Efficient Tone Classification of Speaker Independent Continuous Chinese Speech Using Anchoring based Discriminating Features

Jin-Song Zhang\(^1\), Satoshi Nakamura\(^1\) and Keikichi Hirose\(^2\)

\(^1\)ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Kansai Science City, Kyoto 619-0288, Japan
e-mail: jinsong.zhang@atr.jp, satoshi.nakamura@atr.jp

\(^2\)Dept. of Frontier Informatics, Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0083, Japan
e-mail: hirose@gavo.t.u-tokyo.ac.jp

Abstract
Anchoring based discriminating features were proposed efficient for tone discrimination of Chinese continuous speech, and have been successfully applied before to tone classification of speaker dependent experiment. This paper presents its application to speaker independent tone classification experiments. Furthermore, we made detailed comparison experiments on the efficiencies of three groups of features: the left context dependent, the right context dependent anchoring F0 features, and the conventional F0 features. Experimental results showed that a combination of all three groups achieved a significant improvement of absolute 6.4\% from 82.6\% by the baseline system to 89.0\%. When the three groups of features are used individually, both groups of the anchoring features led to better results than the conventional features, and the left context dependent anchoring features led to the highest performance.

1. Introduction
The fundamental frequency (F0) contours of continuous Chinese speech convey important information in human speech communications: the lexical tones, prosodic phrasing structure, foci and etc.. Robust tone recognition from F0 contours is not only helpful for Chinese speech recognition and understanding systems, but also necessary for intonation function decomposition from the sentential F0 contours. The four Chinese basic lexical tones (referred to as Tones 1, 2, 3 and 4) are usually characterized according to their different F0 contour patterns, i.e., Tone 1 with a high-level, Tone 2 with a mid-rising, Tone 3 with a low-dipping and Tone 4 with a high-falling F0 contour [1]. However, the tonal F0 contours may vary substantially in continuous speech compared with those in isolated syllables, making the tone recognition of continuous speech a much difficult task [2].

Based on the linguistic framework of Tone Nucleus model (Fully summarized in [2]) which divides a syllable F0 contour into tone critical part and articulatory transitions, we have proposed that context dependent anchoring F0 features be efficient for tone discrimination. [3] presented statistically distributional analyses for the anchoring F0 features, showing their different distributions for the four tones. [4] applied the anchoring hypothesis to speaker dependent tone recognition experiment using HMM tonal models, and the results showed significant improvement. [5] formalized the idea of anchoring based discrimination and showed its efficiency in tone prediction of significant tonal contextual F0 variations. The study widened the focus of original tone recognition to more general intonation study and human pitch perception, resulting in rethinking of tone patterns in Chinese and other languages. This paper presents the study of applying anchoring discriminating features to tone classification of speaker independent continuous speech, providing supplementary sound evidences for the proposal of anchoring based tone discrimination.

Following sections are arranged as follows: Section 2 describes the Tone Nucleus model and the anchoring based tone discrimination hypothesis. Section 3 introduces the experimental data. Section 4 presents the experimental results and discussions. Section 5 gives brief conclusions.

2. Anchoring based Tone Discrimination
Anchoring based tone discrimination proposal has taken root in the linguistic framework of Tone Nucleus model [2], which clarifies the tone critical segment from the articulatory transitions in one syllable’s F0s.

2.1. Tone Nucleus Model

<table>
<thead>
<tr>
<th>Sub-syllabic F0 Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>(3)</td>
</tr>
<tr>
<td>Tone Onset</td>
</tr>
<tr>
<td>Tone Nucleus</td>
</tr>
<tr>
<td>(Offset course)</td>
</tr>
</tbody>
</table>

Syllable F0 Contour

Figure 1: Illustrations of the Tone Nucleus model of Chinese syllable F0 contours. F0 segments in parentheses are optional, only the tone nucleus is obligatory.
As illustrated in Figure 1, a syllable F0 contour may be divided into three segments: onset course, tone nucleus and offset course.

- Tone Nucleus: a part of F0 contour that represents pitch targets of the lexical tone. It is the segment containing the most critical information for tonality perception, thus called as the tone-critical segment.
- Onset Course: the asymptotic F0 transition locus to the Tone-onset target from a preceding vocal cords’ vibration state.
- Offset Course: the F0 transition locus from the Tone-offset target to a succeeding vocal cords’ vibration state.

Tone-onset target and tone-offset target indicate the pitch values, which takes either low (L) or high (H) value, at the tone onset and offset, respectively. These pitch values serve as distinctive features characterizing the four basic lexical tones [Table 1].

<table>
<thead>
<tr>
<th>targets</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onset</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>Offset</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>L</td>
</tr>
</tbody>
</table>

Table 1: Pitch target features of the four lexical tones. "H", "L" depict high and low targets respectively.

Among the three segments, the tone nucleus is obligatory for one syllable F0s, the onset and the offset courses are optional. Each F0 segment usually exhibits an asymptotically linear curve and one syllable F0 should contain no more than three F0 segments with quite different slopes. It is also assumed that there should be no consistent relations between the F0 segmental structure and the syllable internal structure, but tone nucleus should reside in the Final portion of a Chinese syllable.

2.2. Anchoring based Tone Discrimination

The basic anchoring based tone discrimination hypothesis [4, 3] was made based on the psycho-acoustic perception findings [6, 7]:

- Relative F0 difference between the offset point of the first lexical tone and the onset of the second lexical tone may be an important discriminating cue for high or low pitch, besides the direct cue of a gliding F0 contour.
- There should be a timing allocation mechanism for the competition effects [7].

Based on this hypothesis, the following acoustic features can be extracted for tone discrimination, as illustrated in Figure 2.

- Left context dependent relative F0s (LCDRF0): the differences between the focused tone and the offset F0 of the preceding tone.
  \[ \Delta h^L_A = \log F_{0A} - \log F_{0B_L}. \]
  \[ \Delta h^L_B = \log F_{0B} - \log F_{0B_L}. \]

Figure 2: Illustrations of the computation of anchoring based tone discriminating features. A and B stand for the onset and offset points of the focused tone, and \( A_L, B_L \) and \( A_R, B_R \) for the corresponding points of the left preceding and the right succeeding tones.

- Right context dependent relative F0s (RCDRF0): the differences between the focused tone and the onset F0 of the succeeding tone.
  \[ - \Delta h^R_A = \log F_{0A} - \log F_{0A_R}. \]
  \[ - \Delta h^R_B = \log F_{0B} - \log F_{0A_R}. \]

With these additional discriminating features, a lexical tone in continuous speech can also be acoustically characterized using the patterns given in Table 2, besides using the flat, rising, dipping or falling F0 patterns.

<table>
<thead>
<tr>
<th>Lexical</th>
<th>LCDRF0</th>
<th>RCDRF0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone</td>
<td>( \Delta h^L_A )</td>
<td>( \Delta h^L_B )</td>
</tr>
<tr>
<td>Tone 1</td>
<td>( &gt; 0 )</td>
<td>( &gt; 0 )</td>
</tr>
<tr>
<td>Tone 2</td>
<td>( \leq 0 )</td>
<td>( &gt; 0 )</td>
</tr>
<tr>
<td>Tone 3</td>
<td>( \leq 0 )</td>
<td>( \leq 0 )</td>
</tr>
<tr>
<td>Tone 4</td>
<td>( &gt; 0 )</td>
<td>( \leq 0 )</td>
</tr>
</tbody>
</table>

Table 2: Left context dependent and right context dependent anchoring based feature patterns for the four basic lexical tones in continuous speech.

The ideas in the Table 2 are: if a pitch target (Table 1) of a tone is high target \( H \), then both the left and right context dependent relative F0s of this point tend to be positive (\( \geq 0 \)). The explanation is that when the pitch values of its neighboring anchoring points are low, \( L \), then the relative F0s tend to be positive according to the anchoring hypothesis. If they are high, \( H \), then the relative F0s tend be \( \approx 0 \). Therefore, \( \geq 0 \) in total. On the other hand, if a pitch target is \( L \), then its related context dependent relative F0s will tend to be \( \leq 0 \). The patterns in Table 2 can be used together with the conventional flat, rising, dipping or falling F0 patterns to make decisions on the tone classification of continuous speech.

3. Tone Classification Data

3.1. Speech Database

As it is not a trivial task to do automatic segmentation of tone nuclei in continuous speech, even for a speaker dependent case [2]. We set this issue aside for future study, and use manually labeled tone nuclei to do the tone classification experiments here. One merit of this Oracle-mode experiment is that the contribution of the proposal
can be clarified to the maximum extent, without suffering
from any influences of bad tone nuclei detection. Furthermore, the baseline system is the tone nucleus based tone
recognizer [2], which also requires accurate tone nucleus segmentation.

The data we used here is the same as those for statistical
distributional analysis in [4]. It has 200 utterances by 20 speakers in the data corpus HKU96: 10 males and
10 females, each speaker with 10 utterances. All the 200
utterances have different text contents and the number
of the four basic lexical tones is 2147. After checking F0
tracking errors, tone labels and phonetic segmentation,
tone nuclei were hand-labeled. We used the data of 16
speakers (8 males and 8 females) as the training data,
and the left 4 speakers as the testing data. In details, there
are 1,710 and 437 tone samples in the training and
testing data respectively.

3.2. Discriminating Features Extraction

Before extraction of the discriminating features, F0 con-
tours in logarithmic scale were normalized with respect
to each speaker’s mean log F0 and standard deviation \( \sigma_{\log F0} \), which are calculated from all the F0 contours of
the speech by each speaker. This normalization is to re-
move F0 range differences of inter-speakers.

\[
z = \log F0_{\text{norm}} = \frac{\log F0 - \log F0_0}{\sigma_{\log F0}}
\]

For each tone, we extracted three groups of F0 fea-
tures:

1. Conventional F0 features:
   - \( z_A \) for tone onset \( A \) point in Figure 2.
   - \( z_M \) for the mid-point in the tone nucleus.
   - \( z_B \) for the tone offset \( B \) point in Figure 2.
   - \( k = (z_B - z_A)/t \), the slope coefficient of a
tonal contour, where \( t \) stands for the duration
   of the tone nucleus in frames.

2. Left context dependent relative z features (LCDRZ):
   - \( \Delta z_A^L = z_A - z_{BL} \). Where \( BL \) is the offset
   point of the preceding tone in Figure 2.
   - \( \Delta z_M^L = z_M - z_{BL} \).
   - \( \Delta z_B^L = z_B - z_{BL} \).

3. Right context dependent relative z features (RC-
   DRZ):
   - \( \Delta z_A^R = z_A - z_{AR} \). Where \( AR \) is the onset
   point of the succeeding tone in Figure 2.
   - \( \Delta z_M^R = z_M - z_{AR} \).
   - \( \Delta z_B^R = z_B - z_{AR} \).

To avoid the influence of fluctuations in F0s, aver-
age over 3 points nearby the estimated points are used
instead of the one-point value. Since anchors may be
missing for a boundary tone or an isolated tone, we use
empirical methods to estimate the missing anchor values.
In this study, the missing anchor value for a boundary
tone equals to the multiplication of a weight of 0.7 and
the average of two neighboring tone nuclei including the
boundary tone itself.

4. Tone Classification Experiments

In the framework of Tone Nucleus model, the tone nucleus
almost keeps full discriminating information for the
tones as well as other prosody information. With the
discriminating features introduced beforehand, we use
Gaussian Mixture model (GMM) as the acoustic mod-
el for the four lexical tones.

4.1. GMM Tonal Acoustic Model

In the GMM model, each feature vector \( \bar{\phi} \) is assumed a
random sampling of a probability density that is a weight-
ed sum of multivariate Gaussian densities:

\[
p(\bar{\phi}|\lambda) = \sum_{i=1}^{M} w_i \cdot b_i(\bar{\phi})
\]

where \( \lambda = (w_i, \mu_i, \Sigma_i, i = 1, \ldots, M) \) is the set of model
parameters. \( M \) is the order of the GMM, \( w_i \) is the
mixture weight and satisfy the condition of \( \sum_{i=1}^{M} w_i = 1 \).
\( b_i(\bar{\phi}) \) are the multivariate Gaussian densities defined by
the means \( \mu_i \) and covariances \( \Sigma_i \). Here, we used diagonal
covariance matrices.

4.2. GMM Tone Classifiers

We have developed 6 GMM tonal models using different
combinations of the above-mentioned features.

- **GMM0**: the acoustic feature vector consists of the
  conventional discriminating features, i.e., \( \bar{\phi} =
  (z_A, z_M, z_B, k) \). This serves as the baseline sys-
tem.
- **GMM1**: the feature vector includes the left con-
text dependent relative z features plus the slope
  coefficient \( k \). \( \bar{\phi} = (\Delta z_A^L, \Delta z_M^L, \Delta z_B^L, k) \).
- **GMM2**: the feature vector includes the right con-
text dependent relative z features plus the slope
  coefficient \( k \). \( \bar{\phi} = (\Delta z_A^R, \Delta z_M^R, \Delta z_B^R, k) \).
- **GMM3**: the feature vector includes the conven-
tional and left context dependent relative z fea-
tures. \( \bar{\phi} = (z_A, z_M, z_B, k, \Delta z_A^L, \Delta z_M^L, \Delta z_B^L) \).
- **GMM4**: the feature vector includes the conven-
tional and right context dependent relative z fea-
tures. \( \bar{\phi} = (z_A, z_M, z_B, k, \Delta z_A^R, \Delta z_M^R, \Delta z_B^R) \).
- **GMM5**: the feature vector includes all the above-
  mentioned features. \( \bar{\phi} = (z_A, z_M, z_B, k, \Delta z_A^L, \Delta z_M^L,
  \Delta z_B^L, \Delta z_A^R, \Delta z_M^R, \Delta z_B^R) \).

4.3. Tone Classification Experimental Results

Figure 3 illustrates the tone classification performances
for the GMM models with 6 Gaussian mixtures. Table 3
gives the confusion matrices of the results of GMM5 and
GMM0.

4.4. Discussions

Results in Figure 3 suggest:

- Even the baseline GMM0 got 82.6% correct rate, which
  is comparatively very high tone recognition
  rate for continuous speech [2], let alone the speaker
Table 3: Confusion matrix indicating tone classification rate in percentages for the testing set of two GMMs: GMM0 and GMM5. Left digit in each cell before / depicts the result of the GMM5 and the right one for the GMM0. T1 - T4 represent the four basic lexical tones.

<table>
<thead>
<tr>
<th>Tone</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>96.2/87.5</td>
<td>2.6/12.5</td>
<td>0/0</td>
<td>1.3/0</td>
</tr>
<tr>
<td>T2</td>
<td>4.6/9.2</td>
<td>88.5/87.4</td>
<td>4.6/2.3</td>
<td>1.1/1.1</td>
</tr>
<tr>
<td>T3</td>
<td>0/1.7</td>
<td>3.3/1.7</td>
<td>91.7/90.0</td>
<td>5/6.6</td>
</tr>
<tr>
<td>T4</td>
<td>5.2/1.9</td>
<td>0.9/0.9</td>
<td>8.2/20.5</td>
<td>85.7/76.7</td>
</tr>
</tbody>
</table>

Figure 3: Illustration of the tone classification performances of the 6 tonal GMM models.

5. Conclusion

We applied anchoring based discriminating features to tone classification of speaker independent continuous speech. The results show the similar tendency as those obtained in speaker dependent experiment. The significant improvements evidenced that they are very efficient for discriminating the four lexical tones in continuous speech. Further study will be made on robust detection of tone nuclei of speaker independent continuous speech.

Acknowledgement: This research was supported in part by the National Institute of Information and Communications Technology.

6. References