An Adaptive Training Program for Tone Acquisition

Chilin Shih[1 2 3], *Hsin-Yi Dora Lu*[1 3], *Lu Sun*[5], *Jui-Ting Huang*[1 4], *Jerry Packard*[2 3]

[1]The Beckman Institute, [2]The Department of East Asian Languages and Cultures
[3]Linguistics, [4]Electrical and Computer Engineering, UIUC
[5]Oracle Corporation
cls@illinois.edu

Abstract

This paper explores ways to incorporate effective teaching methods in computer-aided pronunciation training (CAPT) programs to help second language learners acquire Mandarin lexical tones. It is hypothesized that exaggerated stimuli might help learners to identify relevant acoustic cues, varied stimuli might help them build robust classifications, and an adaptive training program would provide a platform for efficient learning.

We conducted an experiment to compare four training modules: (1) a control group with no tone training, (2) training with similar stimuli, (3) training with varied stimuli in random order, and (4) training with varied stimuli through an adaptive training program. The adaptive group had the best performance: students showed an average 8.2 points improvement on a 100-point scale, or 32% error reduction, after two and a half hours of training.

The experiment results also show that while varied input benefited most students, some students may have been confused by such input. Adaptive training effectively alleviated such confusion. The methodology developed here can apply straightforwardly to the teaching of speech sounds in other languages.

Index Terms: Adaptive training, tone acquisition, tone recognition, difficulty ranking, distance.

1. Introduction

This paper explores ways to incorporate effective teaching methods in computer-aided pronunciation training (CAPT) programs to help second language learners acquire Mandarin lexical tones. It is hypothesized that exaggerated stimuli might help learners to identify relevant acoustic cues, varied stimuli might help them build robust classifications, and an adaptive training program would provide a platform for efficient learning.

We conducted an experiment to compare four training modules: (1) a control group with no tone training, (2) training with similar stimuli, (3) training with varied stimuli in random order, and (4) training with varied stimuli through an adaptive training program. The adaptive group had the best performance: students showed an average 8.2 points improvement on a 100-point scale, or 32% error reduction, after two and a half hours of training.

The adaptive training program incorporated a data bank with more than 7000 sound files covering exaggerated as well as reduced tonal contrasts recorded from multiple speakers. The system automatically advances to a more difficult level when a student has mastered materials at a certain level. Conversely, if a student has trouble, materials from an easier level are presented next.

The experiment results also show that while varied input benefited most students, it may lead to confusion for others. Adaptive training effectively alleviated such confusion.

1.1. Mandarin Lexical Tones

Mandarin Chinese is a lexical tone language, where fundamental frequency (f_0) contours over the domain of a syllable are used to mark lexical contrast. There are four lexical tones in Mandarin, traditionally referred to as tone 1 to tone 4. Changing the tone of a monosyllabic word typically alters the meaning of the word. For example, the syllable *ma1* with tone 1, a high-level tone, means "mother"; *ma2* with tone 2, a rising tone, means "hemp"; *ma3* with tone 3, a low falling or low falling-rising tone, means "horse"; and *ma4* with tone 4, a falling tone, means "to scold".

Tone learning is an important yet challenging task for second language (L2) learners, who need to recognize the characteristic acoustic attributes that distinguish one tone from the others, and to master the prosodic system in order to communicate effectively. This task is difficult for speakers of non-tone languages where the typical function of f_0 in the native language is to convey speech act, discourse and paralinguistic information. Making the switch to use f_0 to encode lexical information requires the recruitment of different cortical areas. Previous studies had shown that tone training programs can be employed to facilitate learners perception and production of tones [1, 2]. In this paper we extend the previous works to compare the performance of different tone training programs.

1.2. Variation and Learning

There are multiple reasons why variation of speech stimuli may be crucial to the learning of new speech sounds. First of all, varied input might help learners to build robust classifications. While it is generally agreed that relative f_0 height and f_0 shapes are the primary acoustic cues for Mandarin tones [3, 4], many acoustic cues are present simultaneously which may affect the perception of tonal categories, such as f_0 turning point, duration, intensity and creakiness [5, 6, 7, 8]. Having access to multiple examples may help learners to evaluate the relative contribution of multiple cues, and to identify the primary ones.

There is considerable tonal variation in natural speech where acoustic attributes change with factors including but not limited to speaker, context, focus, sentence prosody, speaking rate and speaker effort [9, 10, 11]. Being exposed to speech samples rich in natural variation should be helpful to L2 learners.

It may not be feasible to provide varied speech samples from multiple speakers in the classroom due to time and cost constraints. To this end, computer-aided pronunciation training

(CAPT) provides a straightforward platform that can provide ample tonal variation for learners. Though the presentation order of teaching materials may play a role in facilitating learning. If a learner cannot distinguish a given sound contrast, repeated exposure may reinforce sound categories of the native language rather than leading to the learning of new sounds. Under these circumstances, adults learners may be benefited by exaggerated stimuli where differences between sound categories are emphasized [12]. Following this line of argument, varied stimuli may be more accessible to learners if they are presented in some particular order, i.e. exaggerated stimuli first. Another issue that needs to be considered is learner variability. What is easy for one learner may be difficult for another. One way to address this issue is to build an adaptive system that selects materials and adjusts presentation order to suit individual needs.

2. Adaptive Training

Historically, the concept of adaptive testing was developed in Alfred Binets adaptive intelligence test [13], and its implementation actually pre-dated the computer age. Today, adaptive systems are implemented computationally, which adapt to each individual through an iterative process judging the user's capability and selecting materials accordingly. An adaptive system can be used for testing or training. The objective of an adaptive testing program is to find a tester's proficiency level speedily and accurately. The objective of an adaptive training system is to train students to acquire a new skill in the most efficient way.

The components of an adaptive test include a data bank with all items ranked by their difficulty level, and algorithms to decide how to start, to select items, score responses and to terminate the program [14]. Ranking the difficulty level of items in a speech corpus for pronunciation training is a particularly challenging task. For a subject like mathematics, the difficulty level of the items can be assigned, crudely perhaps, by following a blueprint provided by the curriculum. Unfortunately no such shortcut is available for the calibration of speech sounds, especially if they are the spoken variations of the same text.

In commercial adaptive testing systems, the difficulty level of test items are measured by analyzing test taker's answers. A difficult item tends to be answered correctly by fewer people and by test takers who score higher. Collecting such resources are costly and time-consuming, and presuppose a system that is already in operation. That may not be a viable option for the pronunciation training of less commonly taught languages.

As an alternative to bootstrapping an adaptive system without pre-existing resources, we experiment with using speaker-listener distance to encourage speech variation and to assign an initial value representing the difficulty level of the items in the data bank. The assumption is that speech projected over a greater distance may be exaggerated, clearer and easier to understand. Speech produced for someone nearby may be softer, reduced and more difficult to understand [15]. We implemented an adaptive training program of Mandarin tones with the following design.

2.1. Data Bank

The training materials include Mandarin monosyllabic words in 15 minimal sets with 4-way tonal contrasts: *ai, cai, chuan, e, guo, ju, ma, qiao, shu, tong, wang, wei, wu, ya, yin.* These syllables include a range of consonants, vowels and syllable structures in Mandarin, and are real words when combined with 4 tones. Each word was recorded at 11 speaker-listener distance by three native speakers (one male and two females), repeated 4 times, yielding 7920 tokens (15 syllables $\times$ 4 tones $\times$ 11 distances $\times$ 3 speakers $\times$ 4 repetitions). Two native Mandarin speakers screened all tokens by listening to them in random order and identifying the tones. 195 tokens were mis-identified by at least one of the native listeners and were excluded, yielding 7725 tokens in the data bank.

During the recording session, the speaker sat in a soundproofed room with a glass panel. A computer monitor was placed 18 inches in front of the speaker and a Shure 58 microphone beside the monitor. A listener stood outside of the recording room at one of the 11 distances pre-marked on the floor. The first distance level was immediately against the glass panel of the recording room 3 feet away from the speaker. Each increasing distance level was 20 inches further away. The speaker was instructed to speak to the listener while reading prompts from the monitor. The distance between the microphone and the speaker as well as the recording volume were kept constant.

An additional Mandarin syllable database were recorded by the same three speakers with 2 repetitions where the 400 Mandarin syllables were crossed with 4 tones. This database contains about 9000 tokens recorded without a listener, and was used to train the ASR system described in section 5.

2.2. Parameter Settings

Students entered the program at mid point (level 6 in the range of 1-11). If they correctly identified 12 stimuli in a row, they advanced to a more difficult level represented by speech made with a closer distance between the speaker and the listener. On the other hand, if they made two mistakes in a row, they would drop to an easier level represented by more exaggerated speech made with longer distance between the speaker and the listener. Students finished and graduated from the training if they completed 12 items without errors at the most difficult level.

3. Experiment Design

An experiment was conducted to evaluate three tone training programs compared with a control group.

- The Red Team

  Students received no tone training. They received vocabulary training in multiple choice questions where they selected the best English translation corresponding to Chinese words written in characters. No sound inputs nor tone information were provided.

- The Green Team

  Students received training from clearly articulated and acoustically less varied speech files. Training files were selected from recordings made at distance levels 5 and 6.

- The Blue Team

  Students received training on varied materials. Speech files in each training session were selected randomly from the database. Exaggerated, normal and reduced tokens were all included and were presented in random order.

- The Orange Team

  Students in the orange team received training on varied materials through the adaptive training system where the stimuli they received depended on their performance.

	Tone 1	Tone 2	Tone 3	Tone 4	Total
Tone 1	89.9%	5.3%	1.3%	5.0%	8179
Tone 2	4.8%	74.3%	14.4%	3.4%	8358
Tone 3	0.6%	17.8%	82.3%	1.2%	8217
Tone 4	4.5%	3.0%	1.6%	90.7&	8492

Table 1: Learner's tone confusion matrix.

3.1. Subjects

Students at the University of Illinois at Urbana-Champaign who enrolled in Chinese 202, the second semester of Elementary Chinese, were recruited to participate in the tone training experiment. Participants who completed the experiment received extra course credits. There were four class sections and students in each section were randomly assigned to one of four teams.

3.2. Testing and Training Procedure

All students took a pre-test on a Friday, proceeded with training the following week, and returned for the post-test on the Monday following training. The red, green and blue team members were required to complete 5 training sessions while the orange team members may graduate early but they still needed to come back to take the post-test.

The pre-test and post-test were the same but were presented in different random order for each team and for the pre- and post-test. The test consists of 144 items chosen from odd number distance levels following a randomized block design, where equal numbers of stimuli were chosen from each tone, speaker, syllable, and distance level. Each training session for the red, green and blue teams also contains 144 items chosen according to specifications described above. Items used in pre- and post-tests were excluded from training materials.

36 students (12 red, 8 blue, 8 green, and 8 orange) completed the experiment and their answers and scores were used in the analysis.

4. Analysis and Discussion

We analyzed student's answers by combining the pre-test and post-test from all teams, and the training sessions of the blue, green and orange teams. The number of error ranged from 2 to 59 for each session (with 144 items per session). Most students made 15-35 errors. Table 1 presents the learner's tone confusion matrix. Rows 1 to 4 list input tones 1 to 4, respectively. Columns 1 to 4 show learner's perception of these tones. The last column provides the token count of each tone. Confusions of all tone pairs occur and most of the errors were caused by tone 2 and tone 3 confusion.

The red, green, blue and orange team had improvement scores of -4.76, 5.93, 4.09 and 8.21 respectively. Students in the red team, the control group, showed no improvement. In fact, their average post-test scores were worse than the pre-test scores. The green team trained with less varied stimuli had higher performance than the blue team. The orange team trained with the adaptive training program had the best performance, with 8.21 points of improvement on a 100-point scale, or 32% error reduction, after two and a half hours of training.

An ANOVA test was conducted to test the effect of team (red, blue, green, orange) on percentage improvement in tone identification. There was a major effect of team (Sum Sq=0.109, Mean Sq=0.036, F=3.757, $p<0.019$). A Tukey post-hoc test showed that the only significantly different pair is the red team (control) and the orange team (adaptive) ($p=0.028$).

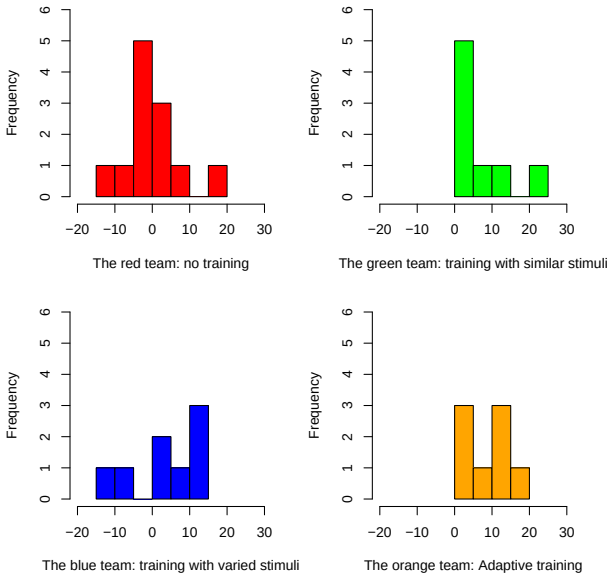

Figure 1: *Histograms of learner's improvement by team.*

Figure 1 shows histograms of learner's improvement on a 100-point scale from the pre-test to post-test. The plots are color-coded by the team's name. Positive scores show improvement. The pre-test and post-test scores of the red team are not significantly different where most learners had improvement scores centered around zero. Everyone improved in the green team where less varied tone stimuli were used. Several learners in the blue team (varied stimuli) had more than 10 points of improvement, but two had worse performance. The orange team had the best performance. Everyone improved as in the green team, and some with big improvement as in the blue team. It appears that the orange program combined the advantages of both the green and the blue team.

5. Automatic Tone Recognition

An automatic tone recognizer classifies the tone type based on a model that was learned from the Mandarin syllable speech database. If the model was trained using a standard tone database, the performance of the recognizer possibly reflected the difficulty level, compared to standard speech, of our distance-varying stimuli. Therefore, we used a standard mono-syllable tone corpus to train an automatic tone recognizer and ran recognition on the distance-varying stimuli.

The recognizer was a neural network with one hidden layer. We extracted several pitch-related features to predict the tone type: the average, maximum, minimum and range of the pitch values within a syllable; the pitch values at the beginning and the last 3 frames and their difference.

The ASR recognition accuracies for different distance levels is compared to the learner's performance during the training sessions in Figure 2. The ASR scores are plotted with black lines and black filled circles, and learner's data in red lines and red open circles. The ASR system out-performed humans for stimuli recorded at every distance levels, and on average was better than humans by 8 points. The effects of the distance level are similar in some areas for the ASR system and human learn-

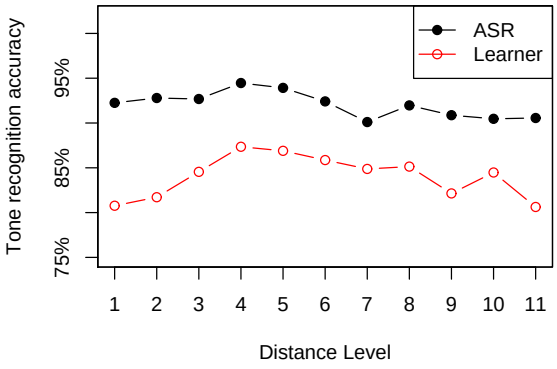

Figure 2: *Recording distance affects the tone recognition accuracy rate for both ASR and human learners, where ASR is leading human at every distance level.*

ers.

Stimuli recorded at distance level 4 had the highest recognition accuracies for both ASR and human learners, indicating that the ASR model trained with the standard tone data matches level 4 best, and this is also the level where the tonal contrasts are easiest for human learners. The degrading performance with decreasing distance levels for both ASR and human learners verifies our assumption that stimuli get more difficult than normal speech as the talker and the listener become closer. Humans apparently have more difficulty than the ASR system in this region, as shown by the steeper rate of decline in accuracy. In the other direction, our tone recognizer also shows mismatch between the standard tone data and the far-distanced stimuli. The degraded performance with increasing distance levels is a little surprising, given previous reports where tone recognition was better under the focus condition [16]. This is probably because hyper-articulation changes the pitch contours, so while tone shapes may be better differentiated under focus or in hyper-articulated speech, they are also distorted and are less similar to the normal tone types produced in clear speech, thus causing problem for ASR systems and human learners alike when the nature of the task was to compare tones across conditions. Again, human learners had steeper decline in performance as distance increases, suggesting that they had more difficulty than the ASR system in their attempt to adapt to exaggerated stimuli.

The ASR results and error analysis of student performance both suggest that the distance level 4 is the easiest and should be used as the entry level of the adaptive training program.

6. Conclusions

One technical challenge in building an adaptive program is to define and to measure *difficulty*. We implemented an approximated, one-dimensional scale using speaker/listener distance as a guide, and evaluated whether speech projected to a distance is clearer, and therefore might be easier for listeners to identify. The experimental results show that both reduced and exaggerated stimuli are difficult, and more so for human learners than for ASR.

Difficulty is a multi-dimensional concept and the mapping relationships from distance to acoustic as well as from acoustic to difficulty are both complex, though with clear patterns that can be used effectively to bootstrap an adaptive training system. Even at its initial stage without theoretically optimal parameter values, the adaptive tone training program was already the most effective training methods being tested in this experiment. Students had the best performance and achieved 32% error reduction with less than two and a half hours of training. Once the system is deployed, continuous usage by learners will provide more data to fine-tune the ranking of items, and the system performance is expected to improve.

7. Acknowledgments

This project is based on work supported by the National Science Foundation under grant numbers IIS-0623805 and IIS-0534133. The opinions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References

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

[2] J. Gandour, D. Wong, L. Hsieh, B. Weinzapfel, D. Van Lancker, and G. Hutchins, "A crosslinguistic pet study of tone perception," *Journal of Cognitive Neuroscience*, vol. 12, pp. 207–222, 2000.

[3] F. Liu, *Si-Sheng Shi-Yan Lu. Experimental report of four tones in Mandarin*. Shanghai: Qunyi Publishing House, 1924.

[4] Y. R. Chao, *A Grammar of Spoken Chinese*. University of California Press, 1968.

[5] X. Shen and M. Lin, "A perceptual study of Mandarin tones 2 and 3," *Language and Speech*, vol. 34, pp. 145–156, 1991.

[6] C. Shih and B. Ao, "Duration study for the Bell Laboratories Mandarin text-to-speech system," in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer, 1996.

[7] D. L. Blicher, R. Diehl, and L. B. Cohen, "Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: evidence of auditory enhancement." *Journal of Phonetics*, vol. 18, pp. 37–49, 1990.

[8] D. Surendran, "Analysis and automatic recognition of tones in Mandarin Chinese," Ph.D. dissertation, The University of Chicago, 2007.

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

[10] ——, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, pp. 55–105, 1999.

[11] G. Kochanski and C. Shih, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–645, 2003.

[12] McCandliss, "Success and failure in teaching the[r]-[l] contrast to japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception," *Cognitive, Affective and Behavioral Neuroscience*, vol. 2, no. 2, pp. 89–108, 2002.

[13] A. Binet and T. Simon, "Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux," *L'Année Psychologique*, vol. 11, pp. 191–244, 1905.

[14] H. Wainer, *Computerized adaptive testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[15] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 272–284, 2002.

[16] D. Surendran, G.-A. Levow, and Y. Xu, "Tone recognition in Mandarin using focus," in *Proceedings of Interspeech 2005*, 2005.