Title

Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations

Journal
Journal of Second Language Pronunciation, 1(1)

ISSN
2215-1931

Authors
Chun, Dorothy M
Jiang, Yan
Meyr, Justine
et al.

Publication Date
2015-04-23

DOI
10.1075/jslp.1.1.04chu

Peer reviewed

Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, *1*(1), 86–114.

Abstract

This paper reports on a study of 35 Mandarin Chinese learners who (1) created pitch curves of their spoken word tones and (2) compared their pitch curves with those of native speakers while practicing pronunciation. Following a pretest, the learners received training for 20-25 minutes weekly over nine weeks and took a posttest. Two types of data analyses were performed. First, native speakers of Mandarin auditorily rated the pretests and posttests. The ratings revealed that learners' pronunciation of tones improved between pretest and posttest. Second, acoustic analyses of the learners' recordings were conducted, and the learners' production was compared with that of native speakers. Results indicated that students' pronunciation of some tones improved in the posttest. The postsurveys indicated that two-thirds of the participants found viewing pitch curves helpful. This study confirms previous research but suggests that acoustic analyses complement auditory analyses with more precise indications of L2 learners' tonal difficulties.

Keywords: pronunciation, tone acquisition, visualization, acoustic analysis, open source software

Acquisition of L2 Mandarin Chinese Tones with Learner-Created Tone Visualizations

1. Introduction

With the increasing importance and popularity of Chinese language education in the U.S. and globally, it is timely to focus on one of the thorniest issues in learning Mandarin, namely mastering the tone system. For learners whose native language is a nontonal language, such as English, as well as for learners whose L1 is a tonal language, e.g., Cantonese or Vietnamese, acquiring the four Mandarin tones (plus a 5th or so-called “neutral” tone) can be very difficult, for different reasons (Hao, 2012). Speakers of nontonal languages must learn to produce correct tones for each syllable in a word, while speakers of related tonal languages might confuse similar yet distinct tones in L1 and L2. Tones are one of the first aspects of the spoken language to be taught, and this has traditionally been done primarily auditorily, where learners ‘listen and repeat’ after native speakers. Learners might also be shown a graphic representation of the four tone contours (see Figure 1). In the digital age, acoustic phonetic software has made it possible both to analyze tone production, and, instead of stylized graphics, to show actual pitch contours of native speakers and of learners (see Figure 2). Research on the use of visualizations of speakers’ pitch curves has shown that such visual input, along with the audio input, can be helpful for learners (Chan, 2003; Hardison, 2004; Molholt & Hwu, 2008; Wang, 2008, 2012). Commercial software that provides learners with immediate and automatic rendering of their pronunciation can be costly (e.g., Tell Me More http://www.tellmemore.com and PinyinPro http://www.pinyinpro.com). Furthermore, the automated speech recognition features are not as refined or precise as they should ideally be. There is thus a need for affordable software to help L2 learners of Mandarin.

This paper reports on a study in which first-year Mandarin learners were taught to create pitch curves of their spoken word tones using an open source program Praat (Boersma & Weenink, 2014; http://www.fon.hum.uva.nl/praat) and to compare them with those of native speakers. Building on an earlier short-term pilot study (Chun, Jiang, & Ávila, 2013), which suggested that visualizations could be helpful for improving tone production, the learners in this study received sustained, systematic training for 20-25 minutes every week over nine weeks for pronunciation of disyllabic words using the visualizations that they had created themselves.

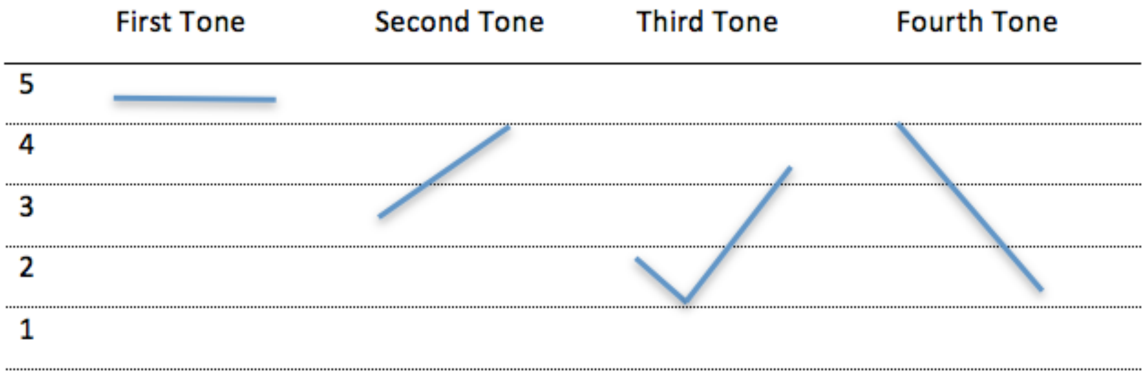


Figure 1. Typical depiction of four Mandarin tones.

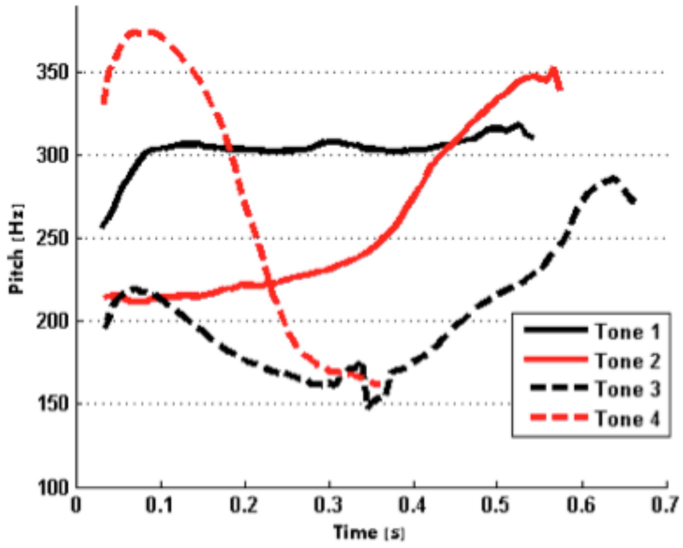


Figure 2. Acoustic representation of four Mandarin tones.

Previous studies of the acquisition of Chinese tones typically focused on either auditory or acoustic analyses of tone production, but not both, though there are a few exceptions. These studies will be discussed in the following section. Our study builds on the work that combines both types of analyses, with the added element of having learners create their own pitch curves in order to visualize their tone production. We are not aware of any studies in which learners were actively involved in using acoustic software. We further triangulate our acoustic and auditory data with student perception data from questionnaires on the usefulness of the visualization tools for mastering Mandarin tones.

2. Previous Studies

Mandarin Chinese tones are phonemic in that words have specific tones assigned to each syllable, and a different tone used with the same syllable (e.g., the syllable *ma*) results in a different meaning (e.g., *ma1* 'mother,' *ma2* 'hemp,' *ma3* 'horse,' *ma4* 'scold,' *ma0* 'question particle'; or with tone diacritics, *mā, má, mǎ, mà, må*). Tones are distinguished and described by both the height and the contour of the pitch. The following subsections summarize the research to date.

2.1. Auditory Analyses

Auditory analyses of L2 learners' tone production typically involve having native speakers listen to learners' spoken Chinese and provide ratings of the quality and accuracy of their tones. Since the four tones are different and distinctive, learners are rated for each separate tone. For example, in Miracle's (1989) study, learners who had been studying Chinese for almost two semesters made both tone contour and tone register (height) errors fairly evenly across all the tones, with the most problematic (though not significant) being the rising Tone 2 and an overall error rate of 42.9%.

Guo and Tao (2008) found that the highest percentage of correct pronunciation in both pretests and posttests was tone 1 (75%/65%), followed by Tone 4 (64%/61%) and Tone 2 (57%/60%). The lowest percentage of correctly produced tones was Tone 3 (48%/57%). He and Wayland (2010) reported similar results in that Tone 1 was best, followed by Tones 4 or 2, and worst was Tone 3.

In an extensive dissertation study of American adult learners of Mandarin Chinese, Sun's (1998) data supported the findings in Miracle's (1989) study that the production of Tone 2 is the hardest and that the majority of subjects cited Tone 2 as the hardest to produce (p. 149).

Hao (2012) investigated the acquisition of Mandarin Chinese tones by L2 learners with different L1s: ten English-speaking and nine Cantonese-speaking learners were asked to mimic tonal stimuli and read words in Mandarin with different tones. The results showed that the Cantonese group (whose L1 has six tones) did not perform significantly better than the English group. Both groups of learners were significantly better at mimicking tones than at reading them. Tones 1 and 4 were produced best when read, and Tones 2 and 3 were produced worst by both sets of learners, regardless of their native languages, reportedly due to the acoustic similarity of Tones 2 and 3 both being rising tones. This suggests that having an L1 that is also a tonal language is not necessarily advantageous in terms of mastering the tones of a different tonal language. In the case of Cantonese learners of Mandarin, Tones 1 and 4 were mapped onto overlapping Cantonese tonal categories. In the case of the English-speaking learners, they are unaccustomed to paying attention to tone contours on each individual syllable in a word or sentence.

To summarize, auditory analyses by native Mandarin speakers who rate L2 learners' production of Mandarin tones have determined that learners have greater difficulty with Tones 2 and 3 than with Tones 1 and 4.

2.2. Acoustic and Auditory Analyses

There are only a few studies that assess L2 tone production with both auditory ratings and acoustic measurements. In a study of eight American students of Chinese who had been studying for one semester, Shen (1989) divided them into "a group of better L2 speakers and a group of worse L2 speakers" based on their performance (p. 35). In contrast to Miracle (1989), Shen found that the differences of rates of errors among the 5 tones were insignificant for the group of better speakers, but significant for the group of worse speakers: Tone 4 had the highest rate of errors and Tone 1 the next highest, while differences among Tones 2, 3, and 0 were insignificant (p. 35), which is the opposite of what auditory studies typically show (see subsection 2.1.). Tonal register (pitch height) constituted the major error for both groups (p. 35), not pitch shape, with the starting-point of tones lower than that of native Mandarin speakers. Shen compared the instrumental evidence with results from auditory analyses of her data and found them to be in agreement (p. 40).

Wang, Jongman, and Sereno (2003) performed both auditory and acoustic analyses of Americans learning Mandarin and found, similar to Shen (1989), that pitch height and pitch contour were not mastered in parallel, with pitch height more resistant to improvement than pitch contour. In the pretest, which was judged by native Mandarin listeners, Tone 1 and Tone 4 were pronounced correctly almost 80% of the time, with Tone 2 slightly over 60% and Tone 3 slightly more than 20%. Following perceptual training, posttest results showed that Tone 1 had the highest percentage of correct production (over 90%), followed by Tone 2 (above 80%), Tone 4 (80%), and Tone 3 (almost 50%). Acoustic analyses of the pretest and post-training tone contours corroborated the auditory analyses, revealing that the post-training contours approximated native norms to a greater degree than the pre-training contours, with Tone 3 the most problematic for learners.

2.3. Learner Training with Visual Input

In a study of two training paradigms for learning Mandarin tones, Wang (2008) reported that one group received perceptual training with only auditory input while the other group received perceptual and production training with both auditory and visual input. After three weeks of training, both groups' production accuracy improved significantly in the posttest as compared with a control group. In a similar study, Wang (2012) found that beginning Mandarin learners with different L1 backgrounds who received six hours of tone training, which consisted of real time display of pitch contours of both native speakers and learners, improved their production of Mandarin tones significantly between the pretests and posttests. A control group that received the same classroom instruction but not the computer-based training did not match this improvement. In both studies, the learners' production was judged by native Mandarin listeners.

A pilot study (Chun et al., 2013) found that L2 Mandarin Chinese learners who had been studying Mandarin for six months pronounced tones well when reading mono- and disyllabic words, showing a ceiling effect in that 83% of the tones were pronounced correctly. Of the 17% of the tones that were *incorrect* in the pretest, after a brief training in which they viewed both native speakers' and their own pitch curves, learners improved on almost 50% of them in the posttest. A larger study with more participants and a longer training period was deemed necessary.

3. Current Study and Research Questions

The current study builds on the pilot study and on previous auditory and acoustic studies of L2 Mandarin tone acquisition. Most previous work used native speaker raters and only Shen (1989) and Wang et al. (2003) included both auditory and acoustic analyses. There is thus a need to corroborate auditory ratings with acoustic data. Our study involves 35 L2 Mandarin Chinese learners who had been learning Chinese for three months and were trained with visualizations of pitch curves over a period of nine weeks. It employs both quantitative and qualitative measures. The research questions are: (1) When learners' tone production in pretests and posttests is rated auditorily by native listeners, does their tone production improve from pretest to posttest? (2) When learners' tone production in pretests and posttests is assessed by acoustic analyses of their production, are the results similar to the auditory assessments of the native listeners? (3) If learners improve their tone production from pretest to posttest as assessed by native speakers, do they report that viewing visualizations of native speakers' pitch contours and comparing them with their own pitch contours is helpful? Based on the results of the study, pedagogical implications will be discussed at the end of the paper.

4. Methodology

4.1. Participants

The participants were 35 students in the second quarter of Chinese language instruction at a large western state university (18 males and 17 females). They ranged in age from 18 to 21 years. When the study began, the students had studied Chinese for at least one quarter, i.e. two and a half months. Twenty-six students spoke English as their native language; 11 spoke a tonal language (6 Cantonese, 2 Chiu Chow, and 3 Vietnamese); the remaining 4 spoke nontonal languages (1 French, 1 Korean and 2 Spanish). Six were bilingual, explaining why the total is greater than 35. All were fluent in English.

It is important to note that the study was carried out in an authentic learning environment, rather than as an experiment. Students in four sections of second-quarter Chinese received the training described below during regular class periods. At the beginning of the quarter, a total of 73 students were enrolled, but only those students who participated in all of the activities (pretests and posttests, weekly training) are included in this study, reducing the number of complete datasets to 35.

4.2. Materials and Procedures

4.2.1. Native Speaker Recordings that Served as Training Materials

Using Praat, a free software package for speech analysis, one male and one female native speaker of Mandarin Chinese each recorded 60 disyllabic words. Each of the 60 words was comprised of two syllables, and all combinations of the four tones and the Neutral Tone were represented (e.g., *ōu zhōu* (T1-T1)*, jīn nián* (T1-T2)*, guān diǎn* (T1-T3)*, yīn yuè* (T1-T4)*, gē bo* (T1-T0)). Appendix A shows the 60 words that the native speakers recorded. Each word was then labeled and numbered, and the digital recordings were put in a weekly training folder for the participants (see Appendix B for the training schedule). The female and male native speakers' recordings were placed in separate folders and labeled accordingly.

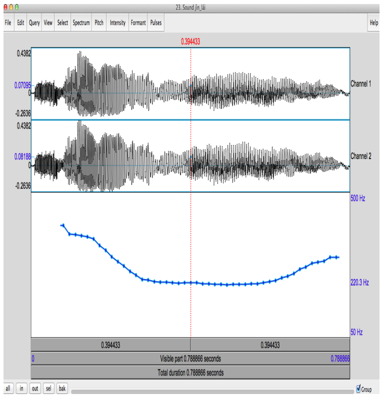


Figure 3: Waveform and pitch curve of a female native speaker saying *jìn lái* 'come in.'

4.2.2. Weekly Student Activities with Praat

Every Friday during the quarter, the students went to a computer lab and had 20-25 minutes during class to practice their tone production. Students were first trained in how to locate the folder containing the native speaker recordings, download, listen to, and then create pitch curves with Praat of the native speaker recordings. Female students used the female native speaker as their model and male students used the male native speaker as their model. They were then instructed how to record themselves saying the same words, listen to their recordings, and create pitch curves of their own recordings. In this way, students could first receive a model input (both visual and auditory) and then create a visualization of this input and compare their pitch curve to that of the native speaker. They could record themselves as many times as they wished until they were satisfied with the sound and visual of their own file. During the first four weeks, students practiced one tone per week, saying words in isolation (see Appendices A and B). During the next four weeks, they did a second round of practice and also had to use the words in sentences.

An example of a native speaker vs. student waveform and pitch curve can be found in Figure 4.

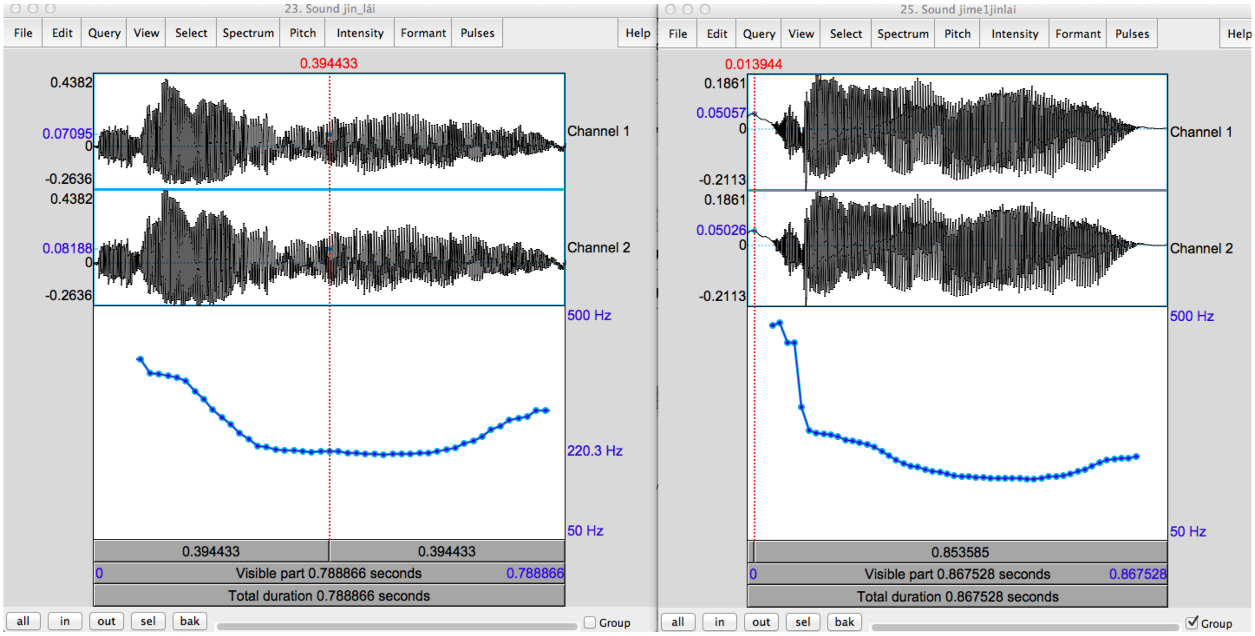


Figure 4: Waveforms and pitch curves of *jìn lái* 'come in' produced by a female native speaker (left) and a female student (right).

4.2.3. Pretest and Presurvey

A pretest and presurvey were administered at the beginning of the quarter. For the pretest, students were asked to record themselves at home reading a list of 60 words and to submit the audio file to the course instructor (see Appendix A). The 60 words were selected based on the position of the tone within each word and were the same words that the native speakers had recorded as training materials. Students were also asked to fill out an online presurvey at home (see Appendix C), which asked for demographic data, including language background, interest in and exposure to Chinese, and the importance of pronunciation to them.

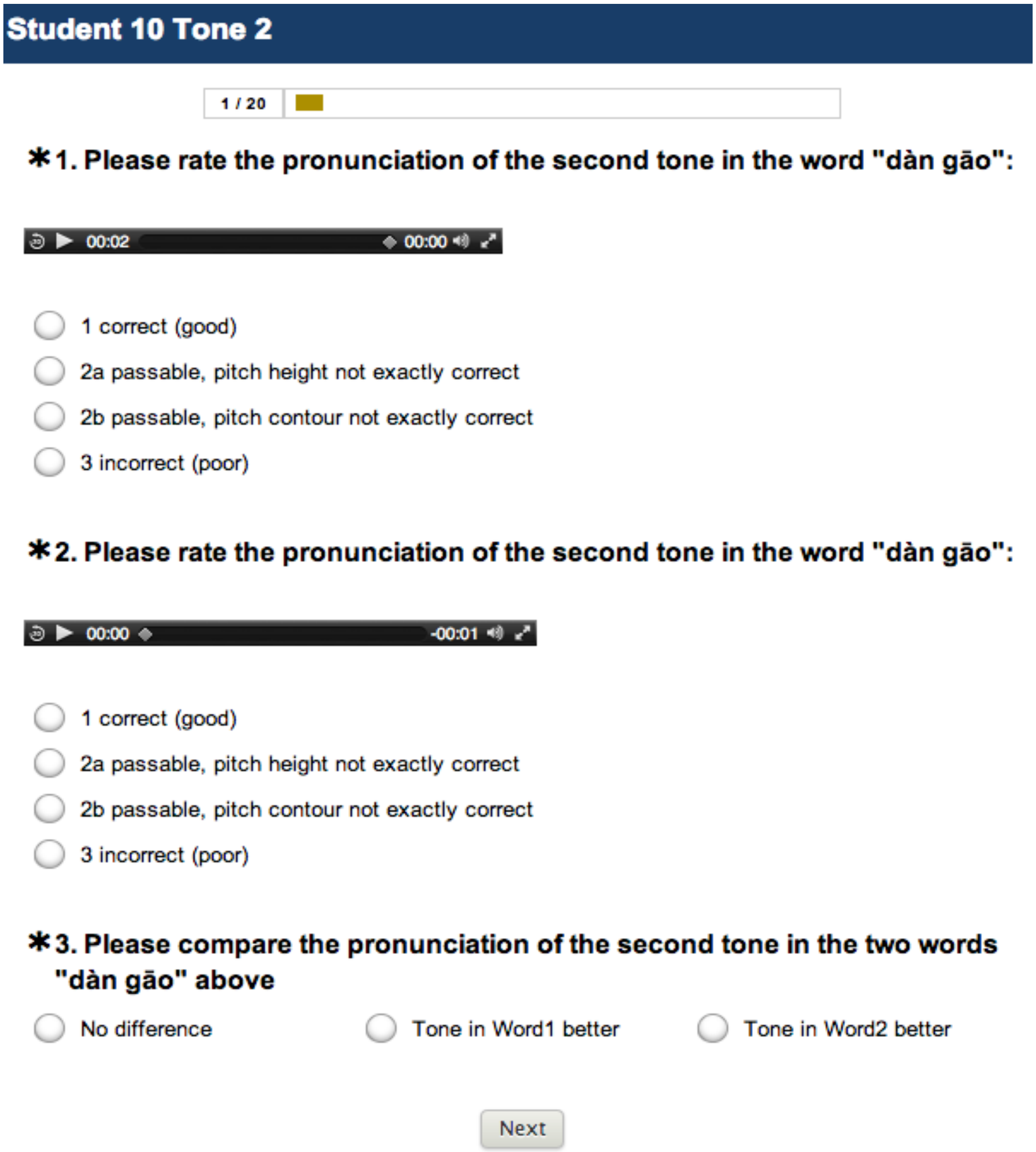
4.2.4. Posttest and Postsurvey

A posttest and postsurvey were conducted at the end of the quarter in order to track the change in the practiced pronunciation of 20 (of the 60) words as well as the attitudes of each individual student. The posttest was conducted during class time in the last week of the quarter and students were asked to record themselves reading the same 60 words into an audio file. Students participated in a short online postsurvey at home, which elicited their opinions about the usefulness of Praat throughout the quarter for improving their pronunciation. Likert-scale items and open-ended questions were used in both pre- and postsurveys (see Appendix D).

4.2.5. Auditory Analyses of the Data

The learners' production data was assessed in two different ways, auditorily and acoustically. For auditory assessment, four native Mandarin speakers were asked to listen to learners' word tones. Of the 60 disyllabic words in the pretest, 20 were selected for auditory analyses representing each possible combination of tones (see Appendix A), and these 20 were presented in pairs (randomized pretest and posttest recordings) to the native speaker raters. Two types of ratings were used, (1) raters' judgments about the correctness of each tone (and in the case of passable but not native-like tones, whether it was the pitch height or the pitch contour that was not exactly correct) and (2) judgments about which of the two randomized pretest and posttest recordings was better (see Figure 5).

The four Mandarin speakers were first trained in the rating process. They were presented with sample learners' recordings and discussed the criteria for evaluation with the researchers. At first, they reported great difficulty in deciding for incorrect tones whether the problem lay with pitch height (e.g., pitch was too high or too low) or with pitch contour (e.g., pitch was level when it should have been rising or falling; pitch was rising when it should have been level or falling, etc.). However, after a training period of an hour with the researchers, during which they listened to many examples, they reached agreement on determining whether a tone had an incorrect pitch height or pitch contour. If both pitch height and pitch contour were incorrect, they rated the tone as "incorrect (poor)." Following the training, the raters individually rated the learners using the online survey software SurveyMonkey, which incorporated audio files and rating options to facilitate the rating process (see Figure 5 for a sample page of the survey).



Student 10 Tone 2

✱1. Please rate the pronunciation of the second tone in the word "dàn gāo":

○ 1 correct (good)
○ 2a passable, pitch height not exactly correct
○ 2b passable, pitch contour not exactly correct
○ 3 incorrect (poor)

✱2. Please rate the pronunciation of the second tone in the word "dàn gāo":

○ 1 correct (good)
○ 2a passable, pitch height not exactly correct
○ 2b passable, pitch contour not exactly correct
○ 3 incorrect (poor)

✱3. Please compare the pronunciation of the second tone in the two words "dàn gāo" above

○ No difference ○ Tone in Word1 better ○ Tone in Word2 better

Figure 5. Sample page of the rating survey in SurveyMonkey.

On each page of the survey, raters were presented with one set of two audio files of the same disyllabic word spoken by the same learner in the pretest and posttest. However, the order of the two audio files was randomized so that the raters would not know whether the audio file was from the pretest or the posttest. Raters first assigned scores separately for each of the two audio files on a scale of 1-3, 1 = correct (good), 2a = passable (pitch height not correct), 2b = passable (pitch contour not correct), and 3 = incorrect (poor). Then the raters were asked to compare the two tonal productions and rate the difference as 0 = no difference, 1 = tone in Word 1 better, and 2 = tone in Word 2 better. They were allowed to listen to the recordings as many times as they wanted. To avoid rater fatigue, two native speakers rated the recordings of 18 students, and the other two native speakers rated the recordings of the remaining 17 students. Thus, for each learner there were two independent sets of ratings. The evaluations from all raters were included in the analysis.

4.2.6. Acoustic Analyses of the Data

Acoustic analyses of the data were then conducted on both the student audio recordings and native speaker audio recordings. The pretests and posttests of the students were acoustically analyzed to track the changes in pitch height and pitch contour. Using Praat, the pitch listings (in Hz) of each syllable were downloaded. The programs Perl and Matlab Version 7.6.0 (R2008a) were then used to obtain graphic representations of the audio files. Below are the steps used for both the native speaker files and student files, following Wang et al. (2003), p. 1037.

1. Individual pitch listings (Hz) were normalized into 5-point scale pitches: a logarithm transformation for each pitch listing was conducted using the formula: T = [(lg X – lg L)/(lg H – lg L)] * 5. (X = any given pitch listing, L = lowest pitch listing by the speaker, H = highest pitch listing by the speaker).
2. The time axis of each token was normalized to (0,1). Formula: (any given time – minimum time)/(maximum time – minimum time).

3. 11 sample points were taken at the normalized timing axis and data at each point were interpolated. For example, at 0, 0.1, 0.2, …1 (i.e., estimate tone pitches using given data).
4. Pitches for each tone were averaged across speakers (native speakers; all students' pretests and posttests; low- and high-proficiency students' pretests and posttests);
5. Tone figures were drawn with the averaged pitches for each group (native speakers; all students' pretests and posttests; low- and high-proficiency students' pretests and posttests).

The graphic representations based on the acoustic analyses of the learners' audio files were then compared to the figures generated from the acoustic analyses of the native speakers' audio files and are discussed in the Results section below (see Figures 11-13).

5. Results

5.1. Auditory Analyses

There were 35 complete data sets, with each data set consisting of 40 syllables in 20 disyllabic words. Each data set was rated by two raters, who provided ratings for each word in the pretest, the posttest and a comparison of the two. A total of 8,400 ratings were made.

5.1.1. Pretest Results

On average, the students' tones (for all 5 tones) were correct approximately 55% of the time in the pretests, suggesting a mild ceiling effect. Among the remaining 45% of the data, 22% of them had problematic pitch height, 11% had problematic pitch contour, and 12% of them were incorrect in both pitch height and pitch contour (see Figure 6). A repeated sample *t*-test revealed that the average percentage of "pitch height incorrect" was significantly higher by 11 points than the average percentage of "pitch contour incorrect" (t (171) = 7.01, $p < .001$). Table 1 displays the means and standard deviations of ratings for individual tones. This means that raters assessed problems with pitch height significantly more often than problems with pitch contour.

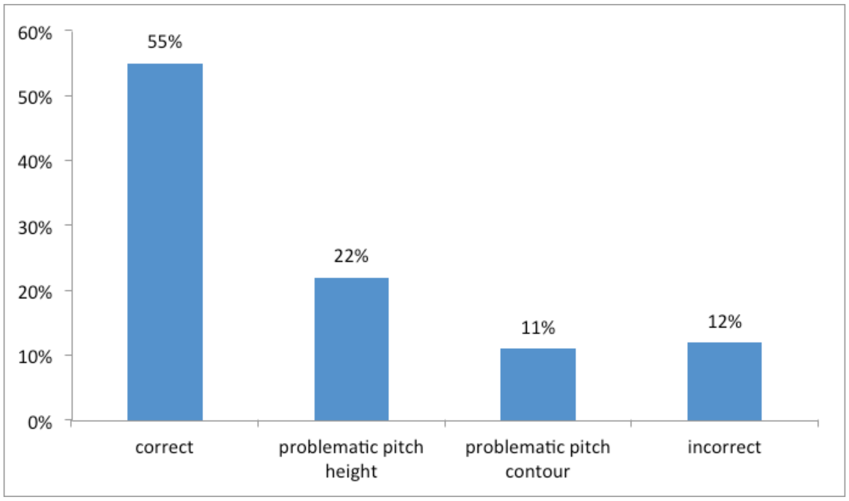


Figure 6. Overall pretest ratings.

Broken down by tone, the raters' results for the pretest for individual tones show that the percentages of "correct" ranged from 45% to 65% (Figure 7). Neutral Tone and Tone 3 had lower percentages of "correct" ratings (45% and 46% respectively), which indicates that Neutral Tone and Tone 3 were more challenging to our English-speaking learners than other tones. Furthermore, the percentages of "pitch height incorrect" were higher than the ones of "pitch contour incorrect" for all 5 tones. This means that raters perceived that learners have more problems with pitch height than with pitch contour when producing Mandarin tones.

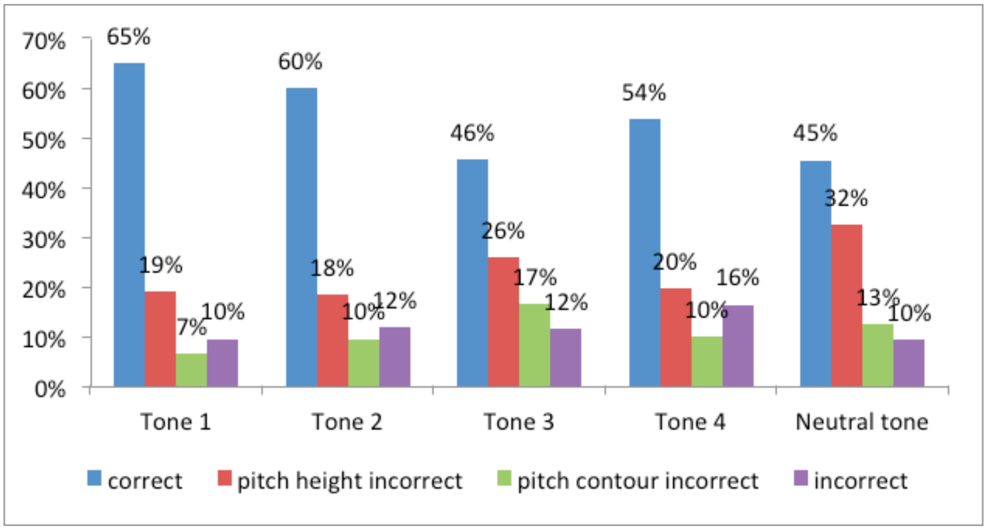


Figure 7. Ratings of each tone in pretests.

5.1.2. Posttest Results of Auditory Analyses

The ratings of the posttests were broken down by tones and are displayed in Figure 8. Table 1 displays the means and standard deviations of ratings for individual tones. Compared to the pretest results, Neutral Tone, Tone 2 and Tone 3 increased by 2% in the ratings of "both correct" in the posttest. Tone 4 made the biggest improvement by 8%. However, Tone 1 dropped by 2% in the "both correct" category. Similar to the pretests, there were more pitch height problems than pitch contour problems in the posttests for all 5 tones. A repeated sample *t*-test found that the average percentage of "pitch height incorrect" was significantly higher by 11 points than the average percentage of "pitch contour incorrect" (t (170) = 7.96, $p < .001$). This indicates that raters judged problems with pitch height significantly more often than problems with pitch contours.

The comparison between the pretest and posttest results shows only a slight difference between the two tests. This may be due to two reasons: First, 55% of students started with correct tones in the pretest. This means that half of the time there was no room for improvement. Second, the possible improvement by some students might be overlooked if we look globally at the entire dataset. Therefore, we conducted further analysis on the remaining 45% "incorrect" data and categorized the students into high-proficiency and low-proficiency groups for a finer-grained analysis. The results are reported in the following section.

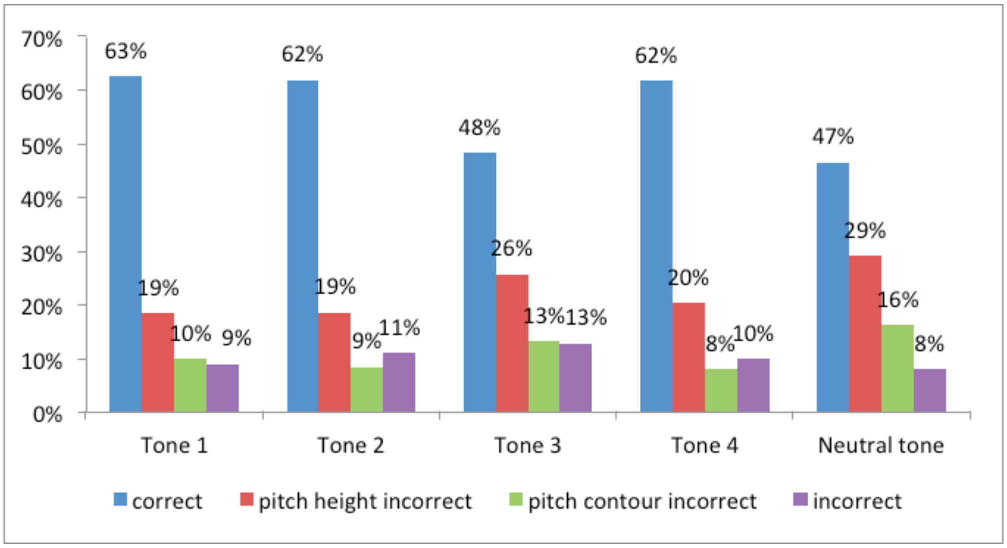


Figure 8. Ratings of each tone in posttests.

Table 1

Statistics (%) of Ratings for Each tone in Pretests and Posttests.

Tone	Rating	N	Pretest *M* (*SD*)	Posttest *M* (*SD*)
1	Correct	35	65(24.22)	63(17.38)
	Pitch height incorrect	35	19(14.00)	19(11.83)
	Pitch contour incorrect	35	7(11.28)	10(11.22)
	Incorrect	35	10(11.10)	9(11.58)
2	Correct	34	60(20.61)	62(17.46)
	Pitch height incorrect	34	18(12.02)	19(11.11)
	Pitch contour incorrect	34	10(8.77)	9(9.76)
	Incorrect	34	12(10.53)	11(10.84)
3	Correct	35	46(24.53)	48(17.22)
	Pitch height incorrect	35	26(17.84)	26(13.30)
	Pitch contour incorrect	35	17(13.81)	13(9.60)
	Incorrect	35	12(11.77)	13(10.45)
4	Correct	35	54(29.01)	62(21.27)
	Pitch height incorrect	34	20(14.02)	20(12.81)
	Pitch contour incorrect	35	10(12.19)	8(8.55)
	Incorrect	35	16(19.99)	10(15.40)

Neutral	Correct	34	45(25.09)	47(22.42)
	Pitch height incorrect	34	32(19.49)	29(16.89)
	Pitch contour incorrect	34	13(14.29)	16(17.10)
	Incorrect	34	10(13.92)	8(10.82)

5.1.3. Improvement Results of the 45% Incorrect Tones

Although the overall improvement was slight when all the data were analyzed together, the remaining 45% of the students who started with incorrect tones could possibly have improved due to the use of pitch curve visualizations. A closer analysis of this 45% shows that 62% of these students did improve from pretest to posttests, 29% remained the same, and 9% experienced a decline (Figure 9).

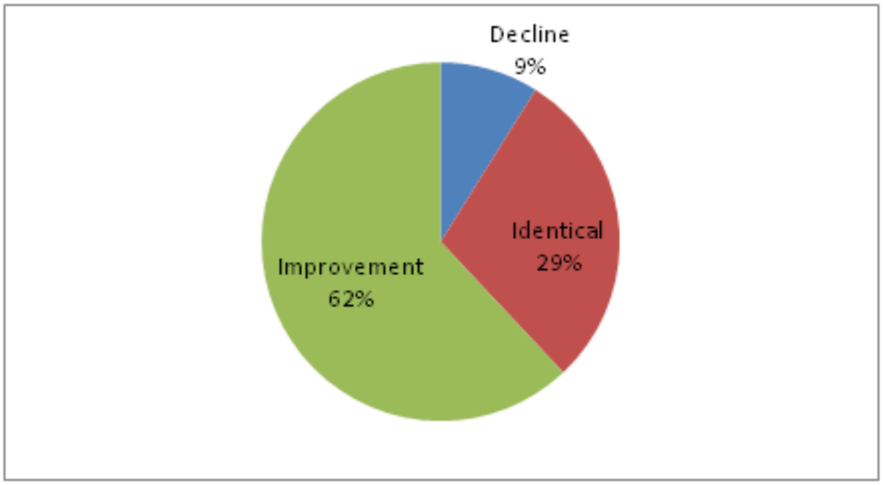


Figure 9. Improvement rating of the remaining 45%

For a more detailed analysis of students' improvement, it is important to look at the specific tones. We broke down the results to analyze whether certain tones would show more improvement than others (Figure 10 and Table 2). The results indicate that improvement was made on all 5 tones. Tone 1 and Tone 4 had the biggest improvement of over 70%. The Neutral Tone showed the least amount of improvement of 42%, and Tone 3 had the second least improvement of 56%. The pretest results indicated that Neutral Tone and Tone 3 were more challenging for the learners before the training, and the posttest results also reveal that they are relatively harder to improve, despite some measure of improvement following training with pitch curve visualization.

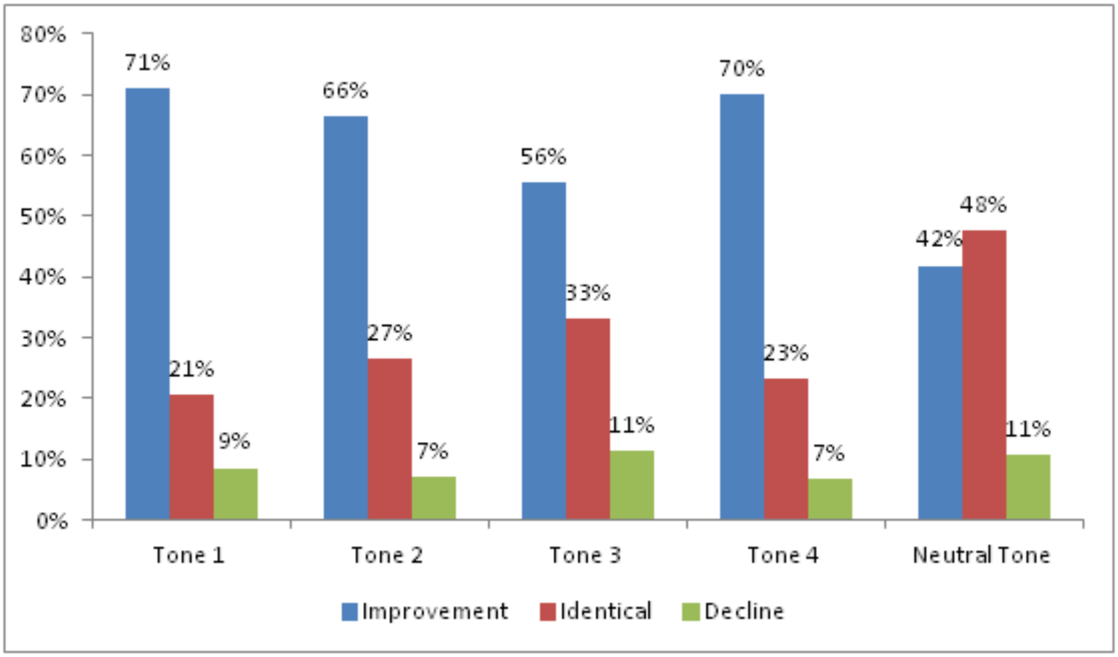


Figure 10. Improvement ratings for each tone.

Table 2

Statistics (%) of Improvement Ratings for Each Tone

Tone	Rating	N	*M*	*SD*
1	Improvement	32	71	28.58
	Identical	32	21	21.03
	Decline	32	9	19.27
2	Improvement	34	66	21.69
	Identical	34	27	20.40
	Decline	34	7	12.46
3	Improvement	35	56	25.88
	Identical	35	33	22.76
	Decline	35	11	11.94
4	Improvement	34	70	23.54
	Identical	34	23	21.09
	Decline	34	7	13.02
5	Improvement	34	42	36.92

Identical	34	48	32.76
Decline	34	11	21.97

5.2. Acoustic Analyses

In this section, graphic presentations of tone production across speakers using their normalized pitch listings are presented, and then the students' production in pretests and posttests is compared to native speakers. Figure 11 shows the average F0 curves for each tone by native speakers (represented by blue lines) and by all students in the pretest (represented by black dotted lines) and the posttest (represented by black solid lines). The horizontal axis indicates 11 sample points at the beginning and at every 10% position of the time duration. The vertical axis indicates the normalized 5-point pitch height scale from 1.5 to 3.5.

Compared to the native speaker, the students' pitch curves in the pretests show both pitch height and contour problems. On average, students began with lower pitch height when pronouncing Neutral Tone, Tone 1, and Tone 4. And they had higher pitch when pronouncing Tone 3. The average starting pitch height of Tone 2 was about the same as the native speakers. With regard to pitch contour, some differences were found in Tone 2 when the students' pitch did not rise as high as the native speakers' and in Tone 4 when the students' pitch did not fall as low as the native speakers.'

Students' pitch height and contour changed in the posttests. For the Neutral Tone, the students' pretests and posttests were almost identical in pitch height, but the average posttest pitch contour was smoother. Students produced higher pitches in the posttests when pronouncing Tone 1 and Tone 2. For Tone 3, the students lowered their pitch in the posttests and matched the native speakers' starting pitch height more closely. For Tone 4, students had the same beginning and ending pitch height in the pretests and posttests. However, their pitch contour changed in the posttests. The steeper drop in the posttests was closer to that of the native speakers.

In sum, students began with either pitch height or pitch contour problems (sometimes both) with all 5 tones in pretests. With the aid of pitch curve visualization, students improved in the posttests for some tones (*e.g.,* pitch height for Tone 3 and pitch contour for Tone 4). However, for Tone 2, students demonstrated a higher pitch height in posttests, making their pitch contours less native-like. Also, there is evidence of overcorrection of pitch height with Tone 1.

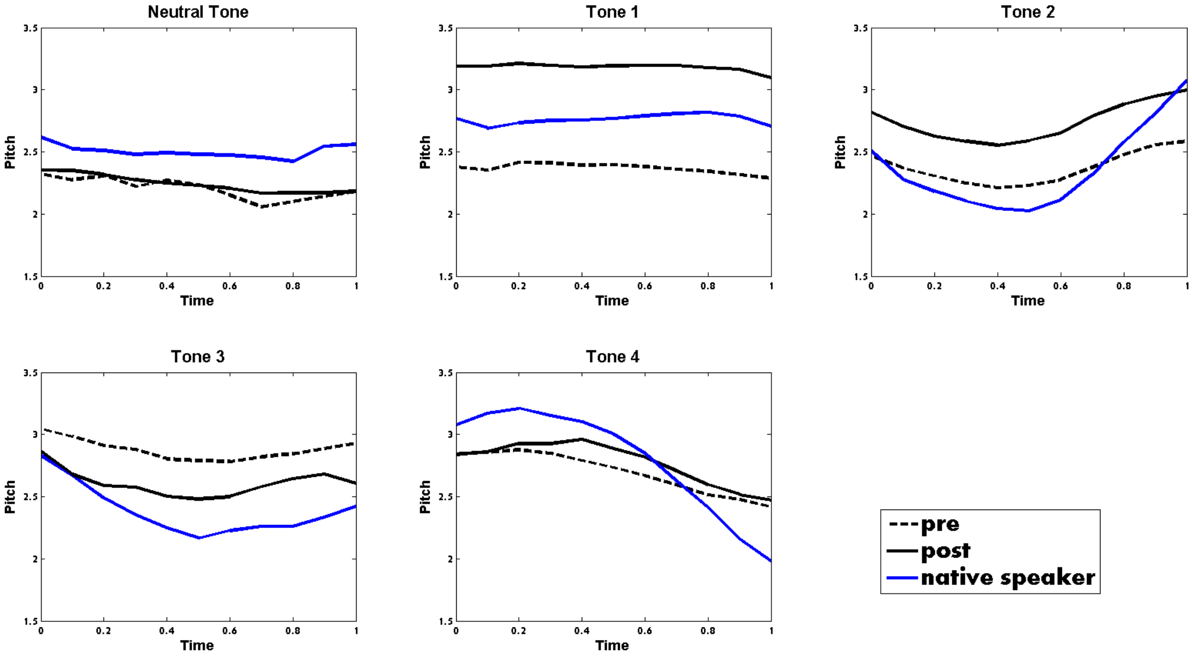


Figure 11. Average F0 curves for each tone by native speakers and students in pretests and posttests.

For the purpose of examining the effect of visualization for students at different proficiency levels, we divided the students into high, medium and low proficiency groups based on their performance in the pretest. Out of the 35 data sets, five students were excluded for having more than 10% missing ratings (due to the poor quality of the recordings). The remaining 30 students were categorized into three proficiency groups (high, medium, and low) based upon the percentages of rating “1” (correct) and rating “3” (incorrect) that they received in the pretest.

Nine students who received the highest percentages of rating “1”, ranging from 67.5% to 95%, and the lowest percentages of rating “3”, ranging from 0 to 6.3%, were deemed “high proficiency,” while eight students who received the highest percentages of rating “3”, ranging from 26.3% to 46.8%, and the lowest rating “1”, ranging from 16.3% to 30%, were categorized as “low proficiency,” and the remaining 13 students were placed in the category “medium proficiency.” The statistics of high, medium and low proficiency students’ ratings in the pretest are presented in Table 3.

Table 3

Statistics (%) of High, Medium and Low Proficiency Students’ Ratings in Pretest.

Proficiency	Rating	*N*	Min.	Max.	*M*	*SD*
High	Correct	9	67.50	95.00	78.38	9.35
	Pitch height incorrect	9	2.50	27.50	13.74	9.27
	Pitch contour incorrect	9	0.00	13.80	4.93	4.90
	Incorrect	8	0.00	6.30	3.32	2.01
Medium	Correct	13	33.80	76.30	54.27	10.18
	Pitch height incorrect	13	10.00	42.50	23.90	9.34
	Pitch contour incorrect	13	3.80	16.30	11.48	3.96
	Incorrect	13	2.60	17.50	10.43	4.73
Low	Correct	8	26.30	43.80	35.55	6.18
	Pitch height incorrect	8	17.50	44.70	26.86	9.25
	Pitch contour incorrect	8	3.90	22.50	15.20	5.88
	Incorrect	8	16.30	30.00	22.48	4.68

When the high-proficiency group (green lines) and low-proficiency group (red lines) were compared separately with the native speaker, refined details are revealed (see Figures 12 and 13). High-proficiency students demonstrated better pronunciation in pretests than low-proficiency students (*e.g.,* Tone 1, Tone 2 and Tone 4, see Figure 12), which corresponds to the raters’ auditory judgments. Also, high-proficiency students’ performance was more stable than the low-proficiency students. In other words, there was less deviation between the pretests and posttests. This might be due to the fact that the high proficiency students' pretests were already closer to the native speakers. Nevertheless, these high-proficiency students demonstrated improvement on some tones in the posttests. For example, their Neutral Tone contour was flatter, the dip in Tone 3 was lower, and their pitch height was closer to the native speakers.'

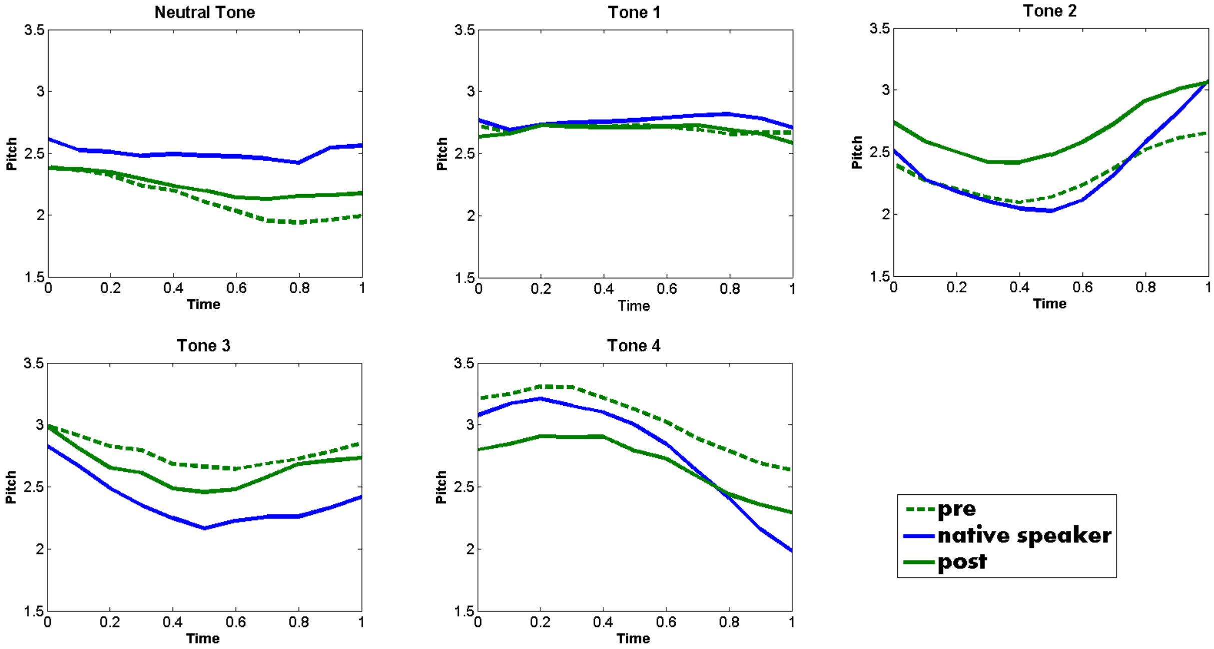


Figure 12. Average F0 curves by native speakers and high-proficiency students in pretests and posttests.

On the other hand, the contrast between pretests and posttests of low-proficiency students was greater, and the improvement in posttests was more obvious (Figure 10). For example, the Neutral Tone contour was consistently level. The pitch heights of Tone 1 and Tone 3 improved, although there was evidence of some overcorrection. Tone 4 had the biggest improvement in pitch height and contour.

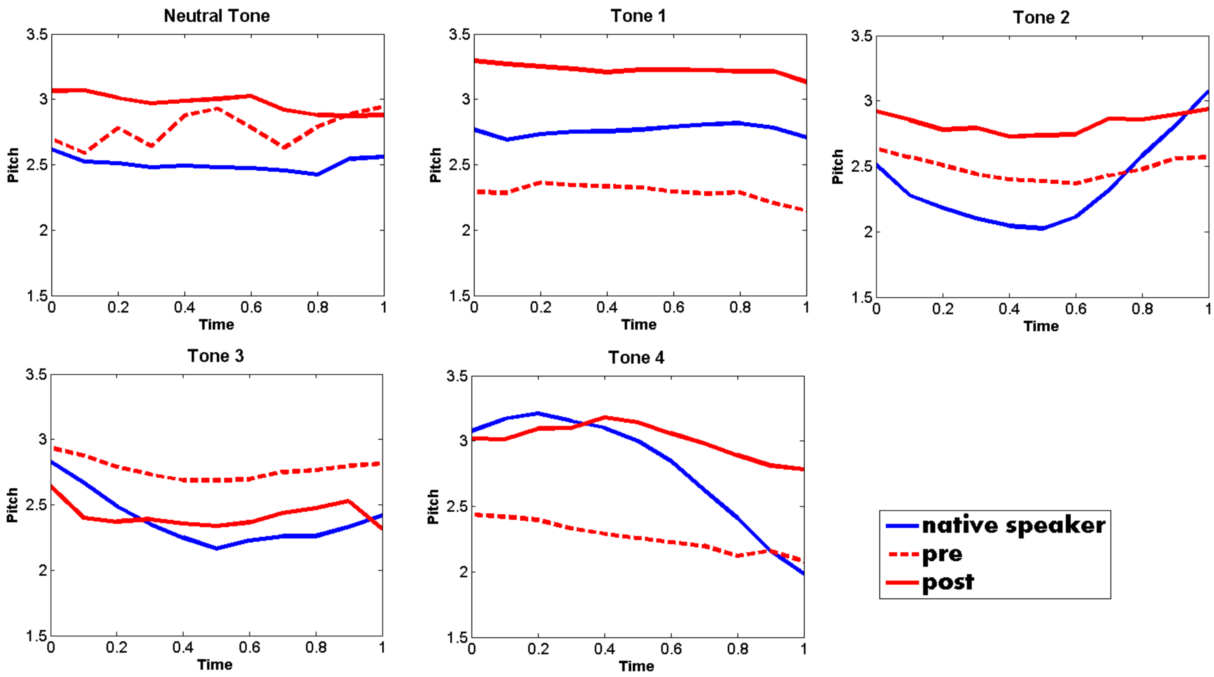


Figure 13. Average F0 curves by native speakers and low-proficiency students in pretests and posttests.

5.3. Student Surveys

Among the 35 participants, one student did not do the postsurvey, resulting in a total of 34 sets of responses. The results of students' postsurveys (Figure 15) indicate that 68% of the participants (23 out of 34) regarded seeing the native speakers' pitch curves as helpful, choosing "agree" or "strongly agree." Similarly, 65% of the participants (22 out of 34) considered viewing their own pitch curves as helpful, and 66% of the participants (21 out of 32 with 2 students skipping this particular item) found comparing their own pitch curves with those of the native speakers' helpful.

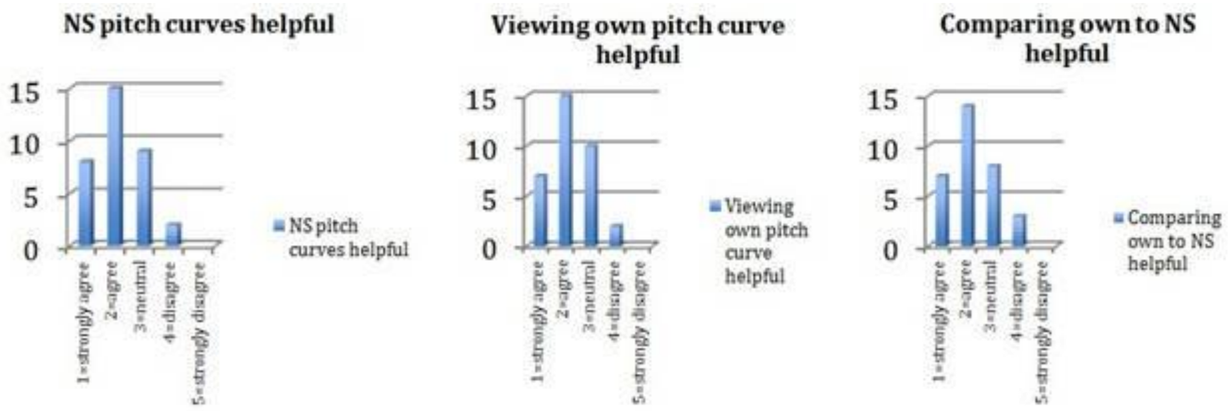


Figure 14. Results of post-surveys.

For a finer-grained analysis, the postsurveys of the 12 participants who had exhibited the greatest improvement between the pretest and posttest were examined, in order to determine whether there was a relationship between perceived usefulness of the visualizations and actual improvement. These 12 participants received higher percentages of ratings indicating that their posttest production was better than their pretest production. Figure 16 shows that these most-improved students had higher percentages of “agreeing” or “strongly agreeing” that viewing and comparing pitch curves was helpful. Specifically, 10 of 12 (83%) found viewing visualizations of native speakers’ pitch contours helpful, 8 of 12 (67%) found viewing their own pitch curve helpful, and 8 of 11 (73%, with one student skipping this item) agreed or strongly agreed that comparing native speakers' pitch contours with their own are helpful. These percentages are higher than the average percentages for the entire group.

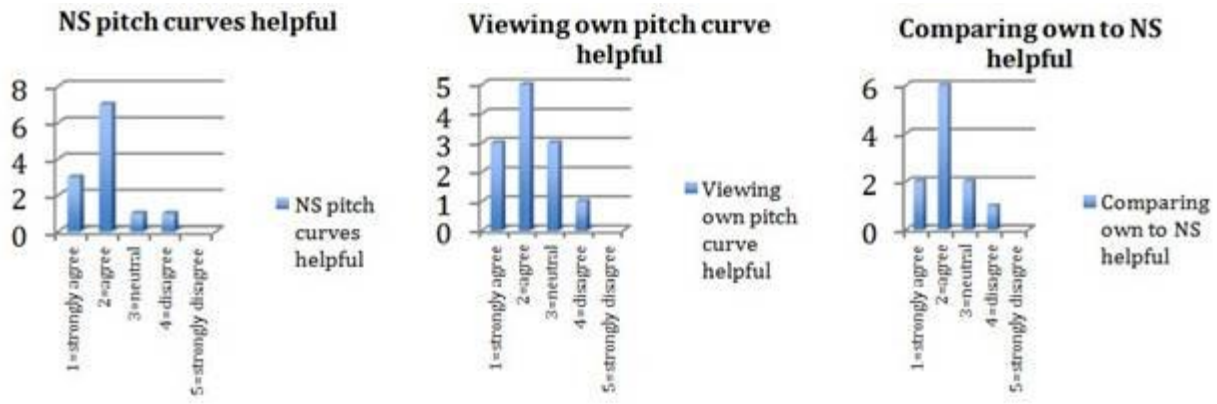


Figure 15. Results of postsurveys of 12 most improved students.

Furthermore, in answering the open-ended question at the end of the postsurvey (Do you have any comments about the project or suggestions for making it more helpful to you?), students made additional positive comments about using Praat to improve their Chinese tone production. For example, one student wrote "I think it was extremely useful to compare tones, not just by ear, but by being able to quantitatively look, see, and improve," and several other participants also further acknowledged the benefit of using this special program by writing "I think that this is a great program. It was very helpful using Praat," and "I thought Praat was a fun learning experience and would like to continue using it." In addition, students suggested making the practice less rushed by having more dedicated class time or using Praat at home, providing more training and instruction, and making the program even more user-friendly.

6. Discussion

6.1. Auditory Analyses

In response to the first question about whether learners' tone production improved when rated auditorily by native listeners, we found that on average, the learners' tones were correct 55% of the time in the pretests, incorrect 12% of the time, and were passable (but not native-like) 33% of the time (with 22% of the tones rated as having problematic pitch height and 11% of the tones rated as having problematic pitch contours).

In the posttests, in line with studies conducted by Miracle (1989), Sun (1998), Shen (1989), Wang et al. (2003), and Wang, Sereno & Jongman (2006), problems with pitch height seemed to negatively impact pronunciation of Tones 0-4 more than pitch shape or contour. In general our findings from the auditory analysis confirm those of past research in that Tone 3 is highly problematic for L2 learners of Mandarin Chinese (Guo & Tao, 2008) and pitch height can be identified as the biggest obstacle in obtaining native like pronunciation of the 5 tones. Many previous studies did not investigate the Neutral Tone, but our study found that learners had as much difficulty pronouncing the Neutral Tone as they did Tone 3.

In addition, of the 45% of the tones that were incorrect in the pretest, Tones 1 and 4 were found to show the most improvement, and Tone 3 and Neutral Tone were most resistant to improvement in the posttest.

6.2. Acoustic Analyses vs. Auditory Analyses

Acoustic data analyses were conducted to serve as a way of triangulating our results and accounting for interrater reliability. Similar to our auditory analyses, the acoustic analyses revealed improvement from pretests to posttests; in addition, problems with pitch height impaired improvement in all five tones. This finding is in line with research conducted by Wang et al. (2003) and Shen (1989) in that pitch height was seen as more problematic when conducting both acoustic and auditory analysis.

In contrast to past research and to our auditory analyses, in our acoustic analyses Tones 3 and 4 were shown to have the most improvement from pretest to posttest: Pitch height improved for Tone 3 and pitch contour for Tone 4. In order to understand this phenomenon more clearly, the high- and low-proficiency groups were examined separately, in order to track their improvement by tone. For the high-proficiency group, their improvements were subtler, as they started out closer to the native speaker norms. Their pitch contours did improve for Tone 3 and Neutral Tone, and their pitch height was closer to that of the native speakers. For the low-proficiency group, the pitch height of Tones 1 and 3 improved in the posttest by moving closer to the initial pitch height of the native speakers, and Tone 4 showed the most improvement in both pitch height and pitch contour.

These findings are important for two reasons: (1) auditory analyses by native speakers may differ from acoustic analyses; (2) auditory analyses often do not provide learners with feedback about what was incorrect about their production, whereas acoustic analyses illustrate directly to the learners whether their pitch levels were too high or too low and whether their pitch contours were similar to or different from the native speakers' contours.

6.3. Learners' Attitudes and Improvement

When reviewing the survey data in relationship to student improvement we found that approximately two-thirds of the participants (66%) found it helpful to view the native speaker pitch curves, their own pitch curves and to compare the pitch curves. This indicates that in general the majority of the students had a positive attitude towards the training using Praat. Of the 12 students who demonstrated the greatest amount of improvement, a somewhat higher percentage (74%) found the weekly practice using Praat helpful for improving their tones.

6.4. Pedagogical Implications and Limitations

The results of this study may provide valuable information for Chinese teachers teaching American L2 students. First, with a well-documented idea of the most challenging and hard-to-improve tones for learners of Chinese, namely, the Neutral Tone and the Tone 3 found in this study, and the tones whose acquisition and production would most likely be facilitated by the use of visualization technology, Chinese language instructors may be able to make better-informed decisions in considering adoption of CALL technology, time allocation for practice of different tones, and focus of instruction. For example, instructors may want to allocate more time to the teaching and practice of Tone 3 and the Neutral Tone rather than allotting equal numbers of exercises to each of the 5 tones. Second, teaching and learning suprasegmental features like tones can sometimes be difficult due to the limitation of instructors' verbal explanations, which might be vague and confusing, and to students' lack of perception of the difference between their tonal production and that of native speakers. Allowing students to view their tone production and compare it to the native speakers provides direct and concrete information about the differences, specifically whether the learners' difficulties lie with nonnative-like pitch height or pitch contours (or both). Finally, our study reveals that two-thirds of the learners found Praat with its visualization feature to be helpful.

Moreover, "effective CALL activities implemented outside class to improve L2 pronunciation will help the instructors to preserve precious instruction time for other tasks" (Wang, 2008, p. 271). If learners can be taught to use tools with visualization features, those who find the tool helpful could create pitch curves on their own and compare them to those of native speakers. In addition, CALL tools can provide new possibilities for more accurate and objective evaluation of language production, which has long been regarded as challenging and problematic.

It is critical, though, not to assume that simply showing students how visualization tools work will lead them to use the tools effectively. Teachers must familiarize themselves both with technological tools as well as the research on these tools so that they can emphasize the benefits of these tools to learners. Teachers must first explain to learners precisely what they must pay attention to when comparing their pitch curves with the native speakers' pitch curves, i.e., they must point out that pitch height and pitch contour are *both* key features of Mandarin tones.

In terms of the study's limitations, although the overall improvement from pretest to posttest does provide some evidence that visualizations may be helpful, especially when combined with the qualitative data of students' perceptions, it is important to note that the study was conducted in an authentic learning environment with intact classes. As such, in order not to disadvantage any learners by not providing them with training, an experimental methodology with a control group was not used. While this may limit any conclusions regarding causality, it is becoming increasingly common in CALL research not to simply compare conditions with technology vs. those without technology, but rather to try to understand which specific features or characteristics of technology can be helpful.

Furthermore, it should be acknowledged that although the native speaker models in this study represented standard pronunciation, there is great dialectal and geographic variation in all languages.

A final limitation is that the present study focuses on the level of isolated disyllabic word production. Tonal production is much more complicated beyond the word level. Even if learners' pitch contour or pitch height is close to the that of the native speakers at the word level, it does not guarantee that they will be able to produce accurate tones at the sentence level.

7. Conclusions

The current study investigated the use of tools for visualization of pitch curves for learning and improving the production of Mandarin Chinese tones by 35 L2 learners, most of whom were native English speakers.

The results of both auditory and acoustic assessments revealed that Tone 3 and the

Neutral Tone were the most difficult, supporting previous studies, while Tone 4 was most often correct. In addition, the mispronunciation of tones lies more in pitch height than in pitch shape or contour. The most important findings are (1) learners in this study showed improvement between pretest and posttest, with variations depending on the specific tone; (2) both the auditory and acoustic analyses have corroborated that learners have the greatest difficulty with certain tones and that specific types of improvements were made for the different tones; (3) when the production data is triangulated with learner perception surveys about the usefulness of the visualizations for improving tone production, the learners who made the greatest improvement were more likely to consider the visualizations helpful.

The new findings of this study that may provide insights into the acquisition and teaching of the five Chinese tones and how technology can aid in pedagogical practices are as follows: First, a comparison of the native speaker raters' auditory analyses of L2 learners' individual tones with the acoustic analyses of the tones revealed that the auditory analyses of where the "problem" lay (with pitch height or pitch contour) did ***not*** always correspond to the acoustic analyses of all 5 tones. This implies that instructors may not always be able to give learners precise feedback about why their tones are problematic or incorrect. However, if learners can see visualizations of the pitch characteristics of their tones, both in terms of pitch height and pitch contour, they then have specific and graphic information about how their tone pronunciation compares with that of native speakers. Second, the combination of auditory and acoustic analyses of the individual tones provide specific new information about which tones are most difficult for learners and the characteristics of the difficulties. This will aid instructors in emphasizing the teaching and practice of the most problematic tones. Third, this study employed learner-created tone visualizations, and did not require instructors to create their own software or analyze huge numbers of learners' pronunciation files. With proper training and explanations for why these technological tools are useful, learners can be encouraged to use these freely available technological tools on their own outside of class.

Future studies should go beyond the word level and be expanded to the sentence and discourse level (see Levis & Pickering, 2004). Since the teacher-talk examples of words and phrases used in this study may not match naturalistic speech, future research should teach learners to use words and phrases (and ultimately sentences) in different contexts (with the caveat that naturalistic speech may exhibit less exaggerated tone heights and contours). A related issue is that there is significant variation among individual native speakers in terms of how far they might deviate from the norm or from neutral speech. Speakers may sound more or less dynamic in their speech, and these individual differences must be taken into account as well. Finally, on the technical side, in response to students' desire for more user-friendly software, we are currently developing a mobile app that records a learner and automatically superimposes the learner's pitch curve on the same screen as the native speaker's curve. Future studies of this type of immediate, direct comparison are planned.

Acknowledgements

We gratefully acknowledge the helpful suggestions of two anonymous reviewers, the support of an Instructional Improvement Grant from our university, and the assistance of Dr. George Michaels, Shu-Chuan Bella Chen, Yiping Chen, Xiao Hu, Jingyu Huo, Xiang Qiu, Qian Wang, Feng Xie, Janet Zhang, Yu Zhang, and all of the students who participated in the study.

References

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.77, retrieved 18 May 2014 from http://www.praat.org/

Chan, M. K. M. (2003). The digital age and speech technology. *Journal of the Chinese Language Teachers Association, 38*(2), 49-86.

Chun, D. M., Jiang, Y., & Ávila, N. (2013). Visualization of tone for learning Mandarin Chinese. In J. Levis & K. LeVelle (Eds.). Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference. Aug. 2012. (pp. 77-89). Ames, IA: Iowa State University.

Guo, L., & Tao, L. (2008). Tone production in Mandarin Chinese by American students: A case study. In M. K. M. Chan & H. Kang (Eds.) *Proceedings of the 20 North American Conference on Chinese Linguistics,* (pp. 123-138). Columbus, OH: The Ohio State University.

Hao, Y-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and nontonal language speakers. *Journal of Phonetics, 40*, 269-279.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology, 8*, 34-52.

He, Y., & Wayland, R. (2010). The production of Mandarin coarticulated tones by inexperienced and experienced English speakers of Mandarin. *Speech Prosody 2010 Conference Proceedings* 100123:1-4.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System 32*(4), 505-524.

Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers' Association, 24*, 49-65.

Molholt, G., & Hwu, F. (2008). Visualization of speech patterns for language learning. In V. M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning* (pp. 91-122). NY: Routledge.

Shen, X-N. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers' Association, 24*(3), 27-47.

Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of Standard Mandarin Chinese*. Honolulu, HI: Second Language Teaching and Curriculum Center.

Wang, X. (2008). Training for learning Mandarin tones. In F. Zhang & B. Barker (Eds.), *Computer enhanced language acquisition and learning* (pp. 259-273). Hershey, PA: IGI Global.

Wang, X. (2012). Auditory and visual training on Mandarin tones: A pilot study on phrases and sentences. *International Journal of Computer-Assisted Language Learning and Teaching, 2*(2), 16-29.

Wang, Y., Jongman, A., & Sereno, J. A. (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.

Wang, Y., Sereno, J., and Jongman, A. (2006). Second language acquisition and processing of Mandarin tone. In Li, P., Tan, L.H., Bates, E., and Tzeng, O.J.L. (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. 1: Chinese). Cambridge, UK: Cambridge University Press