A MULTIPITCH APPROACH TO TONIC IDENTIFICATION IN INDIAN CLASSICAL MUSIC

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

ABSTRACT

The tonic is a fundamental concept in Indian classical music since it constitutes the base pitch from which a lead performer constructs the melodies, and accompanying instruments use it for tuning. This makes tonic identification an essential first step for most automatic analyses of Indian classical music, such as intonation and melodic analysis, and raga recognition. In this paper we address the task of automatic tonic identification. Unlike approaches that identify the tonic from a single predominant pitch track, here we propose a method based on a multipitch analysis of the audio. We use a multipitch representation to construct a pitch histogram of the audio excerpt, out of which the tonic is identified. Rather than manually define a template, we employ a classification approach to automatically learn a set of rules for selecting the tonic. The proposed method returns not only the pitch class of the tonic but also the precise octave in which it is played. We evaluate the approach on a large collection of Carnatic and Hindustani music, obtaining an identification accuracy of 93%. We also discuss the types of errors made by our proposed method, as well as the challenges in generating ground truth annotations.

1. INTRODUCTION

One of the fundamental concepts in Indian classical music is the tonic. The tonic is a base pitch chosen by the performer, and serves as the foundation for the melodic tonal relationships throughout the performance. Every performer chooses a tonic pitch which best allows them to fully explore their vocal (or instrumental) pitch range for a given raga exposition [3]. Consequently, all accompanying instruments are tuned with relation to the tonic chosen by the lead performer.

Since the entire performance is relative to the tonic (corresponding to the *Sa* note of the raga), the lead performer needs to hear the tonic pitch throughout the concert. This is provided by a constantly sounding drone which plays in the background and reinforces the tonic. The drone may be produced by a variety of instruments such as the *Tanpura*, the electronic *Shruti box*, or by the sympathetic strings of an instrument such as the *Sitar* or *Veena*. Along with the tonic, the drone typically produces other important notes in the raga such as the *Pa* (fifth) or the *Ma* (fourth), and slightly less often the seventh (*Ni*), depending on the choice of raga. This drone serves as the reference sound that establishes all the harmonic and melodic relationships during a given performance. Other notes used in the performance derive their meaning and purpose in relation to the *Sa* and the tonal context established by the particular raga [2].

When considering the computational analysis of Indian classical music, it becomes evident that identifying the tonic is a crucial first step for more detailed tonal studies such as intonation [9], motif analysis [13] and raga recognition [1]. This makes automatic tonic identification a fundamental research problem. However, despite its importance in Indian classical music, the problem of automatic tonic identification has received very little attention from the research community to date.

To the best of our knowledge, all previous approaches for automatic tonic identification are based on applying monophonic pitch trackers to the audio recording, meaning they solely use the information proportioned by the predominant melody [16]. In some cases a monophonic pitch tracker is used even though the audio recording contains several instruments playing simultaneously [12]. These approaches have also been fairly restricted in terms of the musical content studied: in [16] only the *Alap* sections of 118 solo vocal recordings are used for evaluation, and in [12] the evaluation material is restricted to *Sampurna raga*. Both approaches also restrict the allowed frequency range for the tonic to a single octave, a limitation which can not be imposed if we wish to devise a single method for tonic identification for both male and female vocal performances.

In this paper we propose a method for tonic identification in Indian classical music based on a multipitch analysis of the audio signal. The motivation for a multipitch approach is twofold: first, the music material under investigation often includes several instruments playing simultaneously. Apart from the lead performer, recordings contain the drone instrument, and may also include other predominant instruments such as the violin, as well as percussive instruments. Second, we know that the tonic is continually reinforced by the drone instrument, an important fact that

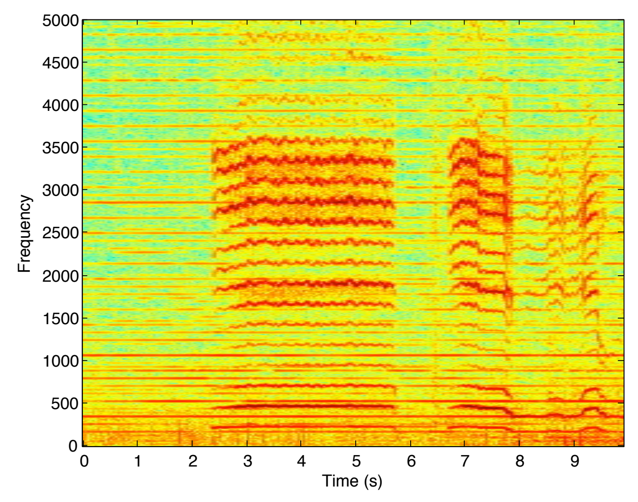

Figure 1. Spectrogram of an excerpt of Hindustani music with two clearly visible types of harmonic series, one belonging to the drone and the other to the lead voice.

is not exploited if we only extract a single pitch estimate for each frame of the recording. To illustrate this point, in Figure 1 we display the spectrogram for an excerpt of Hindustani music [2]. Two types of harmonic series are clearly visible in the spectrogram: the first type of harmonic series, which consist of almost perfectly flat lines, belong to the notes of the drone instrument (playing *Sa* and *Pa*). The second type of harmonic series (which starts roughly at time 2s) belongs to the voice of the lead performer. Evidently, if we only consider the pitch of the lead performer, we loose the pitch information proportioned by the drone instrument which in this case is a better indicator of the tonic pitch.

At the outset of this study, we defined three goals for the method to be developed: first, it should be applicable to a wide range of performances, including both the Carnatic [18] and Hindustani musical styles, male and female singers, and different recording conditions. Second, the approach should identify the tonic pitch in the correct octave, without restricting the allowed frequency range to a single octave. Finally, the approach should be able to identify the tonic using a limited segment of the full recording, and this segment can be taken from any part of the piece.

The structure of the remainder of the paper is as follows. In Section 2 we present our proposed tonic identification method. In Section 3 we describe the evaluation methodology employed in this study, including the music collection used for evaluation and the annotation procedure used to generate the ground truth. Then, in Section 4 we present and discuss the results of the evaluation, and finally in Section 5 we provide some conclusions and proposals for future work.

2. PROPOSED METHOD

The proposed method is comprised of four main blocks: sinusoid extraction, salience function, candidate generation and tonic selection. The first two blocks of the system were originally proposed as part of a predominant melody extraction system [14, 15], and have been adapted here for the task of tonic identification. In the following sections we describe the processing steps involved in each of the four blocks of the system. A block diagram of the proposed method is provided in Figure 2.

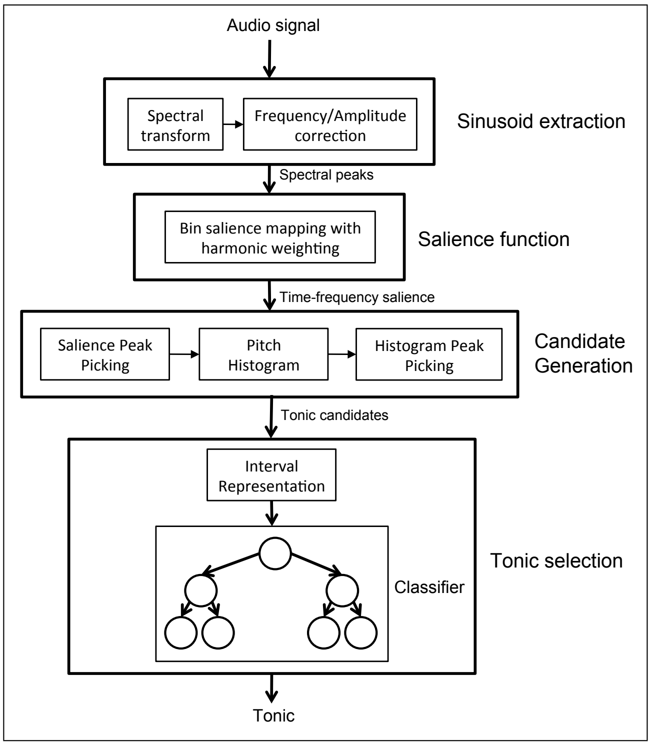

Figure 2. Block diagram of the proposed tonic identification method.

2.1 Sinusoid Extraction

In the first step of the method, we extract sinusoidal components, i.e. spectral peaks, from the audio signal. The sinusoid extraction process is divided into two stages as depicted in Figure 2: spectral transform and sinusoid frequency/amplitude correction.

We start by applying the Short-Time Fourier Transform (STFT) 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 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 Hann windowing function with a window size of 46.4ms, a hop size of 2.9ms and a $\times 4$ zero padding factor, which for data sampled at $f_S = 44.1$kHz gives $M = 2048$, $N = 8192$ and $H = 128$. Given the FFT of a single frame $X_l(k)$, spectral peaks are selected by finding all the local maxima k_m of the magnitude spectrum $|X_l(k)|$.

The location of the spectral peaks is limited to the bin frequencies of the FFT, which for low frequencies can result in a relatively large error in the estimation of the peak

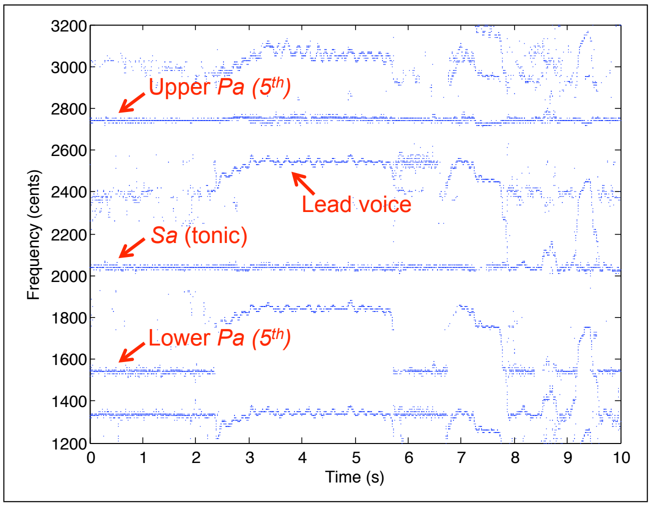

Figure 3. Peaks of the salience function for an excerpt of Hindustani music.

frequency. To overcome this quantisation, in the second stage of this block we apply the approach described in [4], in which the phase spectrum $\phi_l(k)$ is used to calculate the peak's instantaneous frequency (IF) and amplitude, which provide a more accurate estimate of the peak's true frequency and amplitude.

2.2 Salience Function (Multipitch Representation)

We use the extracted spectral peaks to compute a *salience function* – a multipitch time-frequency representation of pitch salience over time. The salience computation is based on harmonic summation similar to [8], where the salience of a given frequency is computed as the sum of the weighted energies found at integer multiples (harmonics) of that frequency. Peaks of the salience function at each frame represent salient pitches in the music recording. Note that whilst the concepts of pitch (which is perceptual) and fundamental frequency (which is a physical measurement) are not identical, for simplicity in this paper we will use these two terms interchangeably.

Our salience function covers a pitch range of nearly five octaves from 55Hz to 1.76kHz, quantized into 600 bins on a cent scale (10 cents per bin). The reader is referred to [14, 15] for further details about the mathematical formulation of the salience function. In Figure 3 we plot the peaks of the salience function for the same excerpt from Figure 1. The tonic (*Sa*) pitch which is played by the drone instrument is clearly visible, as well as the upper and lower fifth (*Pa*), and the pitch trajectory of the voice.

2.3 Tonic Candidate Generation

As explained earlier, the peaks of the salience function represent the pitches of the voice and other predominant instruments present in the recording at every point in time. Thus, by computing a histogram of the pitch values for the entire excerpt, we obtain an estimate of which pitches are repeated most often throughout the excerpt. Though pitch histograms have been used previously for tonic identification [12], they were constructed using only the most predominant pitch at each frame, which means that in many cases the tonal information provided by the drone instrument is not taken into consideration.

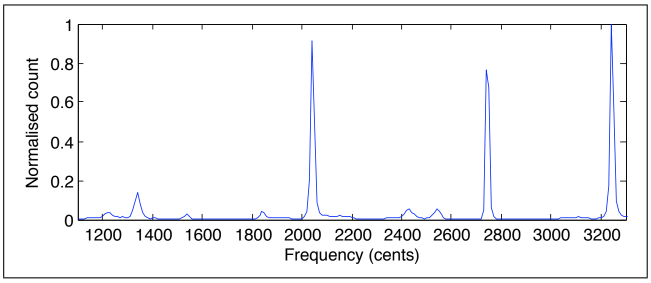

Figure 4. Pitch histogram for an excerpt of Hindustani music.

We start by taking the peaks of the salience function at each frame. Since the frequency range for the tonic pitch selected by singers in Indian classical music is relatively limited, we can reduce the range from which salient pitches are selected. To ensure we cover the complete range for both male and female singers, we consider salient pitches with a fundamental frequency ranging from 110 Hz to 370 Hz. Importantly, note that this range spans almost 2 octaves, meaning the system must be able to identify not only the correct tonic pitch class, but also the octave in which it is played. Within this range, at each frame we take the top five peaks (pitches) of the salience function.

The selected pitches are used to construct a pitch histogram. As the drone is usually weaker than the lead voice, we avoid weighting each peak by its magnitude. The resulting pitch histogram goes from 110 Hz to 370 Hz and has a resolution of 10 cents. Peaks of the histogram represent the most frequent pitches in the excerpt, one of which will be the tonic. In Figure 4 we present the histogram computed from the complete 3 minute excerpt used in the previous examples. The pitch axis is plotted in cents, and the histogram is normalised by the magnitude of its highest peak. For the excerpt under consideration, we note three clear peaks: the tonic *Sa* (2040 cents), the upper *Pa* (2740 cents) and the tonic again, one octave up (3240 cents). This illustrates one of the challenges the system will have to deal with – selecting the tonic at the correct octave. It also highlights another important issue – the peak corresponding to the tonic will not always be the highest peak in the histogram, meaning the (perhaps naïve) approach of selecting the highest peak of the histogram would not provide satisfactory results.

2.4 Tonic Selection

As the tonic will not always be the highest peak of the histogram, we take the top 10 peaks of the pitch histogram p_i ($i = 1 \ldots 10$), one of which represents the pitch of the tonic. As mentioned in the introduction, all other notes present in the musical piece are tuned with relation to the tonic. Bearing this in mind, we hypothesize that the tonic can be identified based on the pitch intervals between the most frequent notes in the recording and their rate of occurrence. For example, in the excerpt in Figure 3, the drone plays the tonic alongside the lower and upper fifth. Thus, a fifth relationship between two frequent notes might serve as a good indicator for the tonic.

In the study of Western music, templates learned from music cognition experiments have been used for the related task of key detection, where a pitch histogram (derived from a symbolic representation of the musical piece) is matched against templates representing the probability of different pitch classes given a certain tonal context [10]. Approaches based on training a classifier to determine the key of a musical piece using chroma features automatically extracted from the audio signal have also been proposed [5]. In this study, we propose a classification approach to automatically learn the best set of rules for selecting the tonic, based on the pitch intervals between the most frequent notes in the piece and their relative rate of occurrence (as indicated by the magnitude of the peaks of the pitch histogram).

We start by annotating for each piece the rank $i = I$ of the tonic (in terms of peak magnitude) out of the top 10 peaks p_i of the pitch histogram. Then, we encode the 10 tonic candidates as the distance (in semitones) between every candidate p_i and the highest candidate in the histogram p_1. This gives us a set of features f_i ($i = 1 \ldots 10$), where f_i represents the distance (in semitones) between p_i and p_1. The features f_i and the annotated rank of the tonic I are used to train a classifier for selecting the tonic. That is, we pose the task of tonic identification as a classification problem where we have 10 classes (10 candidates) and the classifier must choose the rank of the candidate corresponding to the tonic. Note that for all files in our collection the tonic was always amongst the top 10 peaks p_i of the pitch histogram.

For classification we use the Weka data-mining software [7]. We start by performing attribute selection using the *CfsSubsetEval* attribute evaluator and BestFirst search method [6] with a 10-fold cross validation, only keeping features that were used in at least 80% of the folds. The selected features were: f_2, f_3, f_5, f_6, f_8 and f_9. Then, we train a C4.5 decision tree [11] in order to learn the optimal set of rules for selecting the tonic based on the pitch intervals between the tonic candidates. Note that we also evaluated other classification algorithms, namely support vector machines (SMO with polynomial kernel) and an instance-based classifier (k*) [19]. However, the accuracy obtained using the decision tree was significantly higher (6% better than SVM and 5% better than k*), and so for the rest of the paper we will focus on the results obtained using this classifier. Additionally, using a decision tree has the advantage that the resulting classification rules can be easily interpreted and, as shall be seen, are musically meaningful.

The resulting tree is presented in Figure 5. As it turns out, only 3 features are finally used: f_2, f_3 and f_5. Another interesting observation is that the pitch intervals used by the tree for making decisions correspond quite well to the intervals between the notes commonly played by the drone instrument: 5 (i.e. 500 cents) corresponds to the interval between the lower *Pa* and the tonic *Sa*, and 7 (700 cents) to the interval between the *Sa* and upper *Pa*. Note that a distance of 500 cents may also correspond to the distance between the *Sa* and upper *Ma*, which might be a cause for confusion in our system, and we will assess this when we analyse the results.

Examining the rules of the tree, we see that the most important relationship is between the top two peaks of the histogram. When the second highest peak is more than 500 cents above the highest peak, the latter is chosen as the tonic. Examining the data we found that this almost always corresponds to one of two cases – the second peak is either *Pa* (i.e. *Pa* tuning) or *Sa* one octave above the tonic. Branching left, the tree checks whether the highest peak is actually *Pa* (700 cents above the tonic). To confirm this it checks if the third peak is found 500 cents above the highest peak (thus corresponding to *Sa* one octave above the tonic). In this case the highest peak is indeed *Pa*, and the second highest peak is the tonic. Otherwise, we have a case of *Ma* tuning (the second peak is tuned to *Ma*), and the highest peak is the tonic. Similar interpretations can be made for the remaining rules of the tree.

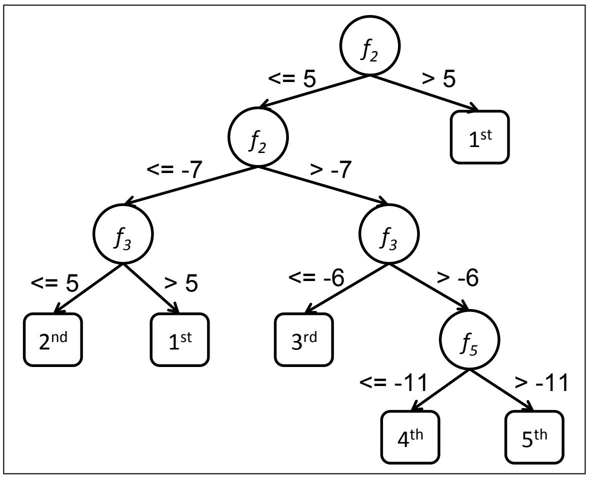

Figure 5. Obtained decision tree for tonic identification.

3. EVALUATION METHODOLOGY

3.1 Music Collection

The music collection used to evaluate the proposed approach was compiled as part of the CompMusic project [17]. It consists of 364 excerpts of Indian classical music including both Hindustani (38%) and Carnatic (62%) music. The excerpts were extracted from 231 unique performances by 36 different artists, including both male (80%) and female (20%) singers. Every excerpt is 3 minutes long, and extracted from either the beginning, middle or end of the full recording (for recordings longer than 12 minutes we are able to extract all 3 excerpts, for shorter recordings a single excerpt from the beginning of the piece was taken). Including excerpts from sections other than the beginning of the piece is important, since in both the Hin-

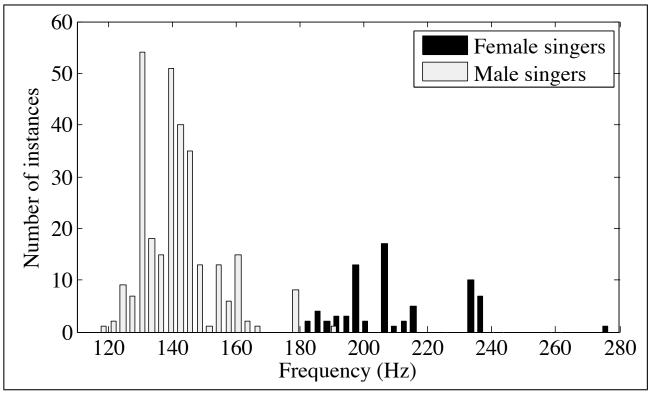

Figure 6. Distribution of tonic frequency for male and female vocal performances in our music collection.

dustani and Carnatic music traditions different sections of a performance can have very different acoustic characteristics. In Figure 6 we display the distribution of tonic frequencies in our collection for both male and female singers.

3.2 Annotation Procedure

The tonic frequency for each excerpt was manually annotated by the authors. To assist the annotation process, we used the candidate generation part of our proposed method to extract 10 candidate frequencies for the tonic in the range of 110 Hz to 300 Hz. The annotator could then listen to the candidate frequencies one by one together with the original recording in order to identify the tonic frequency. Note that for all excerpts in our collection the true tonic frequency was present in one of the 10 candidates provided by the system.

It is worth noting that as part of the annotation process, the listener must determine the octave in which the tonic is played. Since the drone instrument may play the tonic pitch in two octaves simultaneously, the octave of the tonic is determined by the vocal range of the singer rather than the drone instrument directly. Whilst in most cases the correct octave is fairly unambiguous for vocal performances, we encountered a small number of cases in which determining the octave of the tonic was more difficult. In future work, we intend to study the relation between performer and drone instrument in greater depth, as well as conduct listening tests to assess the degree of agreement between listeners when asked to determine the octave of the tonic.

4. RESULTS

We evaluate the proposed classification-based approach using 10-fold cross validation. The experiment is repeated 10 times, and the average results for all 10 repetitions are reported. In Figure 7 we present the classification accuracy obtained for our collection of 364 excerpts, as well as a breakdown of the results based on musical style and gender of the lead performer.

We see that the proposed approach obtains a classification accuracy (hence tonic identification accuracy) of 93% for our complete collection. Importantly, since the allowed

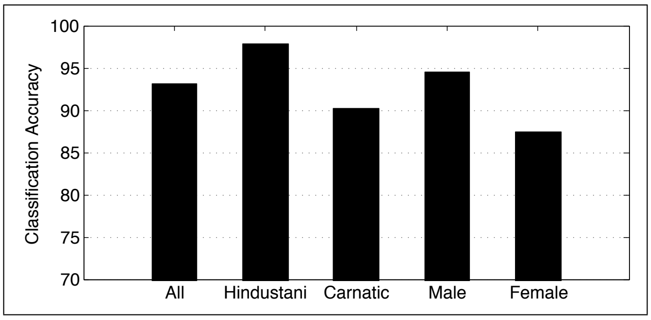

Figure 7. Classification accuracy for the proposed approach. All excerpts 93%, Hindustani 98%, Carnatic 90%, Male 95% and Female 88%.

tonic frequency range spans more than one octave, it means we are correctly identifying not only the pitch-class of the tonic, but also the octave at which it is played. Next, we examine the results depending on the musical style. We see that we have almost perfect classification for Hindustani music (98%), whilst for Carnatic music the performance is somewhat lower (90%). When examining the data, we noted that in the Carnatic excerpts there were more cases where the *Tanpura* was quite weak (in terms of loudness). Consequently, this results in frames where the pitch corresponding to the tonic does not have a prominent peak in the salience function. This in turn means the peak of the pitch histogram which corresponds to the tonic has a fairly low rank, leading to incorrect identification of the tonic.

When considering identification accuracy as a function of the gender of the lead performer, we see that the system performs better for pieces performed by male singers compared to those performed by female singers. A possible cause for this is the different amount of male and female performances in our collection. Since there are considerably more male performances, the rules learned by the system are better suited for identifying the tonic in this type of musical material. Another factor that was identified as influential was the frequency range used to compute the pitch histogram. Whilst our frequency range covers the entire range in which we expect to find the tonic for both male and female cases, for high frequency tonics this range will not include the higher *Sa* one octave above the tonic. As it turns out, the presence of a higher *Sa* is one of the cues used by the system, and for many female excerpts it is outside the range of the pitch histogram. In the future, we intend to experiment with different frequency ranges for the pitch histogram, as well as consider separate ranges for male and female performances to see whether performance can be improved by including this extra piece of information prior to classification.

As a final step in our analysis of the results, we checked what types of errors were the most common in our evaluation. We found that for male singers the most common error was selecting the higher *Pa* or *Ma* as the tonic, whilst for females it was selecting the lower *Pa* or *Ma*. This is understandable, as these are two important notes that are often played by the drone instrument in addition to the tonic. The difference in tonic frequency for males and females, together with the frequency range used for the pitch histogram, explains why for males we erroneously select a higher note, whilst for females we select a lower one. Additionally, for female singers we found that the confusion was often caused due to the use of *Ma* tuning (*Sa* - *Ma* - *Sa*) of the drone instrument. If the higher *Sa* is not present, the *Ma* tuning is equivalent to a rotated version of *Pa* tuning, resulting in the wrong rule being applied.

5. CONCLUSION

In this paper we presented a novel approach for tonic identification in Indian classical music. Our method is based on a multipitch analysis of the audio signal, in which the predominant pitches in the mixture are used to construct a pitch histogram representing the most frequently played notes in the piece. In this way, our representation also captures the notes played by the drone instrument, and not only the pitch of the lead performer. Using a classification approach, we were able to automatically learn the best set of rules for tonic identification given our pitch histogram representation. The resulting decision tree was evaluated on a large collection of excerpts consisting of a wide selection of pieces, artists and recording conditions, and was shown to obtain high tonic identification accuracy. Importantly, the approach is suitable for both Hindustani and Carnatic music, male and female performances, and only requires a short excerpt of the full performance. In addition, the rules learned by the system are easy to interpret and musically coherent.

Following presentation of the results, we discussed the types of errors most commonly made by the proposed tonic identification method, and the main causes for these errors where identified. Finally, we proposed some directions for future work, including a study of tonic octave perception, considering different frequency ranges for the pitch histogram in our proposed method, and devising gender-specific tonic identification approaches.

6. ACKNOWLEDGMENTS

The authors would like to thank Joan Serrà, Emilia Gómez and Perfecto Herrera for their useful comments. This research was funded by the Programa de Formación del Profesorado Universitario (FPU) of the Ministerio de Educación de España and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 267583 (CompMusic).

7. REFERENCES

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

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

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

[4] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 247–252, Montreal, Quebec, Canada, Sept. 2006.

[5] E. Gómez and P. Herrera. Estimating the tonality of of polyphonic audio files: Cognitive versus machine learning modelling strategies strategies. In *5th Int. Conf. on Music Info. Retrieval*, Barcelona, Spain, Oct. 2004.

[6] M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1999.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

[8] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *7th Int. Conf. on Music Info. Retrieval*, Victoria, Canada, October 2006.

[9] 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.

[10] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, 2001.

[11] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[12] T.V. Ranjani, H.G.; Arthi, S.; 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*, pages 29–32, 2011.

[13] J. C. Ross, T. P. Vinutha, and P. Rao. Detecting melodic motifs from audio for Hindustani classical music. In *13th Int. Soc. for Music Info. Retrieval Conf.*, Porto, Portugal, Oct. 2012.

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

[15] 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)*, pages 73–80, Paris, France, Sep. 2011.

[16] R. Sengupta, N. Dey, D. Nag, A. Datta, and A. Mukerjee. Automatic Tonic (SA) Detection Algorithm in Indian Classical Vocal Music. In *National Symposium on Acoustics*, pages 1–5, 2005.

[17] X. Serra. A multicultural approach in music information research. In *12th Int. Soc. for Music Info. Retrieval Conf.*, Miami, USA, Oct. 2011.

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

[19] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, Waltham, USA, 2nd edition, 2005.