A TWO-STAGE APPROACH FOR TONIC IDENTIFICATION IN INDIAN ART MUSIC

Sankalp Gulati, Justin Salamon and Xavier Serra
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
{sankalp.gulati,justin.salamon,xavier.serra}@upf.edu

ABSTRACT

In this paper we propose a new approach for tonic identification in Indian art music and present a proposal for a complete iterative system for the same. Our method splits the task of tonic pitch identification into two stages. In the first stage, which is applicable to both vocal and instrumental music, we perform a multi-pitch analysis of the audio signal to identify the tonic pitch-class. Multi-pitch analysis allows us to take advantage of the drone sound, which constantly reinforces the tonic. In the second stage we estimate the octave in which the tonic of the singer lies and is thus needed only for the vocal performances. We analyse the predominant melody sung by the lead performer in order to establish the tonic octave. Both stages are individually evaluated on a sizable music collection and are shown to obtain a good accuracy. We also discuss the types of errors made by the method.

Further, we present a proposal for a system that aims to incrementally utilize all the available data, both audio and metadata in order to identify the tonic pitch. It produces a tonic estimate and a confidence value, and is iterative in nature. At each iteration, more data is fed into the system until the confidence value for the identified tonic is above a defined threshold. Rather than obtain high overall accuracy for our complete database, ultimately our goal is to develop a system which obtains very high accuracy on a subset of the database with maximum confidence.

1. INTRODUCTION

Tonic is the foundation of melodic structures in both Hindustani and Carnatic music [1, 2]. It is the base pitch of a performer, carefully chosen in order to explore the full pitch range effectively in a given rāg [1] rendition. The tonic acts a reference and the foundation for the melodic integration throughout the performance [3]. That is, all the tones in the musical progression are constantly referred and related to the tonic pitch. All the accompanying instruments such as tablā [2], violin and tānpūrā [3] are tuned using the tonic of the lead performer.

In any performance of Indian art music (in both Hindustani and Carnatic), the tonic is the Sa (also referred as *Ṣaḍja*) *svar* [4] around which the whole rāg is built upon [5] [2, 4]. Other set of svaras used in the performance derive their meaning and purpose in relation to this reference and to the specific tonal context established by the given rāg [3]. Since, the entire performance is relative to the tonic, both the lead artist and the audience need to hear the tonic pitch throughout the concert. A constantly sounding drones instrument at the background of the performance reinforces the tonic pitch. In addition to the tonic pitch (Sa), the drone also produces other pitches like the fifth (Pa), the fourth (Ma) and sometimes the seventh (Nī) with respect to the tonic pitch, depending upon the chosen rāg. Typically the drone is produced by either the tānpūrā, electronic tānpūrā [6] or śruti box [7] for the case of vocal music and by the sympathetic strings of instruments such as sitār [8], sārangī [9] and vīṇā [10] for the case of instrumental performances. The drone acts as a reference of the music to a tonal background, reinforcing all the harmonic and melodic relationships.

The importance of the tonic in Indian art music means identifying the tonic pitch is crucial for many other types of computational tonal analyses such as such as intonation analysis [5, 6], melodic motivic analysis [7] and rāg recognition [8--10]. However, despite its importance in the computational analysis of Indian art music, the problem of automatic tonic identification is not correctly posed and has received very less attention from the research community.

Most of the previous approaches for tonic identification in Indian art music focus on the tonic pitch-class (Sa) identification and discard the octave information which might be useful for many analyses such as intonation analysis [11, 12]. They utilize only the predominant melody information present in the recording. Moreover, the melody extraction is performed using monophonic pitch trackers, even though the music material under consideration is heterophonic in nature. The databases used to evaluate these approaches are quite restricted: in [12] only the ālāp sections of solo vocal recordings are considered, and in [11] only sampūrṇ rāg recordings are considered. In [13], we

[1] http://en.wikipedia.org/wiki/Raga
[2] http://en.wikipedia.org/wiki/Tabla
[3] http://en.wikipedia.org/wiki/Tambura

[4] http://en.wikipedia.org/wiki/Swara
[5] with an exception of the madhyaṁ-śruti songs
[6] http://en.wikipedia.org/wiki/Electronic_tanpura
[7] http://en.wikipedia.org/wiki/Shruti_box
[8] http://en.wikipedia.org/wiki/Sitar
[9] http://en.wikipedia.org/wiki/Sarangi
[10] http://en.wikipedia.org/wiki/Saraswati_veena

proposed an approach that performs a multi-pitch analysis of the audio data in order to utilize the drone sound present in the background of the performance for identifying the tonic pitch. We advanced on some of the issues mentioned above such as identifying the tonic in correct octave and evaluating approach on a sizable database. However, this method works for the vocal performances, as it requires the tonic octave information for training, which is not available for instrumental excerpts.

In this paper, we propose a new method for tonic pitch identification in Indian art music, which divides this task into two stages; first, the tonic pitch-class identification, performed using a multi-pitch analysis and second, the tonic octave identification using the predominant melody information. This enables the method to be used for both vocal and instrumental performances, where the second stage is performed only for the vocal excerpts. The advantage of performing a multi-pitch analysis of the audio to identify the tonic pitch in Indian art music was shown in [13]. It is evident that accompanying instruments, especially tānpūrā provide an important cue for the identification of the tonic pitch. We use the same multi-pitch analysis which was used in [13] to identify the tonic pitch-class. While annotating the excerpts with the tonic pitch it was observed that the decision of the tonic octave is primarily based on the pitch range of the sung melody. This motivates us to analyse the predominant melody present in the vocal performances to identify the tonic octave. As the tonic octave for the instrumental music is not clearly defined as it is for the vocal excerpts, we aim at identifying only the tonic pitch-class for instrumental music [14].

In addition to the specific method, we also present a proposal for a complete system for labelling large databases of Hindustani and Carnatic music with the tonic pitch. The system aims to incrementally utilize all the available data, both audio and metadata to identify the tonic and also estimate a confidence measure for each output.

In Section 2 we describe both the stages of the proposed tonic identification method and Section 3 presents the proposal for a complete system for tonic identification in Hindustani and Carnatic music. In Section 4, we describe the evaluation strategy employed in this work, which includes the database used for the evaluation and annotation procedure followed to generate the ground truth. Subsequently in Section 5 we present and discuss the results of the evaluation. Finally, in Section 6 we provide conclusions and present possible direction for future work.

2. TONIC IDENTIFICATION METHOD

The proposed method divides the task of tonic pitch identification into two stages; tonic pitch-class (Sa) identification and tonic octave estimation as shown in Figure 1. For the instrumental performances only the first stage (S1) is used, whereas for the vocal performances both the stages (S1 and S2) are applied. Subsequent paragraphs describe the method in detail.

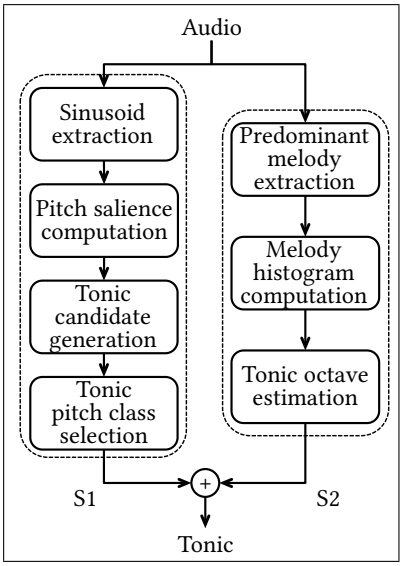

Figure 1. Block diagram of the proposed method. First stage (S1) performs tonic pitch-class identification and second stage (S2) performs tonic octave estimation.

2.1 Tonic Pitch Class Identification

The methodology used for tonic pitch-class identification in this paper is similar to the one used for tonic pitch identification in [13]. Both these methods differ at the candidate selection step, where in the current method we aim at identifying the tonic pitch-class candidate. The proposed method uses a multi-pitch representation of the audio signal to compute the pitch histograms, using which the tonic pitch-class is identified. Following a classification based approach the system automatically learns the best set of rules to select the peak of the histogram that represents the tonic pitch-class.

This stage is comprised of four main processing blocks; sinusoid extraction, pitch salience computation, candidate generation and candidate selection. The first two blocks used in this method, namely, sinusoid extraction and salience function computation (see S1 in Figure 1) are taken from the predominant melody extraction algorithm proposed by Salamon and Gómez in [15].

2.1.1 Sinusoid Extraction

In the first block of the method (S1 in Figure 1), we extract the sinusoidal components of the audio signal. This process is divided into three parts; spectral transform, spectral peak picking and sinusoid frequency and amplitude correction.

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

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

where $x(n)$ is the time domain signal, $w(n)$ the windowing function, l the frame number, M the window length, N

the FFT length and H the hop size. We use the Hamming windowing function with a window size of 46.4 ms, a hop size of 11.6 ms and a $\times 4$ zero padding factor, which for data sampled at $f_S = 44.1$ kHz gives $M = 2048$, $N = 8192$ and $H = 512$ [15].

Given the FFT of a single frame $X_l(k)$, spectral peaks p_i are selected by finding all the local maxima k_i of the magnitude spectrum $|X_l(k)|$. We also apply an energy threshold to discard the low-energy spurious spectral peaks (due to the side-lobes of the window). The energy threshold (T_s) is the calculated as follows:

$$\begin{aligned} T_s &= max(T_r, \alpha), \\ T_r &= E_m + \beta \end{aligned} \tag{2}$$

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

The frequency resolution in STFT is limited by the spectral resolution (number of FFT points), which for a low frequency sinusoid might result in a relatively large error in the estimation of the frequency. To improve the frequency and amplitude resolution of the sinusoids we apply a three-point parabolic interpolation, given by following equation:

$$\begin{aligned} f &= \frac{\alpha - \gamma}{2(\alpha - 2\beta + \gamma)}, \\ y &= \beta - \frac{1}{4}(\alpha - \gamma) f \end{aligned} \tag{3}$$

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

2.1.2 Pitch Salience Computation

The extracted sinusoids are used to compute a salience function, a time-frequency representation indicating the salience of different pitches over time. We use a salience function proposed by Salamon and Gómez in [16], which is based on harmonic summation similar to [17]. In short, the salience of a given frequency is computed as a weighted summation of energy found at all the integer multiples (harmonics) of that frequency. The peaks of the salience function at a given time instance represent the prominent pitches present in that frame. Note that though the two concepts, pitch (which is perceptual) and fundamental frequency (which is a physical measurement) are not identical, for simplicity we use these two terms interchangeably.

The constructed salience function spans a pitch range of 5 octaves, starting from 55 Hz to 1.76 kHz. The frequency values are quantized into a total of 600 bins on a cent scale, where each bin spans 10 cents. The mapping between a given frequency value f_i in Hz to its corresponding bin index $b(f_i)$ is given by:

$$b(f_i) = 1200 \frac{\log_2(f_i / f_r)}{\eta} + 1 \tag{4}$$

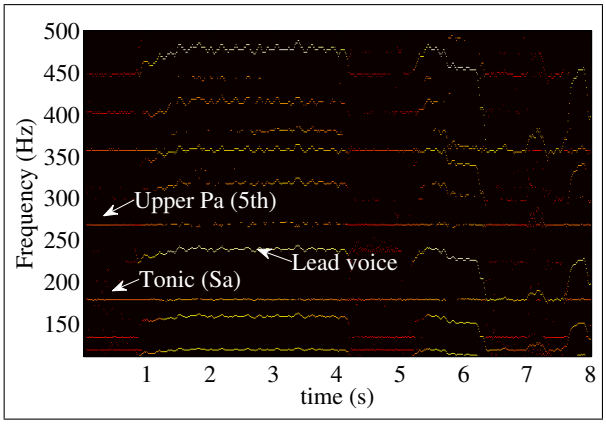

Figure 2. Peaks of the salience function computed for an excerpt in our database. Magnitude of a peak is in logarithmic scale (dB)

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

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

$$S(j) = \sum_{h=1}^{N_h} \sum_{i=1}^{N_p} g(j, h, \hat{f}_i) \cdot (\hat{a}_i)^{\beta} \tag{5}$$

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

Another critical component of the harmonic summation is the weighting function ($g(j, h, \hat{f}_i)$), which defines the weight given to a sinusoid when it is considered as the h^{th} harmonic of the bin j. We use the weighting scheme as follows:

$$g(j, h, \hat{f}_i) = \begin{cases} \cos^2(\delta \cdot \frac{\pi}{2}) \cdot \alpha^{h-1} & \text{if} |\delta| \leq 1 \\ 0 & \text{if} |\delta| > 1 \end{cases} \tag{6}$$

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

In Figure 2, we show the time evolution of the peaks of the salience function computed from an audio excerpt in our database. We notice that the tonic pitch-class (Sa) and fifth (Pa) played by the tānpūrā are clearly visible along with the peaks corresponding to the voice. However, the salience of

the pitch values corresponding to the voice is much higher than those corresponding to the tānpūrā sound.

2.1.3 Tonic Candidate Generation

The process of generating the tonic candidates includes three sub-tasks; detecting peaks of the salience function, computing a pitch histogram using these peaks and extracting candidates as the peaks of the pitch histogram.

We select the peaks of the salience function at each frame to compute a multi-pitch histogram. The peaks of the salience function represent the prominent pitches of the lead instrument, voice and other predominant accompanying instruments present in the audio recording at every point in time. Thus, a histogram computed using these pitch values represents the cumulative occurrences of different pitches at the level of the whole audio excerpt. Though the pitch histograms have been used previously for tonic identification [11], they were constructed using only the predominant melody. Therefore, in many cases the tonal information provided by the drone instrument is not taken into consideration.

We chose a lenient frequency range of 110-370 Hz to select the peaks from the salience function [13]. The selected peaks are used to construct a multi-pitch histogram. We notice that generally the lead voice/instrument is much louder than the drone sound (Figure 2). To normalize this bias towards the dominant source, we drop the saliences of the peaks and consider only their frequency of occurrence. This way a peak that corresponds to the voice has equal weight in the histogram compared to the peak corresponding to the drone.

The tonic pitch-class will not always be the highest peak of the pitch histogram. We therefore consider top 10 peaks of the histogram $p_i(i = 1 \ldots 10)$, one of which corresponds to the tonic pitch-class. We call them tonic pitch-class candidates and store both frequency and amplitude of each of these candidates for every audio excerpt.

2.1.4 Candidate Selection

The candidate which represents the tonic pitch-class is selected based on a template which is learned automatically using a classification based approach. We hypothesize that by learning the interrelationships between the salient candidates, the candidate representing the tonic pitch-class can be selected. This is motivated by the fact that the pitches used in a performance are in relation with the tonic pitch-class and the tānpūrā plays the tonic pitch-class in two octaves. For example, if the two most salient peaks of the pitch histogram are an octave apart, it is highly probable that they correspond to the tonic pitch-class as the drone plays the Sa in two different octaves (lower Sa and Higher Sa).

We compute the distance between every tonic candidate (p_i) and the most salient candidate in the histogram (p_1). This gives us a set of features $f_i(i = 1 \ldots 10)$ (pitch-interval features), where f_i is distance in semitone between p_i and p_1. Another set of features $a_i(i = 1 \ldots 10)$ (amplitude features) include the amplitude ratios of all the candidates with respect to the highest candidate.

We annotate each audio excerpt with a class label (as explained below) and use 20 features (f_i, a_i) to train a classifier in order to predict the class label. In this way the system automatically learns the best set of rules that maximise the class prediction. The strategy for labelling an instance with a class should be such that it allows us to uniquely associate the tonic pitch-class with it, given all the 10 candidates.

The class labels assigned to each instance in this method is the best rank of the tonic pitch-class amongst all the candidates. Note that we use the term `best' to highlight that we select the highest rank of all the candidates corresponding to the tonic pitch-class and since we considered a frequency range of more than one octave, we may have multiple peaks, representing the same pitch class but at different octaves. Theoretically it is a 10 class problem, as the allowed tonic pitch-class rank can go as low as tenth. But after analysing the training data we found that the lowest tonic pitch-class rank was fifth and hence only 5 classes are used in the experiment. Moreover, 98.7% of the instances are labelled with one of the top three classes (first, second, third).

Next, we proceed to select the relevant features for the task at hand. We use the WEKA data-mining software for all the classification related steps [18]. We perform attribute selection using the *CfsSubsetEval* attribute evaluator and BestFirst search method [19] with a 10-fold cross validation option set. We select the features which are used in at least 80% of the folds.

Subsequently, a C4.5 (J48) decision tree is trained using WEKA to learn best set of rules to reliably identify the correct tonic pitch-class candidate [20]. Note that we also tried other classifiers, namely, support vector machine (Sequential Minimal Optimization (SMO) with polynomial kernel) and an instance based classifier K* [21]. However the accuracy obtained by the J48 decision tree was considerably higher and so for the rest of the paper we present our results based on this classifier. Additionally, the advantage of using a decision tree is that the resulting classification rules can be easily interpreted and visualized.

We noticed that the number of instances belonging to each class in our training dataset was highly uneven, which might result into a biased learning, favoring the majority class. To mitigate this effect we also perform instance normalization by repeating the number of instances in minority class. We used the `supervised.instance.Resample' filter in WEKA with `biastoUniformClass' option set to 1 to normalize the number of instances per class [21].

The obtained decision tree is easily interpretable and has musically meaningful rules. For a detailed analysis of the decision tree, we refer to [13, 14].

2.2 Tonic Octave Estimation

In addition to identify the tonic pitch-class we also aim to estimate the octave in which the tonic of the lead performer lies. As the concept of the tonic octave is clearly defined for the vocal artists, we use this stage only for the vocal music performances. The pitch range for the majority of singers lies within three octaves, where the tonic chosen by them is the middle register Sa. The tonic is thus the

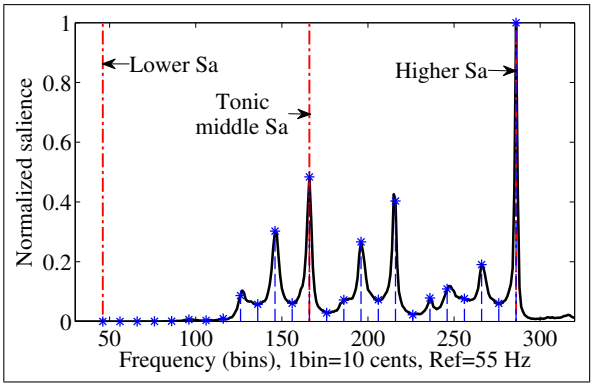

Figure 3. An example of the predominant melody histogram extracted from a song in our database. The red lines mark the tonic pitch-class locations

lowest Sa svar sung by the vocalist (with an exception of the madhyaṁ-śruti case, which is rarely witnessed). This motivates us to analyze the predominant melody contour in order to automatically estimate the tonic octave.

The process of estimating the tonic octave is divided into three steps (see S2 in Figure 1), namely, predominant melody extraction, melody histogram computation and finally octave estimation using the constructed histogram.

The predominant melody extraction is performed using the algorithm proposed by Salamon and Gómez [15], who kindly provided us with an implementation. Their system is specifically designed to extract the pitch contour of the dominant melodic source (lead performer in our case) in a situation where multiple pitched components exist simultaneously in the audio signal. A Vamp plugin to use this method in the SonicVisualizer [11] can be obtained from one of the authors website [12] .

The extracted pitch contour is used to construct the melody histogram. Before computing the histogram the pitch values are converted into a cent scale and quantized into 600 bins with a resolution of 10 cents per bin (Equation 4). An example of a melody histogram is shown in Figure 3. The red lines mark the pitch values (in bins) corresponding to tonic pitch-class (Sa) in different octaves. As can be seen, the tonic pitch corresponds to bin 166 which is the lowest Sa that has non-zero salience in the histogram. We propose two different approaches to estimate the tonic octave using the melody histogram; a rule-based approach (RB) and classification-based approach (CB).

2.2.1 Rule-based Approach

In this approach the tonic octave is estimated by applying a simple rule on the melody histogram. As mentioned earlier, the tonic pitch is the lowest tonic pitch-class used in the melody. Therefore, it would be sufficient to select the lowest tonic pitch-class in the melody histogram, which has a non-zero value. However, in some rare cases the melody extraction algorithm makes octave errors and estimates pitches which are sub-multiples of the true pitch values. This results into a non-zero value in melody histogram at a sub-multiple of the bin corresponding to the tonic pitch, which eventually leads to an error. A solution to this would be to take ratios of histogram values at tonic pitch-class locations in adjacent octaves. As the octave errors are very rare, this ratio would still be maximum at the tonic octave. We calculate the ratio $R(i)$ at every bin corresponding to the tonic pitch-class in different octaves ($i = 1, 2, ..N$) as shown below:

$$\begin{aligned} R(i) &= \frac{h(j_i)}{h(j_{i-1}) + \epsilon}, \\ j_i &= mod(\eta, 120) + 120 \cdot (i-1), \\ i &= 1, 2, 3, 4, 5 \end{aligned} \tag{7}$$

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

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

$$I = \arg\max_i R(i) \tag{8}$$

2.2.2 Classification Based Approach

There are rare cases where the rule-based method is bound to produce erroneous results [14]. Two such interesting scenarios are; the madhyaṁ-śruti case, where the singer may not sing the tonic pitch at all, as the natural fourth (Ma) with respect to the tonic pitch is considered as the Sa svar of the rāg, and the case where the low frequency pitches (mainly for male singers) are not tracked by the melody extraction algorithm. In both these cases the melody histogram values at the bins corresponding to the tonic pitches are very low, which leads to errors.

We handle these cases by adapting a classification based approach and not relying on only the tonic pitch-class locations in the melody histogram. We parametrize the whole histogram and model the lowest octave of the sung melody. The system automatically learns the best set of rules and pitch classes in the melody histogram which are crucial for identifying the tonic octave.

For every tonic pitch-class in different octaves we extract a set of 25 features. These features are the values of melody histogram at 25 equidistant locations spanning two octaves, centered around itself. This gives us a set of 25 features $h_i(i = 1 \ldots 25)$. An example is shown in Figure 3 for a tonic pitch-class at bin number 166. The sampled histogram at 25 equidistant locations centered around 166^{th} bin is marked by blue stars.

Next, we assign a class label to each tonic pitch-class instance in our dataset. We assign a class 'TonicOctave' if the instance is in the tonic octave, else 'NonTonicOctave'. The ground-truth tonic annotations are used for labelling the classes. Thus, by predicting the class ('TonicOctave' or 'NonTonicOctave') of every possible tonic pitch-class in different octaves, we can identify the correct tonic octave.

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

[11] http://www.sonicvisualiser.org
[12] http://www.justinsalamon.com/melody-extraction.html

evaluator and BestFirst search method with a 10-fold cross validation option set [19, 21]. We select the features which are used in at least 80% of the folds. Subsequently, a C4.5 (J48) decision tree is trained using WEKA to learn the best set of rules to predict the class labels.

Note that for computing the melody histogram we used the whole audio file. This is justified, because to find out the lowest tonic pitch-class used in the melody we need to listen to all of it. Otherwise, we have to incorporate the knowledge regarding the tonic pitch range for male and female singers. We also conduct experiments to see the effect of including the information regarding the possible tonic pitch range (110-260 Hz) in the system.

For the practical purposes the tonic pitch-class candidates for which there exists only one possibility of the tonic octave, this second stage of the proposed method can be omitted. For example, if the tonic pitch range for the singers is 110-260 Hz, then for the tonic pitch-class candidates which fall between 130-220 Hz range, there is no need for applying tonic octave estimation, as there is only one possibility. Note that we still perform it as a proof-of-concept.

3. TONIC IDENTIFICATION SYSTEM

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

1. Prevalent methodologies in MIR primarily focus on using only a single type of data source [22]. Most of the approaches either use the available audio data, music scores or the contextual metadata to accomplish certain tasks. Recent efforts towards semantic music discovery combine audio content analysis with social contextual data and metadata [22]. However, there should be more attempts specifically in the area of automatic music description to explore the potential of combining the complementary type of data, to achieve practical solutions with better accuracies.

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

3.

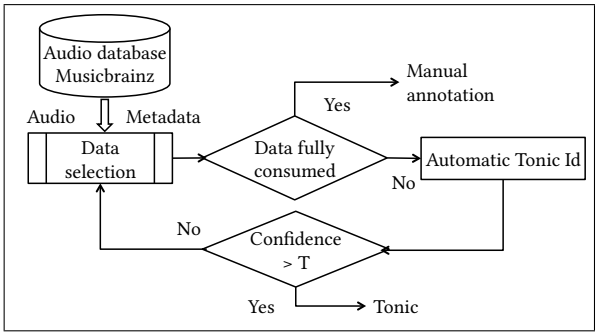

Figure 4. Block diagram of the iterative tonic identification system

Motivated by the aforementioned ideas, the proposed system combines the audio data and the available metadata for the identification of the tonic. Based on the derived confidence measure, the system tries to combine these two data sources to maximise the accuracy in an iterative manner. Figure 4 shows the block diagram of the complete iterative system.

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

4. EVALUATION

The music collection used to evaluate the proposed method is a subset of the musical material compiled as part of CompMusic project [23]. The core database used in this work is comprised of 352 full length audio songs, containing both vocal (237) and instrumental (115) musical pieces. We evaluate both the stages individually on the datasets S1 and S2 derived from the core database (Table 1). The Tonic pitch-class identification stage is evaluated using the dataset S2 containing 540 excerpts and the tonic octave estimation is evaluated using the dataset S1 containing 237 performances. Excerpts in the dataset S2 are 3 minutes long and are extracted from the the start, middle and the end of the full length recording when it is longer than 12 minutes.

Dataset	Size	Len.	Hind.(%)	Carn.(%)	Male(%)	Female(%)	#Usong	#Uartists
S1	237	full	27.4	72.6	75.2	24.8	237	34
S2	540	3 min	36	64	NA	NA	352	54

Table 1. Database description; summary in terms of different constituting components; Hindustani (Hind.), Carnatic (Carn.), male, female, number of unique songs (Usong), number of unique artists (Uartists)

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

Table 2. Performance accuracy of tonic pitch-class identification on dataset S2 with instance normalization.

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

Table 3. Performance accuracy of tonic pitch-class identification on dataset S2 without instance normalization.

Otherwise, only a single excerpt is extracted from the beginning. Table 1 provides statistics of both the datasets (S1 and S2) in terms of different attributes such as number of songs belonging to Hindustani, Carnatic, male and female singers.

The tonic annotations were done by the authors, and later verified by a professional musician. For a detailed description of the procedure followed for annotating music pieces with the tonic pitch, we refer to [14]. We evaluate the first stage of our method in terms of the percentage of the excerpts for which the tonic pitch-class is correctly identified. An output is considered as correct if it is within a bracket of 25 cents from the ground-truth value. For the second stage also the results are reported in terms of the percentage of the excerpts for which the tonic octave is correctly estimated.

5. RESULTS AND DISCUSSION

The performance accuracies for the tonic pitch-class identification stage on the dataset S2 for both with and without normalization are provided in Table 2 and 3. These tables show the performance accuracy (Acc.) on the whole dataset ('full'), as well as the obtained accuracies as a function of different attributes such as Hindustani (Hind.), Carnatic (Carn.), vocal and instrumental (Inst.) music. They also show a breakdown of the total errors made by the system in terms of different types of errors, the octave errors (Oct.Err), the 'Pa' or fifth type errors (Pa Err.), the 'Ma' or fourth type errors (Ma Err.). All other kinds of errors belong to the 'Others' category.

The obtained results for the tonic octave estimation stage

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

Table 4. Performance accuracy of tonic octave estimation stage on the dataset S1 for the rule-based approach. Results shown for both the cases; without imposing any limit on allowed tonic pitch range and constraining it to a limit of 110-260 Hz

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

Table 5. Performance accuracy of tonic octave estimation stage on the dataset S1 for the classification based approach.

for both the approaches (rule-based and classification based) are shown in Table 4 and 5. The evaluation is done both with and without imposing a constraint on the tonic pitch range. In the former case, the allowed frequency range for the tonic pitch was restricted to 110-260 Hz. Note that the results shown are only for the tonic octave estimation stage, evaluated individually using the ground-truth tonic pitch-class information.

The performance of the proposed method is good, with an accuracy of 92.96% for tonic pitch-class identification, without instance normalization. More importantly, the performance is good for not only the vocal excerpts but also for the instrumental excerpts. We see that the performance (76.67%) is inferior when the number of instances are normalized while training the classifier. This can be attributed to the fact that some classes contain a very small number of instances. The increased accuracy for predicting the minority classes does not improve the overall accuracy because a slight decrease in prediction accuracy of the majority classes (because of normalization) causes a greater drop in the overall performance. This hint that for the problem at hand it is better to ignore the specific rare cases than try to learn rules for them.

We also analyse the performance accuracy as a function of different attributes such as for vocal, instrumental, Hindustani and Carnatic excerpts. Table 3 shows the obtained accuracy for the whole database (92.96%), vocal excerpts (94.13%), instrumental pieces (90.3%), excerpts belonging

to Hindustani music (94.39%) and Carnatic music (92.15%). We notice that the performance on the vocal excerpts is better compared to the instrument excerpts. A plausible reason for this difference in performance could be the presence of an accompanying drone instrument. For vocal music, there is always a drone instrument accompanying the lead performer, whereas for the instrumental songs a dedicated drone instrument is absent in some performances.

Further analysing the erroneous cases, we observed that the most frequent error types were selecting the fifth (Pa) or the fourth (Ma) as the tonic or identifying the tonic in another octave. These type of errors are understandable, as Pa or Ma is the secondary pitch-class that is often produced by the drone instrument in addition to the tonic. Moreover, for the male singers the errors were selecting the higher Pa or Ma as tonic, whilst for female singers it was selecting the lower Pa or Ma. This can be attributed jointly to the differences in typical tonic frequencies for male and female singers, together with the frequency range chosen for constructing the pitch histograms.

The accuracy obtained for tonic octave identification is also good, with the classification based approach (96.62%) performing better than the rule-based approach (89.5%) without restricting the tonic pitch range. It is justified as the rule-based approach only considers the melody histogram values at different tonic pitch-class locations, whereas in the classification based approach we densely sample (25 points for two octaves) the histogram to model the lowest octave. We also evaluate both these approaches after incorporating the knowledge of tonic pitch range (110-260 Hz). This considerably improves the performance of the rule-based approach which now achieves an accuracy of 96.2%. The accuracy for the classification based approach also increases to 98.73% but not as significantly as for the former case. We observe that the performance of the classification based approach does not depend a lot on the selected frequency range.

Evaluating the performance of the rule-based approach exposed several interesting cases. It falls short of estimating the correct tonic octave when the song is sung in madhyaṁ-śruti [14], or when the melody extraction algorithm fails to track the low frequency pitches. For a detailed analysis of erroneous cases we refer to [14]. Another interesting observation is that the rule-based approach performs equally well for the performances of male and female singers, and better for Hindustani music compared to Carnatic music. However, the classification based approach performs better for the performances of male singers compared to female singers and for Carnatic music than the Hindustani music. This can be attributed to the predominance of male singers and Carnatic music cases in our database (Table 1).

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented a new approach for tonic identification in Indian art music. Our method divides the task into two stages, where the first stage performs tonic pitch-class identification. In this way, in addition to vocal music the method is also suitable for instrumental music where the concept of tonic octave is not clearly defined. The tonic pitch-class identification is based on a multi-pitch analysis of the audio signal, in which the predominant pitches are used to construct a pitch histogram. The pitch histogram represents the most frequently used pitches in the whole excerpt. We thus utilize the presence of the drone in the background of the recording, which constantly reinforces the tonic pitch. Using a classification based approach the system automatically learns the best set of rules to select the peak of the histogram representing the tonic pitch-class.

We presented two approaches for the second stage of the method, which estimates the tonic octave; a rule-based approach and a classification based approach. In both the approaches we analyze the predominant melody contour to establish the tonic octave. Both the stages are individually evaluated on a sizable database containing a wide variety of music material such as Hindustani and Carnatic music, male and female singers, vocal and instrumental music performances. The method obtains a good accuracy in both the stages. This supports our hypothesis that the drone sound is an important cue to tonic pitch-class identification and tonic octave can be established based on the predominant melody. While performing tonic octave estimation many interesting cases such as madhyaṁ-śruti songs came into light. Along with the results, we also discussed the types of errors most commonly made by the method and plausible reasons for them.

In addition to the the proposed approach we also presented a proposal for a complete iterative system for tonic identification in Indian art music. We briefly discussed the issues which need to be addressed in future in order to incrementally utilize the available metadata in conjunction with the audio data. Specifically, the data selection and the confidence estimation modules are the two important blocks on which we intend to concentrate our efforts on in our future work.

Acknowledgments

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

7. REFERENCES

[1] T. Viswanathan and M. H. Allen, *Music in South India*. Oxford University Press, 2004.

[2] A. Danielou, *The Ragas of Northern Indian Music*. New Delhi: Munshiram Manoharlal Publishers, 2010.

[3] B. C. Deva, *The Music of India: A Scientific Study*. Delhi: Munshiram Manoharlal Publishers, 1980.

[4] S. Bagchee, *NAD Understanding Raga Music*. Business Publications Inc, 1998.

[5] J. Serrà, G. K. Koduri, M. Miron, and X. Serra, ``Assessing the tuning of sung indian classical music," in *Proc. 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[6] G. K. Koduri, J. Serrà, and X. Serra, ``Characterization of intonation in carnatic music by parametrizing pitch histograms," in *13th Int. Soc. for Music Info. Retrieval Conf.*, Porto, Portugal, Oct. 2012.

[7] J. C. Ross, T. P. Vinutha, and P. Rao, ``Detecting melodic motifs from audio for Hindustani classical music," in *Proc. 13th International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012.

[8] P. Chordia, J. Jagadeeswaran, and A. Rae, ``Automatic carnatic raag classification," *Journal of the Sangeet Research Academy (Ninaad)*, 2009.

[9] P. Chordia and A. Rae, ``Raag recognition using pitch-class and pitch-class dyad distributions," in *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[10] G. K. Koduri, S. Gulati, and X. Serra, ``Survey and Evaluation of Pitch-distribution Based Raaga Recognition Techniques," *Journal of New Music Research*, in press.

[11] H. Ranjani, S. Arthi, and T. Sreenivas, ``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*, pp. 29--32, 2011.

[12] R. Sengupta, N. Dey, D. Nag, A. K. Datta, and A. Mukerjee, ``Automatic Tonic (SA) Detection Algorithm in Indian Classical Vocal Music," in *National Symposium on Acoustics*, 2005, pp. 1--5.

[13] J. Salamon, S. Gulati, and X. Serra, ``A Multipitch Approach to Tonic Identification in Indian Classical Music," in *Proc. 13th International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012.

[14] S. Gulati, *A Tonic Identification Approach for Indian Art Music*. (Master's dissertation), Music Technology Group, Universitat Pompeu Fabra, Barcelona, 2012.

[15] J. Salamon and E. Gómez, ``Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759--1770, Aug. 2012.

[16] J. Salamon, E. Gómez, and J. Bonada, ``Sinusoid extraction and salience function design for predominant melody estimation," in *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11), Paris, France*, 2011, pp. 73--80.

[17] A. Klapuri, ``Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*, 2006.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, ``The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10--18, Nov. 2009.

[19] M. Hall, ``Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, University of Waikato, 1999.

[20] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[21] I. H. Witten, E. Frank, and M. A. Hall, *Data mining : practical machine learning tools and techniques*, 3rd ed. Morgan Kaufmann, Jan. 2011.

[22] L. Barrington, D. Turnbull, and M. Yazdani, ``Combining audio content and social context for semantic music discovery," in *Proc. 32nd ACM SIGIR*, 2009.

[23] X. Serra, ``A Multicultural Approach to Music Information Research," in *Proc. 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.