VOCAL MELODY DETECTION IN THE PRESENCE OF PITCHED ACCOMPANIMENT USING HARMONIC MATCHING METHODS

Vishweshwara Rao and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology Bombay,
Powai, Mumbai 400076, India
{vishu,prao}@ee.iitb.ac.in

ABSTRACT

Vocal music is characterized by a melodically salient singing voice accompanied by one or more instruments. With a pitched instrument background, multiple periodicities are simultaneously present and the task becomes one of identifying and tracking the vocal pitch based on pitch strength and smoothness constraints. Frequency domain harmonic matching methods can be applied to detect pitch via the harmonically related frequencies that fit the signal's measured spectral peaks. The specific spectral fitness measure is expected to influence the performance of vocal pitch detection depending on the nature of the polyphonic mixture. In this work, we consider Indian classical music which provides important examples of singing voice accompanied by strongly pitched instruments. It is shown that the spectral fitness measure of the two-way mismatch method is well suited to track vocal pitch in the presence of the pitched percussion with its strong but sparse harmonic structure. The detected pitch is further used to obtain a measure of voicing that reliably discriminates vocal segments from purely instrumental regions.

1. INTRODUCTION

A rough definition of the melody of a song is the monophonic pitch[1] sequence that a listener might reproduce if asked to hum a piece of polyphonic music [1]. This pitch sequence is usually manifested as the F0 contour of the lead or dominant musical instrument in the polyphonic mixture, which is why the term melody detection is used interchangeably with predominant F0 estimation. The recent explosion of music on the internet has led to a surge in CB-MIR research, which in turn has caused a significant increase in interest in the melody detection problem. Although it may seem that melody detection is nothing but the first iteration in the process of multiple-F0 estimation, they differ in that the former, along with the estimation of the predominant F0, is also required to identify locations where the dominant instrument is present in the polyphony [2] (known as the voicing estimation problem for singing).

Different approaches to melody transcription were comparatively evaluated in [1]. It was found that most of the approaches fall into the framework shown in Figure 1. Here, the first block titled multi-pitch extraction identifies a set of candidate pitches that appear to be present at a given time. The melody identification block identifies which of these candidate pitches (if any) belong to the melody line. Finally the raw melody line is post-processed to increase smoothness and to remove spurious notes. The task of identifying whether the lead instrument is actually present at a given instant may be part of the system or be an independent process.

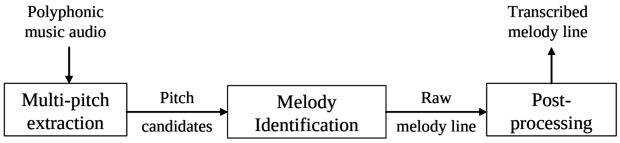

Figure 1: *General framework for melody transcription*

This paper investigates the melody detection problem in the context of Indian Classical Music (ICM), in particular, Indian classical singing. The accompaniment provided to the vocalist is very characteristic, with a continuously present drone and rhythmic percussion that is tonal and relatively strong. In a previous publication [3], a harmonic matching pitch detection algorithm (PDA) was experimentally found to be superior to the time domain autocorrelation PDA for vocal pitch detection in the presence of Indian percussion characterized by a sparse but strongly harmonic spectrum. Harmonic matching PDAs operate in the frequency domain and facilitate the explicit selection and weighting of spectral peaks in the predominant-F0 detection process. In the present work we look more closely at the working of harmonic matching PDAs for the problem of vocal F0 tracking in the presence of sparse tonal interferences. The melody identification and smoothing blocks are combined in a dynamic programming- based (DP) post-processor. Voicing detection is accomplished by a newly proposed pitch-derived feature.

In the next section we review two harmonic matching PDAs that have been developed for musical F0 tracking [4] [5]. While both methods are based on fitting a harmonic sequence to the measured signal spectrum, they differ on the important aspect of the error, or fitness, criterion. In Section 3, we give a brief overview of signal characteristics of the music under consideration. The performances of the two PDAs are compared for predominant F0 detection on simulated signals. The experimental analysis of the pitch accuracies is used to obtain insights on the role of the spectral error criterion. Next, real multi-track recordings are used to validate the results. Finally, a simple method for voicing detection based on the estimated predominant F0 is presented, followed by a discussion on the nature of the voicing errors and a proposed method to reduce these.

[1] Although the term pitch is known to be a perceptual attribute of sound and fundamental frequency, referred to as F0, is considered its physical correlate, both terms are used interchangeably throughout this paper to refer to the physical parameter.

2. PITCH DETECTION BY HARMONIC MATCHING

Harmonic matching PDAs are based on the frequency domain matching of measured spectrum with an ideal harmonic spectrum. Like most of the multi-pitch extractors described in [1], they make use of the magnitude of the short-time Fourier transform (STFT) for the initial processing. For all cases, we use a high-resolution FFT (8192 points at a sampling frequency of 22.05 kHz) computed from a Hamming windowed signal segment of fixed length chosen so as to reliably resolve the harmonics at the minimum expected F0 (four times the maximum expected time period). The two PDAs discussed in this section are the pattern recognition (PR) PDA [5] and the two-way mismatch (TWM) PDA [4]. The PR PDA belongs to the large family of "harmonic sieve" methods based on integrating evidence for every predicted frequency partial from the corresponding energy in the measured spectrum. The TWM PDA uses a less conventional spectral matching error function as discussed later in this section. Both PDAs were originally proposed for the musical F0 tracking problem and have been shown to provide highly accurate pitch detection for clean monophonic voices.

2.1. Pattern Recognition PDA

The PR PDA exploits the fact that for a logarithmic frequency scale, corresponding to musical intervals, a harmonic structure always takes on the same pattern regardless of the value of its F0. Consequently, a pattern recognition algorithm is applied to detect such patterns in the measured spectrum by correlation with ideal spectral patterns for different trial F0 values in the expected pitch range.

2.1.1. Initial Processing of STFT Magnitude

The PR PDA requires the frequency bins of the spectrum to be logarithmically spaced. It was originally recommended to compute the magnitude spectral values at logarithmically spaced frequency locations using cubic spline interpolation at a resolution of 24 points per octave [5]. We use an increased resolution of 48 points per octave.

2.1.2. Cross Correlation Score

For each trial F0, a cross correlation function is computed between the measured magnitude spectrum and an ideal spectrum represented by a logarithmically spaced impulse train of fixed number of components. If *X(f)* and *I(f)* are the signal and ideal magnitude spectra, with *M* frequency bins, respectively, then the cross correlation function is given as

$$C(\psi) = \sum_{f=0}^{M-1} I(f)X(f+\psi) \tag{1}$$

The local peaks in *C(ψ)* provide the pitch candidates for the melody detection, with the magnitude of the peak indicating the pitch strength or salience.

2.1.3. Parameter Tuning

In the original paper [5] the optimal number of components in the ideal spectrum, *I(f)*, was empirically selected for different musical instruments (e.g. flute – 4 components, violin – 11 components). In the present study the optimal number of components for each of the examples used in the subsequent experiments was determined as that which resulted in the optimal performance of the PR PDA for the clean (monophonic) singing voice signal.

2.2. Two-Way Mismatch PDA

The TWM PDA detects the F0 as that which minimizes a mismatch error, which is computed between a predicted harmonic spectral pattern and the measured peaks in the signal spectrum.

2.2.1. Initial Processing of STFT Magnitude

The magnitude spectra are reduced to a set of peaks corresponding to the sinusoidal partials only. The implementation of TWM in [6] achieves this by selecting all local maxima in the spectrum above a fixed threshold below the maximum peak. Due to the large dynamic range of the singing voice spectra, however, it is found that spurious peaks are detected corresponding to the window sidelobes. Additionally, the weak higher harmonics may escape detection. To improve selectivity, we incorporate the more effective measure of local sinusoidality to choose from the local maxima. Sinusoidality measures how closely the shape of a detected spectral peak matches the known shape of the window main lobe [7]. To include the higher harmonics that may undergo some shape distortion due the rapidly varying F0 in ICM we apply a relatively relaxed lower threshold (0.6) on sinusoidality as compared with the recommended threshold (0.8).

2.2.2. TWM Error

The overall TWM error function, for a given trial F0, is computed as shown below.

$$Err_{total} = Err_{p\to m}/N \; + \; \rho Err_{m\to p}/K \tag{2}$$

Here N and K are the number of predicted and measured harmonics respectively. The error $Err_{p\to m}$ is based on the mismatch between each harmonic in the predicted sequence and its nearest neighbour in the measured partials while $Err_{m\to p}$ is based on the frequency difference between each partial in the measured sequence and its nearest neighbour in the predicted sequence. The recommended value [4] [6] of ρ is 0.33. The locations of local minima in Err_{total} are then the possible F0 candidates.

Both of the above errors share the same form. $Err_{p\to m}$, which is the more important of the two [8], is defined below.

$$Err_{p\to m} = \sum_{n=1}^{N}\left[\frac{\Delta f_n}{(f_n)^p} + \left(\frac{a_n}{A_{\max}}\right)\left(q\frac{\Delta f_n}{(f_n)^p} - r\right)\right] \tag{3}$$

Here f_n and a_n are the frequency and magnitude of a single predicted harmonic. Δf_n is the difference, in Hz, between this harmonic and its nearest neighbour in the list of measured partials. A_{max} is the magnitude of the strongest measured partial. Thus an amplitude weighted penalty is applied to a normalized frequency error ($\Delta f/f$) between measured and predicted partials for the given trial F0. Recommended values of p, q and r are 0.5, 1.4 and 0.5 respectively [4] [6]. Higher values of "p" serve to emphasize low frequency region errors.

2.2.3. Parameter Tuning

Unlike originally proposed, here N is not fixed over all trial F0 but is computed as *round*(F_{max}/*F0*), where F_{max} is the upper limit above which the spectral content is not considered useful for voice F0 extraction (we select F_{max} =5 kHz). Additionally, it is found that using ρ = 0.2 favours the target voice fundamental, when the interference is characterized by a few partials only, by placing higher emphasis on $Err_{p\rightarrow m}$.

3. SIGNAL CHARACTERISTICS

A typical Indian classical vocal performance has three components: the voice, which carries the melody, the percussion, which provides a rhythmic framework and the drone.

3.1.1. The Indian Classical Singing Voice

Indian classical singing is characterized by the dominating presence of subtle, but sometimes rapid, modulations in the form of ornaments, embellishments, and pitch slides. Further, unlike Western music, where the scale steps are fixed and grounded in the tempered scale, the location of the scale steps in Indian classical music is variable for different singers and also over different performances by the same singer. This is because at the start of every performance, the singer tunes the tonic location according to his/her comfort level.

In terms of acoustic features, there is a large variability in the spectral distribution of energy across singers, perceived as differences in timbre. However, the locations of significant voice harmonics in the spectrum rarely cross 7 kHz.

3.1.2. Percussion: The Tabla

The *tabla* consists of a pair of drums, one large bass drum, the *bayan*, and a smaller treble drum, the *dayan*. *Tabla* percussion consists of a variety of strokes, often played in rapid succession, each labeled with a mnemonic. Two broad classes, in terms of acoustic characteristics, are: 1. tonal (pitched) strokes that decay slowly and have a near-harmonic spectral structure and 2. impulsive (unpitched) strokes that decay rapidly and have a noisy spectral structure. If we consider the signal of interest to be the singing voice, the local signal-to-interference ratio (SIR) can dip as low as -10 dB around a *tabla* stroke onset.

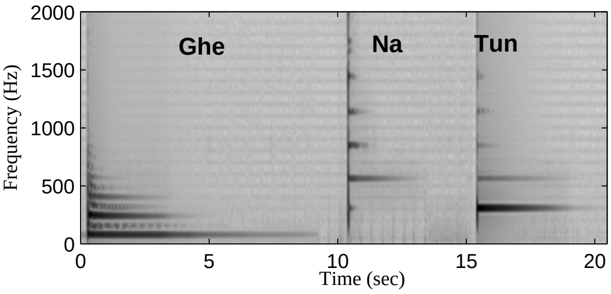

Figure 2: *Spectrogram of three typical tonal strokes (Ghe, Na and Tun) tuned to F0 = 283 Hz.*

From studying an extensive database of *tabla* strokes [9], it was found that while all the impulsive strokes had similar acoustic characteristics, there was a large variability in those of the different tonal strokes. The acoustic features of three typical tonal strokes, associated with the mnemonics *ghe*, *na* and *tun*, are compared by means of narrowband spectrograms in Figure 2. All the three strokes, soon after onset, exhibit harmonics that lie in the same frequency range as those of the singing voice. *Ghe* is produced by the *bayan*, and its harmonics all lie in a low frequency range. In contrast, the pitch of the *dayan* is tuned to the tonic of the singer prior to the performance, with the harmonics of its strokes (*na, tun*) occupying a higher region in the spectrum. The strokes *ghe* and *tun* have a more gradual decay than *na*, which decays quite rapidly, but still much slower than any of the impulsive strokes. The spectrograms of *ghe* and *na* exhibit up to five dominant harmonics for a brief period after the onset. *Tun*, on the other hand, is dominated by a single harmonic, giving it an almost sinusoidal timbre.

3.1.3. Drone: The Tanpura

The *tanpura* is an overtone-rich, stringed instrument, mainly pitched at the singer's tonic and additional strings at the fourth or the fifth, plucked continuously throughout the music performance. Its purpose is to provide an immediate pitch reference to the singer and the listeners. The SIRs for the voice with respect to the *tanpura* range from 20 to 30 dB. In spite of its low strength, the *tanpura* sound is audibly prominent due to the fact that its energy is nearly uniformly spread over harmonic partials throughout the spectrum up to 10 kHz. This leads to frequency bands dominated entirely by *tanpura* partials, thus enhancing its perceived loudness.

4. COMPARATIVE EVALUATION ON SIMULATED SIGNALS

In this section we compare the performances of the TWM PDA and the PR PDA in terms of robustness to sparse tonal interferences using synthetic target and interference signals, based on the signal characteristics described in the previous section. The simulation allows the signal characteristics to be varied systematically with the ground truth pitch known for evaluation. A pitch estimate is computed every 10 ms. The PDAs are operated within the framework of dynamic programming-based (DP) smoothing. DP uses a combination of suitably defined local measurement and smoothness costs into a global cost, which is optimized over a continuous voiced segment. Here the measurement costs are the TWM error (Err_{total}) for TWM and the correlation score (C) for PR, each normalized to lie in the interval [0, 1]. The smoothness cost is derived from the distribution of inter-frame pitch transitions over a training dataset of clean voice pitch contours [3].

4.1. Data

The target signal is a sustained vowel (/a/), generated using a formant synthesizer, at a sampling frequency of 22.05 kHz, with time-varying F0. In order to simulate the F0 variations in Indian classical singing and the typical vocal range of a singer (about 2 octaves), the time variation of the F0 of the synthetic vowel smoothly sweeps ± 1 octave from a chosen base F0 at a maximum rate of 3 semitones/sec. Two different target signals are synthesized using low (150 Hz) and high (330 Hz) values of base F0 respectively. The synthetic vowels have durations of 21 sec in which the instantaneous F0 completes six oscillations about the base F0.

Since the *tabla* is tuned to the tonic of the singer, we can expect interference partials at the harmonics of the tonic. The interference signals for each of the base F0s, are complex tones having 1, 3 and 5 equal magnitude harmonics at F0 equal to the target's base F0. The amplitude envelope of a sequence of *tun*

Table 1: *PA values (in %) for the PR and TWM PDAs, before and after DP, for the low and high target base F0s for increasing number of interference harmonics at fixed SIR.*

Interference	Base F0 = 150 Hz				Base F0 = 330 Hz			
	PR	PR-DP	TWM	TWM-DP	PR	PR-DP	TWM	TWM-DP
None	100.0	100.0	99.6	100.0	100.0	100.0	98.5	100.0
1 harmonic	70.8	68.1	92.5	100.0	69.9	74.2	92.3	100.0
3 harmonics	64.7	63.1	88.7	94.3	66.3	69.1	90.8	97.4
5 harmonics	62.8	61.4	86.9	93.4	65.1	70.2	86.6	93.7

strokes, each of which decays over 2 seconds, is superimposed on the complex tone. This results in 14 strokes over the target signal duration. These complex tones are added to the target signals such that the worst-case local SIR around the onset of any stroke is -10 dB. For each base F0, there are four cases: 1 clean vowel and 3 noisy vowels. Figure 3 shows spectrograms for the target with low base F0, the interference with 5 harmonics at the target base F0 and the mixed signal at -10 dB SIR.

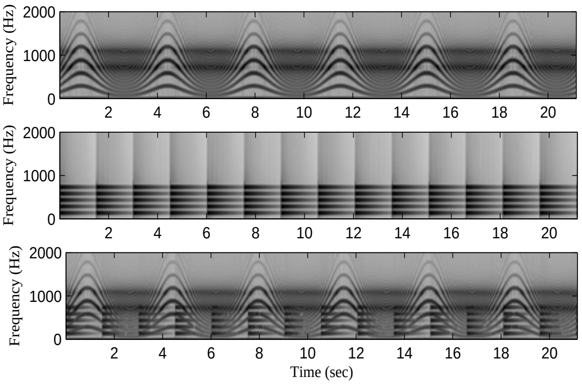

Figure 3: *Spectrograms of the target (top) at low base F0, the interference (middle) with 5 harmonics at the target F0 and the mixed signal (bottom) at -10 dB SIR. The target harmonics vary sinusoidally. The vertical lines in the interference spectrogram mark the onset of each stroke after which the harmonics of the interference decay.*

4.2. Experiment and Results

To keep the comparison between PDAs as fair as possible, the F0 search range is kept fixed for both PDAs for each target signal i.e. from 70 to 500 Hz for the low base F0, and from 150 to 700 Hz for the high base F0. The PDAs only use spectral content below 5 kHz, above which harmonic content in the voice is generally sparse and weak. F0 is estimated every 10 ms resulting in 2013 estimates for each target signal case. The PDA parameter settings for TWM were kept fixed for the low and the high base F0 targets, but for the PR PDA, the number of ideal spectral components is 10 and 6 for the low and high base F0 targets respectively. This is done to achieve optimal performance for the clean target.

The comparison of PDAs, based on the pitch accuracy (PA) values expressed as percentages, before and after DP, appears in Table 1. In the computation of PA, the estimated F0 values are treated as correct if they lie within 50 cents of the known target F0 [1]. We see from Table 1 that for the clean signals both the PDAs display very high PA values and the combination of both PDAs with DP-based post-processing results in 100% accuracy. This indicates that the PDAs are working under suitable parameter settings for the monophonic signals. The addition of a single harmonic, tonal interference, a close approximation of the stroke *tun*, results in a severe degradation of the PA values for the PR PDA but not for TWM, as indicated by row 2 of the Table. In the presence of the sparse tonal interferences, it is clear that the best results, indicated by the highest PA, are obtained for the combination of TWM and DP. Also, in this case we observe that the PA values tend to decrease with an increase in the number of harmonics at constant noise power.

4.3. Discussion

The results of Table 1 indicate two things. 1. At fixed SIR, the TWM PDA is least sensitive to harmonic interference when the number of interference partials is low and, 2. DP-based post-processing is able to correct the majority of errors in the TWM output caused by sparse tonal interferences.

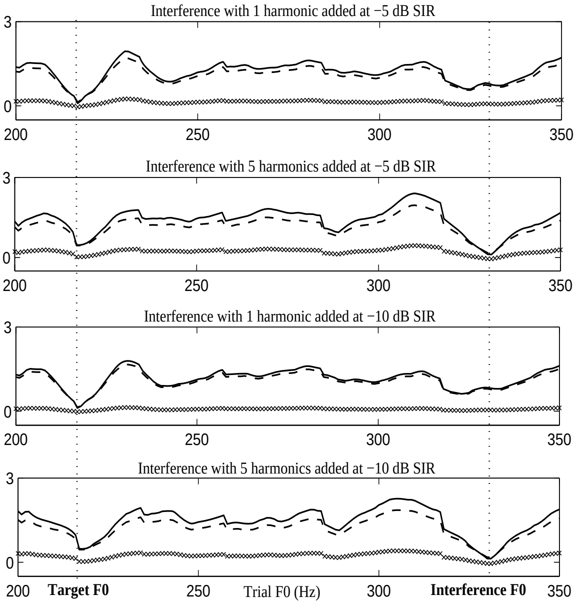

Figure 4: *Plots of Term1 (dashed curve), Term2 (crosses) and Errp→m (solid curve), vs. trial F0 for a single frame for the target at high base pitch for interferences with 1 and 5 harmonics added at -5 and -10 dB SIR*

4.3.1. Robustness to Sparse Tonal Interferences

The robustness of TWM to sparse tonal interferences and its sensitivity to the interference spectral structure can be attributed to the specific form of the TWM error, defined in Equations 2 and 3. $\mathrm{Err}_{p\to m}$ can be viewed as a combination of two terms as shown below.

$$Err_{p\to m} = term1 + term2$$

$$term1 = \sum_{n=1}^{N} \frac{\Delta f_n}{(f_n)^p}; \quad term2 = \sum_{n=1}^{N} \left(\frac{a_n}{A_{\max}}\right)\left(q\frac{\Delta f_n}{(f_n)^p} - r\right) \tag{4}$$

Term1, called the frequency mismatch error, is only affected by location of partials. That is, it is maximum when $\Delta f/f$ is large. *Term2* is affected by relative amplitudes of the partials further weighted by the frequency mismatch error leading to minimum error when $\Delta f/f$ is small and a_n/A_{max} is large. Therefore, for a given trial F0, specific emphasis is placed on the presence of harmonics at the expected frequency locations.

To illustrate the relative significances of the two component terms, consider Figure 4, which displays plots of *term1*, *term2* and $Err_{p\to m}$ against trial F0, for a single frame of a target signal at high base pitch to which are added interferences with 1 and 5 harmonics at -5 and -10 dB SIR. In this frame, the target F0 is 217 Hz while the interference F0 is 330 Hz. For all four cases, we can clearly see that $Err_{p\to m}$ is dominated by *term1* and *term2* is of lesser significance. The dominance of *term1*, which is only affected by partial locations, explains the robustness of TWM to sparse tonal interference.

For the interference with a single harmonic, the global minimum in $Err_{p\to m}$ occurs at the target F0, independent of SIR, and is much lower than the value of $Err_{p\to m}$ at the interference F0. This occurs because all the target harmonics result in low frequency mismatch terms but the numerous missing interference harmonics lead to large frequency mismatch terms irrespective of the overall strength of the interference. As the number of interference and target harmonics become comparable, the value of $Err_{p\to m}$ at the interference F0 decreases in value and the global minimum shifts to the interference F0, again independent of SIR. This occurs because now all the interference harmonics result in lower frequency mismatch.

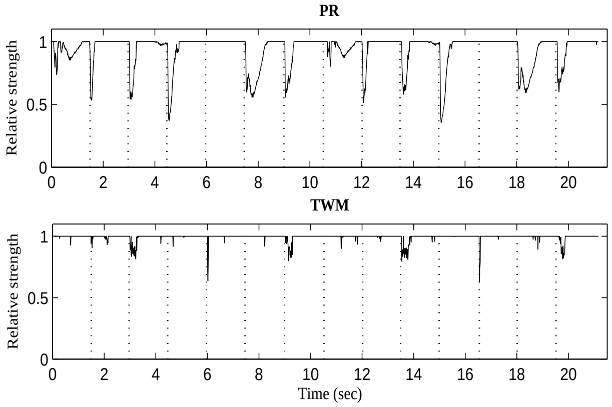

Figure 5: *Relative strength (RS) contours of the target F0 for PR (above) and TWM (below) PDAs for the target at low base F0 added to an intermittent interference with a single harmonic. Vertical dotted lines indicate location of synthetic stroke onset.*

There is a slight increase in the error at the target F0 due to some of the weaker target harmonics becoming distorted by interaction with the interference harmonic lobes in their close vicinity resulting in shifted or suppressed target harmonics. The low PA value of TWM for the case of the target at high base pitch combined with the interference having 5 harmonics is thus caused primarily by the number of interference harmonics, as compared to target harmonics, rather than their strengths.

In contrast, there is no clear trend in the PA values of the PR PDA with an increase in the number of interference harmonics. The PR PDA computes a similarity score that depends on the cross correlation between the actual spectrum and an assumed harmonic spectrum at the trial F0. The overall spectral match at a trial F0 is affected by the energy of the contributing harmonics, independent of whether the overall energy is concentrated in a few strong partials or distributed over several weaker partials.

4.3.2. Suitability to DP-based Post-Processing

In the context of predominant (melodic) F0 extraction, the suitability for dynamic programming-based post-processing is dependent on the quality of the pitch candidates extracted in each frame by way of their local strengths (i.e. measurement costs). The relative strength (RS) of a pitch candidate at the underlying melodic F0 is computed as indicated below

$$RS = 1 - \frac{MC_{tr} - MC_{mf}}{MC_{tr}} \tag{5}$$

where MC_{tr} and MC_{mf} are the measurement costs of the top-ranked and the true melodic F0 candidates respectively.

Since the identical smoothness cost [3] was used for both PDAs, a better performance indicates a superior measurement cost, or equivalently, better RS of the underlying melodic pitch. To confirm this, the RS of the true F0 is computed for each frame using Equation 5. If the target F0 is not present in the list of candidates then its RS is set to 0. Figure 5 displays the melodic pitch RS computed by each PDA across the signal duration for the case of the target with low base pitch and a single harmonic interference. We observe that the RS values of the PR PDA are severely degraded around the onset of most interference strokes (marked by vertical dashed lines). The corresponding degradation in target F0 RS for the TWM is relatively mild. This is consistent with its performance in terms of PA. The higher the relative strength of the candidate at the melodic pitch, the better is the potential for accurate reconstruction by DP-based post-processing, especially when the interference is not continuous.

5. EVALUATION ON REAL MULTI-TRACK MUSIC

In this section we validate the performance of the TWM-DP predominant F0 estimator, which was found to perform with a high degree of accuracy on synthetic signals in the previous section, on real music signals.

Table 2: *PA values (in %) for TWM-DP for Male/Female excerpts for clean voice, voice + tabla, voice+ tabla + tanpura.*

Audio content	Male	Female
Clean voice	99.55	99.76
Voice + *tabla*	99.51	99.51
Voice + *tabla* + *tanpura*	99.28	98.81

5.1. Data and Ground Truth

The voice, *tabla* and *tanpura* signals used here were obtained from multi-track data consists of two 1-minute excerpts from each of two different professional vocal performances (one male singer and one female singer). One excerpt is taken from the start of the performance where the tempo is slow and the other excerpt is taken towards the end of the performance where the tempo is faster and rapid modulations are present in the voice track. Three separate tracks (one for each of voice, *tabla* and *tanpura*) are available for each performance segment. To ensure time-synchrony and acoustic isolation for each instrument the performing artists were spread out on the same stage with considerable distance between them and recorded on separate channels simultaneously.

The availability of the relatively clean voice track facilitates the extraction of ground truth pitch, which is then used to evaluate the accuracy of the TWM-DP PDA on the corresponding polyphonic recording created by mixing at normally expected levels. Three different monophonic PDAs are each applied to the clean voice, and the estimated ground truth pitch at 10 ms intervals is determined by a majority vote. The PDAs used here are YIN [10], SHS [11] and TWM [3]. With each based on essentially different assumptions regarding the underlying signal periodicity, they tend to react differently to the different signal perturbations. For the purpose of evaluation, only the ground truth pitch estimates corresponding to voiced regions (i.e. the sung vowels, which comprise about 97 % of the vocal segments) are considered.

5.2. Experiments

For each voice excerpt, its time-synchronized *tabla* counterpart was added at an audibly acceptable, global SIR of 5 dB. Further the time-synchronized *tanpura* is added to the voice and *tabla* mixture such that the SIR for the voice with respect to the *tanpura* is 20 dB.

The first two rows of Table 2 show the PA values for TWM-DP (with respect to the "ground truth") on the clean voice and the mixture of voice and *tabla*. That the algorithm is robust to *tabla* interference can be clearly inferred by the almost similar, and also high, values of PA for both cases. The PA values of the TWM-DP algorithm on the *tanpura* included signal (Row 3 of Table 2) do not show any significant degradation and are still very high. Even though the *tanpura* signal is spectrally dense, a majority of its partials escape detection during voiced frames because of their very low strength and so are not involved in the TWM error computation.

6. VOICING DETECTION

The TWM-DP PDA produces an estimate of the predominant pitch at every instant irrespective of the underlying signal content. For any useful representation of the melody, it is necessary to find a means to automatically detect frames where the vocal signal is indeed present. An attempt was made at using the TWM error as an indicator of voicing [12], but it was found that the variability of the error was very large as compared to mean. Additionally it is expected that, in the present context, the TWM error takes comparable values in the voiced and unvoiced regions because of the tonal accompaniment (pitched percussion and continuous drone). We propose an alternative measure of voicing based on the signal energy associated with the predominant pitch estimate. Following the overall framework of [13], the output of a frame-level classifier is further smoothened over homogenous segments as determined by boundary detection.

6.1. Implementation

6.1.1. Frame-Level Classification.

Based on the confidence in pitch estimates provided by the high PA values in Table 2, we propose a new, pitch-related feature that is indicative of voicing called pre-dominant F0 harmonic energy. It is defined as the sum of the energies of individual harmonics, in the low frequency region up to 5 kHz, corresponding to the predominant pitch, and is given by

$$HE = \sum_{k=k_{F0}}^{k_{NF0}} \left|X[k]\right|^2 \qquad (6)$$

where *F0* is the detected fundamental frequency and *NF0* is the largest integer multiple of *F0* below 5 kHz. The spectral bin numbers (*k*) corresponding to the closest local maxima, within a 3% neighbourhood, of the expected harmonic location are used in the computation. The HE is normalized by its maximum attained value over a single musical performance. From Figure 6 (bottom), we can see that the HE values are seen to be high for voiced regions and low for instrumental regions. During tonal *tabla* strokes it is possible that the HE values will be high for short durations of time. We can also see the high variability in the TWM error (Figure 6 top) as mentioned in [12]. Clearly HE seems to be a better indicator of voicing than the TWM error.

The MAP classification rule was then applied to the 2-class (vocal/instrumental) problem with the distribution of the HE feature modeled separately for each class by a Gaussian mixture. The HE feature extracted from approx. 22 minutes of hand-labeled recordings of Indian classical vocal performances of various singers is used to train a Gaussian Mixture Model (GMM) with 4 mixtures for each class to account for the observed distinct modes of the underlying signal within each class. Also the prior probabilities of the vocal and instrumental classes were estimated to be 0.7 and 0.3 respectively. The frame-level decisions are then smoothened by grouping frames over boundaries of homogenous regions obtained as described next.

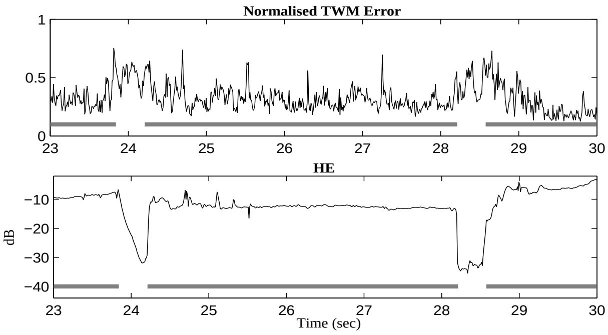

Figure 6: *Plots of normalized TWM error (above) and HE (below) for the same extract of a typical female vocal performance (from Section 5.1). The thick grey horizontal lines under the curves indicate the presence of the voice.*

6.1.2. Automatic Boundary Detection and Grouping

On observing the HE feature we see that while vocal and instrumental regions are reasonably well separated, there are localized fluctuations within regions. On the other hand, boundaries between regions are marked by broad (not abrupt) transitions. A segmentation based on detecting stable transitions in HE feature is expected to help in smoothing frame-level decisions. This can be achieved via a similarity matrix [14], a 2-dimensional representation of how similar each frame is to every other frame. The absolute difference of the HE feature for corresponding frames is smoothened with a 2-d Gaussian difference kernel to obtain a "novelty" score. Peaks in the novelty score above a global threshold correspond to significant changes in the audio content and are picked as potential segment boundaries. A pruning of boundaries is done using a minimum segment duration criteria i.e. if two boundaries are closer than the minimum segment duration threshold then the one with the lower novelty score is discarded. The optimal values i.e. ones that give the best trade-off between true boundaries and false alarms, of the difference kernel duration, the global threshold and the minimum segment duration are empirically found to be 500 ms, 0.15 and 150 ms respectively.

Grouping of frame-level classification decisions over the automatically detected segments is done by a process of majority voting, i.e. the segment assumes the label of that class into which the majority of the frames in that segment have been classified.

6.2. Performance

The voicing detector is tested on the composite data described in Section 5.1. Its performance is evaluated using a measure of overall accuracy, defined as the ratio of the sum of the correctly detected vocal frames and correctly detected instrumental frames to the total number of (vocal and instrumental) frames. At the frame-level stage we see (first row of Table 3) that the overall accuracy is quite high (87.8 %). From Table 4, we can also see that the vocal and instrumental classification accuracies are comparable, which indicates that the voicing detector is not biased towards any one class.

Table 3: *Overall classification accuracies (in %) at the classifier output and after grouping for the composite data with and without pre-processing*

		Overall Accuracy
Before SS	Classifier output	87.80
	After grouping	92.44
After SS	Classifier output	91.88
	After grouping	96.17

Table 4: *Vocal (V) - Instrumental (I) confusion matrix (in %) for the composite data after grouping, before and after pre-processing.*

	Before SS		**After SS**	
	V	I	V	I
V	92.26	7.74	96.55	3.45
I	6.20	93.80	6.60	93.40

It was found that instrumental to vocal errors tend to occur during tonal *tabla* strokes. Vocal to instrumental classification errors are made during unvoiced speech sounds in the middle of sung phrases as well during sung phrase endings where the voice trails off gradually. The grouping process is able to reduce spurious errors caused by tonal *tabla* strokes and unvoiced singing sounds, which is visible in the increased the frame-level classification accuracy post grouping (92.44%). However, the classification errors due to the trailing voice at the ends of sung phrases are more persistent and may not always be corrected by the grouping process. These errors occur when the voice HE becomes comparable to the *tanpura* HE, given that the *tanpura* pitch will be detected during instrumental segments with no percussion. In an attempt to alleviate these errors we propose a method to suppress the *tanpura* partials without significantly attenuating the singing voice.

6.3. Spectral Subtraction-based Pre-Processing

We attempt to exploit the relative stationarity of the *tanpura* partials for suppression. The use of spectral subtraction (SS) [15] was chosen over RASTA processing [16] since it was found that the latter severely attenuated voice harmonics during the steady, held notes. This is attributed to the fact that *vibrato* (F0 modulation around the note frequency) is not a generally prevalent practice in ICM.

Spectral subtraction is a well-known technique for noise suppression in speech communication. It involves the subtraction of an average noise magnitude spectrum, estimated during non-speech regions, from the magnitude spectrum of the noisy signal followed by signal reconstruction from the modified spectrum. In the present case, we exploit the fact that the initial part of most north Indian classical music performances contains at least 4 seconds of only *tanpura*. From this initial *tanpura* segment, an average magnitude spectrum is estimated once only and then subtracted from all subsequent frames in the mixed track. The relatively long segment serves to include the effects of the differently tuned strings plucked in sequence. In the resulting *tanpura*-suppressed signal, it is found that there is no perceptible degradation of the voice while the *tanpura* sound is reduced to a low level of residual noise due to the partial subtraction of its harmonics.

The vocal-instrumental classification was re-executed taking care to now train the classifier with the pre-processed data. From Table 3, we see that there is a significant increase in both frame-level accuracy as well as the accuracy post grouping for the data that has been pre-processed using SS. This increase in classification accuracy is due to the increased separability achieved by the HE feature because of the suppression of *tanpura* energy, resulting in a significant improvement in vocal classification accuracy (as can be seen in Table 4).

7. CONCLUSION

This paper investigates the effectiveness of harmonic matching based PDAs for the pitch tracking of the singing voice in the presence of strong, pitched accompaniment; the case of Indian classical vocal music was considered as a specific example. A harmonic matching PDA can be described by a spectrum computation/peak picking stage and the characteristic spectral fitness measure used in computation of the match between the signal and ideal harmonic spectra. On comparing the TWM PDA and a representative harmonic sieve based PDA, it was found that the specific form of the TWM error function led to greater robustness to strong tonal percussion (represented by signals with sparse, harmonic spectra). A feature based on the energy of the harmonics corresponding to the detected predominant F0 was found to be an accurate indicator of voicing in the same context.

A separate evaluation (not reported here) of the TWM-DP PDA on vocal excerpts from the ISMIR 2004 melody extraction contest dataset [17] indicated high pitch accuracies. The overall PA value for 8 segments of audio (4 female and 4 male, total duration= 2 min 12 sec), in which the voice is dominant, was found to be 92 % using the TWM-DP algorithm. We conclude that while the TWM-DP system is comparable to other predominant F0 detection algorithms when the melodic voice is dominant, it exhibits superior pitch accuracies in the presence of significantly stronger harmonic interference that is spectrally sparse relative to the melodic voice. Specific knowledge of the component spectral envelopes was not utilized in this work but will be considered in future as a way to extend its scope to a wider class of polyphonic signals.

Related audio examples along with their corresponding re-synthesized melodic lines are available for listening at `http://www.ee.iitb.ac.in/daplab/MelodyExtraction/`

8. ACKNOWLEDGMENTS

We would like to thank Mr. D. B. Biswas and Dr. S. Rao of the National Centre for the Performing Arts (NCPA), Mumbai, India for providing us with the multi-track music data used for the experiments reported in this paper.

9. REFERENCES

[1] G. Poliner, et. al., “Melody Transcription from Music Audio: Approaches and Evaluation,” *IEEE Trans. on Audio, Speech and Language Processing,* vol. 15, no.4, pp. 1247-1256, May 2007.

[2] E. Gómez, et. al., “A Quantitative Comparison of Different Approaches for Melody Extraction from Polyphonic Audio Recordings,” Tech. Rep., MTG-TR-2006-01, 2006.

[3] A. Bapat, V. Rao and P. Rao, “Melodic contour extraction of Indian classical vocal music,” in *Proc. Intl. Workshop on Artificial Intelligence and Music (Music-AI ‘07),* Hyderabad, India, January 2007.

[4] R. Maher and J. Beauchamp, “Fundamental Frequency Estimation of Musical Signals using a Two-Way Mismatch Procedure,” *J. Acoustical Soc. America*, vol. 95, no. 4, pp. 2254-2263, 1994.

[5] J. Brown, “Music Fundamental Frequency Tracking using a Pattern Recognition Method,” *J. Acoustical Soc. America*, vol. 92, no. 3, pp. 1394-1402, 1992.

[6] P. Cano, “Fundamental Frequency Estimation in the SMS analysis,” in *Proc. of COST G6 Conference on Digital Audio Effects 1998,* Barcelona, Spain, 1998.

[7] D. Griffin and J. Lim, “Multiband Excitation Vocoder,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223 – 1235, 1994.

[8] G. Peterschmitt, E. Gómez and P. Herrera, “Pitch-based Solo Location,” in *Proc. of MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, 2001.

[9] P. Chordia, “Tabla Stroke Database,” Available at http://ccrma.stanford.edu/~pchorida/tablaStrokes/, Accessed March 10, 2008.

[10] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoustical Soc. America*, vol. 111, no. 4, pp. 1917-1930, 2002.

[11] D. Hermes, “Measurement of pitch by sub-harmonic summation,” *J. Acoustical Soc. America*, vol. 83, no. 1, pp. 257-264, 1988.

[12] C. Sutton et. al., “Transcription of vocal melodies using voice characteristics and algorithm fusion,” in *Proc. Music Information Retrieval eXchange (MIREX)*, 2006.

[13] Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monoaural recordings,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1475-1487, 2007.

[14] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in *Proc. IEEE Intl. Conf. Multimedia and Expo (ICME)*, vol. 1, pp. 452-455, 2000.

[15] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Audio, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[16] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[17] ISMIR 2004 Melody Extraction Contest Test Set http://www.iua.upf.es/mtg/ismir2004/contest/melodyContest/FullSet.zip, Accessed March 19, 2008