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*— The paper addresses the problem of singing voice detection in Indian Classical Music where we have the 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 signals, 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 [1, 2]. This paper targets the problem of detecting singing voice in presence of *harmonium*.

On comparing the spectrograms of a *harmonium* signal to a voice signal (Figure 1) we can see that the harmonics of the *harmonium* appear to be very stable (in frequency) up to high frequencies (5 kHz) but the voice harmonics show increasing temporal fluctuation at higher frequencies. From a production perspective, the *harmonium* pitch is governed by the length and thickness of freely vibrating, metal reeds, which are fixed parameters for a given *harmonium*. So the harmonics for a given *harmonium* pitch are expected to be very stable in frequency. For singing however, it has been observed that the use of *vibrato*, a periodic, sinusoidal modulation of phonation frequency, during sustained notes is common. This will lead to a natural instability in the voice harmonics that increases with increase in frequency. Even for sung notes held without vibrato there will be period-to-period variations in glottal pulse duration, called jitter, which lead to the naturalness in the perception of the voice. The amount of jitter is usually about 0.5 – 1.0 % of the pitch period and so we can expect this jitter to have increasingly larger values at higher harmonics.

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.

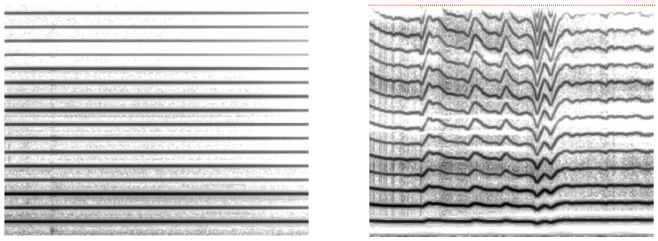

Figure 1: Spectrogram of *Harmonium* (left) and Voice (right) signals shown in the frequency range 0 to 5 kHz over time duration of 4secs.

II. System Description

A block diagram of the vocal segment detection system appears in Figure 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 behavior 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.

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.

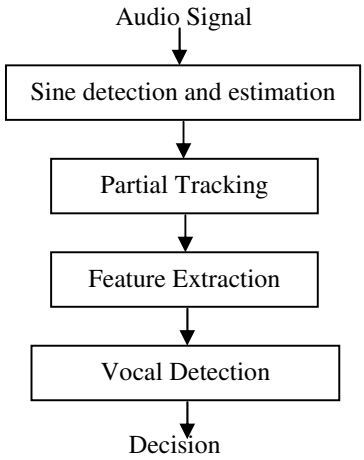

Figure 2: Block diagram of the system

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 Mcaulay and Quatieri [3]. 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

$$s(t) = \sum_{l=1}^{L} A_l(t)\cos[\theta_l(t)] + e(t) \qquad \ldots (1)$$

where, $A_l(t)$ and $\theta_l(t)$ are the instantaneous amplitude and the phase of the l^{th} sinusoid, respectively, and $e(t)$ is the noise component at time t. In above representation L 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. 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 [4] 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 [5] has proposed the use of multi-resolution analysis. 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, Figure 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.

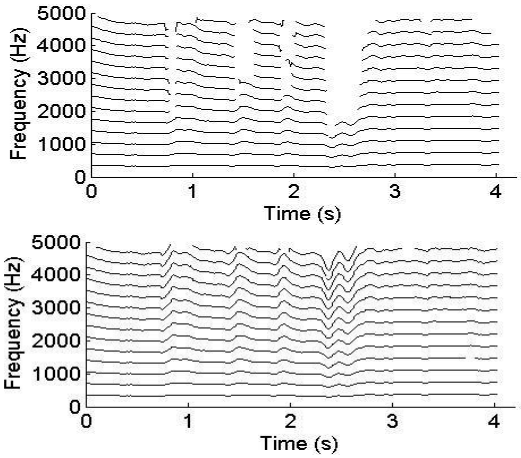

Figure 3: (Top) Tracks of harmonics of a voice signal obtained on performing single resolution analysis using a 50ms window. (Bottom) Tracks for the same signal obtained on performing a multi-resolution analysis using four different windows

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 [6] that phase based methods give more accurate results than magnitude based methods 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 by Griffin and Lim [7] 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 time derivative of its phase. A simple extension of this in discrete domain is to approximate continuous time derivative by finite difference [6, 8 and 9]. For a signal sampled at rate Fs, if φ_1 and φ_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:

$$\hat{f}_0 = \frac{F_s}{2\pi}\frac{\varphi_{2u}(k) - \varphi_1(k)}{R} \quad \ldots (2)$$

In above expression φ_{2u} is the unwrapped version of φ_2. This unwrapping is required since phase can only be estimated modulo 2π. Implementation of phase unwrapping is done by first computing the difference $\varphi_2 - \varphi_1$ and then mapping the obtained 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 [5]. 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:

$$\Delta\kappa(k) < 0.7 * (r+1) \quad \ldots (3)$$

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

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

$$\Delta\kappa(k) = \frac{N}{2\pi R}\, princ\arg\left[\varphi_2(k) - \varphi_1(k) - \frac{2\pi Rk}{N}\right] \ldots (5)$$

where princarg is the principle argument function which maps the phase to $[-\pi\ +\pi]$. In expressions (3) and (4), 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 using N-point FFT and A_{peak} is instantaneous magnitude of the sinusoid. If conditions (3) and (4) are satisfied by bin k, the frequency estimate is obtained as given in equation (2).

Amplitude estimation is done using the method of main lobe correction [6] which uses the estimated frequency value $\hat{f}_0$ to correctly estimate its amplitude. If Δf is the frequency error then amplitude estimate is given by

$$\Delta f = abs\left(\hat{f}_0 - \frac{k}{N}F_s\right) \quad ; \quad \hat{a} = 2\frac{|X(k)|}{\left|W(2\pi\frac{\Delta f}{F_s})\right|} \quad \ldots (6)$$

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 Mcaulay and Quatieri [3] 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 we note that 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 closer but relatively low amplitude peaks [10]. The present cost function (J) is given by

$$J = |(\omega_n^{k+1} - \omega_m^k) * \log(A_n^{k+1} - A_m^k)| \quad \ldots (7)$$

where ω_n^k and A_n^k 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 $k+1$ must lie in a "matching interval" Δ. The Δ chosen for track formation is one semitone, which was arrived at by observing harmonics of voice at different frequencies.

D. Feature description and extraction

Spectrograms of *harmonium* and voice signal are shown in Figure 1. 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 even when voice is held steady. Similar to standard deviation (SD) of harmonic structures [11], 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 of 200ms. For a track with instantaneous frequencies f(n) (where n is the frame number), the standard deviation is given by

$$SD = \left(\frac{1}{M-1}\sum_{n=1}^{M}[f(n) - \bar{f}]\right)^{1/2} \quad \ldots (8)$$

where $\bar{f}$ is the mean of f(n) over M consecutive frequency estimates in the segment considered. To decide whether voice is present in a given 200ms segment, SD of the portion of track within that segment is calculated for all the tracks present and if at least one of the SD value exceeds a certain SD threshold value the segment is declared to be voiced.

III. Experiments and Results

A. Audio signals for experiments

Individually recorded 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/. For our evaluation we have used about 25 seconds each of *harmonium* and voice data. This data consists of different 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.

B. Evaluation criteria

Evaluation of the system in Figure 2 is carried out after different stages. An evaluation is done after sine detection and estimation stage using the criteria of number of true hits vs. false alarms and the average frequency error. For the monophonic signals, the true frequency estimates (i.e. ground

truth data) are obtained from the pitch extracted from an accurate pitch tracking algorithm [1] and its integer multiples up to a maximum frequency of 5000Hz. The estimated sine frequencies are compared with these for evaluation. A sine is said to be true hit if its estimated frequency lies within ±25 Hz of a ground truth frequency, otherwise it is declared to be a false alarm. For every true hit the error in frequency is computed as the difference between ground truth frequency and the estimated frequency. Average error is then computed over all the true hits.

The second evaluation is done after feature extraction stage 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…Nc where *Nc* classes are present. Initial uncertainty in the output class is measured by the entropy.

$$H(c) = -\sum_{c=1}^{N_c} P(c) * \log P(c) \qquad \ldots (9)$$

The average uncertainty after knowing the feature vector *f* is given by

$$H(c/f) = -\sum_{f} P(f) * \left[\sum_{c} P(c/f) * \log P(c/f) \right] \quad \ldots (10)$$

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

$$I(c;f) = H(c) - H(c/f) = \sum_{c,f} P(c/f)P(c) \log \frac{P(c/f)}{P(f)}$$

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.

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 imposing any conditions on the signal, the voice and *harmonium* classes are considered to be equi-probable. 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 the SD.

C. *Experiment 1*

We first compare performance of magnitude and phase based methods after the sine detection and estimation stage when a short analysis window of 20ms is used. Table 1 shows the true hits vs. false alarms and average frequency error using the two different methods for both *harmonium* and voice signals.

The frequency estimates are used to compute the standard deviation from the tracks. Depending on the center 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 2.

Table 1: True hits vs. false alarms and average frequency error (Hz) for magnitude based and phase based methods using a 20ms window

Signal →	Harmonium		Voice	
	Mag	Phase	Mag	Phase
No. Peaks	33113	33113	30414	30414
True Hits	32943	32946	28878	28585
False Alarms	4489	842	6856	2869
Avg Freq Err	1.43	0.72	2.90	1.97

Table 2: MI values for magnitude and phase methods using a short window of 20ms

Detection and estimation method	Low	Mid	High	Overall
Magnitude based	0.54	0.70	0.69	0.63
Phase based	0.71	0.75	0.73	0.68

D. *Experiment 2*

In this experiment, we compare the performance of single resolution analysis versus multi-resolution analysis. A single resolution analysis is performed using a 50ms window and phase method for detection and estimation. A multi resolution analysis is performed for same data using four different window sizes of 20ms, 27ms, 37ms and 50ms. Detection and estimation is again done using phase method. Table 3 shows the number of true hits and false alarms for *harmonium* and voice signals when using different analysis methods. Average frequency error is also shown. Tracks formed in partial tracking stage are used to compute SD feature separately over the three different frequency regions as before. The MI computed is tabulated in table 4.

Table 3: True hits vs. false alarms and average frequency error (Hz) for single resolution and multi resolution analysis methods

Signal →	Harmonium		Voice	
	SingleRes	MultiRes	SingleRes	MultiRes
No. Peaks	33113	33113	30414	30414
True Hits	33021	32979	26902	28556
False Alarms	8463	4352	20853	6301
Avg Freq Err	0.55	0.63	1.33	1.76

Table 4: MI values for single resolution and multi-resolution analysis

Analysis method	Low	Mid	High	Overall
Single resolution	0.68	0.68	0.65	0.65
Multi resolution	0.66	0.72	0.74	0.68

E. *Experiment 3*

The third experiment is performed on a signal obtained from sample by sample mixing of the monophonic voice and *harmonium* signals. The signals are mixed with voice to *harmonium* signal energy ratio of 5dB which is typical in ICM. Tracks are obtained for this mixed data using magnitude and phase methods. These mixed tracks are then manually labeled as either belonging to voice or *harmonium* by comparing them with monophonic tracks data. The labeled tracks are then used as before to compute the SD feature and the obtained MI values are shown in table 5.

Table 5: MI values computed from mixed tracks data using methods of magnitude and phase

Detection and estimation method	Low	Mid	High	Overall
Magnitude based	0.25	0.47	0.58	0.42
Phase based	0.22	0.47	0.62	0.47

F. *Experiment 4*

The above mixed signal is used to demonstrate the classification ability of the SD feature. For detecting presence of voice in the mixed signal the method described in section II.D is used with a SD threshold of 3Hz. For the given test signal, a voice detection accuracy of 100% was observed for all the methods.

IV. Discussion

From Table 1 it is seen that both magnitude and phase method give approximately the same number of true hits but phase method gives smaller number of false alarms. Also the average frequency error is smaller for phase method. These improvements at the detection and estimation stage lead to an increased mutual information value for the phase method as can be seen from Table 2. This increase in MI values suggest that phase based methods are better than magnitude based methods when using a small window for analysis.

From Table 4, 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 Figure 3. This is also reflected in Table 3 where we have an increased number of true hits for voice signal when multi resolution analysis is used. 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 long window for analysis.

The MI values in Table 5 are obtained from analysis of mixed signal containing both voice and *harmonium*. The overall improved MI value indicates that even when voice and *harmonium* are simultaneously present, the phase method can give better discrimination between the tracks. Also observe that the MI values in Table 5 when compared to those in Table 2 have reduced because the partial detection/estimation accuracies are adversely affected when moving from monophonic to polyphonic signals.

In our analysis we have restricted the smallest window used to 20ms which is much larger than smallest window used by Dressler. 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 different amplitudes. To overcome this problem one can do a data adaptive analysis where the analysis window is chosen based on the local nature of the signal spectrum [12].

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 discriminablity. 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 of the system at different stages using the criteria of true hits vs. false alarms, average frequency error and 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 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 as demonstrated by experiment over a small data set. 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

[1] Vishweshwara Rao, S. Ramakrishnan and Preeti Rao, “Singing Voice Detection in North Indian Classical Music”, in Proc. Of the National Conference on Communications (NCC 2008), 2008

[2] S. Ramakrishnan, “Vocal segment extraction for classical music,” *Master’s Thesis*, Indian Institute of Technology Bombay, 2008.

[3] R. Mcaulay and T. Quatieri, “Speech analysis/synthesis based on sinusoid representation,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, pp. 744-754, 1986.

[4] T. Virtanen and A. Klapuri, “Separation of harmonic sound sources using sinusoidal modeling,” in *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP '00)*, Turkey, 2000.

[5] K. Dressler, “Sinusoidal extraction using an efficient implementation of a multi-resolution FFT,” in *Proc. 9th Intl. Conf. on Digital Audio Effects (DAFx-06)*, Montreal, 2006.

[6] F. Keiler and S. Marchand “Survey on extraction of sinusoids in stationary sounds,” in *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, 2002.

[7] D. Griffin and J. Lim, “Multiband Excitation Vocoder,” *IEEE Trans. On Acoustics, Speech and Signal processing*, vol. 36, no. 8, pp. 1223-1235, 1988.

[8] Sean A. Fulop and Kelly Fitz, “Algorithms for computing the time – corrected instantaneous frequency (reassigned) spectrogram, with applications,” J. Acoust. Soc. Am. 119(1), January 2006

[9] G. Wells, “Reading the sines: Sinusoidal identification and description using the short time Fourier transform,” *Music Technology Forum*, University of York, UK, 2004.

[10] S. P. Sira, “Sinusoidal modeling of music signals for audio coding,” *M.Tech dissertation*, Indian Institute of Technology Kanpur, 1999.

[11] Y. Zhang, and C. Zhang, “Separation of Voice and Music by Harmonic Structure Stability Analysis,” in *Proc. of IEEE Intl. Conf. on Multimedia and Expo (ICME 2005)*, 2005.

[12] D. L. Jones and T. W. Parks, “A high resolution data-adaptive time frequency representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, ser. 38, vol. 12, pp. 2127-2135, Dec. 1990.