, P., & Serra, X. (2012). Rāga Recognition based on Pitch Distribution Methods. *Journal of New Music Research*, *41*, 337--350.

Krishna, T. M., & Ishwar, V. (2012). Karṇāṭik Music : Svara, Gamaka, Phraseology And Rāga Identity. In *2nd CompMusic Workshop* (pp. 12--18).

Krishnaswamy, A. (2003). On the twelve basic intervals in South Indian classical music. *Audio Engineering Society Convention*, (p. 5903).

Krishnaswamy, A. (2004). Inflexions and Microtonality in South Indian Classical Music. In *Frontiers of Research on Speech and Music*.

Levy, M. (1982). *Intonation in North Indian Music*. New Delhi: Biblia Implex Pvt. Ltd.

Narmada, M. (2001). *Indian Music and Sancharas in Raagas*. Delhi: Somnath Dhall, Sanjay Prakashan.

Palshikar, G. (2009). Simple algorithms for peak detection in time-series. In *International Conference on Advanced Data Analysis, Business Analytics and Intelligence* (pp. 1--13).

Raman, C. V. (1934). The Indian musical drums. *Journal of Mathematical Sciences, 1*, 179--188.

Rao, T. K. G. (2006). *Varṇasāgaraṁ*. Chennai: Ganamandir Publications.

Salamon, J., & Gomez, E. (2012). Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*, 1759--1770.

Serrà, J., Koduri, G. K., Miron, M., & Serra, X. (2011). Assessing the tuning of sung indian classical music. In *International Conference on Music Information Retrieval* (pp. 263--268).

Serra, X. (2012). Data gathering for a culture specific approach in MIR. In *Workshop on Advances in Music Information Research, World Wide Web Conference* (pp. 867--868).

Shankar, V. (1983). *The art and science of Carnatic music.* Chennai: Music Academy Madras.

Slaney, M. (1998). *Auditory toolbox*. Technical Report.

Subramanian, M. (2007). Carnatic Ragam Thodi – Pitch Analysis of Notes and Gamakams. *Journal of the Sangeet Natak Akademi*, *XLI*, 3--28.

Swathi, D. (2009). *Analysis of Carnatic Music : A Signal Processing Perspective.* Masters' thesis IIT Madras.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, *10*, 293--302.

Vedavalli, R. (2013a). Varnam - the mother of manodharma sangeetam (Part I). *Sruti*, (pp. 61--63).

Vedavalli, R. (2013b). Varnam - the mother of manodharma sangeetam (Part II). *Sruti*, (pp. 59--62).

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

Witten, I., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. *SIGSOFT Softw. Eng. Notes*, *36*, 51--52.

A. Details of music collection

Rāga	Recordings	Duration (minutes)	Artists	Releases/Concerts
Ābhōgi	5	104	4	5
Ānandabhairavi	10	85	10	10
Asāvēri	14	134	13	12
Aṭāṇa	6	42	6	6
Bēgaḍa	10	74	8	8
Behāg	8	47	7	8
Bhairavi	18	411	15	17
Bilahari	7	107	6	6
Dēvagāndhāri	6	42	6	6
Dēvamanōhari	5	53	4	4
Dhanāśri	7	8	2	2
Dhanyāsi	8	141	6	8
Dhīraśankarābharanaṁ	16	367	10	15
Haṁsadhvani	11	95	10	11
Hari kāṁbhōji	8	115	8	8
Hindōlaṁ	8	113	6	6
Jaunpuri	5	17	5	5
Kāpi	7	31	6	6
Kalyāṇi	14	303	11	13
Kamās	19	230	13	13
Kāmbhōji	11	265	10	11
Karaharapriya	9	195	8	8
Kēdāragauḷa	5	58	5	5
Madhyamāvati	10	100	10	9
Mānji	5	49	5	5
Mōhanaṁ	8	127	8	8
Mukhāri	8	81	8	8
Nagasvarāḷi	5	28	5	5
Nāṭakuranji	7	88	6	7
Pantuvarāḷi	17	257	16	15
Pūrvīkalyāṇi	9	177	7	9
Ranjani	5	74	4	4
Rītigaula	5	70	5	5
Sahāna	7	103	6	6
Sauraṣṭraṁ	40	44	9	9
Senchuruṭṭi	7	28	6	6
Ṣanmukhapriya	5	96	5	4
Śrīranjani	5	47	5	5
Śudhdha sāvēri	6	96	6	6
Sindhu bhairavi	6	28	5	5
Suraṭi	9	78	8	9
Tōḍi	27	841	19	21
Vāchaspati	5	46	1	1
Vasanta	6	42	5	5
Yadukula kāṁbhōji	5	54	5	5
45 rāgas	**424**	**5617**	**38**	**62**

Table 13: Detailed statistics of the music collection we use in evaluation for the two intonation description approaches.

B. Performance of peak-detection methods

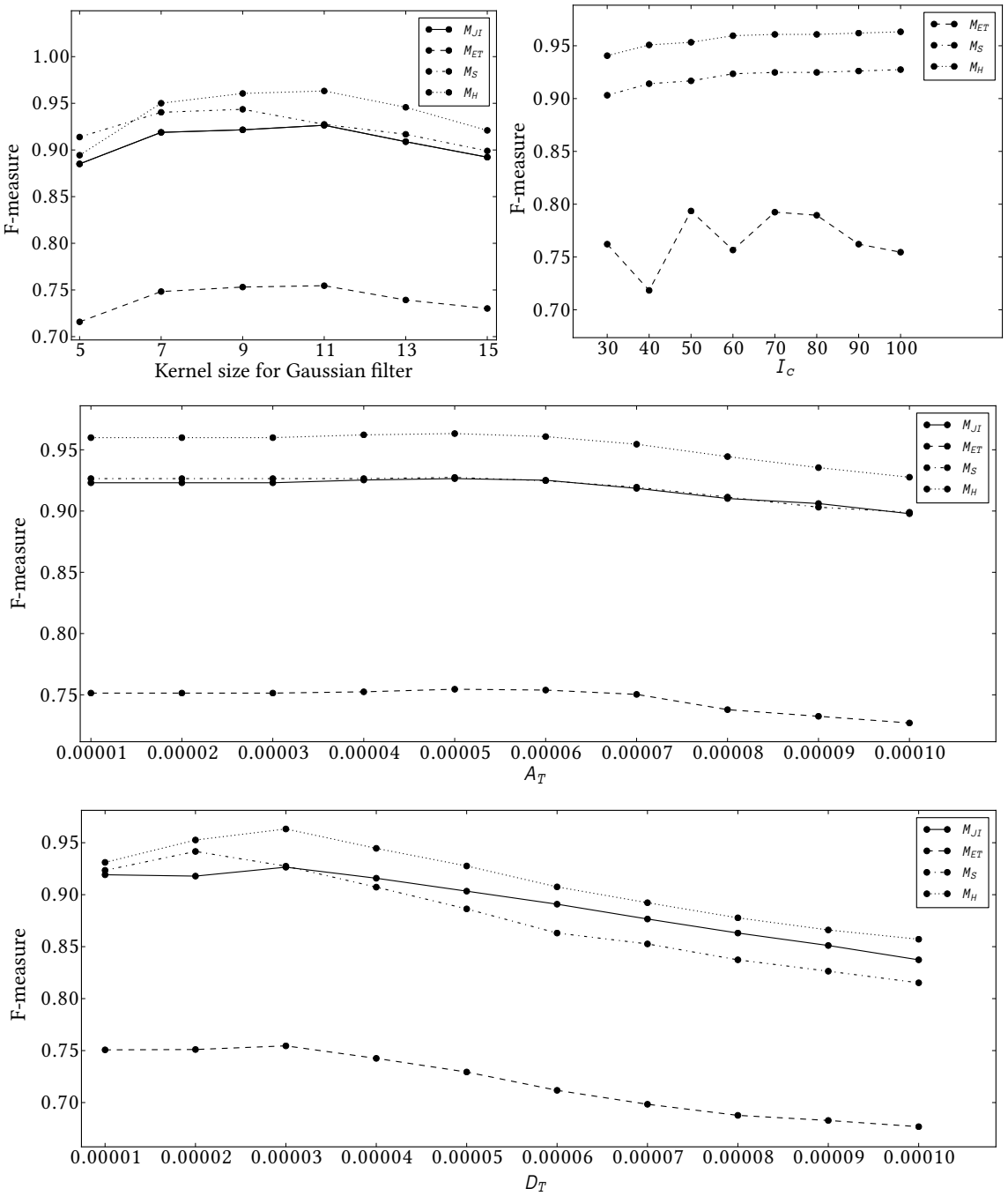


Figure 14: Impact of varying each parameter on the four peak detection methods. Y-axis indicates values of f-measure. and X-axis indicates label and corresponding values for each parameter.

Parameter	Range (step size)
Kernel size for Gaussian filter	5 to 15 (2)
Intervallic constraint (I_C)	30 to 100 cents (10)
Peak amplitude threshold (A_T)	$1.0 \cdot 10^{-5}$ to $1.0 \cdot 10^{-4}$ ($1.0 \cdot 10^{-5}$)
Valley depth threshold (D_T)	$1.0 \cdot 10^{-5}$ to $1.0 \cdot 10^{-4}$ ($1.0 \cdot 10^{-5}$)

Table 14: Range of values of each parameter over which grid search is performed to obtain the best combination of parameters.