Measuring the Appropriateness of Lexical Tones in Bisyllabic Chinese Words

Akira Ishida

Department of Computer Science and Media Engineering,
University of Yamanashi, Yamanashi, Japan
ishida@esi.yamanashi.ac.jp

Abstract

A method to classify Mandarin Chinese bisyllabic words based on the appropriateness of their tones was developed. Overall classification accuracy was approximately 86 percent. Depending on the tone combination, accuracy ranged from 64 to 100 percent. The proposed method might assist non-native learners acquiring tone pronunciation skills.

1. Introduction

Lexical tones in Mandarin Chinese exhibit fairly clear pitch and duration patterns in isolated syllables, but these patterns often become indistinct in connected speech -- an important process in natural, running speech[1]. Non-native learners need to transition from saying lexical tones in isolation to producing them with pitch contours appropriate to context[2].

This paper describes a computational method to disambiguate appropriate and inappropriate pitch patterns in Mandarin bisyllablic words. The following sections describe the speech data (section 2), the algorithm (section 3), and the evaluation experiment (section 4).

2. Speech Data

Speech data for training and testing were collected in two steps: (a) record non-native speech, and (b) have it graded by native speakers.

2.1 Recording non-native speech

Given that Mandarin has 4 lexical tones, bisyllabic words theoretically have 16 tone combinations. Some combinations do not occur in native speech due to tone sandhi (i.e., phonological rules that alter tones in specific contexts). However, in non-native speech, learning tone sandhi rules is as important as learning how to say the tones. Furthermore, phonology and phonetics pertaining to tone are often taught together in Mandarin language classrooms. Hence, all 16 underlying tone combinations were used to collect speech. For each tone combination, 10 words were chosen for a total of 160 words. Table 1 shows a partial list of words, some having different underlying tone combinations that collapse into identical tone combinations if spoken correctly.

The 160-word set was read aloud by 5 adult native speakers of Japanese learning Chinese as a foreign language. The learners had recently started learning Chinese and were somewhat accustomed to pinyin but had limited knowledge of pronunciation, vocabulary and syntax.

Speech was digitally recorded at 48 kHz, 16-bit mono using a Shure SM-10A microphone and a DAT recorder. As utterances improperly recorded were discarded, a total of 763 words were obtained.

Table 1. Partial list of Chinese bisyllabic words. As Chinese has 4 lexical tones, 4 x 4 = 16 tone combinations are theoretically possible for bisyllabic words. Some combinations do not occur in native speech due to phonological rules.his is an example of a table

	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	dīdā 滴答	āiháo 哀号	sīkǎo 思考	bāobì 包庇
Tone 2	yúbō 余波	bómó 波膜	bómǔ 伯母	bówù 博物
Tone 3	fǎyī 法衣	bǎnyá 板牙	bǎotǎ 宝塔	bǎwò 把握
Tone 4	tàkān 踏勘	bèifú 被服	dàdǐ 大抵	dàqì 大气

2.2 Grading by native speakers

Adult native speakers of Chinese (6 male, 6 female) listened to all bisyllabic words and graded each word's tones as correct or incorrect (forced-choice decision). The graders were instructed to ignore mispronunciations at the segment level, and to focus on tone only.

Depending on the number of "correct" responses given by native graders, each non-native utterance was placed in correct or incorrect bins at three levels (Table 2).

Table 2: Tone accuracy divided into correct and incorrect bins at three levels. Depending on the number of "correct" responses given by native graders, each non-native utterance was placed in correct or incorrect bins at three levels. Level one is the strictest; level three is relatively lax.

	correct	incorrect
Level 1	12	11-0
Level 2	12 or 11	10-0
Level 3	12-10	9-0

Level one is the strictest; the graders must be unanimous if the learner's utterance is to be judged correct. Levels two and three are progressively less strict; level three requires a 3-quarters majority or better for the learner's utterance to be judged appropriate.

2.3 Grading by native speakers

To model the pitch contours of tones, we measured the difference in pitch values at the beginning and end of each syllable (Figure 2). We also investigate other measurements afterwords.

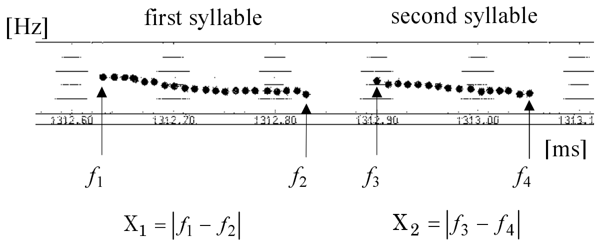

Figure 2: Pitch contours of tones measurements. The difference in pitch values at the beginning and end of each syllable was measured. Here X_1 *and* X_2 *are the pitch differences for the first and second syllables respectively.*

2.4 Classification algorithm

Given an utterance whose tone appropriateness is unknown, the correct/incorrect bins created in section 2.2, and the F0 differences measured in section 2.3, we compute the Mahalanobis distances between the F0 differences of the unknown utterance and those in the correct and incorrect bins (Figure 2). The smaller distance indicates the unknown utterance's tone correctness.

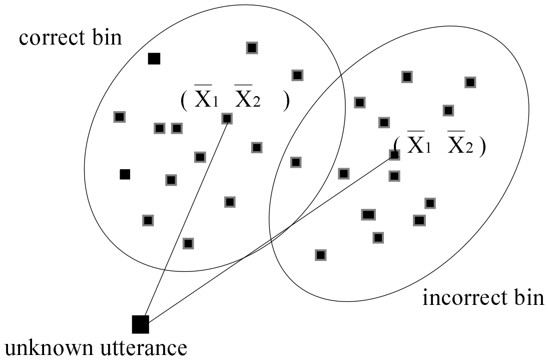
correct bin
$(\overline{X}_1\ \overline{X}_2\ )$
$(\overline{X}_1\ \overline{X}_2)$
incorrect bin
unknown utterance

Figure 2: Determining the correctness of an unknown utterance. Given an utterance whose tone appropriateness is unknown, the Mahalanobis distances between the F0 differences of the unknown utterance and those in the correct and incorrect bins are computed. The smaller distance indicates the unknown utterance's tone correctness.

Mahalanobis distance is defined as follows:

$$D^2_{(a,b)} = \left(a - \overline{X}_1 \quad b - \overline{X}_2\right) \cdot S^{-1} \cdot \begin{pmatrix} a - \overline{X}_1 \\ b - \overline{X}_2 \end{pmatrix}$$

where a and b are X_1 and X_2 of unknown utterance. $\overline{X}_1$ and $\overline{X}_2$ are averages, and S is variance-covariance matrix of utterances in the bin.

3. Evaluation Experiment

The algorithm's results were compared with the native graders' using the following evaluation metric:

$$agreement\_factor = \frac{number\_of\_correctly\_predicted}{total\_number\_of\_words}$$

3.1 Grading strictness effects

The effect of grading strictness was studied in closed and open experiments.

In the closed experiment, all 763 utterances were used for both training and testing. In the open experiment, 505 utterances from 4 out of 5 speakers were used for training, and 158 utterances from the remaining speaker were used for testing.

Figure 3 shows results of closed and open experiments at three levels of grading strictness. While the closed experiment naturally outperforms the open experiment at all three levels of grading strictness, the gap between closed and open experiments shrinks progressively as grading strictness is relaxed.

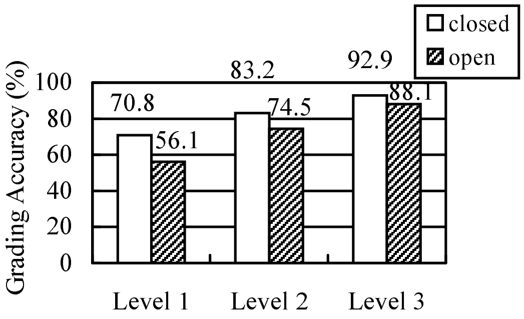

Figure 3: Results of closed and open experiments at three levels of grading strictness. Percentages indicate agreement between native graders and the algorithm (higher values are better).

3.2 Jackknife open experiment

As the third level of grading strictness yielded best performance, a jackknife open experiment was run using 4 out of 5 speakers for training, and the remaining speaker for testing. Results are shown in Figure 4.

Grading accuracy differed depending on the learner. The average agreement factor was 85.7 percent.

Another factor affecting grading accuracy was the tone combination being graded. Table 3 shows grading accuracy for each tone combination at the third level of grading strictness. While tone 3 followed by tone 2 yielded 100 percent accuracy, tone 4 followed by tone 4 yielded 64 percent. Because all utterance of tone 1 followed by tone 1 and tone 1 followed tone 3 were graded correct in this level, no incorrect bin was created. The correctness of these utterances could not be determined.

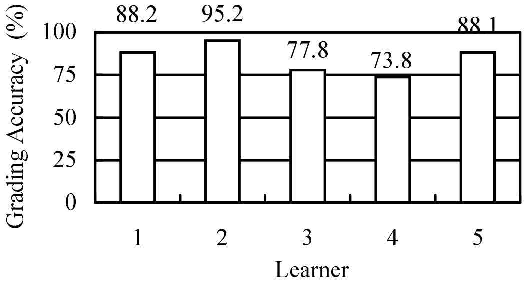

Figure 4: Results of jackknife open experiments at the third level of grading strictness. Percentages indicate agreement between native graders and the algorithm (higher values are better). Average agreement was 85.7 percent.

Table 3: Grading accuracy for each tone combination at the third level of grading strictness. Accuracy differs depending on the tones of the first and second syllables. Percentages indicate agreement between native graders and the algorithm (higher values are better).

tone 1 x 1	---	tone 1 x 2	90.0 %
tone 1 x 3	94.4 %	tone 1 x 4	---
tone 2 x 1	82.5 %	tone 2 x 2	82.0 %
tone 2 x 3	81.3 %	tone 2 x 4	92.3 %
tone 3 x 1	96.0 %	tone 3 x 2	100.0 %
tone 3 x 3	75.0 %	tone 3 x 4	77.6 %
tone 4 x 1	95.0 %	tone 4 x 2	86.0 %
tone 4 x 3	87.2 %	tone 4 x 4	63.6 %

3.3 Analysis

Experimental results indicate that, while the proposed method is a good first-order approximation of bisyllabic tone contours, a more robust method of modeling tone contours is needed. For instance, the proposed method treats tone 2 and tone 3 similarly because they may both involve pitch rises (Figure 5). Considering tone duration and/or energy may improve performance.

Dealing with ambiguous responses is another issue. Some of the learners' utterances were not judged unanimously by native speakers. The disparity stems partly in the judges' inability to tone intelligibility from tone naturalness (the former ideally being a measure of how well the utterance might be understood, and the latter being how likely a native speaker might say the utterance in the way it was said). A simpler reason is that some learners' utterance were borderline. Force-choosing between right and wrong may be inappropriate for such utterances. Adding intelligible, natural utterances (perhaps spoken by natives) may help disambiguate appropriate vs inappropriate speech.

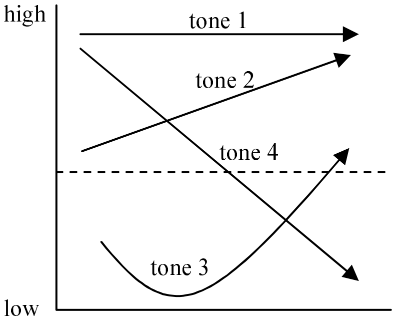

Figure 5: Stylized representations of tone contours. Tones 2 and 3 may exhibit pitch rises, hindering discrimination. Actual tone contours are more complex, as tones may differ markedly in duration.

3.4 Other Measurements

We investigate other measurements instead of subtracted value for tone 2 and 3 combinations as follows:

1) logarithm

$$X_1 = \left|\log f_1 - \log f_2\right| \ , \ X_2 = \left|\log f_3 - \log f_4\right|$$

2) value of cent

$$X_1 = 1200 \times \frac{\log f_1}{\log f_2}, \ X_2 = 1200 \times \frac{\log f_3}{\log f_4}$$

3) gradient

$$X_1 = \left|\frac{f_1 - f_2}{t_1}\right|, \ X_2 = \left|\frac{f_3 - f_4}{t_2}\right|$$

where t_1 and t_2 are duration of the first and second syllable respectively.

4) logarithm gradient

$$X_1 = \left|\frac{\log f_1 - \log f_2}{t_1}\right|, \ X_2 = \left|\frac{\log f_3 - \log f_4}{t_2}\right|$$

Figure 6 shows grading accuracy of each measurement in jackknife open experiment. In this figure, gradient achieves 90% of accuracy in tone 2 x 3. However, the original measurement is superior in the other combinations,

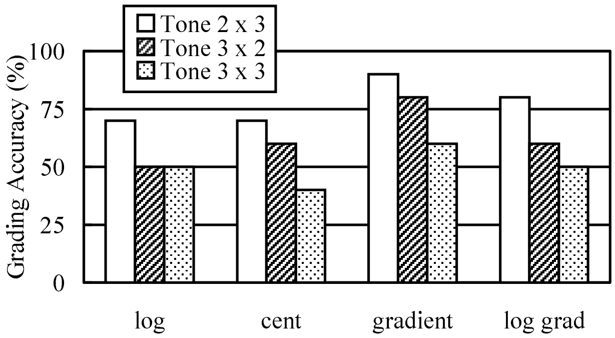

Figure 6: Grading accuracy of other measurements in jackknife open experiment.

4. Conclusion

A method to classify Mandarin Chinese bisyllabic words based on the appropriateness of their tones was developed. Overall classification accuracy was approximately 86 percent. Depending on the tone combination, accuracy ranged from 64 to 100 percent. The proposed method might assist non-native learners acquiring tone pronunciation skills.

References

[1] Y. Zheng amd M. Yanagida,(2000), Study in Reformation Phenomena of Tone Patterns and Their Rule in Chinese Trisyllabic words, *Tech. Rep. of IEICE.* SP2000-117, pp.1-8

[2] R.-S. Yu (1993), Pronunciation of Chinese – An Approach to Typical Mispronuncations of Japanese Learners –, *Journal of Nagasaki Wesleyan College*, Vol.16, pp.93-109