Improving singing voice detection in presence of pitched accompaniment

N. Santosh, S. Ramakrishnan, Vishweshwara Rao, and Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai 400076, India
Email: {nsantoshv, vishu, ramakrishnan, prao}@ee.iitb.ac.in

Abstract—This paper addresses the problem of singing voice detection in Indian Classical Music where we have presence of strongly pitched accompaniment. Visual observation of spectra suggests that the temporal fluctuation of the higher harmonics is a strong characteristic of the singing voice. Standard deviation of the frequency tracks, obtained from sinusoidal modeling of the harmonic signal, is proposed as a feature for the classification. In order to reliably compute the feature, accurate frequency estimation is required. We observe that phase based methods provide for superior sinusoid detection and estimation when compared to magnitude based methods. Further, the superiority of multi resolution spectrum analysis over conventional single resolution analysis is demonstrated for real musical signals.

I. INTRODUCTION

Music Information Retrieval (MIR) related applications (such as melody extraction and artist identification), often require the accurate detection of vocal segments in audio. In the context of Indian Classical Music (ICM), the simultaneous presence of other pitched, harmonically rich, spectrally overlapping musical instruments pose challenges to the vocal segment detection problem. Two of these instruments are the tanpura (drone) and the harmonium (secondary melodic instrument). The high signal-to-interference ratio (SIR) of the voice with respect to the tanpura (20-30 dB) and the large spectral spread of the tanpura signal cause individual harmonics to have too low an energy to cause a significant problem. On the other hand the strength of the harmonium partials can be comparable to those of the voice. The presence of a loud harmonium has been known to reduce the classification accuracy of a vocal segment detection system based on spectral shape based features [Vishu08, Rama08].

This paper investigates whether this difference in the frequency stability of harmonics for the voice and the harmonium can be exploited to classify each of the harmonics as belonging to voice or harmonium. Apart from vocal detection, such an approach can be very useful in source separation.

![Figure 1: Spectrograms of Harmonium (left) and Voice (right) signals shown in the frequency range 0 to 5 kHz over time duration of 4 secs.](image)

II. SYSTEM DESCRIPTION

A block diagram of the vocal segment detection system appears in Fig.2. The major stages are sine detection and estimation in each frame (i.e. every 10 ms) from the short-term spectrum of the windowed signal, linking sinusoids across frames to form tracks as in sinusoidal modeling of signals, computing a suitable feature that captures the temporal behaviour of each track across a fixed duration. Finally, a decision on whether singing voice is present at a given time is made considering the characteristics of all tracks present at the time.

1 The harmonium falls under the class of instruments that are incapable of continuous pitch variations and can only make discrete pitch jumps. Other instruments in this category are keyed instruments like the piano, the accordion and wind instruments like the flute and the saxophone.
Final detection accuracy can be enhanced by improving any or all of the blocks of the system. In the present work we focus on the problem of reliable and accurate sinusoid detection and estimation. Alternate methods are investigated and comparatively evaluated for their ability to discriminate voice partials from harmonium partials in a representative database.

A. Spectral representation

In order to capture the frequency fluctuation of harmonics it is necessary to first derive an intermediate representation for the time evolution of each harmonic. One such representation that is widely in use is the sinusoidal model, originally proposed by [McAul86]. Here individual harmonics are represented by tracks whose amplitude and frequency values vary with time.

For example a signal $s(t)$ in this representation is modeled as

$$s(t) = \sum a_i(t) \cos[\theta_i(t)] + \epsilon(t) \quad \ldots (1)$$

where, $a_i(t)$ and $\theta_i(t)$ are the instantaneous amplitude and the phase of the $i^{th}$ sinusoid, respectively, and $\epsilon(t)$ is the noise component at time $t$. In above representation R harmonics have been used.

Typically the amplitude and frequency parameters are estimated via the Short Time Fourier Transform (STFT) of the windowed signal at regularly spaced time instants. The STFT window duration has an important influence on the accuracy of the estimates. Usually it is constrained by the lowest frequency component present in the signal which is to be analyzed. To be able to analyze signals with low pitch, a large window is generally preferred. However, as observed earlier, non-stationary signals like voice have fluctuating pitch or fundamental frequency. The fluctuations become more pronounced in higher frequency harmonics because of the multiplying effect. In order to detect higher partials of harmonic sounds that exhibit frequency or amplitude modulation, Virtanen and Klapuri [Vir00] use a very large overlap between adjacent frames. But this still has the problem of main lobe distortion because of the non-stationary nature of the signal within a single frame of analysis. To overcome this problem Dressler has proposed the use of multi-resolution analysis [Dres06]. In multi-resolution analysis different window sizes are used at different frequency bands. A larger window is used in low frequency regions where good frequency resolution is desired and a short window used in high frequency regions where good time resolution is required to capture the fluctuations in harmonics.

To illustrate the advantage of multi-resolution analysis on real musical signals, Fig. 3 shows plots of harmonic tracks obtained using single resolution analysis and multi resolution analysis. Clearly the fast varying frequency components are missing in single resolution analysis due to missed detections of the sine peaks at the corresponding time instants.

B. Detection and estimation of sinusoids

For sinusoidal modeling we need to detect the sinusoids present and estimate their parameters. This is done by first picking peaks in the spectrum. A peak is said to exist in the spectrum if it has higher amplitude than its immediate neighbors. Accuracy of these peaks is limited because both the signal and its spectrum available to us are sampled. For a signal sampled at 22050 Hz computing 8192-point FFT gives a resolution of 2.69 Hz. Also many spurious peaks are identified which do not correspond to true harmonics in the signal. Further processing of identified peaks is required to overcome these problems. Available methods can be categorized as magnitude spectrum based methods and phase spectrum based methods. It has been observed that phase based methods give more accurate results than magnitude based methods [Kei02] for detection and estimation of sinusoids.

One of the widely used magnitude spectrum based methods is main lobe matching. To detect the presence of a sinusoid at the identified peak location, the criterion of sinusoidality, based on window main-lobe matching, as defined in [Gri88] is used. A relaxed sinusoidality threshold of 0.6 is used here so
as to not omit any true peaks. Although this increases the number of spurious peaks, it is expected that many of these will not survive the partial tracking stage. Further to increase the accuracy of the peak frequency and amplitude we use parabolic interpolation over a three point neighborhood i.e. for a better estimate of the frequency value of the \( k \)-th bin, the frequencies and amplitudes of the \( k-1 \), \( k \) and \( k+1 \) bins are used to arrive at parabolically interpolated value of frequency and amplitude.

Phase spectrum methods are based on the fact that frequency of an ideal continuous time sine wave can be obtained from its phase by finite difference [Kei02, Wells04, and Fulp06]. For a signal sampled at rate \( F_s \), if \( \Phi_1 \) and \( \Phi_2 \) are phases of DFT of two consecutive frames separated by \( R \) samples, then actual frequency at bin location \( k \) can be estimated as follows:

\[
\hat{f}_k = \frac{F_s}{2\pi} \frac{\Phi_2(k) - \Phi_1(k)}{R} \quad \ldots (2)
\]

In above expression \( \Phi_{unw} \) is the unwrapped version of \( \Phi_1 \). This unwrapping is required since phase can only be estimated modulo \( 2\pi \). Implementation of phase unwrapping is done by first computing the difference \( \Phi_2 - \Phi_1 \) and then mapping the result to \([-\pi, \pi]\). This will give errors if \( R \) is large. We use \( R = 1 \) in our implementation.

Sinusoidal detection used in phase based method is the weighted bin offset criteria suggested by Dressler [Dres06]. The main idea used here is that for a true harmonic actual peak location should be within certain vicinity of the detected bin peak location and instantaneous frequency of neighboring bins should be close to actual frequency. These conditions can be expressed as follows:

\[
\Delta k(k) < 0.7 \cdot (r + 1) \quad \ldots (3)
\]

\[
|\Delta k(k) - \Delta k(k \pm 1)| < 0.4 \cdot \frac{A_{peak}}{|X(k)|} \quad \ldots (4)
\]

where \( \Delta k(k) \) is the fractional bin offset of actual frequency location from the detected peak location \( k \) and is given by:

\[
\Delta k(k) = \frac{N}{2\pi r} \text{princarg} \left[ \frac{\Phi_2(k) - \Phi_1(k) - 2\pi rk}{N} \right] \quad \ldots (5)
\]

where princarg is the principle argument function which maps the phase to \([-\pi, \pi]\). In expression 3, the parameter \( r \) is the resolution parameter taking values 1, 2, 3 and 4 for successively smaller windows. \( X(k) \) is the DFT spectrum for a single frame and \( A_{peak} \) is instantaneous magnitude of the sinusoid.

Amplitude estimation is done using the method of main lobe correction [Kei02] which uses the estimated frequency value \( \hat{f}_k \) to correctly estimate its amplitude. If \( \Delta f \) is the frequency error then amplitude estimate is given by

\[
\Delta f = \frac{\text{abs} \left( \hat{f}_k - \frac{k}{F_s} \right)}{2} \quad \ldots (6)
\]

where \( W \) is the window spectrum.

C. Partial tracking

Once the peaks of the spectrum have been selected and their parameters determined they are tracked from one frame to the next. This process, called partial tracking (PT), helps to isolate the stable partials in the sound and in a sense represents the core of sinusoidal modeling technique.

In the present implementation of the partial tracking algorithm, the same algorithm as proposed by [McAul86] is used, with a single modification. In the original algorithm, conflicts between peaks to be picked for continuation of a peak in the previous frame are resolved by means of a cost function, which is made up the frequency difference between peaks in adjacent frames. However it was observed that the closest frequency to a track in a given frame is not always the one through which the track must pass. The amplitude must also be given consideration in the tracking decision as without this the high amplitude peaks that correspond to genuine partials may be completely missed in favour of other relatively low amplitude peaks [Sira99]. The present cost function (J) is given by

\[
J = \left| \log (A_{n+1} - A_{n}) \right| \quad \ldots (7)
\]

where \( \omega_o^n \) and \( A^n \) are the frequency and amplitude respectively of the \( n \)-th peak in the \( k \)-th frame. Also for a given peak in frame \( k \), competing peaks for track continuation in frame \( k+l \) must lie in a “matching interval” \( \Delta \). The \( \Delta \) chosen for track formation is one semitone, which was arrived at by observing harmonics of voice at different frequencies. Further all tracks whose durations are lesser than 60 ms are pruned out because it is assumed that such short duration tracks must have been formed by linking spurious peaks in the magnitude spectrum.

D. Feature description and extraction

Sinusoidal model representations of harmonium signal and of a relatively stable singing voice signal are shown in Fig. 4. Contrast between the instability of the voice tracks, especially at higher frequencies, and the almost straight-line nature of the harmonium tracks is clearly visible. Similar to standard deviation (SD) of harmonic structures [ZhaZha05], it appears here that the SD of the voice tracks would be greater than that of harmonium tracks and it should be possible to distinguish the two from each other. In polyphonic signals however, it was observed that there were cases where a harmonium track in the close vicinity of a voice track got linked to the voice track during track formation, and using the SD of the entire track to group the track would not be advisable. Instead here the SD is computed over short-duration segments (500 ms) with a hop of 100 ms. For a track with instantaneous frequencies \( f(n) \) (where \( n \) is the frame number), the standard deviation is given by

\[
SD = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} [f(n) - \bar{f}]^2} \quad \ldots (8)
\]
where \( \bar{f} \) is the mean of \( f(n) \) where we have \( N \) consecutive frequency estimates in the segment considered.

![Figure 4: Tracks of harmonium (top) and voice (bottom). Harmonium tracks can be seen to be very stable while those of voice show fluctuations, especially at higher frequencies.](image)

### III. EXPERIMENTS AND RESULTS

#### A. Evaluation criteria

Evaluation is carried out using the Shannon Mutual Information (MI) between the SD feature and the output class. We first give a brief introduction to MI. Let probabilities of the different classes be \( P(c) \), \( c = 1, 2, \ldots, N_c \) where \( N_c \) classes are present. Initial uncertainty in the output class is measured by the entropy.

\[
H(c) = -\sum_{c=1}^{N_c} P(c) \log_2 P(c) \quad (9)
\]

The average uncertainty after knowing the feature vector \( f \) is given by

\[
\begin{align*}
H(c/f) &= -\sum_{c=1}^{N_c} P(c) \left[ \sum_{c=1}^{N_c} P(c/f) \log_2 P(c/f) \right] \\
&= -\sum_{c=1}^{N_c} P(c) \log_2 P(c/f) - H(f) \quad (10)
\end{align*}
\]

where, \( P(c/f) \) is the conditional probability for the class \( c \) given the input vector \( f \). In general the conditional entropy is less than or equal to the initial entropy. The amount by which uncertainty is reduced after knowing the feature vector is mutual information

\[
I(c:f) = H(c) - H(c/f) = \sum_{c=1}^{N_c} P(f/c)P(c) \log_2 \frac{P(f/c)}{P(f)} \quad (11)
\]

Higher the value of mutual information for a given feature better is the discrimination achieved between classes using that feature. In the current context the final mutual information values are normalized with respect to the maximum entropy so that its maximum value is one and minimum value zero.

In our problem we have two classes namely harmonium and voice and a single dimensional feature vector of standard deviation. To compute MI, we need to know the distributions of class and feature. Without assuming any conditions on the signal, the voice and harmonium classes are considered to be equiprobable. We now obtain the conditional probability mass functions for the SD feature.

For each harmonium and voice class, from the tracks data we compute SD values and obtain the histogram for each class. A Gaussian Mixture Model (GMM) is fitted to each histogram in order to obtain smooth distributions. The GMM model is a continuous function while for computing mutual information we need mass function (here we are implicitly assuming that the SD feature is discrete). This is obtained by sampling the continuous time distribution and normalizing so that it sums to one. In our implementation we have used two mixture Gaussian models for modeling track variances.

#### B. Audio signals for experiments

Pure monophonic harmonium and singing voice signals are used in our experiments. Harmonium recording consists of some individual notes and an octave played continuously. The voice signal consists of single sung vowel /a/. It consists of both types of singing i.e. with the singer holding the pitch steady on a note, and passages where the pitch varies rapidly. All signals are sampled at the rate of 22.05 kHz. For our evaluation we have used about 25 seconds each of harmonium and voice data.

#### C. Experiment 1

We first compare performance of magnitude and phase based methods at the sine detection and estimation stage when a short analysis window of 20ms is used. The sine frequency estimates so obtained are used to compute the standard deviation feature values of each track that is formed by partial tracking. Depending on the mean frequency, the tracks are separated into 3 frequency regions – low (0 – 1000 Hz), mid (1000 – 2500 Hz) and high (2500 – 5000 Hz). The MI of the standard deviation feature is measured separately in each region, and the results appear in Table 1. We note the superiority of the phase based method.

<table>
<thead>
<tr>
<th>Detection and estimation method</th>
<th>Low</th>
<th>Mid</th>
<th>High</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude based</td>
<td>0.58</td>
<td>0.73</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td>Phase based</td>
<td>0.71</td>
<td>0.77</td>
<td>0.74</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 1: MI values for magnitude based and phase based methods using a short 20ms window

#### D. Experiment 2

In this experiment, we compare the performance of single resolution analysis versus multi-resolution analysis in the sine detection and estimation stage. A single resolution analysis is performed using a 50 ms window and using the phase method for detection and estimation. Mutual Information computed is shown in table 2. Next we perform a multi resolution analysis for same data using four different window sizes of 20ms, 27ms, 37ms and 50ms. Detection and estimation is done using phase method. Tracks thus formed are used to compute SD feature separately over the three different frequency regions as before. The results appear in Table 2.
Results in Table 1 suggest that phase based methods are better than magnitude based methods when using a small window for analysis. This is because using a short window and performing main lobe matching gives more errors in frequency estimation. Fig.5 shows tracks of a harmonium signal as estimated using phase and magnitude methods. Main lobe matching gives large amount of spurious variations in the estimate which are inherently not present in the harmonium tracks.

From Table 2, we can see there is an increase in MI value for mid, high and overall frequency bands when going from a single resolution analysis to multi resolution analysis. This is because multi resolution analysis allows us to capture fast frequency variations present in higher harmonics of voice which cannot be done in single resolution analysis as previously seen in Fig.3. Also observe that increase in MI value is highest for high frequency region. This is expected since maximum degradation occurs in this region while using a single long window for analysis.

In our analysis we have restricted the smallest window used to 20ms which is much larger than smallest window used by Dressler [Dres06]. Also, to analyze even higher frequency components (>5000 Hz) a further reduction of window size may be required. Smallest window used is usually constrained by the minimum separation between consecutive partials in a signal. Smaller the window, wider is the main lobe causing increased interference between adjacent windows. This problem is even more pronounced if the neighboring partials are of very different amplitudes. This depends purely on the characteristics of the signal under consideration. To overcome this problem one can do pre-processing so as to ‘whiten’ the spectrum and then perform multi resolution analysis on it.

V. CONCLUSION

Detecting the occurrence of singing voice in the presence of strongly harmonic accompanying instruments such as the harmonium is a challenging problem. Purely spectral timbre based features do not provide the needed discriminability. In this work, we have proposed a feature that characterizes the temporal fluctuation of the estimated frequencies of individual short duration harmonic tracks. An evaluation using the mutual information measure on a representative database reveals that the standard deviation feature has the potential to discriminate between the two with reasonable accuracy. Phase based sine detection and estimation methods are observed to perform better than a magnitude based method. Multi-resolution spectral analysis is provides better frequency estimates and therefore is better able to capture frequency variation of the signal harmonics as demonstrated by the improvement in mutual information value of the feature. These improvements in mutual information are expected to influence the vocal detection accuracy achieved by the overall system. Future work includes further enhancement of the remaining system blocks, research on further features that could complement the current feature and the evaluation of performance on larger data sets comprised of polyphonic audio.

REFERENCES


<table>
<thead>
<tr>
<th>Analysis method</th>
<th>Low</th>
<th>Mid</th>
<th>High</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single resolution</td>
<td>0.72</td>
<td>0.71</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>Multi resolution</td>
<td>0.71</td>
<td>0.74</td>
<td>0.74</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 2: MI values for single resolution and multi-resolution based analysis using phase based method for detection and analysis

Figure 5: Harmonium tracks obtained using a short window of 20ms. (Top) tracks obtained from main lobe matching method show variations in their estimates which are not inherently present in the signal. (Below) phase based method gives more realistic stable estimates.