Annotation and Features of Non-native Mandarin Tone Quality

Mitchell Peabody, Stephanie Seneff

Spoken Language Systems, CSAIL, MIT
{mizhi,seneff}@csail.mit.edu

Abstract

Native speakers of non-tonal languages, such as American English, frequently have difficulty accurately producing the tones of Mandarin Chinese. This paper describes a corpus of Mandarin Chinese spoken by non-native speakers and annotated for tone quality using a simple *good*/*bad* system. We examine inter-rater correlation of the annotations and highlight the differences in feature distribution between native, good non-native, and bad non-native tone productions. We find that the features of tones judged by a simple majority to be bad are significantly different from features from tones judged to be good, and tones produced by native speakers.

Index Terms: computer aided language learning, tone evaluation, Mandarin, Chinese

1. Introduction

Computer Aided Language Learning (CALL) systems in foreign language classrooms are becoming increasingly popular for assigning coursework to students, supplementing material learned in traditional language teaching classrooms, providing opportunities for practice in non-threatening contexts, and allowing self-study of foreign languages. CALL frameworks take many forms and are guided by considerations of technical practicality and language learning pedagogy.

Pronunciation assessment is a major component for speech based CALL systems, where students are given feedback on the quality of their pronunciation of the target foreign language. Assessing the pronunciation of foreign language learners is difficult because while their speech may be intelligible enough to native speakers, it may also be so heavily accented that understanding it can be taxing for the native speaking participant [1, 2]. The challenge for language teachers is to strike a balance between overly-criticizing students for pronunciation mistakes and not providing any pronunciation guidance at all.

This implies that an assessment engine for pronunciation must be able to tell when speech is so poorly pronounced that a human language teacher would point it out, but not necessarily when that speech is merely accented. This is a considerable challenge due to the fact that even non-native speech considered well-produced typically has more pauses, slower rate of speech, and much greater phonetic variation than native speech.

Tonal languages, such as Mandarin Chinese, pose additional problems because correct pronunciation depends not only on the phonetic realization of the target word, but also on the tone production for a given word. The correct realization of a tone depends on such factors as left and right contexts, anticipatory and carryover effects [3], intonation, and tone sandhi rules [4]. Furthermore, non-native speakers may not easily remember the lexical tone for a particular word.

We believe that speech technology can be effective in helping students learn the tones of Chinese words, as well as to understand how to properly express tonal aspects in production. Our research into CALL uses computer games that students interact with using their voice. These systems simulate conversational partners with which students are required to participate in dialogues centered around small domains and problems [5]. These dialogues are dynamically generated and enable the student to both understand and produce spontaneous speech in the target language, thus exercising their communication skills.

These games are presented on the web using the WAMI toolkit [6] but currently lack the facility for pronunciation assessment. We envision integrating pronunciation assessment into our CALL games such that students are given targeted feedback on their pronunciation errors. We make the assumption that only the most egregious pronunciation errors need to be pointed out and that it is unnecessary and counter-productive to point out every single pronunciation problem.

In order to accomplish the task of assessing non-native tone quality, we need to determine the features that may be useful for distinguishing *good* from *bad* non-native tone productions. This paper utilizes two corpora to quantify differences in tone features between native, non-native tones judged as *good* and non-native tones judged as *bad*.

2. Background

In general, speakers of a non-tonal language who are learning Mandarin as a foreign language have difficulty both perceiving and producing tone (see, for example [7]). A tonal language uses pitch, the perception of fundamental frequency (F_0), to lexically distinguish tones.

In Mandarin Chinese every syllable is marked with a tone. Syllables in Chinese are composed of two parts: an initial and a final. The initial is either a consonant or the null initial (silence). The final portion of the syllable is composed of vowels and possibly a post-vocalic nasal, and is also the tone bearing unit.

Mandarin has five official tones. The fifth tone is often referred to as the neutral tone and is often deemphasized. Mandarin tones are mainly distinguished by shape, though there are other perceptual cues such as duration [8] and amplitude [9]. Some tone languages, such as Cantonese, have tonal contrasts that depend on the pitch height (register) of the contour [10]. Although Mandarin tones tend to be produced at different pitch registers (Tones 1 and 4 at high registers, Tone 2 at mid to low registers, and Tone 3 at low registers), pitch register is not critical for tone identification in Mandarin.

Recent work has involved the training of non-native speakers in the perception and production of Mandarin tones. Wang [11, 12] examined the effects of perceptual training on speakers' ability to produce Mandarin tones in an isolated set of Chinese words. Prior work by Leather [7] looked at the use of visual feedback on the ability of non-native speakers with-

out any prior perceptual training to produce the four tones of a single Chinese initial-final pair.

3. Methodology

This work makes use of two corpora, both in the flight domain: the Yinhe [13] corpus consisting of $5,218$ spontaneous sentences from native Mandarin speakers, and ftgame, an annotated corpus of data collected during student practice sessions with our Flight Translation Game [14]. This section concerns the annotation procedure and pitch normalization algorithm used for the later analysis.

3.1. Annotation

Utterances from the ftgame corpus were transcribed using the pinyin romanization of Chinese. A total of $2,073$ utterances from 8 speakers, 2 female, 6 male, were transcribed. The speakers were all students of Mandarin Chinese with levels of study from under 1 year to approximately 4 years. Utterances that included English, disfluencies, or partial words were excluded. The remaining $1,702$ utterances, containing $14,845$ syllables, were selected for annotation.

The annotation was performed by 6 native Mandarin speakers from Taiwan using a web-enabled annotation interface [15]. Each annotator independently made a binary (*good* or *bad*) judgement on the tone quality for each syllable. The annotated corpus contained a total of $89,070$ judgements.

3.2. F_0 Normalization

The major preceptual cue for distinguishing Mandarin tones is pitch shape, for which F_0 is the primary feature. We extract the F_0 using the pitch extraction algorithm detailed in [16]. Differences in the mean F_0 of speakers require that the F_0 of the data be normalized in order to make meaninful shape comparisons.

The normalization process, which is an extension of the method discussed in [17], has three main steps. First, the declination for each utterance is removed. Second, each F_0 in the utterance is scaled by a factor computed based on the utterance mean F_0 and a globally computed corpus mean F_0. Finally, a logarithmic value for each F_0 is computed and normalized to place the pitch on a common scale.

The intonation of a sentence and the lexical tones both contribute to F_0, and the effects of the two are not easily separated. Our approach can be seen as attempting to subtract the contribution of intonation to F_0 in order to gain access to the shape of the lexical tone. Although some research [18] has found that the rate of sentential downdrift in Chinese utterances depends on sentence length and can be modeled using exponential decay functions, Wang [19] found that a simple linear declination model still produced a significant improvement in tone classification. For simplicity, we adopted this method for removing the intonation contour from the utterances.

Each F_0 in the utterance is scaled by a constant factor to remove F_0 differences due to individual voice characteristics and gender. This makes the utterance mean F_0 equal to a fixed global value.

The final step of the normalization process is to use Equation 1 to compute a log for each F_0 value based on a method commonly used for Mandarin tone studies [12, 20, 21].

$$T(x) = 5\frac{\lg x - \lg L}{\lg H - \lg L} \quad (1)$$

where H and L are the highest and lowest F_0 over all the tones after declination is removed and the utterance has been scaled. This places the F_0 on a common 5-pt scale for Mandarin originally proposed by [22], and allows for direct comparison of contour shapes.

	1	2	3	4	5	6
1	1	0.45	0.59	0.45	0.53	0.44
2		1	0.51	0.50	0.51	0.50
3			1	0.42	0.53	0.46
4				1	0.57	0.52
5					1	0.53
6						1

Table 1: Cohen pairwise κ-statistics.

4. Results and Analysis

The analysis of the data is divided into two parts: correlation and feature analysis. We analyzed the *good*/*bad* judgements of the six annotators to confirm that they could provide consistent assessments. The feature analysis uses these assessments to examine differences between native, non-native *good*, and non-native *bad* tone productions.

4.1. Annotator Correlation

The annotation corpus was analyzed to assess the degree the annotators agreed with one another on the quality of the non-native tones. The average, pairwise raw percentage agreement was 91.1%; however, this agreement is primarily due to the fact that raters marked most syllables as having *good* tone quality.

We believe this is due to two main reasons: (1) the default rating for a syllable in our system is *good* and (2) we instructed the annotators to mark a tone as *bad* only if it would be a tonal mistake they would point out to a student learning Chinese. This was a deliberate design decision based on our approach to point out only the worst errors.

Better correlation statistics are the pairwise Cohen κ-statistic [23] and the Fleiss κ-statistic [24], which measure the amount of agreement when the probability of chance agreement is removed. Cohen κ is computed between two raters, while Fleiss κ is computed for multiple raters. Table 1 summarizes the Cohen κ-statistics. The Fleiss κ-statistic for all raters was 0.515. Overall, these correlations indicate a moderate amount of agreement among the raters, using the scale proposed by [25].

4.2. Tone Features

Previous work [17] that examined the differences between native and non-native productions of tones made the implicit assumption that all tones produced by non-native speakers were *bad*. This research further breaks down the non-native speech into syllables native speakers judged as *good* and those that were judged as *bad*, and examines differences in easily extracted F_0 features.

Specifically, we examine those features found to be perceptually important for tonal contrasts in native speech (sec. 2). We focus on tone shape and duration. The annotators did not always agree that tones were poorly produced. Since our eventual goal is to provide assessments of the worst tones, we restricted our analysis to tones with all *good* assessments or at least three *bad* assessments. The intuition is that if a tone was truly bad, more annotators would mark it as *bad*. This allows sharper contrasts to be seen in the analysis.

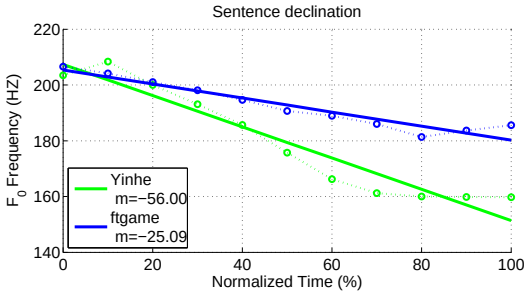

Figure 1: Pitch contours for our native and non-native corpora demonstrating differing sentential declination.

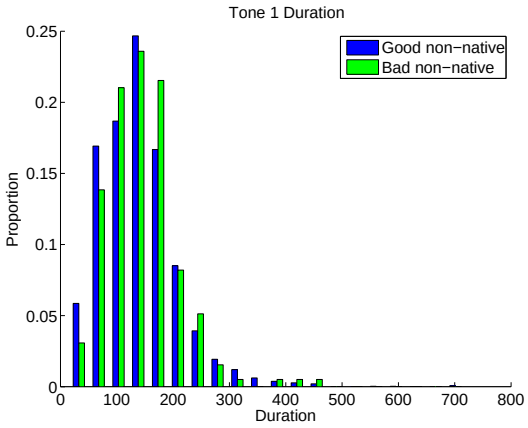

Figure 2: Duration (ms) of Tone 1 syllables.

First, we examine the slope of the linear regression computed as part of the normalization process. We found that our native and non-native utterances have different declination patterns. Figure 1 shows plots of the F_0 for utterances in the Yinhe and ftgame corpora. Linear regressions on both datasets show that the native speakers exhibit a steeper decline over the normalized length of an utterance than the non-native speakers.

We hypothesize that the non-native speakers differ in their intonation pattern for three reasons. First, intonation patterns from English have been shown to interfere with production of Mandarin tones [26]. Second, users interacting with the Yinhe system mostly issued commands. Students playing the Flight Translation Game were given English language prompts and told to provide a translation into Chinese. The uncertainty in correctly accomplishing this task may have induced the student to have a final intonation rise after the 80% time mark, as if asking for confirmation. Third, non-native speakers tend to have a slower rate of speech in both the duration of the syllables and pauses between them. This may manifest itself in the F_0 as minor "pitch resets" throughout their sentences and contribute to a slower declination.

Duration plays a role in tonal contrasts for Mandarin speakers. Tones 1 and 4 tend to have the shortest durations, Tone 2 is in the middle, and Tone 3 is usually the longest [8]. Duration of phonetic segments has been found in other studies (see for example [27]) to correlate well with human judgements of non-native pronunciation quality. Our measurements indicate that non-native speakers tend to produce syllables that are 58.7% longer in duration than those of native speakers: 146.2ms vs 92.1ms. We expected that duration would be a salient feature

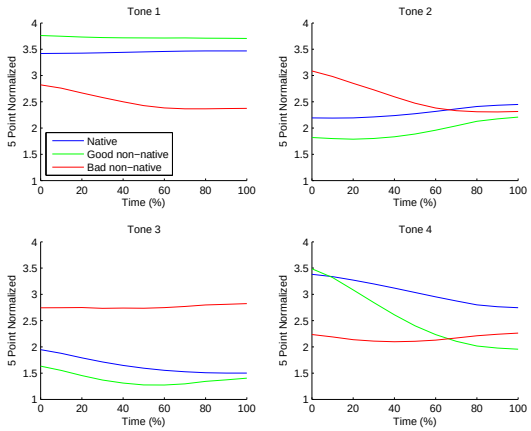

Figure 3: Mean log F_0 for tones.

for tone quality perception; however, as Figure 2 shows (representative of all tones), this is not the case - non-native speakers are consistently slower for both good and bad productions.

The key contrasting feature for Mandarin tones is the shape of the pitch contour. Figure 3 shows normalized F_0 contours for Tones 1-4. The blue and green lines represent the native and *good* non-native tones, and the red lines represent the *bad* non-native tones.

Our first observation is that in all four tones, *bad* contours are completely separated from the native contours in terms of normalized pitch register and that the *good* contours are very close to the native counterparts. Work in [17] found that, while native speakers are remarkably consistent in the relative ordering of the tone registers (Tone 1 and 4 are rendered at the highest pitches, then Tone 2, and then Tone 3), the non-native speakers are not as consistent. These results show that this inconsistency is due to *bad* productions of the tones.

Our second observation is that the shapes of the contours for the native and *good* productions are almost identical. The *bad* productions, in contrast, are completely different. Tone 1 is produced at a high register with a flat slope; the bad productions are produced at a lower register with a negative slope. Tone 2 is produced at a lower pitch register with a flat initial slope that becomes slightly positive after the 30% time mark; the bad productions are initially produced at a high register with a sharply negative slope that is the complete opposite. Tone 3 starts at a lower pitch register with a slight negative slope that eventually flattens out by the end of the syllable; the bad productions are produced at a medium register and have a completely flat slope.

The one anomally is tone 4. Although both the native and *good* contours start out at a high register (as expected), the non-native production has a much sharper negative slope than the native production. They both start at almost the same register, have the same general shape, and both contrast strongly against the shape and register of the *bad* contour. The difference in the slope can be explained by noting that non-native speakers may experience more difficulties producing tone 4 due to native language interference [28]. Our own research indicates that non-native productions of tone 4 were much more likely to be rated as *bad* than the other tones.

5. Conclusions and Future Work

This paper presents our research quantifying non-native Mandarin tone production errors based on native-speaker assessmentx. We showed that the annotators had a moderate agree-

ment with one another on a binary *good*/*bad* decision for the tone quality of each syllable. We then used these assessments to compare and contrast the features of good tone productions and bad tone productions and found that they manifest themselves mainly in the shapes and height of the pitch contours. We plan to use these features to implement assessment algorithms which will be incorporated into CALL games for students learning Mandarin.

6. Acknowledgements

This research was supported by Information Technology Research Institute (ITRI) in Taiwan. We would like to thank Hsien-Cheng Liao for coordinating the annotation effort in Taiwan and for his patience while bugs were removed from the annotation system.

7. References

[1] F. Hinofotis and K. Bailey, "American undergraduates' reactions to the communication skills of foreign teaching assistants," in *On TESOL '80*, J. Fisher, M. Clarke, and J. Schacter, Eds., Washington, DC, 1980, pp. 120–133.

[2] F. Zhang, "Exploring computer-based browsing systems in the teaching of pronunciation," in *Applied Languages Curriculum Design Conference for the 2001 4th Southern Technical Institutes and Schools of Taiwan, Republic of China.* KaoHsiung: Fortune Institute of Technology, 2001.

[3] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, pp. 61–83, 1997.

[4] M. Chen, *Tone Sandhi: Patterns Across Chinese Dialects*. Cambridge, UK: Cambridge University Press, 2000.

[5] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *Proc. of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, 2004.

[6] A. Gruenstein, I. McGraw, and I. Badr, "The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces," in *Proceeding of the International Conference on Multimodal Interfaces*, Chania, Greece, October 2008.

[7] J. Leather, "Perceptual and productive learning of Chinese lexical tone by dutch and english speakers," in *New Sounds 90*, J. Leather and A. James, Eds., University of Amsterdam, April 1990, pp. 72–97.

[8] J. M. Howie, *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge University Press, 1976.

[9] D. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, pp. 25–47, 1992.

[10] J. Gandour, "Tone dissimilarity judgements by Chinese listeners," *Journal of Chinese Linguistics*, vol. 12, pp. 235–260, 1984.

[11] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, "Training american listeners to perceive Mandarin tones," *Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3649–3658, December 1999.

[12] Y. Wang, A. Jongman, and J. A. Sereno, "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," *Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1033–1043, February 2003.

[13] C. Wang, J. R. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue, "Yinhe: A mandarin chinese version of the galaxy system," in *Proc. Eurospeech '97*, Rhodes, Greece, Sept. 1997, pp. 351–354.

[14] C. Wang and S. Seneff, "A spoken translation game for second language learning," in *Proceedings of Artificial Intelligence in Education*, 2007.

[15] A. Hawksley, "An online system for entering and annotating non-native Mandarin Chinese speech for language teaching," Master's thesis, Massachusetts Institute of Technology, 2008.

[16] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 887–890.

[17] M. Peabody and S. Seneff, "Towards automatic tone correction in non-native Mandarin," in *ISCSLP*, ser. Lecture Notes in Computer Science, Q. Huo, B. Ma, C. E. Siong, and H. Li, Eds., vol. 4274. Springer, 2006, pp. 602–613.

[18] C. Shih, "Declination in Mandarin," in *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, G. K. A. Botinis and G. Carayannis, Eds., Athens, Greece, 1997, pp. 293–296.

[19] C. Wang, "Modeling of Mandarin tone and intonation for improved speech recognition," May 2004, preprint submitted to IEEE TSA.

[20] D. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann, and K. Scherer, "Evidence for the independent function of intonation countour type, voice quality, and f_0 range in signaling speaker affect," *Journal of the Acoustical Society of America*, vol. 78, pp. 435–444, 1985.

[21] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech Communication*, vol. 6, no. 4, pp. 343–352, 1987.

[22] Y. R. Chao, *Mandarin Primer*. Cambridge: Harvard University Press, 1948.

[23] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[24] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[25] R. J. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[26] Y. Wang, A. Jongman, and J. A. Sereno, *The Handbook of East Asian Psycholinguistics*. Cambridge University Press, 2006, vol. 1, ch. L2 Acquisition and Processing of Mandarin Tone.

[27] C. Cucchiarini, H. Strik, D. Binnenpoorte, and L. Boves, "Pronunciation evaluation in read and spontaneous speech: A comparison between human ratings and automatic scores," in *Proceedings of the Fourth International Symposium on the Acquisition of Second-Language Speech*, 2002, pp. 72–79.

[28] X. Shen, "Toward a register approach in teaching Mandarin tones," *Journal of Chinese Language Teachers Association*, vol. 24, pp. 27–47, 1989.