DISCRIMINATING CHINESE LEXICAL TONES BY ANCHORING F0 FEATURES

Jin-Song Zhang¹, Satoshi Nakamura¹ and Keikichi Hirose²

¹ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho, Soraku-gun Kyoto 619-0288, Japan
 e-mail: jzhang@slt.atr.co.jp, nakamura@slt.atr.co.jp
²Dept. 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

In this paper we present some new features, called as anchoring F0 features including left-context-dependent and right-context-dependent ones, for discriminating Chinese lexical tones of continuous speech. The features are calculated based on a hypothesis that auditory tone perception possibly depends on the shift of the judgement boundary of tone targets by the anchoring effects from the neighboring tone targets. Statistical distribution analyses on continuous speech data by 20 speakers (10 males and 10 females) showed that the proposed anchoring F0 features are statistically efficient for tone discrimination, and the left-context-dependent F0 anchoring feature is more efficient than the F0 height for discriminating the pitch value of tone onset. Moreover, higher efficiency of the left-context-dependent anchoring features than the right-context-dependent ones indicates that carry-over effect plays a more important role than the anticipation effect for tone discrimination.

1. INTRODUCTION

Robust discriminating features are necessary for efficient classification of the linguistic categories, in the case of either human auditory perception or automatic speech recognizer. The four Chinese basic lexical tones (referred to as Tones 1, 2, 3 and 4) are usually characterized according to their different patterns of fundamental frequency (hence F0 contours, 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]. Therefore, the conventional methods for discriminating the four tones are based on the absolute F0 heights (either in Hz or F0 in logarithmic) and the slope coefficients of F0 contours. F0 heights are helpful for discriminating between tones of different register levels, such as between Tones 1 and Tone 3 or between Tone 4 and Tone 3 (not always). F0 slope coefficients are helpful for discriminating between tones with different F0 inclinations, such as Tone 1 (flat), Tone 2 (rising) and Tone 4 (falling).

However, it is well known that it is not easy to find a discriminating feature for a given linguistic category that is constant and always free of context. This is the case with the lexical tones, since tonal F0 contours may vary substantially in continuous speech as compared with those in isolated syllables due to a number of factors, such as physiological constraints and high-level intonation functions. For example, a high-pitch tone in a latter position of an utterance may own a lower F0 contour than a low-pitch tone in a former position, the distinctive slopes of some tones may be altered due to tonal coarticulations in continuous speech [2, 3], and a sentential focus may enlarge the F0 ranges of the focused tones and suppress those of the following tones [4]. These complex variations in F0 contours degrade the discrimination efficiency of the conventional F0 features, as evidenced by the much lower tone recognition performance for continuous speech than for isolated syllables [5, 6].

Although tonal F0 variations degrade the efficiencies of the F0 height and the F0 slope coefficients for tone discrimination, they never change the tonalities so much as they change from the standard tonal F0 patterns. In fact, auditory perception experiments showed that human beings are able to perceive the purported underlying lexical tones with high consistency despite of the substantial F0 variations, when provided the tonal context [2]. This fact suggested that there exist other discriminating cues in the tonal context besides the F0 height and F0 slope coefficients. A reasonable modeling of such discriminating cues will not only further our understanding about the tones, but also help us to develop more robust tone recognizers.

This paper reports our investigations on the discriminating efficiencies of some relative F0 differences for the four tones. The study was triggered by the findings from the auditory perception studies, where context stimuli, usually referred to as anchors [7, 8, 9], have long been known to exert substantial influences on the perception of a target stimulus. For the Chinese lexical tones, we made a hypothesis that the onset and the offset targets of a tone might serve as anchors for the discrimination of its neighboring tone targets. And we suggested that the relative F0 difference to the anchors, called as anchoring F0 features, be efficient discriminating features for the tones. This proposal has once been successfully applied to tone recognition of one female speech [11], and is furthered in this study to show its generality across speakers by collecting evidence from continuous utterances of multiple speakers’ data (20 speakers). Statistical distribution analyses showed clearly the discriminating efficiencies of those anchoring features for the four basic lexical tones.
2. ANCHORING EFFECT AND ANCHORING F0 FEATURES

2.1. Anchoring effects

Two kinds of anchoring effects on the perception of a stimulus have been known in auditory perception studies: one is called contrast and the other as assimilation [7].

- Contrast effect indicates a shift of the category boundary toward the anchor.
- Assimilation effect indicates a shift in the opposite direction.

Previous studies have revealed that anchoring effects tend to be more likely contrast than assimilation when the inter-stimuli-interval (ISI) exceeds 10ms and the stimulus become more complex [7]. This means that if there is any anchoring effect on the perception of the lexical tones, it may be more likely the contrast effect, since lexical tones are carried by the speech signal with complex frequency components and the ISI is usually much longer than 10ms. The possible contrast effect on lexical tone perception hints that the relative difference between a pair of anchor and stimulus should carry important discriminating information.

2.2. Pitch target patterns of the four basic tones

Instead of characterizing the four basic lexical tones by the well-known tonal F0 patterns, we adopt the 2-target pitch patterns for them. Each lexical tone is specified by a pair of distinctive pitch targets in its onset and offset, taking either the value of low(L) or high(H) as in Table 1 [3].

<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: Target features of the four lexical tones. "H," "L," depict high and low targets respectively.

The 2-target pitch tonal patterns changed the tone discrimination task from making a one-of-four decision to making one-of-two decisions for two targets. More importantly, it highlights the possibility of anchoring effects on tone discrimination. When we view the neighboring targets as anchor and stimulus, the relative position in the acoustic cues between the anchor and the stimulus should play an important role in the pitch target discrimination if there exist some contrast effects on the tone discrimination. In other words, a "close" position in the acoustic domain to the anchor easily cue the judgment of the same category as the anchor for the stimulus target, whereas a certain distance to the anchor is needed to cue a different judgment.

2.3. Anchoring F0 features

F0 and pitch are considered as different concepts here in that F0 is the physical cue of the perceptual dimension of pitch. Although their relationship is usually regarded as rather straightforward: high F0 cues high pitch and low F0 cues low pitch, we suggest a second way for F0 to cue pitch from the view of contrast anchoring effect. After the category boundary for H or L pitch shifted to the anchor's F0 range, relative F0 difference between the anchor and stimulus targets become an important cue for the judgement of the stimulus one. For examples, a lowered F0 boundary for H judgement by an anchor target may easily cue more H responses as long as the relative F0 difference is positive enough, even though the F0s are low in absolute height, and vice versa.

We consider contrast anchoring effects in two directions: one is the carryover from the left anchor of the tone offset of the preceding tone, and the other is the anticipation from the right anchor of the tone onset of the succeeding tone. Therefore, we have two kinds of relative F0 differences as the anchoring F0 features:

- Relative F0 difference to the left anchor for the carryover influence.
- Relative F0 difference to the right anchor for the anticipation influence.

If the anchoring F0 features are efficient in discriminating the four lexical tones, we would expect that there should be different distributions of relative F0 difference features for the H and L targets.

3. DATA PREPARATION

3.1. Material

In order to show that the anchoring F0 features are robust for tone discrimination of both speaker and context independent speech, we randomly selected 10 continuous utterances per speaker for both the 10 male and 10 female speakers in the data corpus HKU06. All the 200 utterances have different text contents and the number of the lexical tones total up to 2332. After checking F0 tracking errors, tone labels and phonetic segmentation, we hand-labeled the tone nuclei of all the tones according to the definition of tone nuclei. (Tone nucleus model, a F0 segmental structure model we proposed formerly to offer a consistent account of the F0 variations in a syllable F0 contour in various phonetic context, provides a reliable framework to manually identify tone onset and offset of each lexical tone in the continuous speech [10].) As a result, we got 2147 samples for the four basic lexical tones, excluding the 185 samples of the neutral tone from the analysis. Since the database includes punctuation marks, which are usually pronounced in Tone 4, the sample set is not well balanced for the four tones, with much more samples for the Tone 4 as listed in Table 2.

<table>
<thead>
<tr>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>412</td>
<td>450</td>
<td>265</td>
<td>990</td>
<td>2147</td>
</tr>
</tbody>
</table>

Table 2: Distribution of the sample set used for the analysis.

To mitigate any possible influences arising from the unbalanced sample set, we averaged over all the tokens of each speaker to get a set of features per tone, and merged the 80 sets of features from the 20 speakers to form a tone-balanced set. All the analyses will be carried out on both the 2147 sample set (hereinafter, the total set) and the 80 tone-balanced set (the balanced set).
3.2. Anchoring F0 features extraction

Before extraction of the anchoring F0 features, F0 contours in logarithmic scale were normalized with respect to each speaker’s mean log F0, log F0, and standard deviation, σ_log F0, which are calculated from all the F0 contours of the speech by each speaker. This normalization is to remove F0 range differences of inter-speakers.

\[ z_0 = \frac{\log F0_{onset} - \log F0_{offset}}{\sigma_{log F0}} \]

For each tone, two z values are collected to represent the absolute F0 heights for the onset and the offset targets in the normalized scale. Since fluctuations cannot be avoided during F0 detection, the average of 3 points near the onset or offset is used instead of one-point value.

- \( z_0 \) for the tone onset,
- \( z_1 \) for the tone offset.

Relative differences of the two z values to the preceding anchor, i.e., the tone offset of the preceding tone, and the succeeding anchor, i.e., the tone onset of the succeeding tone, are calculated to represent the anchoring F0 features in the normalized scale.

- \( z_i^L = z_i - z_0 \) of the preceding tone, \( i = 0, 1 \): left context dependent anchoring features.
- \( z_i^R = z_i - z_0 \) of the succeeding tone, \( i = 0, 1 \): right context dependent anchoring features.

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. The missing anchor value for an isolated tone is assumed to be 0.

4. STATISTICAL DISTRIBUTION STUDIES

We reported here three statistical studies over the collected data from the two sample sets. The first one is to check the mean values of the six features, including \( z_i, i = 0, 1 \) and their anchoring correspondents, for the four basic tones in order to achieve an overall impression about the feature distributions. The second one is to study the different distributions of the six features for the H and L onset targets by analyses of variance (ANOVA). Features with discriminating efficiency for the different onset targets should show quite different statistical distributions. The third one is to investigate the discriminating efficiency of the six features for the H and L offset targets, taking the similar way of the second study.

4.1. Mean values for the four tone groups

Table 3 and 4 list the mean values of the six features calculated from the total sample set and the balanced set, for each group of the four basic tones. Interpretation of these values can be illustrated by Figure 1. We can see:

- Since \( z_0 \) and \( z_1 \) represent the absolute F0 heights in the normalized scale, differences in their means between the four tones indicate (albeit partially when variance is not considered) the efficiency for the F0 height to discriminate the four tones.
- Mean distributions of the left and right anchoring features generally keep the similar difference as the \( z_0 \) and \( z_1 \) for the four tones. This indicates that the anchoring features possess the capability to discriminate the tones when influence from F0 variance is ignored.
- Another notable phenomenon is that Tone 3 and Tone 4 mainly differed from each other in the register level since both own a falling mean contour. However, the difference in the values of \( z_0 \) and \( z_1 \) is less than those in the \( z_i^L \) and \( z_i^R \). This means that anchoring features should be possibly more efficient for discriminating these two tones.
- There are no obvious differences for the values in the two tables.

<table>
<thead>
<tr>
<th>Tone</th>
<th>( z_0 )</th>
<th>( z_1 )</th>
<th>( z_i^L )</th>
<th>( z_i^R )</th>
<th>( z_0^L )</th>
<th>( z_0^R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>0.801</td>
<td>0.807</td>
<td>0.883</td>
<td>0.848</td>
<td>0.872</td>
<td>0.875</td>
</tr>
<tr>
<td>Tone 2</td>
<td>0.801</td>
<td>0.807</td>
<td>0.883</td>
<td>0.848</td>
<td>0.872</td>
<td>0.875</td>
</tr>
<tr>
<td>Tone 3</td>
<td>0.801</td>
<td>0.807</td>
<td>0.883</td>
<td>0.848</td>
<td>0.872</td>
<td>0.875</td>
</tr>
<tr>
<td>Tone 4</td>
<td>0.801</td>
<td>0.807</td>
<td>0.883</td>
<td>0.848</td>
<td>0.872</td>
<td>0.875</td>
</tr>
</tbody>
</table>

Table 3: Mean values of the six features for the four tone groups in the total sample set.

<table>
<thead>
<tr>
<th>Tone</th>
<th>( z_0 )</th>
<th>( z_1 )</th>
<th>( z_i^L )</th>
<th>( z_i^R )</th>
<th>( z_0^L )</th>
<th>( z_0^R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>0.824</td>
<td>0.874</td>
<td>0.913</td>
<td>0.876</td>
<td>0.896</td>
<td>0.898</td>
</tr>
<tr>
<td>Tone 2</td>
<td>0.824</td>
<td>0.874</td>
<td>0.913</td>
<td>0.876</td>
<td>0.896</td>
<td>0.898</td>
</tr>
<tr>
<td>Tone 3</td>
<td>0.824</td>
<td>0.874</td>
<td>0.913</td>
<td>0.876</td>
<td>0.896</td>
<td>0.898</td>
</tr>
<tr>
<td>Tone 4</td>
<td>0.824</td>
<td>0.874</td>
<td>0.913</td>
<td>0.876</td>
<td>0.896</td>
<td>0.898</td>
</tr>
</tbody>
</table>

Table 4: Mean values of the six features for the four tone groups in the balanced sample set.

Figure 1: Illustration of the mean values of the features for Tones 1-4. T1, T2, T3 and T4 for the four basic lexical tones Tone 1, Tone 2, Tone 3 and Tone 4.

4.2. Statistics with respect to the onset target

Taking the tone onset target as the independent variable and the six features as the dependent variables, we carried out ANOVA analyses and got the results in Table 5 for the total sample set and in Table 6 for the balanced set.

By summarizing the results in Tables 5 and 6, we get:

- In both the two tables, it is quiet evidential that \( z_0 \), \( z_i^L \) and \( z_i^R \) have very different distributions for the H and L onset targets. This means they should be efficient in discriminating between the two targets.
- Since pitch target is highly correlated with F0, it is not surprising that \( z_0 \) is efficient in discriminating the H and L onset targets.
Table 5: Statistics with respect to the tone onset target for the total sample set.

<table>
<thead>
<tr>
<th>Tone onset</th>
<th>mean value</th>
<th>l</th>
<th>0.704</th>
<th>0.600</th>
<th>0.300</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.474</td>
<td>-1.473</td>
<td>-1.472</td>
<td>-1.471</td>
<td>-1.470</td>
<td>-1.469</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 6: Statistics with respect to the tone onset target for the balanced set.

- However, the $z_2^T$ got the highest F0 ratios of between-group variance and within-group variance in the two tables. This indicates that left-context-dependent onset anchoring feature is even more efficient than the absolute F0 height for discriminating an onset target in context.
- Although there are less evidences that $z_1$ distributes differently for the different onset targets, there are statistical evidences that left-context-dependent or right-context-dependent $z_1$ features become more differently distributed than the absolute F0 height for the different onset targets. This also indicates the improved discriminating efficiency by the anchoring features.

4.3. Statistics with respect to the offset target

Similarly, by using the offset target as the independent variable we get ANOVA analyses results for the total sample set in Table 7 and the balanced set in Table 8. Based on the

<table>
<thead>
<tr>
<th>Tone offset</th>
<th>mean value</th>
<th>l</th>
<th>0.704</th>
<th>0.600</th>
<th>0.300</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.474</td>
<td>-1.473</td>
<td>-1.472</td>
<td>-1.471</td>
<td>-1.470</td>
<td>-1.469</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 7: Tone offset based group statistics over the 2147 sample set.

<table>
<thead>
<tr>
<th>Tone offset</th>
<th>mean value</th>
<th>l</th>
<th>0.704</th>
<th>0.600</th>
<th>0.300</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.474</td>
<td>-1.473</td>
<td>-1.472</td>
<td>-1.471</td>
<td>-1.470</td>
<td>-1.469</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 8: Tone offset based group statistics over the tone balanced sample set.

Tables 7 and 8, we may see:

- In both the two tables, it is quiet evident that $z_1$, $z_1^T$ and $z_1^R$ should be efficient in discriminating between the two offset targets. The efficiency of $z_1$ is not surprising, whereas the efficiencies of $z_1^T$ and $z_1^R$ indicate that anchoring $z_1$ features help to discriminate the tone offset targets.
- It is consistent in the two tables that the discriminating efficiencies of the three features for the H and L offset targets have the same order: $z_1 > z_1^T > z_1^R$.

- The order can be interpreted as two meanings: one is that anchoring features are less efficient than the absolute F0 height for discrimination of offset targets. The second is that left-context dependent offset anchoring feature is more efficient than the right-context dependent anchoring feature.
- Although it is evident that $z_0$ and its anchoring features distribute differently for the L and H offset targets in the total sample set, it is not evident in the balanced set. A check of the mean values showed that $z_0$ in the total set is 2.9 times of that of the balanced set. Surely this resulted from the more Tone 4s in the total sample set.

5. DISCUSSION

By summarizing the statistical distribution analyses, we say that the proposed anchoring F0 features are efficient for discriminating between H and L tone targets, therefore efficient for discriminating the four basic lexical tones. When the target position is considered, statistical results showed that anchoring features are more efficient for discriminating the pitch values of onset target than those of offset target. As for the influences from left or right context, left-context-dependent anchoring features consistently perform better than the right-context-dependent features. One anchoring F0 features, the left-context-dependent onset anchoring feature $z_1^T$ is even more efficient than the absolute F0 height for discriminating the onset target. These results indicate that carry-over effect play a more important role than the anticipation effect for Chinese tone discrimination.

6. REFERENCES