Mandarin Tone Acquisition through Typed Interactions

Mitchell Peabody, Stephanie Seneff, and Chao Wang
Spoken Language Systems Group
MIT Computer Science and Artificial Intelligence Laboratory
The Stata Center, 32 Vassar Street, Cambridge, MA 02139 USA
{mizhi,seneff,wangc}@sls.csail.mit.edu

Abstract

This research aims to assess whether typed natural language interaction with a computer using tone-marked Mandarin pinyin can lead to improved tone production in L2 learners of Chinese. A method for correcting lexical tone errors in free form typed pinyin sentences is presented. A Web-based framework that allows students to enter sentences in pinyin and receive feedback on lexical tone mistakes was developed. To assess the feasibility of our approach, typed and acoustic data were collected from a small number of volunteers. Our plans to assess students' performance and possible extensions to this research are also discussed.

1 Introduction

Numerous studies have been done on the influence of perceptual training on the ability of adult foreign language learners to produce the sounds of a target language. A well-known example is the ability of Japanese learners of English to learn to perceive the distinction between /l/ and /r/ and to subsequently properly produce those phonemes (Bradlow and Pisoni, 1997).

Recent work has involved the training of non-native speakers in the perception and production of Mandarin tones. Wang (Wang et al., 1999; Wang et al., 2003) examined the effects of perceptual training on speakers' ability to produce Mandarin tones in an isolated set of Chinese words. Prior work by Leather (Leather, 1990) looked at the use of visual feedback on the ability of non-native speakers without any prior perceptual training to produce the four tones of a single Chinese initial-final pair.

Our general interest is in the area of designing CALL systems that will allow users to engage in spoken dialog with a computer to improve their fluency in a foreign language (Seneff et al., 2004). In particular, accurate tone production for native English speakers learning Mandarin is known to be a difficult task. We believe that speech technology can be effective in helping students learn the tones of individual Chinese words, as well as to understand how to properly express tonal aspects in production. For example, pitch contours could be displayed, or their utterances could be processed to repair erroneous tones.

In this paper we describe experiments to determine if a relationship exists between a student's knowledge of lexical tone in Mandarin and their ability to produce that tone. We are motivated by the fact that native English speakers do not use F_0 to lexically distinguish meaning in spoken language, and the observation that English speaking students of Mandarin seem to place insufficient emphasis on learning the lexical tone of a word. If students are coerced through a series of exercises to correct gaps in lexical tone knowledge, this knowledge could lead to measurable improvements in tone usage.

To address this issue, we have devised a two-phase drill exercise that involves both typing and speaking. While the primary goal of the exercise is to help the student master the language usage of the related lesson's topic, our interest here is in assessing its utility for tone acquisition. We have selected the weather as the focus topic initially, mainly because we have available considerable prior resources for multilingual dialog interaction in the weather domain (Wang et al., 2000). The exercise has been designed to promote acquisition of competence in the phonetic and linguistic aspects of the domain in the first phase, with emphasis shifting towards acquiring knowledge of the tones in the second phase. A comparison between the tone productions in the recorded utterances solicited at the end of each phase can quantify any improvements that could then be attributed to the explicit feedback on typed tone errors provided in the second phase.

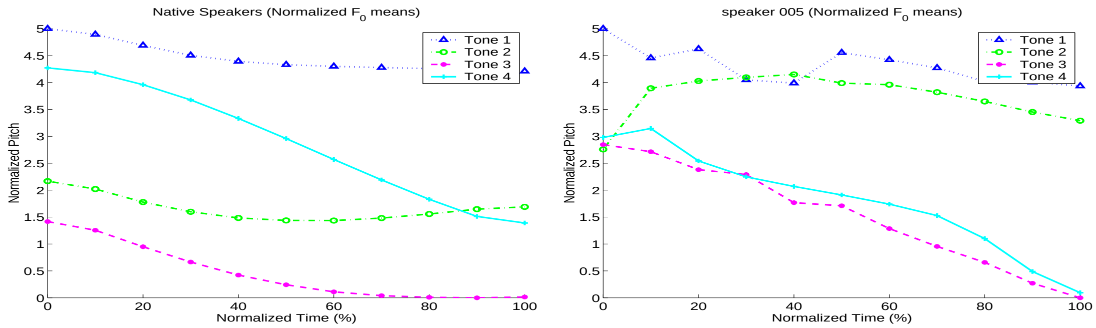
2 Technology Development

In this section, we first describe our methodology for extracting pitch contours associated with each tone, and show that it can provide a quantitative assessment of the tone proficiency of individual speakers. Subsequently, we describe our Web-based interface, which allows users to practice constructing sentences in pinyin, where errors in tone are automatically corrected. Finally, we give some details on the technology behind the tone-correction capability.

2.1 Tone Analysis

We analyzed utterances spoken by native Mandarin teachers and their English-speaking first-year students during an oral examination[1]. A corpus of 2065

[1]Data provided by the *Defense Language Institute* (DLI).



(a) Native Mandarin. (b) Native English

Figure 1: Averaged tone F_0 contours for utterances spoken by native and non-native speakers.

utterances spoken by 4 Mandarin teachers, and 4658 utterances spoken by 20 students, was transcribed in pinyin, and a pitch extraction algorithm (Wang and Seneff, 2000) was applied to track F_0.

The pitch contours were normalized for time, measured at 10% intervals, and for pitch value on a range from 0 to 5 according to a formula commonly used for analysis of F_0 values in Mandarin speech:[2]

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

where H and L are the highest and lowest F_0 for a given speaker.

We found that the tone contours for the four teachers were highly consistent and predictable, but the students' tone profiles varied widely and typically bore little resemblance to the teacher targets. Plots for four teachers and one student are shown in Figure 1. These results convinced us that tone production is a difficult task for native English speakers.

2.2 Web-based Interface

In order to facilitate our experiments, we developed a Web-based framework, as illustrated in Figure 2. We rely heavily on the pinyin representation of Mandarin, which encodes tone via a numeral entered at the end of each syllable (1 - high steady, 2 - rising, 3 - low rising, 4 - falling, 5 - neutral). Students are asked to solve 10 scenarios involving inquiries about specific weather events (rain, wind, temperature, etc.) in a specified city (Taipei, San Francisco, etc.) on a specified day (tomorrow, Tuesday, etc.). The system can optionally provide explicit feedback on any tone mistakes. It monitors the student's progress throughout the session, producing cumulative scores on the number of tries needed to solve the scenario and the total number of tone errors. During the exercise, the student can type a phrase or sentence in English, and the system will provide a Mandarin translation (Wang and Seneff, 2004), but with tones omitted.

[2] see (Rose, 1987; Wang et al., 2003).

2.3 Lexical Tone Correction

To provide feedback on lexical tone errors in typed sentences, a method for detecting and correcting such errors is needed. For this, we use the TINA natural language parser (Seneff, 1992), which has the capability to parse a word graph specified as a finite state transducer (FST). Normally, the FST arcs are labeled with confidence scores produced by the recognizer. This FST is searched to find the highest scoring parse, considering both confidence scores and linguistic probabilities. Our technique expands each typed pinyin syllable for all possible tone variations, which results in a graph representation of the input string, as illustrated in Figure 3. Heuristic scores for alternative tones are attached on the arcs in the graph, which is controlled by a global weight (currently set at 0.50) for the original input tone. The rest of the weights are distributed evenly among the other possible variations.

3 Evaluation and Experiments

3.1 Tone Error Correction Evaluation

The described technique for lexical tone correction would be of little practical value if it were unstable or untrustworthy. A simple test of its accuracy is to perturb the tones of an utterance set and let the algorithm attempt to correct them.

For this purpose, a grammar was trained on 8567 Mandarin sentences in the weather domain. We then randomly perturbed up to nine tones in each sentence. The altered sentences were then run through the tone correction algorithm. All sentences were recovered perfectly by the parser, indicating that, within this limited domain, the error-correction algorithm is very effective.

3.2 Experimental Conditions

Each of our two-phase drill exercises was partitioned further into two stages, a type-in stage and a recording stage. In the first phase, the student types Mandarin sentences into the Web page as illustrated in Figure 2, to solve the prompted 10 scenarios. Immediately after the type-in exercise, students are asked to speak queries from the same series of prompts,

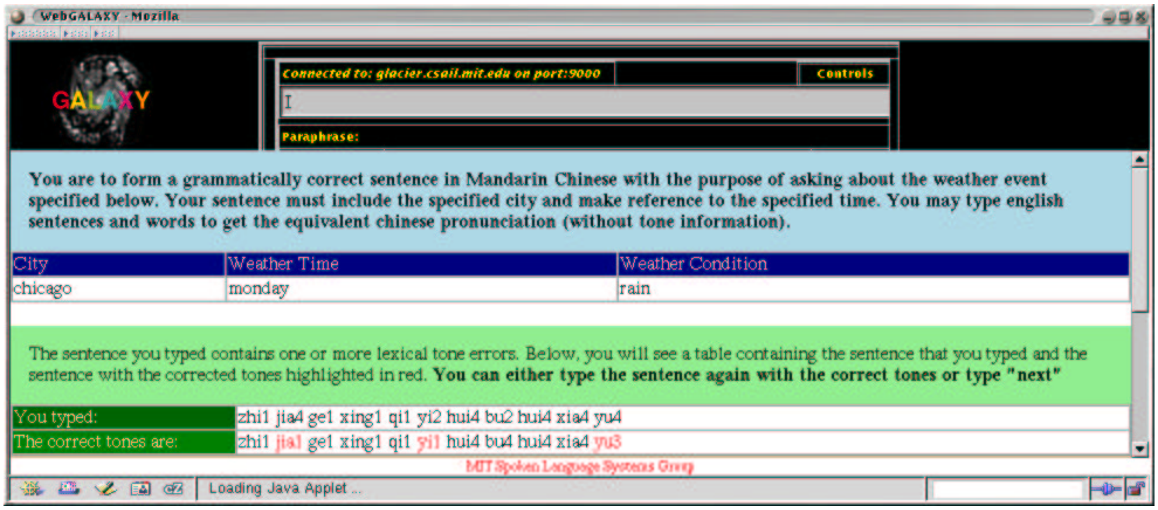


Figure 2: Screen dump of Web-based interface for the exercise during the second phase.

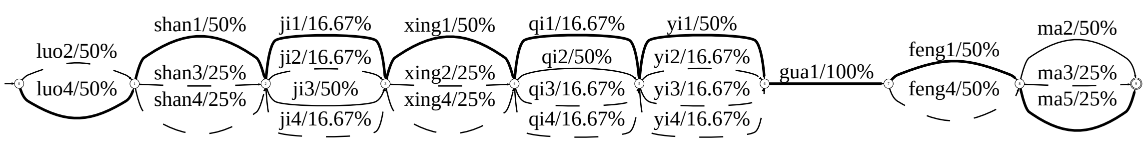


Figure 3: Example tone-expanded FST for the sentence "luo4 shan1 ji1 xing1 qi1 yi1 guan1 feng1 ma5" entered with six tonal errors, as "luo3 shan1 ji3 xing1 qi2 yi1 gua4 feng2 ma2." Bold line: corrected path. Dashed line: other alternative tones hypothesized by the system.

which are recorded and subsequently analyzed for tone production quality. The second phase differs from the first phase only in that, after the student types each question, feedback is given as to the location and nature of any incorrect lexical tones. The scenarios are generated randomly and usually differ between the two sessions.

Participants in the data collection so far have been volunteer students from the Chinese program at MIT. With the exception of two American students who had over 5 years of experience studying Chinese, the students are in the first year of study. All students were in their early to mid-twenties.

3.3 Effects on tone quality

Currently, not enough data has been collected to make any general claims about the effects of lexical tone knowledge on tone production. The data collection is ongoing, and we only discuss here the process we are using to assess students' tone productions.

A pitch contour was extracted for each utterance using the pitch tracking algorithm described in (Wang and Seneff, 2000). Context independent statistical models were built for each of the 4 lexical tones (the neutral tone is excluded from analysis) based on the pitch contours before and after feedback was given during the typed phase. Corresponding context-independent models of native speech were built from the DLI corpus mentioned previously. We have created plots like the ones shown in Figure 1, but have not yet obtained quantitative measures of tone accuracy.

4 Discussion

It is premature to make any claims about the efficacy of forcing students to remember lexical tone. We are assuming that the students are capable of producing tones if they know the lexical tones on a symbolic level. However, we observed that many subjects at the beginning level of Chinese did not have adequate training in tone production. One control condition we should examine is how well they can produce tone when *reading* a familiar text. This would provide a reference for their tone production proficiency given "perfect" knowledge of the tone.

The approach taken for the automatic correction of lexical tone errors had unforeseen benefits. For example, it is often the case that new domains in TINA are tested using a typed interface. For non-native speakers involved in domain development, an error on a single tone normally resulted in a failed parse. Now corrections can automatically be made to the input, alleviating the developer's cognitive load.

Our contours of the native models differ from those found in (Wang et al., 2003; Leather, 1990). However, the F_0 for those studies came from words spoken in *isolation*. The models in this study were pooled over instances extracted from all contexts within *continuous* speech. Co-articulation and overall sentential prosody would both influence the F_0 contours (Xu, 2004), and would likely account for the differences.

5 Future work

Data will continue to be collected in order to make valid generalizations about the effects of lexical tone knowledge on tone production. Duration statistical models will also be examined. As more data are collected, it will be possible to build statistical models of F_0 contours that take into account the contextual effects and to examine phenomena seen in native Mandarin, such as F_0 down-drift across a sentence (Wang and Seneff, 1998).

Computer Aided Pronunciation training through the use of automatic speech recognition has been approached in a number of ways (Witt, 1999; Eskenazi, 1999; Neri et al., 2001; Franco et al., 1999). For the most part, these studies have concentrated on the assessment and scoring of segmental properties of non-native speech at the phoneme, word, and sentence level.

In contrast to the previously mentioned studies, the acoustic data that was collected during this experiment will be used to develop tone scoring techniques as well as segmental scoring techniques for automatic pronunciation assessment. A correlation of these scores with native speaker perceptions of tone quality, combined with segmental feedback techniques following those developed in (Kim et al., 2004) could be used to tailor a system suitable for use by students learning Mandarin.

Since our system parses the utterances typed by students, we should be able to detect and characterize syntax errors from both the typed and spoken data. For instance, it was observed that, among the non-native speakers, the "hui4...ma5?" and "hui4 bu2 hui4" constructs were rarely employed, while they were normally used by native speakers. Such syntax errors could be modeled along with disfluencies common in non-native speech to develop language tutoring capabilities.

Clearly, our system would be improved if users could have access to spoken examples of the utterances they type in at the Web page. For this purpose, we have available to us a high quality concatenative speech synthesis framework, called Envoice (Yi et al., 2000). We would need to create an appropriate corpus by recording the voice of a native speaker. We plan to augment typed queries with automatically generated examples of synthetic speech that the user would be able to play by clicking at the Web interface.

Another future component would provide the capability to transform the student's F_0 contour to match that of the synthetic speech, using phase vocoder techniques as described in (Tang et al., 2001). The student could then listen to speech that preserved their own voice quality but exhibited prosodic contours appropriate for native speech.

Acknowledgments This research is supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

A. R. Bradlow and D. B. Pisoni. 1997. Training japanese listeners to identify english /r/ and /l/: Iv. some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101(4):2299–2310.

M. Eskenazi. 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology*, 2(2):62–76.

H. Franco, L. Neumeyer, M. Ramos, and H. Bratt. 1999. Automatic detection of phone-level mispronunciation for language learning. In *Proceedings of Eurospeech 99*.

J. Kim, C. Wang, M. Peabody, and S. Seneff. 2004. An interactive English pronunciation dictionary for Korean learners. Submitted to ICSLP'04.

J. Leather. 1990. Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In Jonathan Leather and Allan James, editors, *New Sounds 90*, pages 72–97, University of Amsterdam.

A. Neri, C. Cucchiarini, and H. Strik. 2001. Effective feedback on L2 pronunciation in ASR-based CALL. In *Proceedings of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Converence*, pages 40–48, San Antonio, Texas.

P. Rose. 1987. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, 6(4):343–352.

S. Seneff, C. Wang, and J. Zhang. 2004. Sopken conversational interaction for language learning. In *These proceedings*.

S. Seneff. 1992. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86.

M. Tang, C. Wang, and S. Seneff. 2001. Voice transformations: From speech synthesis to mammalian vocalizations. In *Proc. Eurospeech 2001*, Aalborg, Denmark.

C. Wang and S. Seneff. 1998. A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition. In *Proc. ICSLP'98*, pages 635–638, Sydney, Australia.

C. Wang and S. Seneff. 2000. Robust pitch tracking for prosodic modeling in telephone speech. In *Proc. ICASSP*, Istanbul, Turkey.

C. Wang and S. Seneff. 2004. High-quality speech translation for language learning. In *These Proceedings*.

Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno. 1999. Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, 106(6):3649–3658.

C. Wang, D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue. 2000. MUXING: A telephone-access Mandarin conversational system. In *Proc. ICSLP'00*, pages 715–718, Beijing, China.

Y. Wang, A. Jongman, and J. A. Sereno. 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, 113(2):1033–1043.

S. M. Witt. 1999. *Use of Speech Recognition in Computer-assisted Language Learning*. Ph.D. thesis, University of Cambridge.

Y. Xu. 2004. Understanding tone from the perspective of production and perception. Submitted to Language and Linguistics.

J. Yi, J. Glass, and I. Hetherington. 2000. A flexible, scalable finite-state transducer architecture for corpus-based concatenative speech synthesis. In *Proc. of the 6th ICSLP*.