Presented at National Conference on Communications, NCC 2003, Jan 31 –Feb 2, IIT Madras 2003

RETREIVING PITCH OF THE SINGING VOICE IN POLYPHONIC AUDIO

Saurabh Kumar Shandilya and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai 400076

saurya@ee.iitb.ac.in prao@ee.iitb.ac.in

ABSTRACT

This paper explores the extraction of melodic pitch contour from the polyphonic soundtrack of a song. The motivation for this work lies in developing automatic tools for the melody-based indexing of the database in a music retrieval system. The melody is assumed to be carried by the singer's voice accompanied mainly by percussive instruments. This scenario is typical of a large class of Indian movie songs. The challenges raised by this application are presented. A pitch detection method based on a perceptual model is shown to be a promising approach to the tracking of voice pitch in the presence of strong percussive background.

1. INTRODUCTION

The advent of Internet has made it feasible for people to share not only thoughts but also emotions and music. The soul of music is melody and that of melody is pitch. So the importance of pitch detection in music related software is significant. In this paper we address the problem of pitch detection associated with the singing voice in the context of the polyphonic recordings of Indian movie songs.

In the field of speech processing, pitch detection algorithms have always evoked much research interest. Every new application demands a new approach. Our problem can be viewed as that of estimating the time-varying pitch of a singing voice in the presence of orchestral background. This work is motivated by the requirements of query-by-humming based music retrieval systems. One such system under development at IITB [1] targets a database of Indian film songs. In the interest of building an indexed database, as well as enabling its continuous augmentation with new songs, it would be very valuable to have available a tool that automatically extracts the needed pitch information from original polyphonic sound tracks. The problem is especially important for our target database since, unlike Western music corpora (e.g. [2]), accurate MIDI versions of Indian music are very scant or non-existent.

The real challenge to pitch detection pops up in the form of the polyphonic nature of the songs. Various instruments keep coming and disappearing at intervals along with the main singing voice. Computer scientists view this problem as that of CASA (Computational Audio Scene Analysis) [3] where they are primarily interested in identifying the various objects (instrument voices) that contribute to the audio scene at any time. This step may be useful to solving the problem of estimating features of interest from specific voices but is not essential [3,4,5]. Early works in the field were too specific but later on the impetus shifted towards a general approach which can give results independent of type and number of instruments being played. While most of the work has been done under the name of multiple-pitch-detection, where the sources are assumed to be harmonic, the scenario is not exactly same in our case as we will see soon.

In this paper we test the application of a perception model based pitch detection algorithm and investigate the advantages that it offers against conventional pitch detection algorithms. But before that we present some interesting facts about the genre of Hindi movie songs, which help us in defining our problem better.

2. SOURCE BASED CLASSIFICATION

At a coarse level, audio segments in a soundtrack can be classified into three classes-silence, orchestral background only, and background plus singing voice. The melody of the song is typically carried in the time-varying pitch of the singing voice. We therefore focus on the this class, which further can be characterized by the particular type of background instrument: percussion or non-percussive. Percussive sounds may have strong partials but are not strictly harmonic. From a study of songs in the database of our interest, we have found that *segments containing singing voice almost always contain purely percussive instruments in the background.* These instruments contribute to the rhythm but not to the melody of the music. So, the problem boils down to finding accurate pitch-contour of the voice in the presence of percussive sounds. Further, in the paradigm of Indian music, the percussions belong to one two categories [6]: membranous percussions coming from instruments with struck membranes e.g. drum*, tabla, pakhawaj, dhol, nall,* etc., and non-membranous percussions coming from instruments which have solid resonators e.g. *jal-tarang, manjira, ghatam, chimpta, ghungaru* etc.

Percussive sounds belonging to the class of membranous percussions are characterized by sharp increase in instantaneous energy followed by an exponential decay. These sounds are extremely dynamic in nature and they disturb the harmonic structure of the singing voice when they occur. In the frequency domain these onsets are detected by a smearing across frequency. Percussions of these kinds don't pose a big problem as the effect dies down very quickly, and pitch estimate may go wrong for only a small fraction of the duration of the note. These deviations can be rectified by median smoothing the pitch values over note duration (assuming the notes are separated by some method). But the percussions of non-membranous variety leave a trail of strong partials in some regions that decay only slowly. This can cause the pitch estimate to go haywire over a significant portion of the note, especially when the energy of the voice signal is not high relative to the background.

3. PITCH DETECTION ALGORITHMS

A pitch detection algorithm works in three stages: (i) preprocessing, which may be linear or non-linear filtering; (ii) estimation of candidate pitch-values from the pre-processed signal; (iii) post-processing to reject / modify spurious pitch values. Candidate pitch values are generated by either pure time-domain methods or spectral domain methods. Autocorrelation function (ACF) based method is the most widely used time-domain method. A related algorithm in this is the Rabiner-Gold parallel processing algorithm, which extracts pitch by running decision logic on several derived time-domain parameters [7]. Spectral domain methods look for harmonicity in the frequency domain to find out pitch; i.e. they look for partials in specific spectral locations, or for the intervals between partials. The above methods do not explicitly take into account any model of pitch perception by humans. Since we can easily identify a tune even in the presence of background percussions, it is reasonable to expect that a study of the pitch perception mechanism can offer significant clues, which may help in extracting the melody information that is buried in a percussive background.

In 1991, Meddis and Hewitt proposed a theory for the subjective perception of pitch [8]. This model is essentially based on detecting the dominating spectral interval by observing the beats between frequency partials. The theory incorporates several distinct functional blocks of the hearing mechanism. The acoustic signal entering the ear undergoes a filtering equivalent to outer/middle ear transformation. In the inner ear, the basilar membrane acts as a bank of bandpass filters with bandwidth increasing with center frequency. The output of each filter represents the motion of basilar membrane and the next stage involves conversion of this motion to neural signals. The model hypothesizes that the probability of spike occurrence is proportional to the permeability of the hair cell membrane. Permeability in turn is related to the instantaneous amplitude of the filtered signal. At the next stage, for each filtered channel the histogram of time intervals among spikes (not just successive spikes) is generated. A running ACF is one practical way to simulate this. Finally these autocorrelation functions along different channels are summed up to give summary autocorrelation function, which is examined for peaks to detect the pitch period. So the important conclusion here is that the process of pitch perception involves processing in individual bands. These bands are also represented best using psychoacoustically derived filters.

4. IMPLEMENTATION OF THE M-H MODEL ALGORITHM

Borrowing the important keys from the above described theory we have implemented an algorithm, which imitates the behavior of human ear in a simplistic way. The input audio signal in 16 bit PCM format, sampled at the rate of 44.1 kHz is passed through an auditory filter-bank as obtained from VOICEBOX [9]. We have used 27 standard "gammatone" filters with center frequencies ranging from 123 Hz to 5.636 kHz or 4 to 30 on the 'equivalent rectangular bandwidth' scale (a slight modification of the critical band scale) [10]. In each frequency band the amplitude envelope of the filtered signal is obtained. This can be achieved by using standard demodulation technique. First the signal in each channel is half-wave rectified. Then the half-wave rectified signal is low pass filtered to allow frequencies up to 700 Hz, assuming that this is the maximum expected pitch frequency. Envelope calculation by half-wave rectification is consistent with the Meddis-Hewitt ear model, according to which probability of spike generation is proportional to instantaneous amplitude, and low-pass filtering is justified by the arguments relating to refractory periods.

In each band an ACF is calculated. Keeping in view the range of 150-700 Hz in which we are expecting the pitch to lie in, we have used a 20 ms frame and frame-shift of 50%. We again had a choice here between ACF and circular ACF and we have used both the approaches. An ACF assumes the signal to be zero outside the frame while circular ACF assumes the pattern to be repeating outside the window of size N.

$$acf(k,\tau) = \sum_{i=0}^{N-\tau-1} y(k+i) * y(k+i+\tau) \quad \ldots(1)$$

$$acf_c^{cir}(k,\tau) = \sum_{i=0}^{N-1} y_c(k+i) * y_c(k+\mathrm{mod}(i+\tau,N)) \quad \ldots(2)$$

Where k and τ are position of window and correlation lag respectively and y is the input signal. The final task is more analytical in nature as we have 27 ACF arrays each of them containing information about frequency content in each band. Lots of conclusions can be drawn regarding pitch-values, harmonicity, note onset etc. Constraining ourselves to pitch-detection we build a summary autocorrelation function which is a summation of various ACFs (across the channels) raised to power "p". We are currently using a power of 2.

$$ACFsq_{summary} = \sum_c (ACF_c)^p \quad \ldots(3)$$

This causes the underlying pitch to reinforce peaks at corresponding locations.

5. EVALUATION

We have compared the accuracy of the pitch-estimates obtained by this method with that obtained by the traditional autocorrelation based method. Traditional method involved calculation of ACF using Eqn.1, with window size of 20ms and shift of 50%. The position of the most prominent secondary peak is taken as the estimate of pitch period.

Evaluation of a pitch detector can be done by comparing with manual estimates (observe waveform, spectrum) or by synthesizing a sound and listening. The typical pitch errors in the context of this paper are gross (not fine) and therefore can be easily detected by listening. A comparison to the original song was performed by listening to the resynthesized melody using estimated pitch values and local energy estimates (after smoothing out the narrow bursts of energy due to percussive strikes). The resynthesis program generates tones of fundamental frequency equal to the specified pitch values (or its integral multiple, which is equivalent to playing same note on higher octaves).

To test the algorithm we have used a number of audio segments, which had different percussive instruments ranging from drum to *chimpta* (non-membranous percussion). In order to provide insight into the problem and the different approaches, we discuss in detail one example taken from the film song *"jai jai shiv Shankar kaata lage naa kankar"* from an old Hindi movie which had *chimpta* and *manjira* in the background. Fig. 1 shows a spectrogram of the second phrase with manually marked syllable-notes. Fig. 2 shows the energy contour and pitch contours as obtained from the two different pitch detection methods. The raw autocorrelation based method failed to detect the correct pitch at certain instants e.g. during note /kan/. Figure 3 shows how the wrong peak was picked up. This has happened at a place where the syllable (the nasal "kan") was weak and percussion was strong. Low energy of the sung syllable is evident from the low-levels of energy contour near frames 150-165 (See Fig. 2(a)). So the interference from background affects the low-energy voice regions strongly. The gist of this is that the relative strength of voice and percussion energy plays a crucial role. The M-H method appears to be more robust, successfully retrieving the perceptually correct pitch. Also in portions where either a note is starting/ending or a percussion instrument has just been struck, the M-H method fares much better than the traditional ACF pitch detector. Audio clips and test results are available at [11].

The failure of ACF based approach can be explained by the fact that presence of strong partials from the percussion perturbs the voice pitch peak pattern due to additional temporal modulation.

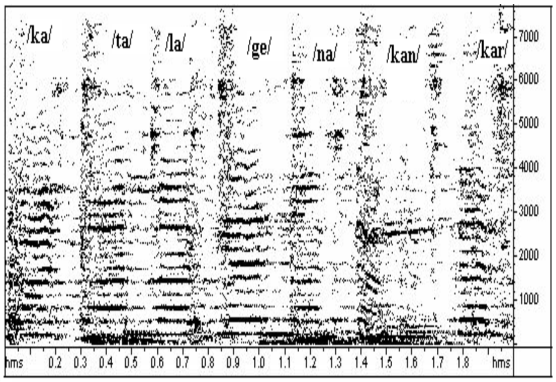

Figure 1: Spectral view of audio fragment *"kata lage na kankar"*

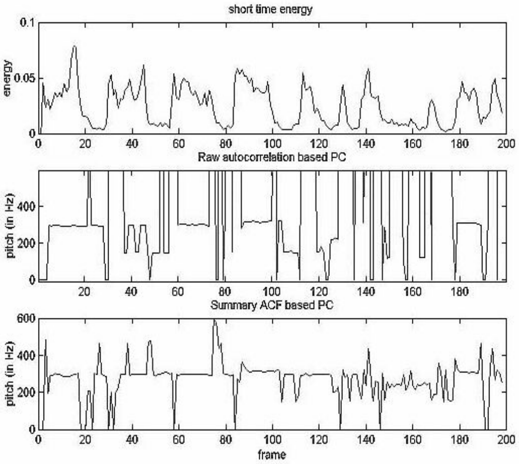

Figure 2: (a) Short time energy variation for song of Fig. 1; (b) pitch contour obtained by raw ACF based method (c) pitch contour obtained by M-H method.

On the other hand, the M-H method does independent processing in separate bands, so corruption of harmonic content by a few strong partials is reduced. In any given frequency band, the presence of two or more harmonics leads to beats of frequency equal to the difference of two harmonics, which is the fundamental frequency itself. Envelope detection followed by band-pass filtering is a step towards obtaining this beat frequency. A good harmonic structure as is found in singing voice strengthens the peak in the ACFs at a lag equal to pitch period, when individual ACFs are combined. Special care has to be taken when the fundamental frequency is really high as in the case of female singers. It may happen that one band may comprise only one partial and no beats are generated, hence envelope detection followed by low pass filtering may be misleading. In such cases i.e. when fundamental frequency turns out to be higher than 500 Hz, the value should be cross checked with the summary ACF for the filtered signals on which envelope detection has not been performed and the lower of the two values should be chosen.

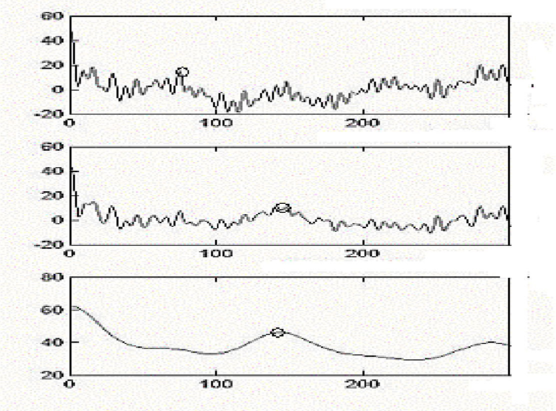

Figure 3: (a) Raw circular ACF of a frame of note /kan/ over non-redundant portions only (Circles indicate peak used for computation of pitch) (b) Summary ACF for filtered-only signals (c) Summary ACF for envelope of filtered signals.

6. CONCLUSIONS

We have investigated the application of a perception based pitch-detection algorithm to the problem of melody retrieval from the polyphonic audio of Indian film songs. It was shown that the traditional ACF based method could fail, esp. in the presence of non-membranous percussive sounds. The Meddis-Hewitt pitch model based method seems to be promising for pitch detection especially when the energy related with the singing voice is low. Its higher robustness may be attributed to the independent bandwise processing for ACF peaks. Work is in progress on refining this algorithm and testing its applicability to a variety of audio samples. The incorporation of prior knowledge also needs to be considered due to its potential in improving performance further by enabling the tuning of algorithm parameters. If the nature of percussion (e.g. spectral properties) is known, calculation of summary ACF can be restricted to uncorrupted bands only. This would suggest a better way of combining the ACFs. Development of decision logic for presence of singing voice will take us yet another step ahead in the direction of obtaining note boundaries and hence note-contours.

REFERENCES

1) M. A. Raju and P. Rao, *Building a Melody Retrieval System*, Proc. NCC, IITB, 2002; also *TANSEN: A qbh based music retieval system* , submitted to NCC 2003.
2) Tuneserver: http://name-this-tune.com
3) M. Goto and S. Hayamizu, *A Real-time Music Description System: Detecting Melody and Bass Lines in Audio Signals*, Workshop on CASA, pp. 31-40, Aug-99
4) A. Klapuri, *Wide-band pitch estimation for natural sound sources with inharmonicities, 106th Audio Engineering Society Convention*, Munchen, Germany, 1999
5) A. Klapuri, T. Virtanen, J. M. Holm, *Robust multipitch estimation for the analysis and manipulation of polyphonic music signals*, Proc. COST-G6 Conference on Digital Audio Effects, Italy, 2000.
6) Indian instruments by David Courtney http://chandrakantha.com/articles/indian_music/instruments.html
7) B. Gold and L. Rabiner, *Parallel processing techniques for estimating [pitch periods of speech in time domain*, JASA, Vol. 46(2), 1969.
8) R. Meddis and M. J. Hewitt, *Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,* JASA, Vol. 89, June 1991
9) Voicebox home page, Deptt. Of Electrical and Electronics Engg., Imperial College, London. www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, August 2001.
10) Global Index, J.O.Smith, CCRMA, Stanford University, http://www-ccrma.stanford.edu/~jos/bbt/
11) www.ee.iitb.ac.in/uma/~saurya/ncc.htm