CHAPTER 4

TONIC DETECTION

4.1 THE CONCEPT OF TONIC

An important concept in Indian classical music is the tonic or *Adhara Shadjam* denoted by *Sa*. It is the base pitch selected by an artist and it serves as the foundation on which the artist builds his performance. The entire performance will be relative to the tonic. A performer chooses a tonic which is most comfortable for him/her to fully explore their vocal (or instrumental) pitch range [15]. All accompanying instruments are tuned in relation to the tonic chosen by the lead performer. The tonic pitch needs to be heard throughout the concert. This is provided by a constantly sounding '*tanpura*' (drone) or an electronic *sruthi* box which plays the note *Sa* in the background and thereby reinforces the tonic. The *'tanpura'* or *sruthi* box also produces other important notes such as the *Pa* (fifth) or the *Ma* (fourth), and slightly less often the *seventh* (Ni). The drone produces 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 [13], [44].

4.2 THE NEED FOR AUTOMATIC TONIC DETECTION

For the computational analysis of Carnatic music, accurately identifying the tonic is a very essential first step. It serves as the foundation for more detailed studies such as raga recognition and classification. This makes automatic tonic identification a fundamental

research problem. However, despite its importance in the computational analysis of Carnatic music, the problem of automatic tonic identification has not much been explored by the research community.

Most of the previous approaches for tonic detection are more or less performed manually. For example in [11], tonic identification was done manually by tuning an oscillator and noting the value in Hz. Even when attempts were made to automatically identify tonic, those efforts were restricted to monophonic audio recordings. Such methods used information corresponding to the predominant melody only [45]. Hence those methods are not of much use when the audio recording contains several instruments playing simultaneously [41]. These approaches were also fairly restricted in terms of the musical content studied. For example in [45], only the *Alap* (humming) sections of 118 solo vocal recordings are used for evaluation, and in [41] the evaluation material is restricted to only *sampoorna* raga.

In our work, tonic detection based on a multi-pitch approach was needed to extract tonic information from polyphonic recordings. Since our dataset consisted of actual concert recordings, the music material included several instruments playing simultaneously. Apart from the lead performer, recordings contained the drone instrument which continually reinforces the tonic, other accompanying instruments such as violin, *mridangam* etc. The presence of more than one instrument will create more than one pitch track. This fact had to be incorporated into our tonic detection method and efficient methods to extract multiple pitches needed to be developed instead of a single pitch estimate for

each frame of the recording. An earlier attempt in this area is detailed in [18].

We devised a novel approach for automatic tonic detection based on the following peculiar properties of musical notes.

4.2.1 Critical Bands and Dissonance

Critical bands refer to a very important concept in the study of frequencies of musical scales. When sound enters the ear, it causes vibrations on the basilar membrane within the inner ear. Different frequencies of sound cause different regions of the basilar membrane to vibrate. This is how the brain discriminates between various frequencies. However, if two frequencies are close together, there will be an overlap of response on the basilar membrane. That is, vibrations are caused by both frequencies. In such case, the frequencies cannot be distinguished as separate frequencies. Instead an average frequency is heard. If the two frequencies are 440 Hz and 450 Hz, for example, we will hear 445 Hz. If the lower frequency is kept at 440 Hz and the higher one is raised slowly, then there will come a point where the two frequencies are still indistinguishable and there is just a roughness to the total sound. This is called dissonance. It would continue until finally the higher frequency would become distinguishable from the lower. At this point, further raising the higher frequency would cause less and less dissonance. When two frequencies are close enough to cause the roughness or dissonance described above, they are said to be within a critical band on the basilar membrane. For much of the audible range, the critical band around some central frequency will be stimulated by frequencies within about 15% of that central frequency [27].

In the study of musical notes and musical scales, critical bands play an important role. Two frequencies that stimulate areas within the same critical band on the basilar membrane will produce dissonance which is undesirable in music.

4.2.2 Consonance

The opposite of dissonance is consonance. That is pleasant sounding combinations of frequencies. In the example discussed in the previous section, if the 450 Hz is replaced with an 880 Hz (2 x 440 Hz), we can hear excellent consonance. This especially pleasant sounding combination comes from the fact that every crest of the sound wave corresponding to 440 Hz would be in step with every other crest of the sound wave corresponding to 880 Hz. So doubling the frequency of one tone always produces a second tone that sounds good when played with the first. This interval between two frequencies is called a diapason. Not only that 440 Hz and 880 Hz sound so good together, they also sound the same. Similarly, if the frequency of the 880 Hz tone is increased to 1760 Hz (2x880Hz or 4x440Hz), it sounds the same as when the frequency of the 440 Hz tone is increased to 880 Hz. This feature has been exploited in various musical systems to build musical scales by using an arbitrary frequency and another frequency, exactly one diapason higher, as the first and last notes in the musical scale [27]. As mentioned above, frequencies separated by one diapason not only sound good together, but they sound like each other. That is why an adult and a child or a man and a woman can sing the same song together simply by singing in different diapasons. And they do this naturally, without even thinking about it [27]. The same applies to a vocalist and his/her supporting instrumentalist like, for example, a Carnatic vocalist and the

accompanying violinist. The base pitch of violin is usually an octave higher than that of vocal. But since they are separated by one diapason, the audiences get the feeling that they are playing in unison.

4.3 PROPOSED METHOD FOR AUTOMATIC TONIC DETECTION

The concept of critical bands and consonance discussed in the previous sections has been used as the underlying principle of our tonic detection method [51]. Our method calculates the tonic based on a sample taken from the recording under study. This sample may contain either the vocal part or a supporting instrument like violin or a combination of both. Normally the base frequency of vocal and a supporting instrument like violin will have a difference of one diapason. That is, the frequency of a note generated from the violin will be twice the frequency of the same note generated by the vocalist. However, as mentioned above, due to consonance, two notes separated by a diapason will sound alike. This is the reason a violinist and a vocalist are able to perform in unison. Based on this fact, we hypothesized that the tonic identification can be independent of the medium of performance. That is, we can identify the tonic from the sound of violin or from the sound of vocalist or from a combination of these two. In all these cases, the detected tonic can be used to identify the raga from the violin portion as well as from the vocal portion or from a combination of these two. We also hypothesized that, since the tonic is medium independent, it can also be used to identify raga from portions containing polyphonic music, for example, from portions where the sound of *mridangam* or some other accompanying instrument is also present. Our experiments have successfully proved this hypothesis. This is a major advancement from earlier works where the tonic was found either by tuning an oscillator

and noting the value in Hz [11] or by categorizing instruments as either male or female and asking explicitly for the base frequency [23].

Our method for tonic detection is based on a multi-pitch analysis of the audio signal, in which all the predominant pitches in the input piece were used to construct a pitch histogram representing the most frequently played notes in the piece. The method automatically detected the tonic in two different ways: grouping by zero and pitch histogram. Also, it captured the notes played by the drone instrument and other accompanying instruments along with the pitch of the lead performer. That is, it was medium independent. It needed only a short excerpt of just 1 or 2 second duration as input. The method was tested successfully on a large collection of excerpts consisting of a wide range of ragas, artists and recording conditions, and was shown to obtain high tonic identification accuracy. It is comprised of seven blocks (Figure 4.1): input signal pre-processing, framing, decomposition into filter-bank channels, summation, peak picking, pitch estimation and tonic selection.

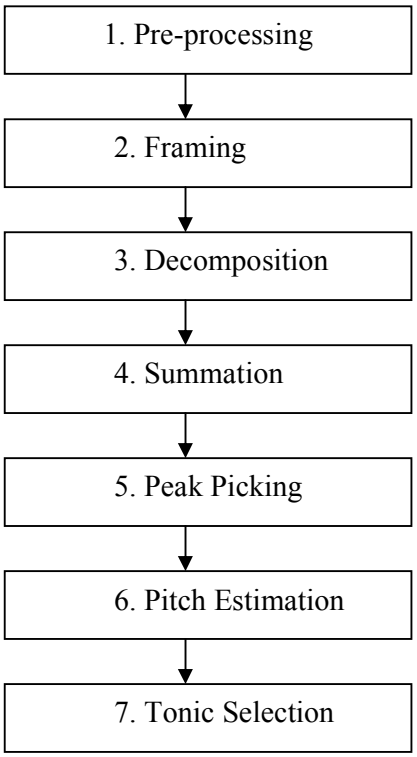

Figure 4.1

Block diagram of our tonic detection method

Blocks 1 to 6 constitute the first phase, called pitch estimation phase, of our tonic identification method. In this phase, the pitch values are estimated. Phase II, called tonic selection phase, identifies the tonic from these pitch values.

4.3.1 Phase I: Pitch Estimation

Analyzing the lower amplitude portions of the input recording (stored as a 'wav' file with a sampling frequency of 44.1 KHz) like the ending portions of *raga visthara* (elaboration of a raga accompanied by the *tanpura* and sometimes the violin), a small piece lasting about only 1 or 2 seconds was chosen for tonic detection. The musical signal

contained in this piece was first decomposed with a frame size of 25 ms. Then each frame was decomposed into channels ranging from low frequency channels to high frequency channels using a bank of filters each one selecting a particular range of frequency values (Figure 4.2). This transformation models an actual process of human perception, corresponding to the distribution of frequencies into critical bands in the cochlea. This enables to study each of these channels separately.

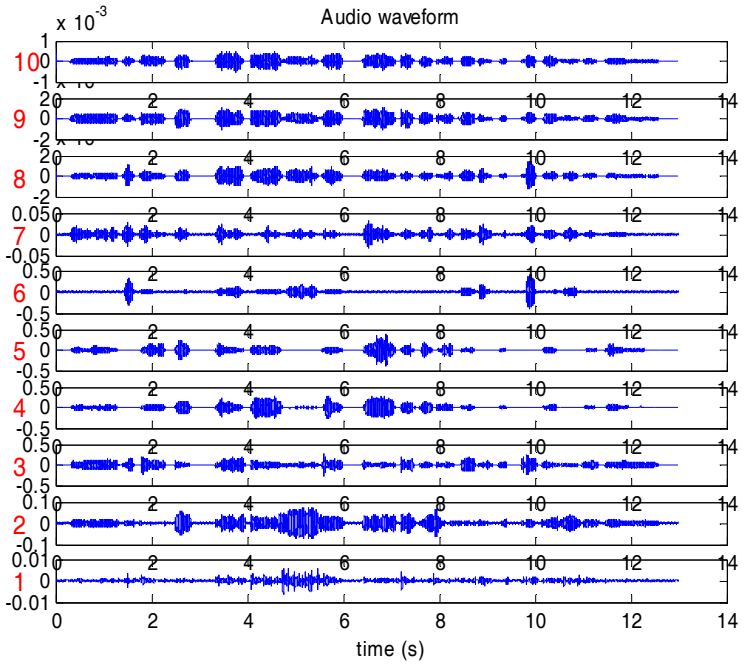

Figure 4.2

Decomposition into frequency channels

Frequency estimation was done using autocorrelation method. If we take a signal x, such as, for instance, the signal in the following audio file,

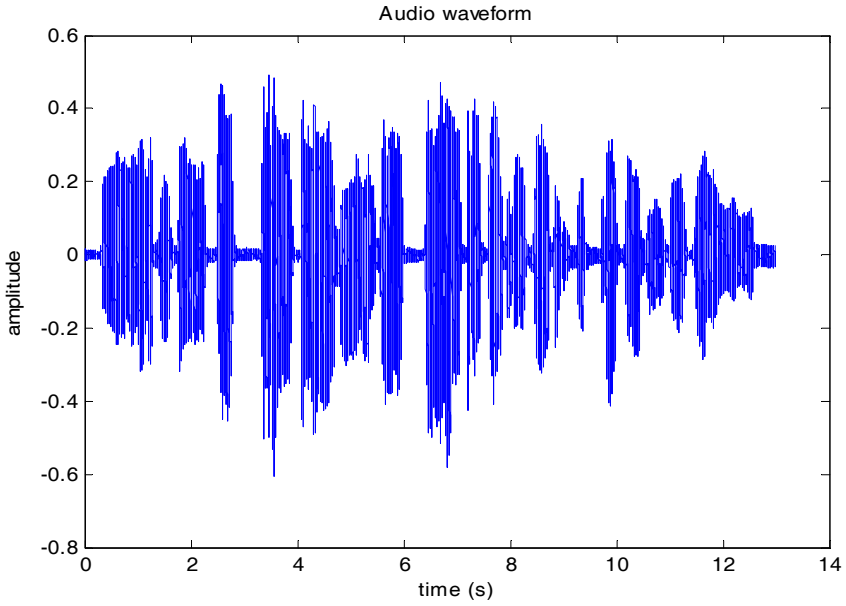

Figure 4.3

A sample audio waveform for tonic detection

The autocorrelation function is computed as

$$R_{xx}(j) = \sum_{n} x_n \overline{x}_{n-j}$$

For a given lag j, the autocorrelation $R_{xx}(j)$ was computed by multiplying point par point the signal with a shifted version of it of j samples. The result was the curve in Figure 4.4

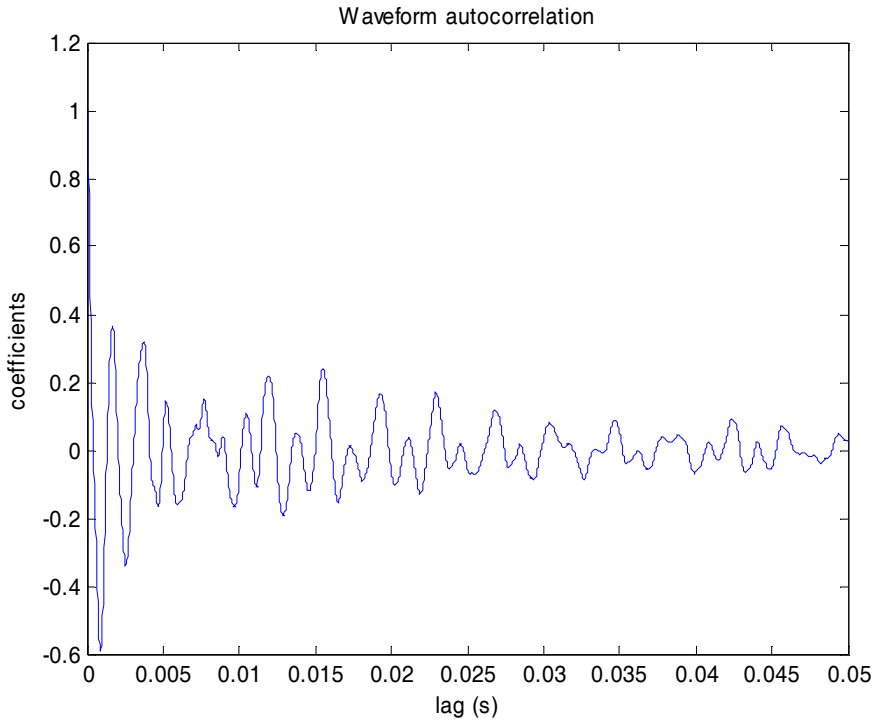

Figure 4.4

Resultant waveform of autocorrelation computation

When the lag j corresponds to a period of the signal, the signal is shifted to one period ahead, and therefore is exactly superposed to the original signal. Hence the summation gives very high value, as the two signals are highly correlated.

Since the input audio waveform was decomposed into channels using a filter-bank, after envelope extraction, these channels are summed back. After that, Peaks (or important local maxima) are detected. Then pitches are estimated to obtain frequencies in Hz. Note that pitch is not estimated for each separate channel. The channel decomposition is used

solely for the preliminary computations (the computation of autocorrelation function): channels are summed back before pitch extraction. Zero, one or several pitches can be detected for one frame. It depends on the peaks found in the autocorrelation function.

Figure 4.5 shows the graph where extracted pitch values (frequencies) are plotted against temporal location of beginning of the frames.

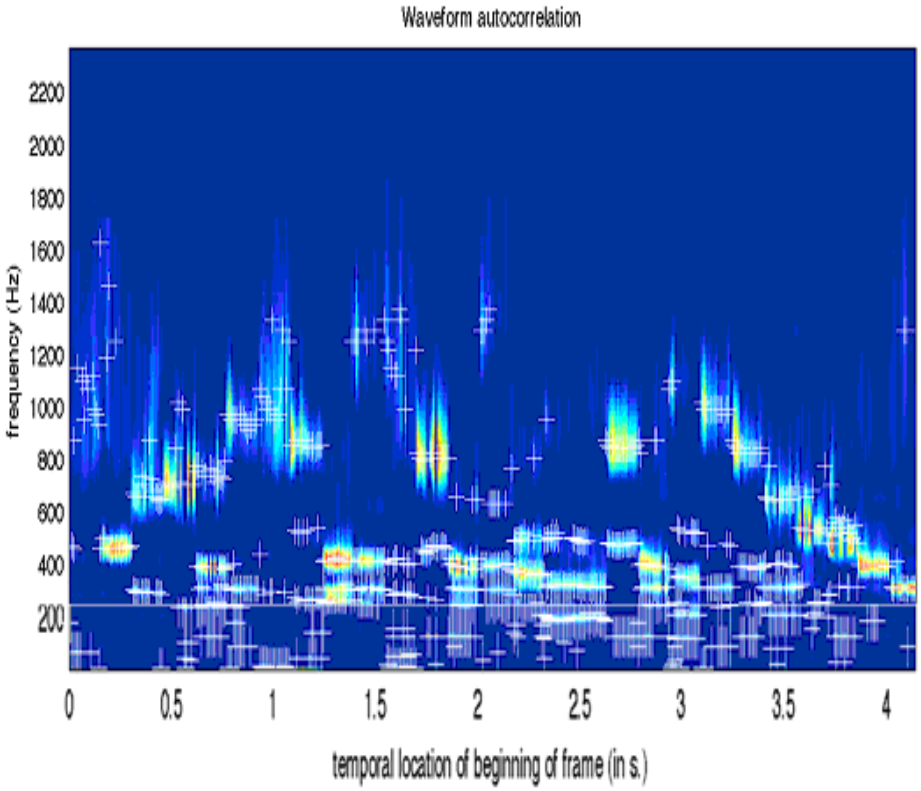

Figure 4.5

Graph of frequency Vs temporal location of beginning of frames

4.3.2 Phase II: Tonic Selection

In order to cover the entire range of tonic values that can be chosen by a Carnatic music performer, male or female, vocalist or instrumentalist, the theoretical range of frequency values is approximately 100-400 Hz. However, in our study it was purposefully set as 20-2000 Hz in order to practically test the authenticity of the range 100-400 Hz. Our studies reinforced the above theoretical range of 100-400 Hz authentically.

Tonic selection was done in two different ways: grouping by zero and histogram.

4.3.2.1 Grouping by Zero

The extracted frequencies included groups of nonzero frequency values separated by zeros. The musical piece may contain frequencies other than the tonic frequency indicating, probably, the presence of other notes. Hence, as a criteria for separating the tonic frequency, it was assumed that more than one zero value coming together indicated a note boundary. That is, when more than one zero occurred together, it indicated the gap between two notes. So the nonzero frequencies up to that point represented a note. In order to fix the correct frequency of the note, all the nonzero frequencies up to that point were considered as a group and analyzed. Most of these frequencies were having only slight differences in their values and hence an average of these frequencies seemed to be the immediate choice for the frequency of the note.

However, it was observed that there existed some very high and very low frequencies among these extracted frequencies. This could be

due to the various noises that can occur during a real performance. Due to the presence of these highly variant frequencies, the average differed highly from most of the frequencies. Obviously, average was not a good choice. In order to obtain a frequency value that represented most of the extracted frequencies in the group and to filter out the highly variant abnormal frequencies, another statistical measure median was chosen. Median of the frequencies in the group was computed and it was fixed as the frequency of the candidate (probable) tonic from that group. Similarly, candidate tonics from subsequent groups were also obtained. The process terminated when extracted frequencies in all groups were examined and candidate tonics from all groups were obtained. The result is an array containing the resultant candidate tonic values. Majority of these candidate tonics were almost the same with only negligible differences. However, possibility of at least a single extreme value could not be avoided. Hence, the median of these candidate tonics was taken as the tonic of the audio recording under analysis.

4.3.2.2 Pitch Histogram

As a cross-validation experiment, we also identified the tonic using histogram of the pitch values. The peak values among the generated frequencies 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 obtained an estimate of which pitches are repeated most often throughout the excerpt. Though pitch histograms have been used previously for tonic identification [41], they were constructed using only the most predominant pitch at each frame whereas we included each and every generated pitch values in the histogram computation. Peaks of the

histogram represented the most frequent pitches in the excerpt [Figure 4.6]. The peak corresponding to the tonic may not always be the highest peak in the histogram. So median of the most prominent peaks was taken as the tonic.

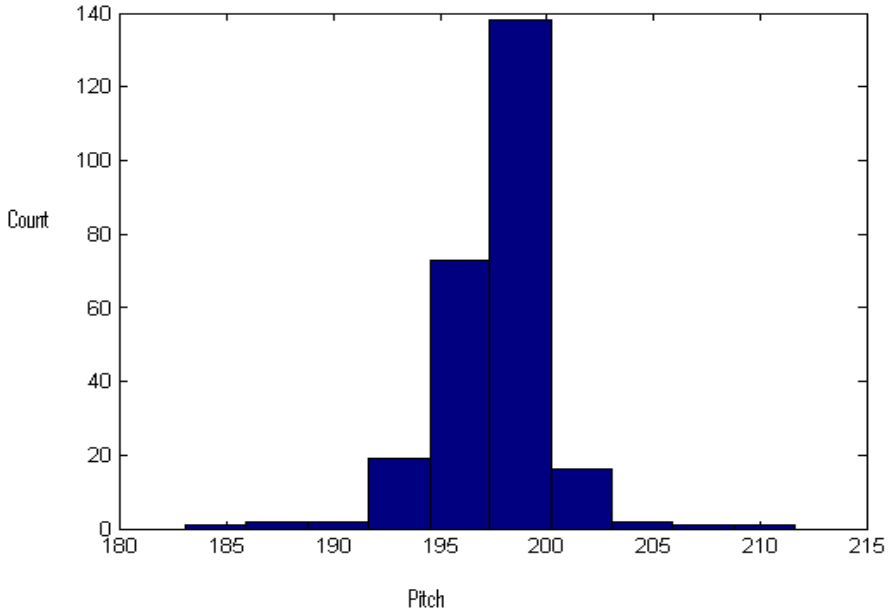

Figure 4.6

Histogram of pitch values

Figure 4.6 shows the histogram computed from a performance in the raga *Mayamalavagowla* by the great M S Subbalakshmi. The pitch axis is plotted in Hz, and the histogram is normalized by the magnitude of its highest peak. For the excerpt under consideration, we can see four peaks: between 190 and 205. Taking the median of the four peaks, we get the tonic value as 198 Hz.

4.4 SUMMARY

A novel method for tonic detection we developed has been presented in this chapter. This method is based on a multi-pitch analysis of the audio signal, in which all the predominant pitches in the input piece were used to construct a pitch histogram representing the most frequently played notes in the piece. The tonic was computed in two different ways: grouping by zero and pitch histogram. Our method also captured the notes played by the drone instrument and other accompanying instruments along with the pitch of the lead performer. That is, it was medium independent. It needed only a short excerpt of just 1 or 2 second duration as input. The method was tested successfully on a large collection of excerpts consisting of a wide range of ragas, artists and recording conditions, and was shown to obtain high tonic identification accuracy. Detailed test results and discussions are given in Chapter 7.