OBJECTIVE EVALUATION OF A MELODY EXTRACTOR FOR NORTH INDIAN CLASSICAL VOCAL PERFORMANCES

Vishweshwara Rao and Preeti Rao
Digital Audio Processing Lab, Electrical Engineering Department,
IIT-Bombay, Powai, Mumbai – 400 076.
{vishu, prao}@ee.iitb.ac.in

Abstract

Obtaining accurate melodic contours from polyphonic music is essential to several music-information-retrieval (MIR) applications and is also useful from a musicological perspective. The presence of tabla and tanpura accompaniment in north Indian classical vocal performances, however, degrades the performance of common pitch detection algorithms (PDAs) that are known to provide accurate results when presented with monophonic singing voice. Recently, a melody extraction algorithm designed to be robust in the presence of the accompaniment was proposed. In this work, the same melody extraction algorithm is tested on actual professional vocal performances, facilitated by the availability of time-synchronized multi-track recordings of the voice, tabla and tanpura. The results indicate that the new method is indeed very robust to the presence of accompaniment clearly overcoming the limitations posed to melody extraction by the widely used monophonic PDAs.

Keywords: Melody extraction, Pitch detection, Indian classical music

1. Introduction

Obtaining accurate melodic contours from polyphonic music is important to several music-information-retrieval (MIR) applications apart from being essential in musicological studies. In north Indian classical vocal music the melody is carried by the voice. What makes the task of melody extraction particularly challenging is the typical polyphonic setting in north Indian classical music where the voice is accompanied by a drone such as the *tanpura*, providing the fixed tonic, and rhythm provided by a percussive instrument, capable of producing pitched-sounds, such as the *tabla*. The presence of the accompaniment leads to degradation in the performance of common pitch detection algorithms (PDA) which typically work well on monophonic music. Due to these factors, musicological research involving voice pitch analysis is often constrained to using specially created monophonic voice recordings [1] [2].

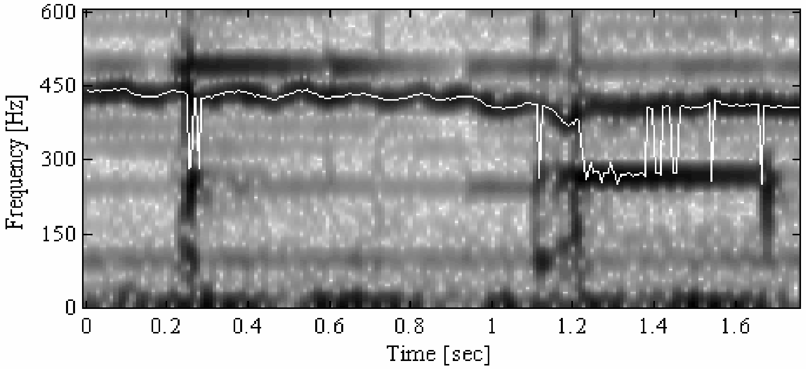

Figure. 1. Pitch contour *(white line)* as detected by a modified ACF PDA[3] superimposed on the zoomed in spectrogram, of a segment of Indian classical music that contains a female voice and a drone throughout and *tabla* strokes in some regions. [4]

An illustration of the degradation caused by *tabla* percussion is seen in Fig. 1, which shows the melodic contour estimated by a modified ACF PDA [3] with recommended parameter settings. The estimated melodic contour is superimposed on a spectrogram of the signal, a segment from a classical vocal recording. In this segment, the sequence of *tabla* strokes is as follows: impulsive stroke (0.22 sec), impulsive stroke (1.15 sec), tonal stroke (1.2-1.7 sec), and impulsive stroke (1.7 sec). The impulsive strokes appear as vertical narrow dark bands. The tonal stroke (associated with the mnemonic '*Tun*') is marked by the presence of a dark (high intensity) horizontal band around 290 Hz, which corresponds to its fundamental frequency. The other, relatively weak horizontal bands correspond to *tanpura* partials. We note that all the strokes degrade the performance of the PDA, which is otherwise able to accurately track the pitch of the voice in the presence of the *tanpura*, as indicated by the region where the melodic contour overlaps with the dark band in the spectrogram corresponding to the voice fundamental frequency (between 0.4 and 1 seconds). While the errors due to the impulsive strokes are localized, and so may be corrected by known smoothing techniques such as low-pass/median filtering, the tonal stroke causes errors that are spread over a long segment of the melodic track. These latter errors are not just voice octave errors but interference errors i.e. when the melody estimated is actually the interference pitch, indicated by the lower dark band present temporarily between 1.2 and 1.7 seconds.

In a recent publication [4], we had proposed a melody extraction algorithm for polyphonic recordings of north Indian classical singing, which was robust to *tabla* interference. The algorithm was evaluated for pitch estimation accuracy on a wide range of simulated voice and *tabla* signals. For real audio data (available recordings of vocal performances), however, the accuracy of the algorithm was measured only qualitatively by subjective listening comparisons of a melody synthesized from the estimated contour with the original recording. Objective evaluation was not possible since there was no practical way to obtain the "ground truth" melodic pitch values for real performance data. However, the recent availability of excerpts of time-synchronized multi-track recordings of individual instruments (voice, *tabla*, *tanpura*) has made the measurement of ground truth pitch possible, and new evaluation results are presented here. In the next section, we provide a brief overview of the PDA and propose a pre-processing technique for *tanpura* suppression. Following that we present pitch accuracy results of the proposed algorithm on the polyphonic recordings.

2. Melody extraction and pre-processing

2.1. Melody extractor

The melody extractor proposed in [4] is based on pitch tracking using the two-way mismatch algorithm (TWM) [5] followed by a post-processing operation of dynamic programming (DP)-based smoothing [6]. The TWM PDA is an example of a frequency domain PDA that detects the fundamental frequency (F0) as that which best explains the measured partials of the signal i.e. that which minimizes a mismatch error, which is computed between a predicted harmonic spectral pattern and the spectral peaks detected in the signal. Here the analysis window length used is the minimum length required to resolve the harmonics of the minimum expected F0 and pitch estimates are generated every 10 ms. Since we are interested in the voice F0 and significant voice harmonics are mainly present below 5 kHz we only consider spectral content uptil 5 kHz. The choice of the TWM PDA over other PDAs was based on the knowledge that the TWM algorithm gives more weight to signals whose harmonics have greater spectral spread and the spread of the significant voice harmonics (5 kHz) is greater than that for the tonal *tabla* strokes (2 kHz). At every analysis time instant (frame) the TWM PDA outputs a list of F0 candidates and associated reliability or confidence values. These are then input into a smoothing technique based on DP.

The operation of DP can be thought of as finding the globally optimum path through a state space, where, for a given frame, each state represents a possible F0 candidate. With each state are associated two costs, the measurement and smoothness costs. The measurement cost is derived from the reliability value of a particular pitch candidate. The smoothness cost is the cost of making a transition from a particular candidate in the previous state to a particular candidate in the current state. A local transition cost is defined as the combination of these two costs over successive frames. An optimality criterion to represent the trade off between the measurement and the smoothness costs is defined in terms of a global transition cost, which is the cost of a path passing through the state space, by combining local transition costs across a singing spurt. The path, or F0 contour, with the minimum global transition cost, for a given singing spurt, is then the estimated melodic contour. The use of DP-based smoothing seems favorable to voice pitch extraction since the voice signal, over a singing spurt, is expected to be continually present as compared to the *tabla* signal, which is intermittent. This favorability will hold as long as the reliability values of the voice pitch candidates during *tabla* strokes are significant.

2.2. Spectral subtraction-based pre-processing

As discussed in the previous section, the combination of the TWM PDA with DP-based smoothing is inherently relatively robust to *tabla* (percussive) accompaniment. Although the *tanpura* sound is audibly quite prominent relative to the singer's voice, its energy is spread over a very large number of partials throughout the spectrum up to 10 kHz, and its overall strength is very low. As such the performance degradation caused by the *tanpura* to most PDAs is much less than that caused by the *tabla*. However, the *tanpura* (drone) is known to cause occasional pitch estimation errors.

Here we propose a pre-processing method for *tanpura* suppression that makes use of spectral subtraction [7], which is a well known technique for noise suppression in speech communication. As its name implies, it involves the subtraction of an average noise power spectrum, estimated during non-speech regions, from the power spectrum of the noisy signal. The enhanced signal is reconstructed from the modified magnitude spectrum and original phase spectrum of the noisy signal. The assumptions made are that the noise is additive and stationary to the degree that its spectrum in the non-speech regions is similar to that during speech.

The application of spectral subtraction in the context of *tanpura* suppression exploits 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 and then subtracted from all subsequent frames in the mixed track. Such a long segment is used to average out the effects of plucking strings tuned to different F0s. The resulting spectrum is subjected to half wave rectification and finally, the signal is reconstructed using the overlap-add (OLA) method. In the resulting *tanpura*-suppressed signal 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.

3. Evaluation

The multi-track test 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 *taans* 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 by common monophonic PDAs. The ground truth pitch is then used to evaluate the accuracy of the proposed TWM-DP pitch detection algorithm on the corresponding polyphonic recording created by mixing at normally expected levels.

3.1. Ground truth computation and evaluation metric

The procedure for computation of the ground truth melodic contour is as follows. First the F0 contours are extracted from the clean voice tracks using a combination of three different PDAs, each of which is known to independently perform well on monophonic signals, and DP-based smoothing. The PDAs used here are YIN [8], SHS [9] and TWM [5]. The three PDAs result in three pitch contours for a single excerpt with pitch estimated every 10 ms. The PDAs are each based on essentially different assumptions regarding the underlying signal periodicity and hence tend to react differently to the different signal perturbations. At each time instant, a pitch estimate is labeled as the ground-truth pitch if two out of the three estimated pitches are in concurrence (i.e. the two pitch estimates are within 3% of each other). 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.

In order to objectively evaluate the accuracy and robustness of the proposed algorithm to *tabla* and *tanpura* accompaniment, we use a measure of pitch accuracy (PA) as the evaluation metric. Here PA, for each excerpt, is computed as the percentage of frames for which the pitch estimate of the proposed algorithm and the ground-truth are in concurrence.

3.2. Results

Voice + *tabla*: For each voice excerpt, its time-synchronized *tabla* counterpart was added at an audibly acceptable, global signal-to-noise ratio (SNR) of 5 dB. The first two rows of Table 1 show the comparison between the PA values for the proposed algorithm (with respect to the "ground truth") on the clean voice and the mixture of voice and *tabla* respectively. 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.

Voice + *tanpura*: In the case of the *tanpura*, the audibly acceptable global SNR with respect to the voice was found to be 20 dB. Row 3 of Table 1 shows the PA values of the proposed algorithm on the mixture of voice with *tanpura*. There is some degradation in PA when compared to row 1. Row 4 of Table 1 shows the PA values of the proposed algorithm on the mixture of the voice and *tanpura* after spectral subtraction. There is a noticeable improvement in accuracy as compared to row 3.

Table 1. PA values for the proposed algorithm (TWM + DP) for each of the four excerpts for clean voice, voice + *tabla*, voice + *tanpura* and voice + *tanpura* after spectral subtraction (SS)

Audio content	**Male Pt. 1**	**Male Pt. 2**	**Female Pt. 1**	**Female Pt. 2**
Clean voice	99.14 %	99.32 %	99.66 %	99.85 %
Voice + *tabla*	99.14 %	97.94 %	99.31 %	99.74 %
Voice + *tanpura*	98.94 %	95.56 %	98.56 %	96.99 %
Voice + *tanpura* (SS)	99.06 %	97.68 %	99.42 %	99.56 %

Voice + *tabla* + *tanpura*: Table 2 shows the pitch accuracies obtained on the finally mixed recordings. The mixed recordings are obtained by combining the time-synchronized voice, *tabla*, at 5 dB SNR, and *tanpura*, at 20 dB SNR before and after spectral subtraction so that the mixed signal sounds like the typical recording of a vocal performance. To place the performance of the proposed algorithm in perspective, Table 2 also provides the PA values obtained by a well-known and commonly available PDA (ACF) [3]. The results clearly demonstrate the superiority of the proposed algorithm for melody extraction from typical north Indian classical music recordings.

4. Summary

This paper evaluates a recently proposed melody extractor for north Indian classical vocal performances. The objective evaluation of pitch accuracy was facilitated by the availability of time-synchronized multi-track recordings of the voice, *tabla* and *tanpura* from professional vocal performances. The proposed melody extractor was evaluated for only *tabla* and only *tanpura* accompaniment as well as for both together added to the voice. The high values of pitch accuracy reported indicate that the proposed pitch tracker, in conjunction with spectral subtraction-based pre-processing, will accurately extract melodies from typical north Indian classical vocal performances where the accompanying instruments are the *tabla* and the *tanpura* only. Further work involves the investigation of the effect of a secondary melodic instrument, such as a *harmonium*, on the proposed melody extraction algorithm. The eventual goal is the automatic transcription of north Indian classical music including all melodic parts as well as detection and labeling of *tabla* strokes.

Table 2. PA values for ACF + DP and TMW + DP for each of the four excerpts for clean voice, voice + *tabla* + *tanpura* and voice + *tabla* + *tanpura* after spectral subtraction (SS)

PDA	**Audio content**	**Male Pt. 1**	**Male Pt. 2**	**Female Pt. 1**	**Female Pt. 2**
ACF+DP	Clean voice	99.01 %	98.49 %	99.84 %	99.85 %
	Voice + *tabla* + *tanpura*	49.14 %	44.09 %	50.46 %	73.84 %
	Voice + *tabla* + *tanpura* (SS)	48.25 %	43.75 %	50.17 %	72.31 %
TWM+DP	Clean voice	99.14 %	99.32 %	99.66 %	99.85 %
	Voice + *tabla* + *tanpura*	97.88 %	94.42 %	97.66 %	95.93 %
	Voice + *tabla* + *tanpura* (SS)	98.84 %	96.98 %	99.11 %	99.42 %

Acknowledgements

We would like to thank Mr. D. B. Biswas and Dr. S. Rao of the National Centre for the Performing Arts (NCPA), Mumbai for providing us with the individual voice, *tabla* and *tanpura* track excerpts, which were used for the experiments reported in this study.

References

[1] Datta, A. et. al. (2002) “Studies on identification of Raga using short pieces of Taan: A signal processing approach,” *Journal of the ITC Sangeet Research Academy (SRA)*, vol. 16, pp. 63-74.

[2] Chordia, P. (2006) “Automatic Raga classification of sarod and vocal performances using pitch-class and pitch-class dyad distributions,” *Journal of the ITC Sangeet Research Academy (SRA)*, vol. 20, pp. 65-82.

[3] Boersma, P. (1983) “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proc. of the Institute of Phonetic Sciences*, Amsterdam, vol.17, pp. 97-110.

[4] Bapat, A., Rao, V. and Rao, P. (2007) “Melodic contour extraction for Indian classical vocal music,” *Proc. of Music-AI (International Workshop on Artificial Intelligence and Music) in IJCAI*, 2007, Hyderabad, India.

[5] Maher, R. and Beauchamp, J. (1994) “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254-2263.

[6] Ney, H. (1983) “Dynamic programming algorithm for optimal estimation of speech parameter contours,” *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-13, no. 3, pp. 208-214.

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

[8] de Cheveigné, A. and Kawahara, H. (2002) “YIN, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917-1930.

[9] Hermes, D. (1988) “Measurement of pitch by sub-harmonic summation,” *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257-264.