Acoustic Features for Robust Classification of Mandarin Tones

Hongbing Hu\textsuperscript{1}, Stephen A. Zahorian\textsuperscript{1}, Peter Guzewich\textsuperscript{1}, Jiang Wu\textsuperscript{1}

\textsuperscript{1}Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, 13902, USA
{\{hongbing.hu, zahorian, peter.guzewich, jiang.wu\}@binghamton.edu}

Abstract

For applications such as tone modeling and automatic tone recognition, smoothed $F_0$ (pitch) all-voiced pitch tracks are desirable. Three pitch trackers that have been shown to give good accuracy for pitch tracking are YAAPT, YIN, and PRAAT. On tests with English and Japanese databases, for which ground truth pitch tracks are available by other means, we show that YAAPT has lower errors than YIN and PRAAT. We also experimentally compare the effectiveness of the three trackers for automatic classification of Mandarin tones. In addition to pitch tracks, a compact set of low-frequency spectral shape trajectories are used as additional features for automatic tone classification. A combination of pitch trajectories computed with YAAPT and spectral shape trajectories extracted from 800ms intervals for each tone results in tone classification accuracy of nearly 77%, a rate higher than human listeners achieve for isolated tonal syllables, and also higher than that obtained with the other two trackers.

Index Terms: pitch tracking, tone classification, Mandarin Chinese, fundamental frequency

1. Introduction and Background

Accurate fundamental frequency ($F_0$) (commonly referred to as pitch)—the terms pitch and $F_0$ are used interchangeably in this paper) tracking in speech remains an elusive goal, especially for noisy and/or band-limited speech, typically the scenarios where reliable pitch tracking would be most useful. Good results have been reported by Talkin in RAPT where a normalized cross correlation function is used [8]. High accuracy pitch tracking results have also been obtained by the YIN algorithm, which uses a modified version of the autocorrelation method [2]. Probably the most widely used tool for pitch tracking is the speech analysis program PRAAT [1] because it provides fairly reliable tracking and is readily available. Since about 1980, several pitch trackers have been developed and several studies have been done to evaluate these trackers [5, 7]. Our own tool for pitch tracking is named YAAPT for “Yet Another Algorithm for Pitch Tracking” [12].

For automatic recognition of tones in tonal languages such as Mandarin, robust all-voiced pitch tracking is especially important, as pitch is widely considered as the most important acoustic correlate of a tone [1][9].

In this paper, we first summarize and illustrate the YAAPT method in the remainder of this section. Section 2 introduces several modifications motivated by the desire to improve automatic tone classification and describes a method for computing spectral temporal features, which are effective in addition to pitch for use in tone classification. The evaluation results of several experiments, which illustrate the effectiveness of YAAPT and the additional features, are reported in Sections 3 and 4. For control purposes, experimental results obtained with YIN and PRAAT pitch trackers are also given.

The main signal processing steps in YAAPT are illustrated in Figure 1 [12]. For each frame of speech, multiple pitch candidates are computed using the normalized cross correlation. A smoothed pitch track is computed from the spectrogram of the squared signal; to some extent the squaring restores the fundamental, which is likely to be missing from band-limited speech such as telephone speech. All $F_0$ candidates, both time domain and frequency domain, as well as an unvoiced candidate, are assigned merit values and the highest overall merit path is determined using dynamic programming. More details, as well as illustrations of the various steps involved, are given in [12]. All three of these trackers have settings to minimize “Gross” error (large errors in the voiced sections of speech) or “Big” error, which takes into account both large errors in the voiced speech regions, and voiced/unvoiced decision errors. Unfortunately, neither of these minimum error cases is best suited for computing pitch tracks for Mandarin tone classification.

2. Algorithms

2.1. YAAPT improvements

The most significant change is the introduction of additional post-processing techniques to refine the final pitch tracks, especially for the case when the track is intended to be all voiced. With the previous settings, as given in [12], optimized for minimum gross error, visual inspection of computed pitch
tracks showed apparent abnormalities, especially in the interpolated pitch values through unvoiced regions. Nevertheless, the gross error values for YAAPT were low, since the estimated pitch values in actual unvoiced regions were not considered in the error calculation.

![Pitch Tracking Original vs Revised](image)

**Figure 2:** Illustration of YAAPT F0 tracking. Original (Blue), revised (Red), ground truth reference (Black).

To improve YAAPT, the algorithm was changed and now always determines the minimum big error track with voicing decisions, even if finally an all voiced (minimum gross error) track is desired. Heuristics are then incorporated to identify and eliminate pitch values which appear to be in error due to pitch halves or doubles. If a track with minimum gross error is desired, post processing then includes cubic polynomial interpolation through the unvoiced regions using a filtered version of the calculated track. This method was empirically determined to work effectively at reducing error and producing a smooth track. Figure 2 above depicts the ground truth pitch track, the former YAAPT track, and the YAAPT track with the modifications introduced in this paper.

The last of the modifications to YAAPT was a code refinement to improve the processing time and accuracy. One of the more significant of these modifications was to change an inner loop for the spectral harmonic correlation calculation to reduce computational time. Two other changes to this section of code helped improve overall performance by more accurately calculating the spectral track, even with a shorter FFT length for spectral calculations. These changes corrected for possible frequency misalignment between temporal and spectral pitch candidates, which depended on the frequency resolution (FFT length). Consequently, a shorter FFT length can be used, decreasing computational time, while not significantly degrading performance. These code refinements decreased overall computation time by around 25% and decreased error rates by small percentages.

2.2. Additional spectral temporal features useful for tone classification

In our initial work with Mandarin tone recognition [10], we observed that the four primary Mandarin tones (High, Rising, Falling, Dipping) were also relatively apparent from inspection of the low frequency region of the spectrogram. Therefore, global spectral shape trajectories, computed with a small number of spectral Discrete Cosine Transform Coefficients (DCTCs) each of which is encoded with several Discrete Cosine Series Coefficients (DCSCs), appeared to be a relatively effective approach for computing tone features. The details of DCTC and DCSC calculations are given elsewhere [13]. Summarizing briefly, DCTCs are coefficients of a cosine-like basis vector expansion of speech log magnitude spectra, where the cosine basis vectors are modified to take into account a mel-like frequency scale. A DCTC representation of speech spectra is a smoothed representation, with degree of smoothing determined by the number of DCTCs used. A DCSC encoding of any feature over time (such as a DCTC term or pitch) is a cosine basis vector expansion over time, but with the cosine basis vectors modified to give more resolution near the center of the time interval and less resolution near the endpoints of the interval. In our work, the time resolution of a DCSC representation was determined by a Kaiser window, with the degree of resolution variation determined by the Kaiser constant. The DCTCs/DCSCs are similar to MFCCs and delta/acceleration terms [11], but more general and flexible.

### 3. Experimental Evaluations of Pitch Tracking Accuracy

In order to evaluate the accuracy of YAAPT for pitch tracking accuracy, pitch tracks were computed from two databases, the Keele pitch database [6] and a Japanese database [2], for which ground truth pitch tracks are available. The Keele database contains 10 sentences, each about 30 seconds long, with each sentence spoken by a different British speaker. Both studio quality and telephone versions of the speech were used. The Japanese database consists of 30 utterances by 14 male and 14 female speakers, resulting in a total of 840 utterances.

The pitch tracks were computed and compared using YAAPT, YIN, and PRAAT for both full bandwidth and telephone and/or simulated telephone conditions. Tests were done with clean versions of the speech and also at 5 dB SNR levels with additive white noise and additive babble noise. For YAAPT and PRAAT, tracks were computed both for an all-voiced condition and a condition for which the tracker made voiced/unvoiced decisions. For YIN, the track is always considered to be all-voiced, so that was the only case tested.

Results in terms of Big Error and Gross Errors are given in tables 1, 2, and 3 for clean speech, white noise at a 5 dB SNR, and babble noise at a 5 dB SNR.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Clean</th>
<th>W-5</th>
<th>B-5</th>
<th>Clean</th>
<th>W-5</th>
<th>B-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAAPT</td>
<td>3.1</td>
<td>3.4</td>
<td>7.9</td>
<td>4.6</td>
<td>6.4</td>
<td>28.2</td>
</tr>
<tr>
<td>PRAAT</td>
<td>5.2</td>
<td>7.8</td>
<td>17.3</td>
<td>11.2</td>
<td>14.3</td>
<td>29.8</td>
</tr>
<tr>
<td>YIN</td>
<td>3.0</td>
<td>4.6</td>
<td>14.8</td>
<td>21.0</td>
<td>27.3</td>
<td>38.5</td>
</tr>
<tr>
<td>YAAPT</td>
<td>6.1</td>
<td>8.1</td>
<td>21.7</td>
<td>14.0</td>
<td>16.8</td>
<td>43.8</td>
</tr>
<tr>
<td>PRAAT</td>
<td>8.7</td>
<td>19.9</td>
<td>34.2</td>
<td>15.4</td>
<td>21.3</td>
<td>47.5</td>
</tr>
</tbody>
</table>

Table 1. Big and Gross errors (%) with the Keele database
Five feature conditions were tested in conjunction with each pitch tracking method:

1. Spectral trajectory features only (35 features): computed with 5 DCTC terms each encoded with 7 DCSC terms, from a frequency range of 50 to 800 Hz. These particular conditions are consistent with observations of spectrograms that indicate tonal information is most easily observed in the low frequency region over segments longer than 100 ms.

2. “Raw” pitch trajectories (P) (7 features): encoded with 7 DCSC terms.

3. Normalized pitch trajectories (NP) (7 features): also each encoded by 7 DCSC terms. The normalization is accomplished by first computing the mean and standard deviation of the pitch over the entire sentence from which each tone segment is extracted. These mean values are then subtracted from pitch values in each segment, and the resultant values divided by the standard deviation.

4. A combination of feature sets 1 and 2. (42 features)

5. A combination of feature sets 1 and 3. (42 features)

For the clean full bandwidth conditions, the errors are small and fairly similar for all three pitch tracker methods. However, for most of the noisy or band-limited cases, YAAPT results in lower error rates than for the other two trackers. For example, for the case of Keele telephone speech, and additive white noise, the gross error for YAAPT is under 10%, whereas for the other two pitch trackers, it is over 20%. Note that for comparable cases tested, error values are quite similar to those obtained in [12]; although YAAPT was “improved,” the changes are more apparent by visual inspection of the tracks. Big and Gross error figures changed very little.

### 4. Experimental Evaluations of Tone Classification

Although YAAPT gives lower error rates than either YIN or PRAAT, it was still not clear which tracker would be the most effective for Mandarin tone classification. Therefore a series of tone classification experiments, comparing the three trackers, was performed.

The database used was the Shanghai region portion of RASC863 [4]. Only the four prominent tones of Mandarin (H, R, F, D) were used. Tone labels supplied with RASC863 were considered as ground truth. A multilayer feed-forward neural network classifier, trained with back propagation, was used for classifying tones from a combination of pitch and/or DCTC/DCSC spectral features. The number of network inputs ranged from 7 to 42, as described below, depending on the feature set under evaluation. In all cases, the network had 50 hidden nodes in the first hidden layer, 25 nodes in the second hidden layer, and 4 output nodes (one for each of the four tones). The overall configuration of the network (with two hidden layers with sigmoidal nonlinearities and number of nodes mentioned) was determined from pilot tests. A total of 1539 sentences were used for training; 670 sentences were used for testing.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Telephone</th>
<th>Telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W-5</td>
<td>B-5</td>
</tr>
<tr>
<td>Gross</td>
<td>YAAPT</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>PRAAT</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>YIN</td>
<td>14.0</td>
</tr>
<tr>
<td>Big</td>
<td>YAAPT</td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td>PRAAT</td>
<td>16.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Studio</th>
<th>Simulated telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean</td>
<td>W-5</td>
</tr>
<tr>
<td>Gross</td>
<td>YAAPT</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>PRAAT</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>YIN</td>
<td>1.7</td>
</tr>
<tr>
<td>Big</td>
<td>YAAPT</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>PRAAT</td>
<td>7.1</td>
</tr>
</tbody>
</table>

For the clean full bandwidth conditions, the errors are small and fairly similar for all three pitch tracker methods. However, for most of the noisy or band-limited cases, YAAPT results in lower error rates than for the other two trackers. For example, for the case of Keele telephone speech, and additive white noise, the gross error for YAAPT is under 10%, whereas for the other two pitch trackers, it is over 20%. Note that for comparable cases tested, error values are quite similar to those obtained in [12]; although YAAPT was “improved,” the changes are more apparent by visual inspection of the tracks. Big and Gross error figures changed very little.
We hypothesize that YAAPT is superior to both YIN and PRAAT for Mandarin tone classification primarily because of the better interpolation through unvoiced regions as illustrated in Figure 4, where PRAAT and YIN can be seen to exhibit large anomalies compared to YAAPT, primarily in the unvoiced regions. Although details are not given here, due to length constraints, the previous version of YAAPT (as in [12]) resulted in tone classification accuracies typically 1% to 7% lower than for the YAAPT results reported here.

Figure 3: Tone classification accuracy for features based on YAAPT (top), YIN (middle) and PRAAT (bottom).

5. Conclusions

This paper presents several modifications to YAAPT including a smooth interpolation of pitch through unvoiced regions with the interest of improving pitch modeling for Mandarin tones. The experiments demonstrate that YAAPT has lower errors, especially for noisy bandlimited speech, than either YIN or PRAAT pitch trackers. The YAAPT features, when combined with DCTC/DCSC features to capture spectral-temporal trajectories, are also shown to be more effective than either YIN or PRAAT pitch features.

The YAAPT algorithm is available at http://www.ws.binghamton.edu/zahorian/yaapt.htm as a MATLAB function, along with a user guide and recommendations for parameter settings. We have begun a series of character recognition experiments with continuous Mandarin to more thoroughly compare the effects of different pitch algorithms.

6. Acknowledgements

This research is sponsored by the Air Force Research Laboratory under agreement number FA87501210093. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.
7. References


