Vocal Melody Extraction in the Presence of Pitched Accompaniment in Polyphonic Music

Vishweshwara Rao and Preeti Rao

***Abstract*—Melody extraction algorithms for single-channel polyphonic music typically rely on the salience of the lead melodic instrument, considered here to be the singing voice. However the simultaneous presence of one or more pitched instruments in the polyphony can cause such a predominant-F0 tracker to switch between tracking the pitch of the voice and that of an instrument of comparable strength, resulting in reduced voice-pitch detection accuracy. We propose a system that, in addition to biasing the salience measure in favor of singing voice characteristics, acknowledges that the voice may not dominate the polyphony at all instants and therefore tracks an additional pitch to better deal with the potential presence of locally dominant pitched accompaniment. A feature based on the temporal instability of voice harmonics is used to finally identify the voice pitch. The proposed system is evaluated on test data that is representative of polyphonic music with strong pitched accompaniment. Results show that the proposed system is indeed able to recover melodic information lost to its single-pitch tracking counterpart, and also outperforms another state-of-the-art melody extraction system designed for polyphonic music.**

***Index Terms*—Fundamental frequency estimation, music information retrieval (MIR), music transcription, predominant pitch detection.**

I. Introduction

AUTOMATIC melody extraction from polyphonic music is an area of research that has received considerable attention in the past decade. A rough definition of the melody of a song is the monophonic pitch sequence that a listener might reproduce if asked to hum a segment of polyphonic music [1]. This pitch sequence is usually manifested as the fundamental frequency (F0) contour of the lead musical instrument in the polyphonic mixture (often called the predominant F0). Since the lead instrument for several genres of music, especially popular music, is the human singing voice, the focus of this paper is on voice pitch contour extraction in polyphony.

Poliner *et al.* [1] provide a comprehensive review of state-of-the-art in melody extraction. The majority of melody extraction algorithms reviewed by them adopt the "understanding-without-separation" paradigm as described by Scheirer [2] and by-and-large adhere to a standard framework as depicted in Fig. 1. Here, a short-time, usually spectral, signal representation is extracted from the input polyphonic audio signal. This is then input to a multi-F0 extraction block whose goal is to detect candidate F0s and associated salience values. The melody identification stage attempts to identify a trajectory through the F0 candidate-time space that best represents the melody of the song. The voicing detection block identifies whether the melody is active or silent at each time instant. There also exist melody extraction algorithms that do not follow the above paradigm such as those that attempt to first segregate the melodic source and then track the pitch of the extracted "monophonic" source [3], [4].

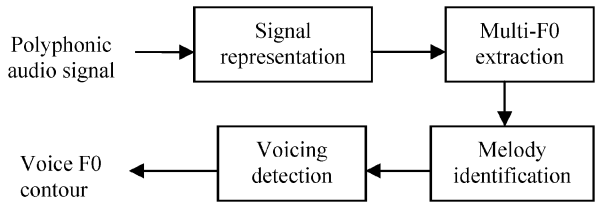


Fig. 1. Block diagram of a typical melody extraction system.

Both problems, *viz*. melodic pitch detection and voicing detection, continue to be far from solved for use in practical automatic music transcription systems operating on large datasets across music of various genres and styles. However, the applicability of available algorithms can be extended considerably by employing semi-automatic approaches tailored for specific applications [5]. For instance, the manual marking of vocal segment (sung phrase) boundaries is much easier than manual detection of the frame-by-frame voice pitch. Therefore a semi-automatic melody extraction framework that exhibits a highly accurate automatic voice pitch tracker for polyphony and allows user control for vocal segment detection can be a valuable transcription tool. Consequently in the present work, we have chosen to primarily focus on accurate vocal-pitch tracking (first three blocks of Fig. 1) without touching upon voicing detection.

A major cause of errors in vocal-pitch tracking for state-of-the-art polyphonic melody extraction systems is the presence of strong, pitched accompaniment [6]–[9]. In this paper, we propose a new system for vocal melody extraction that demonstrates increased robustness to pitched accompaniment when compared to another state-of-the-art melody extraction system. Our system utilizes a spectral harmonic-matching pitch detection algorithm (PDA) followed by a dynamic programming-based (DP) optimal path finding technique that tracks pitch within certain melodic smoothness constraints. The major contribution of this paper is in the design of a system that demonstrates superior voice-pitch tracking accuracy in polyphonic music particularly when pitched accompaniment is strong. The novel aspects of the system include the separation of the F0 candidate selection and salience computation steps, the joint tracking of two F0 contours by a DP algorithm with a harmonic-relationship constraint on F0 pairing and the final identification of the voice pitch contour from the dual-F0

Manuscript received July 17, 2009; revised January 02, 2010. Date of publication March 15, 2010; date of current version September 08, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India (e-mail: vishu_rao@iitb.ac.in).

Digital Object Identifier 10.1109/TASL.2010.2042124

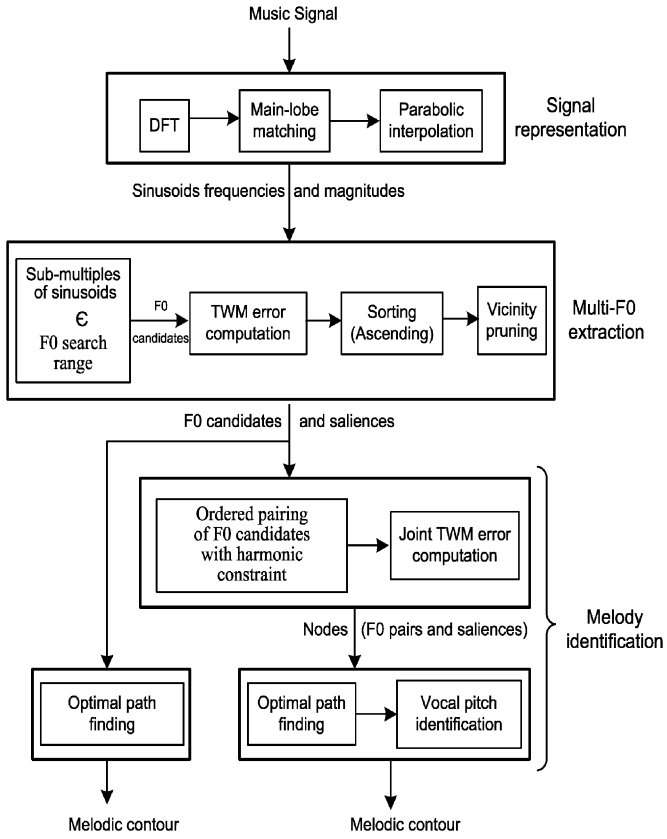


Fig. 2. Block diagram of proposed system.

tracking output using a voice-feature that exploits the temporal instability (in frequency) of voice harmonics.

Our melody extraction system does not use machine learning-based algorithmic modules as is done in many other systems [10], [11]. The performance of such systems is highly dependent on the diversity and characteristics of the training data available. In contrast, our system provides a limited set of parameters that may be tuned, if needed, based on higher level knowledge of the signal properties. This also allows for easy integration into a semi-automatic melody extraction framework.

Fig. 2 shows a detailed block diagram of the proposed system of which the three main modules are the computation of a suitable signal representation, multi-F0 extraction and melody identification. Sections II–IV describe each of these stages. In the design of each stage, we attempt to exploit characteristics specific to the singing voice and increase robustness to pitched accompaniment. The F0 candidate salience measure used in our multi-F0 extraction module is more dependent on the frequency extent of harmonics rather than their strengths. The melody identification module contains novel enhancements to an existing dynamic programming-based framework that involve harmonically constrained F0 candidate pairing and the simultaneous tracking of these F0 candidate pairs. The subsequent selection of the voice pitch is done using a new feature based on voice-harmonic instability. Section V comparatively evaluates the performance of our system with another state-of-the-art system on different music datasets. The last section presents the conclusions.

II. Signal Representation

It is well known that the frequency-domain analysis of a pitched musical note results in a set of harmonics at near-integer multiples of an F0. In order to enhance the pitched content of the polyphonic audio, in this module we extract a sparse signal representation in the form of a set of harmonic frequencies and amplitudes.

A. Sinusoid Identification by Main-Lobe Matching

For a stationary periodic sound, sinusoidal components in the magnitude spectrum will have a well-defined frequency representation, i.e., the transform of the analysis window used to compute the Fourier transform. In order to detect sinusoids from the short-time magnitude spectrum of a signal we use a measure of closeness of a local spectral peak's shape to the ideal sinusoidal peak, as proposed by Griffin and Lim [12]. This criterion is defined as the mean square difference between the local spectrum and the window main lobe.

The above main-lobe matching method for sinusoid identification has been chosen over computationally simpler methods, such as fixed amplitude thresholding [13] and amplitude-envelope-based thresholding [14], keeping the polyphonic context of the audio in mind. The detection of harmonics of the melodic F0 is critical to the overall performance of the melody extraction system. The use of amplitude thresholding methods may miss these harmonics in the vicinity of strong, interfering harmonics from pitched accompaniment. On the other hand, even relatively weak sinusoidal components will be detected by the chosen method. In the results of separate unreported experiments the performance of the three methods (fixed amplitude thresholding, amplitude-envelope-based thresholding and main-lobe matching) was compared using mixtures of simulated voice and pitched interference signals. Different signal conditions such as steady and varying voice pitch, different strengths, and different numbers of interference harmonics were experimented with. It was found that the best tradeoff between correctly detected sinusoids and false alarms was consistently obtained using the main-lobe matching technique.

B. Implementation

First, we compute a high-resolution, discrete Fourier transform $S_W(\omega)$ of a 40-ms-long Hamming windowed signal frame. Then all three-point local maxima are detected in the windowed magnitude spectrum $|S_W(\omega)|$. We only consider the spectral content below 5 kHz since significant voice harmonics are not usually found above this limit. Next, an error value for the mth local maximum is computed as [12]

$$\varepsilon_m = \frac{1}{2\pi}\int_{a_m}^{b_m} \left[|S_W(\omega)| - |A_m|\,|E_W(\omega)|\right]^2 d\omega \qquad (1)$$

where $|E_W(\omega)|$ is the precomputed Hamming window spectrum centered at the mth local peak frequency, $[a_m, b_m]$ is the frequency interval over the width of the window main lobe in the neighborhood of the mth peak. For a 40-ms-long Hamming

window, the width of the window main-lobe is 100 Hz. $|A_m|$ is a scaling factor that will minimize ε_m and is given by

$$|A_m| = \frac{\int_{a_m}^{b_m} |S_W(\omega)|\,|E_W(\omega)|\,d\omega}{\int_{a_m}^{b_m} |E_W(\omega)|^2\,d\omega}. \quad (2)$$

Now a normalized sinusoidality measure (S) is defined as

$$S = 1 - \frac{\varepsilon_m}{\frac{1}{2\pi}\int_{a_m}^{b_m} |S_W(\omega)|^2\,d\omega}. \quad (3)$$

Those peaks for which S lies above a given threshold are labeled as sinusoidal components. Although a strict threshold on sinusoidality (S) was originally proposed (0.8) [12], we use a more relaxed threshold (0.6) in order to detect melodic F0 harmonics that may be distorted due to voice-pitch modulations, such as vibrato, while still maintaining a high side-lobe rejection. The sinusoid frequency and amplitude estimates are further refined using parabolic interpolation.

III. Multi-F0 Extraction

The goal of this module is to reliably detect potential F0 candidates in each frame with their associated salience values, given the sinusoid frequencies and amplitudes output by the previous module. In the present context, there are two requirements of the multi-F0 extraction module: 1) the voice-F0 candidate should be reliably detected; 2) the salience of the voice-F0 candidate should be relatively high compared to those of instrument-F0 candidates. Our approach to F0 candidate identification and salience computation is described next, followed by the implementation details of the complete multi-F0 extraction module.

A. F0 Candidate Identification

As described by de Cheveigné [15], multi-F0 estimation approaches that adopt an iterative "estimate-cancel-estimate" mechanism or jointly estimate multiple F0s in a single step are superior to the simple extension of a single-F0 algorithm to estimate multiple F0s, e.g., identifying the largest and second largest peak in an autocorrelation function (ACF). Both the iterative and the joint estimation approaches are of significantly higher computational complexity than the last category, i.e., a single voice (monophonic) algorithm applied to find multiple F0s.

Our multi-F0 extraction module, which falls under the last category, is able to reliably extract voice-F0 candidates (as will be shown in Section V-C.) without having to resort to the iterative or joint estimation approaches. This is achieved by the distinct separation of the F0 candidate identification and salience computation parts as follows. Rather than computing a salience function over a range of trial F0 and then picking F0 candidates as the locations of maxima of this function, we first identify potential candidates by an independent method that selects all integer sub-multiples of well-formed detected sinusoids and only compute the salience function at these candidates. This ensures that the voice-F0 candidate will be selected (and therefore actively considered in the next stage of melody identification) even if a single well-formed higher harmonic of the voice-F0 is detected.

B. F0 Candidate Salience Computation

The salience of an F0 candidate (hypothesized pitch) can be computed using a pitch detection algorithm (PDA). Most PDAs can be classified as spectral (spectral pattern matching) or temporal (maximization of a correlation-type function). These approaches have been shown to be equivalent, i.e., minimizing the squared error between the actual windowed signal spectrum and an idealized harmonic spectrum is analytically equivalent to maximizing an autocorrelation function of the windowed signal [12], [16]. The above PDAs fall under the "harmonic sieve" category [15]. The resulting salience functions put strong emphasis on the high amplitude portions of the spectrum, and thus are sensitive to the presence of interference containing strong harmonics.

Recent multi-F0 extraction systems, including those proposed by Tolonen and Karjailenen [17], Li and Wang [18], and Klapuri [8], have used a spectro-temporal approach to F0 candidate estimation. This involves the computation of independent correlation functions on multiple frequency channels (such as a multi-band ACF), usually motivated by an auditory model. Although this may overcome the high amplitude interference problem by allowing the weight of each channel to be adjusted to compensate for amplitude mismatches between spectral regions, such an approach requires suitable decisions to be made on the frequency bands and associated channel weighting. Such decisions may again be dependent on the nature of accompanying instruments.

Our choice of salience function is the Two Way Mismatch (TWM) error, as described originally by Maher and Beauchamp [19] which, to the best of our knowledge, has not been previously explored in the context of melody extraction from polyphony. The TWM PDA qualifies as a spectral method for pitch detection. However it is different from the "pattern matching" methods (i.e., those that minimize the squared error or maximize the correlation between the actual and idealized spectra) in that it minimizes an unconventional spectral mismatch error which is a particular combination of an individual partial's frequency deviation from the ideal harmonic location and its relative strength. As described by Maher and Beauchamp [19], this error function was designed to be sensitive to the deviation of measured partials/sinusoids from ideal harmonic locations. Relative amplitudes of partials too are used in the error function but play a significant role only when the aforementioned frequency deviations are small [19, Sec. II-A]. Unlike the multi-band ACF, the TWM error computation does not require decisions on bands and weighting but is primarily dependent on the number of predicted harmonics (N) for a given F0. The choice of N depends on the known or assumed spectral characteristics of the signal source. We have tuned N for biasing the error function in favor of spectrally rich musical sources, such as the singing voice, by predicting harmonics upto 5 kHz (a previously used upper spectral limit for dominant harmonics of typical melody lines [7]). Although other parameters are used in the computation of the TWM error

function for monophonic signals [19], these are unchanged in our implementation.

We have previously confirmed, using simulated and real signals, that in the presence of strong, spectrally sparse, tonal interferences, the melodic (voice-F0) candidate was indeed detected with significantly higher salience on using the TWM PDA as compared to other harmonic matching or correlation-based PDAs [20], [21]. This was attributed to the dependence of TWM error values on the frequency extent of harmonics as opposed to the strengths of the harmonics (which is the case with most "harmonic-sieve" based methods [15].) This has the advantage that F0s belonging to the spectrally-rich singing voice (having gentler roll-off than common pitched accompanying instruments such as the piano and the flute [22]), are characterized by lower TWM errors, i.e., better salience.

C. Implementation

Probable F0 candidates are first identified as submultiples of the frequencies of well-formed sinusoids i.e., those having a sinusoidality (S) greater than 0.8. Candidates that do not lie within the F0 search range (from 80 to 500 Hz) are ignored.

For each of the above detected F0 candidates, the corresponding salience is computed as the normalized Two-Way Mismatch (TWM) error [19]. The TWM error Err_{TWM}, for a given trial F0 f, is a weighted sum of two errors, the predicted-to-measured error $Err_{p\rightarrow m}$ and the measured-to-predicted error $Err_{m\rightarrow p}$, as shown as follows:

$$Err_{\mathrm{TWM}}(f) = \frac{Err_{p\rightarrow m}(f)}{N} + \rho\frac{Err_{m\rightarrow p}(f)}{M} \tag{4}$$

where N and M are the number of predicted and measured harmonics, respectively, and ρ is a weighting factor. $Err_{p\rightarrow m}$ is based on the mismatch between each harmonic in the predicted sequence and the nearest measured partial while $Err_{m\rightarrow p}$ is based on the mismatch between each partial in the measured sequence and the nearest predicted harmonic. Both of these share the same form. $Err_{p\rightarrow m}$ is defined as follows:

$$Err_{p\rightarrow m} = \sum_{n=1}^{N}\left[\frac{\Delta f_n}{(f_n)^p} + \left(\frac{a_n}{A_{\max}}\right)\left(q\frac{\Delta f_n}{(f_n)^p} - r\right)\right] \tag{5}$$

where 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 neighbor 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 between measured and predicted partials for the given trial F0. p, q, and r are independent parameters. Note that here we use a value $\rho = 0.1$. This gives lesser weight to $Err_{m\rightarrow p}$ and leads to Err_{TWM} being almost the same as $Err_{p\rightarrow m}$, since $Err_{m\rightarrow p}$ for single-F0 values will be comparatively unreliable in the presence of harmonics of another pitched source.

Finally, F0 candidates are sorted in ascending order of Err_{TWM} and weaker candidates (having higher Err_{TWM}) that lie in the close neighborhood (25 cents) of a stronger candidate are erased. This is done so as to include only the most relevant F0 candidates in the final list since the candidate identification process will typically result in a large number of F0 candidates clustered around (sub)-multiples of the F0s of pitched sound sources. Only the top ten candidates and their corresponding normalized Err_{TWM} values are chosen for further processing.

IV. Melody Identification

In the melody identification module, most algorithms utilize F0 salience values from the multi-F0 extraction module, and impose smoothness constraints to identify the melodic trajectory. Two approaches have been widely used. The first involves finding an optimal path through the F0 space over time by dynamically combining F0 salience values (also called measurement cost) and smoothness constraints (also called smoothness cost) using methods either based on the Viterbi algorithm [23] or dynamic programming (DP) [24], [25]. The second approach applies variants of the partial tracking (PT) algorithm, used in classical sinusoidal modeling [13], to forming multiple F0 trajectories/contours through the F0 candidate space over time. The final melodic contour is usually chosen as that trajectory with the greatest accumulated salience/energy.

In our melody identification module, we use the DP framework. The use of DP in melody extraction is attractive since it finds the *optimal* path by combining trajectory forming and melodic contour identification in one computationally efficient, *global* framework, i.e., a black box that outputs a single F0 contour given suitably defined local and smoothness costs. PT, on the other hand, first forms trajectories using *local* frequency proximity and subsequently identifies melodic tracks or fragments. Here multiple trajectories may be formed through different musical instrument F0 contours and their (sub)-multiples leading to a densely populated track space [6].

We next describe the application of DP in our melody identification module (referred to as single-F0 tracking). Situations in which this module is expected to suffer from irrecoverable degradations are then identified. Enhancements, in terms of dual-F0 tracking, within the DP framework are then described that may enable the retrieval of previously inaccessible melodic information. Finally we describe a voice-pitch identification framework that uses a novel feature that enables the identification of the voice pitch from the dual-F0 tracking output by detecting temporal frequency-instability in the voice harmonics. Both the single- and dual-F0 tracking approaches are shown in the melody identification stage of Fig. 2.

A. Application of DP in Single-F0 Tracking

The application of DP for single-F0 tracking in our melody identification module is quite straightforward. The local measurement cost for each pitch candidate is given by the normalized TWM error of the F0 candidates obtained in the multi-F0 extraction stage. The smoothness cost must reflect the characteristics of typical voice pitch transitions and is designed based on the following musical considerations. Since musical pitches are known to be logarithmically related, such a cost must be symmetric in log-space. Smaller pitch transitions must be assigned a near zero penalty since these are especially common over short durations (such as the time between consecutive analysis time instants). Improbable, very large pitch transitions can all be penalized by a fixed ceiling value. We found that a Gaussian cost function (W) described in log-space satisfies the above requirements and is defined as follows.

$$W(p,p') = 1 - e^{\frac{-(\log_2(p')-\log_2(p))^2}{2\sigma}} \tag{6}$$

where p and p' are the F0 candidates in the previous and current frames. A value of $\sigma = 0.1$ results in a function that assigns very low penalties to pitch transitions below two semitones. Larger rates of pitch transition (in the 10-ms frame interval chosen in this work) are improbable even during rapid singing pitch modulations and are penalized accordingly.

The pitch tracking accuracy of the system with the single-F0 tracking approach to melody identification was an entry in the audio melody extraction task at the 2008 and 2009 Music Information Retrieval Evaluation eXchanges (MIREX).[1] It was found to demonstrate high pitch accuracy, especially for the vocal music datasets. Particular instances of pitch error have been identified where the output melodic contour switches between voice and instrument F0s. This occurs when some accompanying instrument in the polyphony is comparable in salience to the voice.

B. Shortcoming of Single-F0 Tracking

The above melodic identification module may output an (partially) incorrect melody when either the measurement and/or the smoothness costs are in favor of the accompanying instrument F0 rather than the melodic F0. The bias in measurement cost occurs when an accompanying, pitched instrument has a salience comparable to that of the voice. This may cause the output pitch contour to incorrectly identify accompanying instrument F0 contour segments as the melody. [An example of such an occurrence is seen in Fig. 4(a)]

Smoothness costs are normally biased towards musical instruments which are capable of producing sustained, stable-pitch notes. It is well known that the human voice suffers from natural, involuntary pitch instability called jitter in speech and flutter in singing [26]. Further in singing, pitch instability is much more emphasized in the form of voluntary, large, pitch modulations that occur during embellishments and ornaments such as vibrato. So the presence of stable-pitch instruments, such as most keyed instruments, e.g., the piano and accordion (especially when the voice pitch is undergoing rapid and large modulations) could also lead to incorrect identification of the melodic fragments. Such errors are more likely to occur when the F0s of the voice and instrument intersect since at the point of intersection, the F0 candidates for both sources are one and the same with a single salience.

In cases of incorrect melodic identification for PT-based approaches, the recovery of the actual melodic tracks may still be possible based on the assumption that correct melodic fragments have been formed but not identified. DP, on the other hand, is forced to output only a single, possibly "confused," contour with no mechanism for recovering the correct melodic F0s. This information may be retrieved if DP is extended to tracking multiple F0 contours simultaneously.

C. Dual-F0 Tracking

Here we describe an enhancement to the DP formulation that simultaneously tracks two F0 contours (hereafter referred to as dual-F0 tracking) with the aim to better deal with accompanying pitched instruments. We restrict ourselves to tracking only two pitches simultaneously on the realistic assumption that in vocal music, there is at most only one instrument which is more dominant than the voice at any time [18].

The closest previous related work is that of Every and Jackson [27] who had designed a DP framework to simultaneously track the pitches of multiple speakers. The singing/music scenario is very different from speech. The design of the measurement and smoothness cost functions therefore require completely different considerations. We use the joint TWM error as the measurement cost of the F0 candidate pair. We have also used a novel harmonic relationship constraint to avoid the tracking of an F0 candidate and it's multiple since this would defeat the purpose of using DP to track the F0 of multiple distinct sources.

1) Implementation: We extend our previously described single-F0 tracking DP algorithm to track ordered F0 pairs called nodes. The additional F0 members of the node help to better deal with the accompanying pitched instrument(s). If we consider all possible pairs of F0 candidates the combinatory space becomes very large (number of permutations of F0 pairs formed from ten F0 candidates is ${}_{10}P_2 = 90$) and tracking will be computationally intensive. More importantly, we may end up tracking an F0 and its (sub)-multiples rather than two F0s from separate musical sources. Our method to overcome this is to explicitly prohibit the pairing of harmonically related F0s during node generation. Specifically, two local F0 candidates (f_1 and f_2) will be paired only if

$$\min_k \left(|f_1 - k.f_2|\right) > T; \quad k.f_2 \in [F_{\text{low}}, F_{\text{high}}] \tag{7}$$

where $k.f_2$ represents all possible multiples and submultiples of f_2, T is the harmonic relationship threshold, and F_{low} and F_{high} are the lower and upper limit on the F0 search range. Using a low threshold (T) of 5 cents does not allow F0s to be paired with their multiples but allows pairing of two distinct source F0s that are playing an octave apart, which typically suffer from slight detuning especially if one of the F0 sources is the singing voice [28].

The measurement cost of a node is defined as the jointly computed TWM error of its constituent F0 candidates [29]. In the interest of computational efficiency the joint TWM error for two F0 candidates, f_1 and f_2, is computed as follows:

$$Err_{\text{TWM}}(f_1, f_2) = \frac{Err_{p\to m}(f_1)}{N_1} + \frac{Err_{p\to m}(f_2)}{N_2} + \rho\frac{Err_{m\to p}(f_1, f_2)}{M} \tag{8}$$

where N_1 and N_2 are the number of predicted partials for f_1 and f_2 resp. and M is the number of measured partials. The first two terms in (8) will have the same values as during the single-F0 TWM error computation in (4). Only the last term, i.e., the mismatch between all measured partials and the predicted partials of both F0s (f_1 and f_2), has to be computed. Note that here we use a larger value of ρ (0.25) than before. This is done so as to reduce octave errors by increasing the weight of $Err_{m\to p}$ thereby ensuring that Err_{TWM} for the true F0 pair is lower than that of the pair that contains either of their respective (sub)-multiples.

The smoothness costs between nodes are computed as the sum of smoothness costs between the constituent F0 candidates, given previously in (6). A globally optimum path is finally computed through the node-time space using the DP algorithm. Two pitch contours are available in this minimum cost node-path.

[1]These annually held evaluations provide a framework for formally assessing different music information retrieval (MIR)-related systems for a wide variety of tasks, including melody extraction, on common test data-sets.

D. Voice-Pitch Identification

The melody identification module is required to output a single-F0 contour from the dual-F0 DP stage as the final melodic contour. One possible approach to solving the above problem would be to adopt a source discrimination approach, as proposed by Marolt [30] which attempts the unsupervised clustering of melodic fragments using timbral features. In such an approach the selection of the final contour after clustering is still unresolved.

Recent experiments have validated the voice-instrument discriminative ability of a feature that is indicative of the relative instability of the voice pitch as compared to keyed instrument notes [31]. (The relative difference in the harmonic instability of voice harmonics as compared to most keyed instrument harmonics was previously mentioned in Section IV-B.) This feature is called Sinusoidal Track Harmonic Energy [STHE]. A detailed description of the feature implementation can be found in [31]. Here we provide a brief description of the feature computation within a framework used to choose the final voice-F0 contour.

Although melodic smoothness constraints are imposed in the dual-F0 tracking system, each of the output contours cannot be expected to faithfully track the F0 of the same source across silence regions in singing or instrument playing. Therefore, choosing one of the two output contours as the final output is unreliable. Rather we rely on the continuity of these contours over short, non-overlapping windows and make voice-F0 segment decisions for each of these "fragments."

1) Implementation: Each of the dual-F0 output contours is divided into short-time (200 ms long) non-overlapping F0 fragments. We then identify sinusoids (from the output of the signal representation module described in Section II) that are in the 100-cent neighborhood of the multiples of the F0 within each fragment (with an upper limit on the vicinity fixed at half the main lobe width, i.e., 50 Hz). Using these sinusoids we build harmonic sinusoidal models for each of these F0 fragments using the partial tracking algorithm described by Serra [32]. The linking cost between sinusoids and constructed partial tracks is based purely on frequency proximity of sinusoids and is fixed at 200 cents. For each of these sinusoidal models we next prune/erase tracks whose standard deviations in frequency are below a specified threshold (here 2 Hz), indicating stability in frequency. The total energy of the residual signal within the analysis window is then indicative of the presence of vocal harmonics. The fragment with the higher energy is therefore selected as the final voice-F0 fragment.

V. Experimental Evaluation

In this section, we present an experimental evaluation of our single- and dual-F0 tracking algorithms, hereafter referred to as the TWMDP system, as compared to another state-of-the-art singing voice melody extraction system on three different sets of polyphonic vocal music. This other algorithm is the one proposed by Li and Wang [18], hereafter referred to as the LIWANG system, who have made their program code available on the internet.

The LIWANG system initially processes the signal using an auditory model and correlation-based periodicity analysis, following which different observation likelihoods are defined for the cases of 0, 1, and 2 (jointly estimated) F0s. A hidden Markov model (HMM) is then employed to model, both, the continuity of F0 tracks and also the jump probabilities between the state spaces of 0, 1, or 2 F0s. The 2-pitch hypothesis is introduced to deal with the interference from concurrent pitched sounds. When two pitches are output the first F0 is labeled as the predominant (voice) pitch [18]. The LIWANG system has been previously shown to be superior to those of Rynnanen and Klapuri [33], Klapuri [34] and Wu, Wang, and Brown [35], for detecting the pitch of the singing voice in polyphonic audio. It should be noted that, unlike the TWMDP system, the LIWANG system also makes a voicing decision.

TABLE I
Description and Durations of Each of the Testing Datasets

Dataset	Description	Vocal (sec)	Total (sec)
1	Li & Wang data	55.4	97.5
2	Examples from MIR-1k dataset with loud pitched accompaniment	61.8	98.1
3	Examples from MIREX'08 data (Indian classical music)	91.2	99.2
TOTAL		**208.4**	**294.8**

A. Data Description

The durations of each of the three datasets used in this evaluation are shown in Table I. Here total duration refers to the length of the entire audio, and vocal duration refers to the duration for which voiced utterances are present. All the audio clips in each of the datasets are sampled at 16 kHz with 16-bit resolution.

The first dataset, provided by Li and Wang, consists of the same audio examples as used by them for the evaluation of their predominant F0 extraction system in [18]. This set consists of 25 clips from ten songs that include both male and female singers. Five of these songs belong to the rock genre and the other five belong to the country music genre. The clean vocal and accompaniment tracks of these songs were extracted from karaoke CDs using de-multiplexing software [18].

The second dataset consists of a subset (13 clips) from the MIR-1k database [36]. The accompaniment in the MIR 1-K dataset is again extracted from karaoke CDs of Chinese pop songs. The time-aligned vocal tracks have been recorded separately by amateur singers. These 13 clips have been selected based on the presence of strong pitched accompanying instruments such as acoustic guitar, piano, harmonica, and accordion. Again this dataset includes both male and female singers.

The third dataset consists of excerpts from two North Indian classical vocal performances, sung by a male and female, respectively. These also form part of the MIREX'08 dataset, which was provided by us. These performances consist of the voice, tonal percussion, a drone and a secondary melodic instrument called the *harmonium*, very similar to the accordion. The harmonium accompaniment is meant to reinforce the melody sung by the singer. Since, in this genre of music, each vocal performance is a complete improvisation without the presence of a musical score, the instrumentalist attempts to follow the singer's pitch, resulting in frequent F0 collisions. The individual monophonic tracks of different instruments were obtained by ensuring

TABLE II
TWMDP SYSTEM PARAMETERS

PARAMETER	VALUE
Frame length	40 ms
Hop	10 ms
F0 search range	80 – 500 Hz
Upper limit on spectral content	5000 Hz
Single-F0 TWM param. (p, q, r & ρ)	0.5, 1.4, 0.5 & 0.1
Dual-F0 TWM param. (p, q, r & ρ)	0.5, 1.4, 0.5 & 0.25
Std. dev. of smoothness cost (σ)	0.1
Harmonic relationship threshold	5 cents

acoustic isolation between the instrument-performing artists by spreading them out on the same stage with considerable distance between them.

For all datasets, the ground-truth voice pitch was computed from the clean vocal tracks using an independent PDA, here the YIN PDA [37], known to be very accurate for monophonic signals, followed by DP-based postprocessing and manual correction of octave and voicing errors.

B. Experimental Setup

In the following evaluation, we first compare the performance of the TWMDP single-F0 tracking melody extraction system with the LIWANG system using the first dataset. The mixed voice and accompaniment tracks used in this experiment, at signal-to-accompaniment ratios (SARs) of 10, 5, 0, and −5 dB, were obtained directly from Li and Wang. Next we compare the performance of the TWMDP single and dual-F0 tracking systems for all three datasets. The second and third datasets however are particularly representative of the kind of polyphonic scenario where the TWMDP dual-F0 tracker is expected to show significant improvement. The voice and accompaniment in these cases are mixed at SARs where both the voice melody and instrument pitch are clearly audible. This results in SARs of 10 dB and 0 dB for the second and third datasets, respectively. Almost none of the clips in first dataset contain strong pitched accompaniment.

A fixed set of parameters for the TWMDP system is used for the entire experiment (shown in Table II). Also, code provided to us by Li and Wang for the LIWANG system is compiled without making any modifications. In the interest of fairness the same F0 search range (80–500 Hz) as used by the LIWANG system is also used by the TWMDP system. Both systems provide a pitch estimate every 10 ms.

The multi-F0 extraction module of the TWMDP system (described in Sections II and III) is separately evaluated in terms of percentage presence of the ground-truth voice pitches in the F0 output candidate list. Percentage presence is defined as the percentage of voiced frames that an F0 candidate is found within 50 cents of the ground truth voice-F0.

For the evaluation of the complete single- and dual-F0 melody extraction systems, the metrics used are pitch accuracy (PA) and chroma accuracy (CA) [1]. PA is defined as the percentage of voiced frames for which the pitch has been correctly

TABLE III
PERCENTAGE PRESENCE OF GROUND-TRUTH VOICE-F0 IN F0 CANDIDATE LIST OUTPUT BY MULTI-F0 EXTRACTION MODULE FOR EACH OF THE THREE DATASETS

DATASET	PERCENTAGE PRESENCE OF VOICE-F0 (%)	
	TOP 5 CANDIDATES	TOP 10 CANDIDATES
1	92.9	95.4
2	88.5	95.1
3	90.0	94.1

detected, i.e., within 50 cents of a ground-truth pitch. CA is the same as PA except that octave errors are forgiven. Only valid ground-truth values, i.e., frames in which a voiced utterance is present, are used for evaluation. These evaluation metrics are computed with the output of the TWMDP single- and dual-F0 tracking systems as well as with the output of the LIWANG system.

For dual-F0 tracking evaluation two sets of metrics are computed. The first is a measure of whether the correct (vocal) pitch at a given instant is tracked by *at least* one of the two contours, and is called the Either-Or accuracy. This metric is an indicator of melodic recovery by the dual-F0 tracking system. The second set of metrics is computed on the final single contour output after vocal pitch identification. Comparison between these two sets of metrics will be indicative of the reliability of the system for vocal pitch identification.

C. Results

Results for the evaluation of the multi-F0 extraction part of the TWMDP system for all three datasets appear in Table III. For dataset 1, we have used the 0 dB mix. The percentage presence of the voice-F0 is computed in the top five and top ten candidates, respectively, as output by the multi-F0 extraction system. It can be seen that the voice-F0 is present in the top ten candidates about 95% of the time thus supporting the choice of TWM error for use in the salience function.

Fig. 3(a) and (b) compares the performance of the LIWANG system with the TWMDP single-F0 tracking system for different SAR mixes of dataset 1 in terms of pitch and chroma accuracy, respectively. The TWMDP system is clearly superior to the LIWANG system. The relative difference in accuracies increases as the SARs worsen.

Finally, Table IV compares the performance of the LIWANG, TWMDP single- and dual-F0 tracking systems. Here too we have used the 0-dB mix for dataset 1. The percentage improvements of the TWMDP single- and dual-F0 tracking systems over the LIWANG system (treated as a baseline) are provided in parentheses. It should be noted that the accuracies of the LIWANG system under the "single-F0" and "dual-F0 final" headings are the same, since their vocal pitch identification mechanism just labels the first F0 of the two output F0s (if any) as the predominant F0. Again here we can see that the TWMDP single-F0 accuracies are significantly higher than the LIWANG accuracies. For datasets 2 and 3, in which a strong pitched accompaniment was often present, the use of the dual-F0 approach

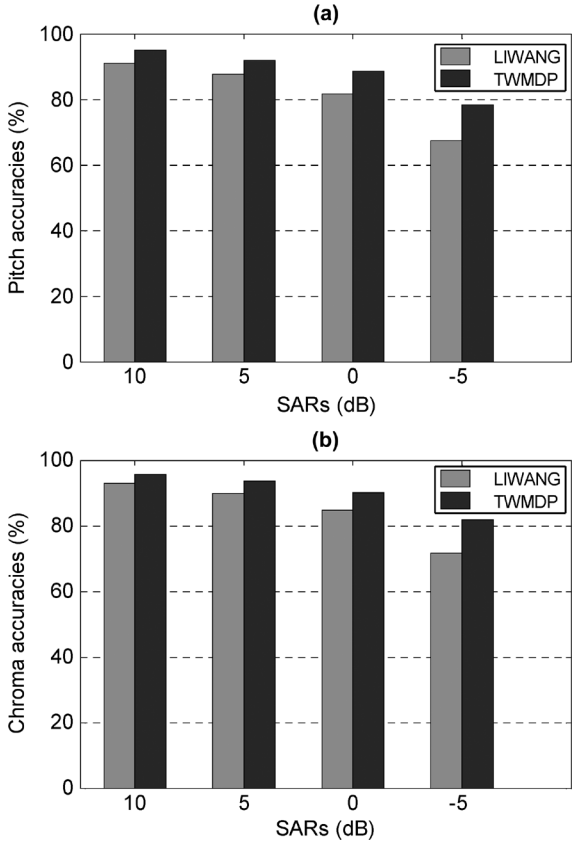


Fig. 3. (a) Pitch and (b) chroma accuracies for LIWANG and TWMDP single-F0 tracking systems for Dataset 1 at SARs of 10, 5, 0, and −5 dB.

TABLE IV
PITCH ACCURACIES (PA AND CA) OF TWMDP SINGLE- AND DUAL-F0 TRACKING SYSTEMS FOR ALL DATASETS. THE PERCENTAGE IMPROVEMENT OVER THE LIWANG SYSTEM IS GIVEN IN PARENTHESES

		TWMDP (% IMPROVEMENT OVER LIWANG)		
DATASET		SINGLE-F0	DUAL-F0	
			EITHER-OR	FINAL
1	PA (%)	88.5 (8.3)	89.3 (0.9)	84.1 (2.9)
	CA (%)	90.2 (6.4)	92.0 (1.1)	88.8 (3.9)
2	PA (%)	57.0 (24.5)	74.2 (-6.8)	69.1 (50.9)
	CA (%)	61.1 (14.2)	81.2 (-5.3)	74.1 (38.5)
3	PA (%)	66.0 (11.3)	85.7 (30.2)	73.9 (24.6)
	CA (%)	66.5 (9.7)	87.1 (18.0)	76.3 (25.9)

in the TWMDP system results in further significant improvement over the single-F0 system.

D. Discussion

1) Melodic F0 Recovery for TWMDP: From the results in Table IV it is observed that for all datasets the Either-Or pitch accuracy of the TWMDP dual-F0 tracking system is higher than that of the single-F0 system indicating that some of the melodic contour information, lost by the latter, has been recovered. Errors in the output of the single-F0 tracking system were observed when some pitched accompanying instrument in the polyphony is of comparable strength to the singing

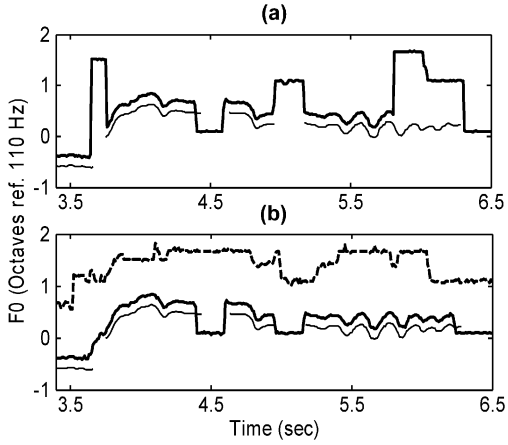


Fig. 4. Example of melodic recovery using the dual-F0 tracking approach for an excerpt of an audio clip from dataset 2. Ground truth voice-pitch (thin) are offset vertically for clarity by −0.2 octave, (a) single-F0 output (thick) and (b) dual-F0 output (thick and dashed). Single-F0 output switches from tracking voice to instrument pitch a little before 6 s. Dual-F0 contours track both, the voice and instrument, pitches in this region

voice. At these locations, the single-F0 pitch contour very often tracks the pitch of the accompanying instrument rather than the singing voice. The dual-F0 tracking approach alleviates the bias in the single-F0 system measurement cost towards such locally dominant pitched accompaniment by including another pitch trajectory in the tracking framework, which deals with the instrument F0, thereby allowing the continuous tracking of the voice-F0. The dual-F0 tracking approach also aids melodic recovery around F0 collisions between the voice-F0 and an instrument-F0 because of the faster resumption of tracking the voice-F0 around the collision by any one of the two contours in the dual-F0 system output.

The Either-Or accuracy for datasets 2 and 3 is significantly higher than the single-F0 tracking accuracies but this is not the case for dataset 1 where the difference is much smaller. As mentioned before the presence of strong pitched accompaniment in dataset 1 was rare. This indicates that the dual-F0 tracking approach is particularly beneficial for music in which strong, pitched accompaniment is present but may not provide much added benefit otherwise.

An example of melodic recovery by the dual-F0 tracking approach is shown in Fig. 4. This figure shows the ground truth pitch contour (thin) along with the F0-contours output by the single-F0 (thick), in Fig. 4(a), and dual-F0 (thick and dashed), in Fig. 4(b), tracking systems for an excerpt of an audio clip from dataset 2. The F0s are plotted in an octave scale using a reference frequency of 110 Hz. The ground truth pitch is offset vertically by −0.2 octaves for clarity. It can be seen that the single-F0 contour switches over from tracking the voice pitch to an instrument pitch (here acoustic guitar) around 6 s. However one of the contours output by the dual-F0 tracking is able to track the voice-pitch in this region since the other contour is actively tracking the guitar pitch in this region.

It is possible that the simultaneous tracking of more than 2 F0s may lead to even better melodic recovery. However such an approach is not expected to result in as significant an improvement in voice-pitch tracking accuracy as the improvement resulting in the transition from single- to dual-F0 tracking. This hypothesis is based on our premise that in vocal music the voice

is already the "dominant" sound source. On occasion, an accompanying instrument may be more locally dominant than the voice. However, we feel that the chances that two pitched instruments are simultaneously of higher salience than the voice are relatively small.

2) Comparison of TWMDP and LIWANG Algorithms: From Fig. 3 and Table IV it is seen that the TWMDP algorithm consistently, and in most cases significantly, outperforms the LIWANG algorithm. The relatively lower performance of the LIWANG system could be for various reasons. One of these could be their multi-F0 extraction module, which applies a correlation-based periodicity analysis on an auditory model-based signal representation. Such multi-F0 extraction methods require that voice harmonics are dominant in at least one channel to ensure reliable voice-F0 detection, though not necessarily high salience. Previous studies indicate that such multi-F0 extraction algorithms often get confused in two-sound mixtures, especially if both sounds have several strong partials in the pass-band [8], [38], and may not even detect a weaker sound F0 [17]. Another cause of inaccurate pitch output of the LIWANG algorithm is the limited frequency resolution, especially at higher pitches, caused by the use of integer-valued lags.

Although the LIWANG system incorporates a two-pitch hypothesis in its implementation (as described previously) and therefore has potential for increased robustness to pitched interference, its final performance for datasets 2 and 3, which are representative of such accompaniment, is significantly lower than that of the TWMDP dual-F0 tracking system. This is due to multiple reasons. For dataset 2 the lower final accuracy of this system is due to a lack of a sophisticated vocal pitch identification stage. The Either-Or accuracies for this dataset are higher than those of the TWMDP system indicating that the voice pitch is indeed present in one of the two output pitches but is not the dominant pitch and so is not the final output. For dataset 3 it was observed that the LIWANG system tracks an F0 and it's multiple rather than F0s from separate sources, which leads to lower Either-Or and final accuracies.

3) Voice-Pitch Identification: The voice-pitch identification method used in the TWMDP dual-F0 tracking system does lead to increased accuracies when compared to the single-F0 tracking system. However, the final accuracies are still below the Either-Or accuracies. This indicates that some errors are being made and there is potential for further improvement in voice pitch identification. Currently, we are using only a single temporal feature for voice pitch identification. We could, in the future, additionally exploit the temporal smoothness of timbral features such as MFCCs.

4) Errors Due to F0 Collisions: Collisions between voice and instrument pitches often causes the dual-F0 tracking output contours to switch between tracking the voice and instrument pitch contours. This is explained as follows. Around the collision, one of the contours tracks a spurious F0 candidate. If this contour is the one that was previously tracking the instrument pitch then a contour that was tracking the voice pitch may now switch over to tracking the smoother instrument pitch. This will cause discontinuities in both the contours which will result in nonhomogenous fragment formation during the voice-pitch identification process, which in turn degrades the voice-pitch identification performance. This is indicated by the larger differences between the Either-Or and final accuracies of the dual-F0 tracking system for dataset 3, which is replete with F0 collisions, as compared to dataset 2. Further, even melodic recovery may be negatively affected since the resumption of voice-pitch tracking may be delayed after a collision.

The use of predictive models of F0 contours, similar to those used for sinusoidal modeling in polyphony [39], may be investigated to ensure F0 continuity of the contours output by the dual-F0 tracking system across F0 collisions. To avoid the negative effects of spurious candidate tracking at the exact F0 collision location care would have to be taken to ensure that both contours be assigned the same F0 value at that location.

VI. Conclusion

In this paper, we have investigated voice pitch contour extraction in polyphonic music with a focus on improving pitch accuracy in the presence of strong pitched accompaniment. The major contribution of this paper is the combination of some novel and some existing methods in the design of an ultimately novel system that demonstrates superior voice-pitch tracking accuracy in polyphonic music especially in the presence of strong pitched accompaniment. The novel aspects involve the separation of the F0 candidate selection and salience computation into two distinct steps, the joint tracking of two F0 contours by the DP algorithm with a harmonic-relationship constraint on F0 pairing and the final identification of the voice pitch contour from the dual-F0 tracking output using a voice-feature that exploits the temporal instability (in frequency) of voice harmonics. The existing methods involve the use of a main-lobe matching method for the identification of sinusoids from the short-time magnitude spectrum of a signal and the TWM error as the F0 candidate salience measure. Both of these methods however have *not* previously been used in the context of melody (voice-pitch) extraction from polyphonic music. We have justified our choice of these methods both analytically and experimentally.

On evaluation using datasets with strong pitched accompaniment, it was found that the single-F0 tracking system made pitch tracking errors caused by the output pitch contour switching between tracking the voice and instrument pitches. The dual-F0 tracking approach, which dynamically tracks F0-candidate pairs generated by imposing specific harmonic relation-related constraints and then identifies the voice-pitch from these pairs, retrieves significant quantities of voice pitches. For this same test data it is also shown that the performance of the proposed single- and dual-F0 tracking algorithms is significantly better than another contemporary system specifically designed for detecting the pitch of the singing voice in polyphonic music. It is also shown that our multi-F0 extraction system reliably detects the voice-F0 in polyphony without having to adopt computationally complex iterative or joint F0 estimation approaches. Representative examples from the evaluation datasets used in this paper and the corresponding synthesized pitch contours are available at http://www.ee.iitb.ac.in/daplab/DualF0TrackingResults.

Acknowledgment

The authors would like to thank the National Centre for the Performing Arts (NCPA) Mumbai, and Y. Li and C.-L. Hsu

for sharing their multitrack audio data used in the experiments reported in this paper. They would also like to thank the team at IMIRSEL and GSLIS, University of Illinois at Urbana-Champaign, for conducting the MIREX evaluations. The authors are grateful to the anonymous reviewers and associate editor for their valuable comments which greatly helped improve the quality of this paper.

REFERENCES

[1] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[2] E. D. Scheirer, "Machine-listening systems," Ph. D. dissertation, Mass. Inst. of Tech., Cambridge, 2000.

[3] M. Lagrange, L. Martins, J. Murdoch, and G. Tzanetakis, "Normalised cuts for predominant melodic source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 278–290, Feb. 2008, (Special Issue on MIR).

[4] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 169–172.

[5] Y. Wang and B. Zhang, "Application-specific music transcription for tutoring," *IEEE Multimedia*, vol. 15, no. 3, pp. 70–74, Jul.–Sep. 2008.

[6] R. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic music signals," *Comput. Music J.*, vol. 30, no. 4, pp. 80–98, Winter, 2006.

[7] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, 2004.

[8] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[9] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 17–20.

[10] G. Poliner and D. Ellis, "A classification approach to melody transcription," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[11] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search," in *Proc. IEEE Int. Conf. Audio Speech, Signal Process.*, Toulouse, France, 2006, pp. 253–256.

[12] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 8, pp. 1223–1235, Aug. 1988.

[13] J. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[14] M. R. Every, "Separation of musical sources and structure from single-channel polyphonic recordings," Ph.D. dissertation, Dept. Electron., Univ. York, York, U.K., 2006.

[15] A. de Cheveigné, "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D.-L. Wang and G. J. Brown, Eds. New York: Wiley-IEEE Press, 2006.

[16] J. Wise, J. Capiro, and T. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 418–423, Oct. 1976.

[17] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[18] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monoaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, May 2007.

[19] R. Maher and J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Amer.*, vol. 95, no. 4, pp. 2254–2263, Apr. 1994.

[20] V. Rao and P. Rao, "Vocal melody detection in the presence of pitched accompaniment using harmonic matching methods," in *Proc. 11th Int. Conf. Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[21] A. Bapat, V. Rao, and P. Rao, "Melodic contour extraction of Indian classical vocal music," in *Proc. Int. Workshop Artif. Intell. Music (Music-AI'07)*, Hyderabad, India, Jan. 2007.

[22] J. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *J. Acoust. Soc. Amer.*, vol. 92, no. 3, pp. 1394–1402, 1992.

[23] G. Forney, "The viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.

[24] B. Secrest and G. Doddington, "Postprocessing techniques for voice pitch trackers," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Dallas, TX, 1982, pp. 172–175.

[25] H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans. Syst., Man. Cybern.*, vol. SMC-13, no. 3, pp. 208–214, Apr. 1983.

[26] P. Cook, "Pitch, periodicity and noise in the voice," in *Music, Cognition and Computerized Sound*, P. Cook, Ed. Cambridge, MA: MIT Press, 1999, pp. 195–208.

[27] M. Every and P. Jackson, "Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm," in *Proc. Interspeech'06*, Pittsburgh, PA, Sep. 2006.

[28] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 766–778, May 2008.

[29] R. Maher, "Evaluation of a method for separating digitized duet signals," *J. Audio Eng. Soc.*, vol. 38, no. 12, pp. 956–979, 1990.

[30] M. Marolt, "On finding melodic lines in audio recordings," in *Proc. 7th Int. Conf. Digital Audio Effects*, Naples, Italy, 2004.

[31] V. Rao and P. Rao, "Singing voice detection in polyphonic music using predominant pitch," in *Proc. InterSpeech'09*, Brighton, U.K., 2009.

[32] X. Serra, "Music sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. London, U.K.: Swets & Zeitlinger, 1997.

[33] M. Rynnanen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. Int. Conf. Music Inf. Retrieval*, Victoria, BC, Canada, Oct. 2006.

[34] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

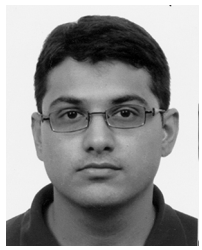
[35] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[36] C.-L. Hsu and R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1k dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[37] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[38] P. Rao and S. Shandilya, "On the detection of melodic pitch in a percussive background," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 378–390, Apr. 2004.

[39] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1625–1634, Jul. 2007.



Vishweshwara Rao, received the B.E. degree in electronics engineering from Mumbai University, Mumbai, India, in 2002, and the M.S. degree in music engineering from the University of Miami, Miami, FL, in 2004. He is currently pursuing the Ph.D. degree in electrical engineering at the Indian Institute of Technology, Bombay.

His current research interests include content analysis of Indian music and music information retrieval. Mr. Rao is also a proficient Tabla player

Preeti Rao, received the B.Tech. E.E. degree from the Indian Institute of Technology (IIT), Bombay, in 1984, and the M.S. and Ph.D. degrees from the University of Florida, Gainesville, in 1987 and 1990, respectively.

She taught at IIT Kanpur from 1994 until 1999, and has since been on the faculty at IIT Bombay where she is currently a Professor in the Department of Electrical Engineering. She has also held visiting positions at Hitachi CRL in Tokyo and at IPO, Center for User-System Interaction, Eindhoven, The Netherlands. Her research interests are in speech and audio signal processing with a focus currently on music content analysis and retrieval and acoustic–phonetic approaches to speech recognition.