# THE USE OF VISUAL FEEDBACK TO TRAIN L2 LEXICAL TONE: EVIDENCE FROM MANDARIN PHONETIC ACQUISITION

**Alexis Zhou,** Purdue University
**Daniel J. Olson,** Purdue University

Acquisition of L2 lexical tone has proven to be difficult for L1 speakers of non-tonal languages, resulting in potential issues for intelligibility, comprehensibility, and accentedness. Recent studies have suggested that visual feedback is a useful method for tone training. However, the potential for generalizability of improvement from the word level to the phrasal level has yet to be systematically investigated. This study explores the use of visual feedback to teach tone to L1 English learners of Mandarin. Four L1 English–L2 Mandarin beginning-level learners participated in a visual feedback paradigm with a pretest, intervention, posttest design. Stimuli included disyllabic words in isolation and embedded in phrases. Results suggest improvement in tone production following the visual feedback paradigm for words in isolation, although with different outcomes for different tones. Furthermore, there was little suggested generalizability to the phrasal level, potentially owing to greater crosslinguistic interference from L1 phrasal level intonation.

**Cite as:** Zhou, A., & Olson, D. J. (2023). The use of visual feedback to train l2 lexical tone: Evidence from Mandarin phonetic acquisition. In R. I. Thomson, T. M. Derwing, J. M. Levis, & K. Hiebert (Eds.), *Proceedings of the 13th Pronunciation in Second Language Learning and Teaching Conference,* held June 2022 at Brock University, St. Catharines, ON.

## INTRODUCTION

Tonal languages (e.g., Mandarin) use lexical tone, which can be described as the systematic manipulation of fundamental frequency (f0) at the syllable level (Singh & Fu, 2016; Yip, 2002), to create phonemic contrasts. Native speakers of non-tonal first languages (L1), like English, sometimes experience difficulties acquiring lexical tone in a second language (L2) (Chen, 1974), as they are not familiar with the f0 characteristics of tones (e.g., height, contour). Difficulties in tone acquisition can result in issues in comprehensibility, intelligibility, and accentedness (for a discussion of these terms see Munro & Derwing, 1995). Moreover, as non-tonal L1s like English employ prosodic features, such as intonation at the phrasal level, production of phrases is a likely area where cross-linguistic prosodic transfer will impact L2 lexical tone production (Chen, 1997). The current study focuses on one method for training L2 Mandarin lexical tone, specifically visual feedback, and examines its efficacy for both isolated words and words in phrases.

### Visual Feedback for L2 Lexical Tone Training

As acquisition of L2 lexical tone has proven particularly challenging for learners at both the syllable and phrasal levels, several instructional methods have been used. These methods include more traditional approaches that focus on matching each tone with a physical representation, such as a gesture (Morett & Chang, 2015) or color (Dummit, 2008). The use of a visual feedback (VF) paradigm to train L2 tone has also been growing in popularity (Chen, 2022; Chun et al., 2013; Chun et al., 2015; Wang, 2008, 2012). This method encourages learners to become aware of differences between their own productions and those of a native speaker, as described by Schmidt's (1995) 'noticing hypothesis'. VF uses visualizations of speech (i.e., acoustically), articulatory movements, or both, for training features of L2 pronunciation, often using both learners' and native speakers' speech (Offerman & Olson, in press). While research has shown the efficacy of visual feedback in training both segmental features (Chun, 2007; Offerman & Olson, 2016) and phrasal level prosody in non-tone languages (Jiang & Chun, 2021), significantly less research has addressed lexical tone.

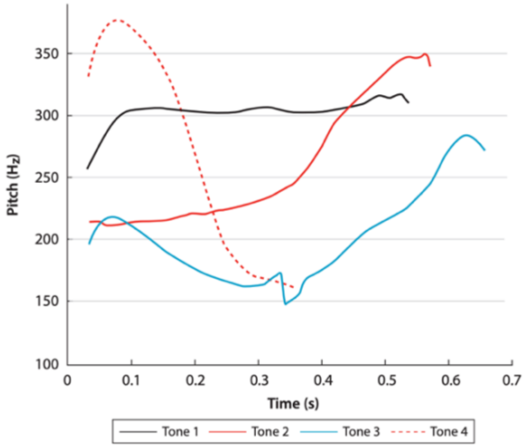
Most previous studies on VF and lexical tone have focused on word-level production. At the monosyllabic word level, Wang (2008) found improvement in the perception and production of Mandarin tones by learners with different L1 backgrounds after VF, specifically in comparisons of participants' pitch contours with those of native speakers. Success has also been found at the disyllabic word level (Chen, 2022; Chun et al., 2015). Chun et al. (2015), using a paradigm similar to Wang (2008), found significant improvement in tone production, albeit with varying levels of improvement for different tones (see also Chen, 2022).

Moving beyond the word level, two pilot studies have explored VF training of L2 lexical tone at the phrasal level (Chun et al., 2013; Wang, 2012). Although these studies suggest that VF may be useful for L2 tone training, results at the word and phrasal level were collapsed. The generalizability of VF training for segmental features from the word to phrasal level has been shown in previous research (e.g., Offerman and Olson, 2016). However, the potential for generalization of gains from the word to phrasal level has yet to be fully explored. A systematic analysis differentiating the effects of VF training at the word and phrasal levels for L2 lexical tone training is still needed.

## Mandarin Tone

Standard Mandarin has four tones, a high-level pitch (Tone 1), a high-rising pitch (Tone 2), a low-dipping pitch (Tone 3), and a high-falling pitch (Tone 4) (Chao, 1948). Mandarin also has what is called a Neutral Tone (also described as allophonic variation), which changes pitch depending on the previous tone (Zhang, 2017). English-speaking students are typically exposed to tones through *pinyin*, the standard romanization system that uses diacritics (e.g., mā, má, mǎ, mà) or numbers (e.g., ma1, ma2, ma3, ma4) to label the tones. These representations can be misleading, because the diacritics do not reflect the acoustic productions of these tones (Figure 1; Chun et al., 2015). Considering the interaction between lexical tone and utterance-level intonation, while there is some debate about how tones are produced at the phrasal level, most researchers agree that tones keep their tonal value in multi-word utterances, with adjustments to the top and bottom limits of the pitch range (Lin & Wang, 1992).

**Figure 1.**
*Acoustic representations of the Mandarin tones. From Chun et al. (2015: 87).*



*© John Benjamins. Reprinted with permission.*

Prior research has shown some variability in the ease or outcomes of acquisition across the different tones by non-tonal L1 speakers. Broadly, recent literature has shown that Tone 1 and Tone 4 may be easier to acquire than Tone 2 and Tone 3 (Tao & Guo, 2008). Tone 3 is typically reported as the most difficult tone to acquire, as this tone has the highest allophonic variation and goes through Tone 3 sandhi (i.e., when the distinction between Tone 2 and Tone 3 is neutralized in the Tone 3–Tone 3 context (Hao, 2012)).

## Research Questions

The current study addressed the following research questions:

1. Does visual feedback serve to improve the production of L2 lexical tone, defined as both contour shape and f0 height, by non-tonal L1 speakers?
2. Do improvements in L2 lexical tone production following visual feedback at the word level generalize to improvement in L2 tone production at the phrasal level?

To address these questions, the current study used a VF paradigm to train lexical tone at the disyllabic word level to L1 English–L2 Mandarin learners. Based on previous findings (Chun et al., 2015; Wang, 2008), improvement was expected at the word level following a VF paradigm, although with some variability in outcomes by tone (e.g., greater improvement for Tone 1 and Tone 4 followed by Tone 2 and Tone 3; Chun et al., 2015; Tao & Guo, 2008). It was expected there would also be improvement at the phrasal level, with the same overall improvement pattern by tones.

# METHODS

## Participants

A total of 4 (3 female, 1 male) L1 English–L2 Mandarin learners ($M_{age}$ = 21.5, *SD* = 2.18) in their second semester of Chinese study at a large American public university participated in this pilot study. One participant reported speaking two other non-tonal languages with family (Marathi and Hindi) but was retained for final analysis. The course that participants were recruited from was designed to train beginning listening, speaking, reading, and writing skills in Mandarin. Standard Mandarin speakers ($M_{age}$ = 25, *SD* = 1) were also recruited to record the stimuli to be used in the analysis (*n* = 2, male).

## Materials

Stimuli for the current study consisted of 60 unique disyllabic Mandarin words divided into three sets; one set was produced as isolated words (20 repeated at the pretest and posttest), and two sets were produced as words embedded at the end of novel phrases (20 at the pretest, 20 at the posttest). Stimuli were presented to participants with Chinese characters, romanizations, and the words' English translations. All words were controlled for tone, resulting in each possible combination of the four lexical tones plus the Neutral Tone (e.g., T2–T1, T2–T2) each being represented by three different target words (see Table 1). Due to the highly variable nature of the Neutral Tone, it was excluded from analysis.

**Table 1**
*Example Stimuli for Tone 1*

| Combination | Characters | Romanization | Translation |
|---|---|---|---|
| T1–T0 | 心思 | xīnsi | thoughts |
| T1–T1 | 开工 | kāigōng | go into operation |
| T1–T2 | 天文 | tiānwén | astronomy |
| T1–T3 | 经理 | jīnglǐ | manager |
| T1–T4 | 医院 | yīyuàn | hospital |

Stimuli were also controlled for familiarity through the use of a study completed by a separate group of second semester Mandarin learners (*n* = 8). Those learners demonstrated broad familiarity with the meaning and pronunciation of each syllable of the disyllabic compound words, with each character scoring 5 and above on a 1–7 Likert scale, adapted from Auer et al. (2000). On the scale, 1 meant "I've never seen this character and I don't know its pronunciation" and 7 meant "I know this character and am confident of its pronunciation". Example 1, below, shows a sample phrasal level stimulus.

(1) Characters: 我喝了一杯咖啡。
Romanization: Wǒ hēle yībēi kāfēi.
Translation: I drank one cup of coffee.

## Procedures

A pretest, training (four sessions, one per tone), posttest design was implemented in a Mandarin classroom during regular class time. Following Chun et al. (2015), each training session contained three parts, a pre-recording which occurred before class at home (approx. 10 min.), visual inspection and comparison, which occurred during class with the experimenter (approx. 10 min.), and a re-recording, which occurred after class at home (approx. 10 min.). The pretest, posttest, and four training sessions took place over a 7-week period (approx. two hours total).

For the pre-recording portion of training (one target tone per session), participants recorded 15 words in isolation and 15 phrases (distinct from pretest/posttest stimuli) in Praat (Boersma & Weenink, 2022) using their own personal computers, for a total of six productions of each tone combination. Participants then created stylized pitch contours of five disyllabic words in isolation, one for each possible tone combination for the tone under study.

Participants were instructed on how to view their pitch contours (in the pre-recording worksheets) by using the "Draw visible pitch contour" function in Praat, giving them visualizations of their shapes and relative heights of their tones without the exact pitch values. Using a guided inductive approach, participants answered a series of questions about their own productions and how their productions compared to those of native Mandarin speakers. Questions focused on both contour shape and relative f0 heights (see Example 2).

(2) How would you describe the first line that corresponds to the tone of your production of “知”? Describe if it is rising, falling, or flat. Also describe its relative height.

In the re-recording portion of the training, participants re-recorded the list of words and phrases from the pre-recording, using their own personal computers. The posttest recording was conducted in week 7.

### Data Analysis

The acoustic analysis paralleled that of Chun et al. (2015) and Wang et al. (2003). The disyllabic stimuli for the native speakers’ and participants’ productions were isolated using Praat (Boersma & Weenink, 2022) and input into Matlab (v.R2022a). F0 listings (Hz) were extracted at 10 millisecond intervals for each syllable. Contours containing more than 20 f0 measurements were down-sampled to 20 data points. Productions returning less than 20 f0 measurements were used if they met the following criteria: They contained at least 10 points and had existing points at the beginning and end of the syllable. Duration was then normalized to start from 0 (start) to 1 (end). F0 standardization was performed to control for the different genders, pitch ranges, and speaking rates of participants, along with syllable contexts, by converting the f0 values to their logarithms using the formula below (Rose, 1987), where H and L represent the maximum and minimum f0 for a given speaker and X represents each individual f0 measurement in a contour. The resulting T value can range from 0 to 5, corresponding to a 5-point pitch scale (Chao, 1948).

$$T = 5 \times \frac{\log(X) - \log(L)}{\log(H) - \log\,(L)}$$

The ‘native norm’ (see Wang et al., 2003) was calculated using the data extracted from the two native speakers’ productions, specifically by averaging the productions across each tone using the calculated T values.

Statistical analysis followed Wang et al. (2003). Overall deviation scores (i.e., average distance between native T-scores and participant produced T-scores across all f0 points in a contour) were calculated for the pretest and posttest for each tone (scores closer to 0 represent more native-like production).

For L2 learners, a total of 1610 syllables were included in the final analysis (812 in words, 798 in phrases), with ~7% eliminated due to missing or noisy data (40 stimuli × 2 syllables × 2 sessions (pretest vs. posttest) × 3 repetitions of each stimulus × 4 participants – Neutral Tone = 1728). For native speakers, a total of 595 syllables (204 in words, 391 in phrases) were included in the final analysis, with ~8% eliminated (60 stimuli × 3 repetitions × 2 participants – Neutral Tone = 648).

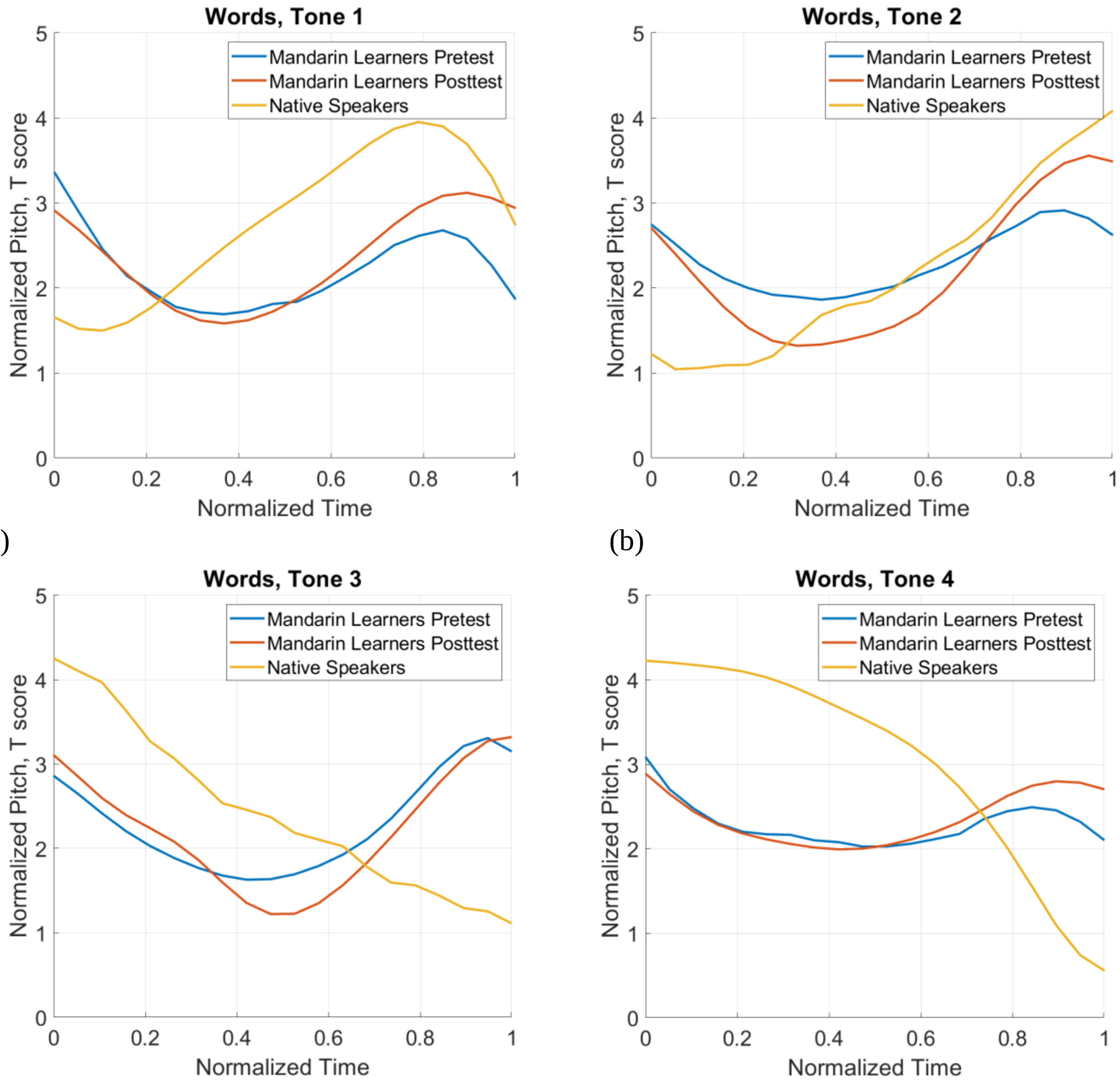
## RESULTS

### Words in Isolation

The mean pretest and posttest productions of the participants compared to the native norm for the words in isolation (for each tone) can be seen below in Figure 2 (a)–(d).

**Figure 2.**

*Participant pitch contour averages compared to native norm for words in isolation*



(a)

(b)

(c)

(d)

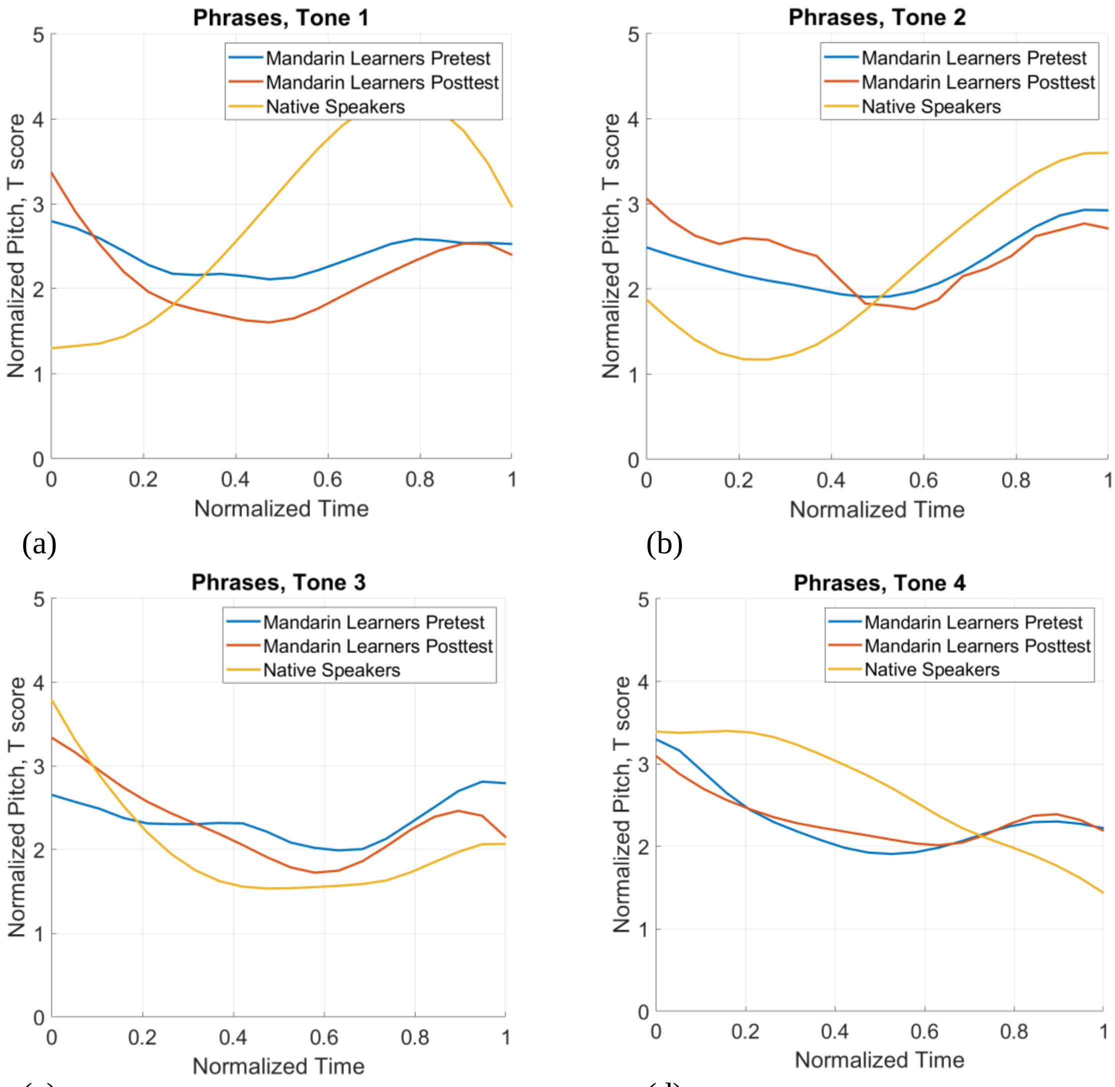
Deviation scores for Tone 1 suggest the posttest productions (deviation score: -0.42) were more native-like than pretest productions (deviation score: -0.55). Analysis of Figure 2a shows an improvement in contour shape, specifically in pitch height in the second half of the contour from pretest to posttest. For Tone 2, analysis of deviation scores suggests that the posttest productions (deviation score: 0.02) were more native-like than pretest productions (deviation score: 0.14). Analysis of Figure 2b shows an improvement in the contour shape (lower dipping in the middle, higher rise at the end), with notable improvement in pitch height in the second half of the contour from pretest to posttest. For Tone 3, the pretest (deviation score: -0.14) and posttest (deviation score: -0.22) contours had similar shapes. Deviations from the native norm were made in both contour shape and height for both the pretest and the posttest. The pretest (deviation score: -0.73) and posttest (deviation score: -0.65) contours were also similar for Tone 4. Deviations from the native norm were made in both contour shape and height, as the participants' contours failed to show the final fall produced in the native-speaker norms.

## Words in Phrases

The mean pretest and posttest productions of the participants compared to the native norm for the words embedded in phrases (for each tone) can be seen in Figure 3 (a)–(d), below.

**Figure 3.**

*Participant pitch contour averages compared to native norm for words in phrases*



(a)

(b)

(c)

(d)

Analysis of deviation scores for Tone 1 suggests that productions at the posttest (deviation score: -0.64), were less native-like than at the pretest (deviation score: -0.48). Analysis of the time series graph in Figure 3a shows that the posttest contours dipped noticeably lower than the native norm, and that the starting pitch height was higher than the native norm and the pretest. For Tone 2, analysis of Figure 3b shows that the pretest (deviation score: 0.11) and posttest (deviation score: 0.19) contours started higher and ended lower than the native norm, suggesting that the pretest and posttest are similar. Analysis of deviation scores for Tone 3 suggest that posttest productions (deviation score: 0.29) may have been more native-like than pretest productions (deviation score: 0.33). Analysis of Figure 3c shows that the posttest productions appear closer to the native norm in pitch height and were closer to the native speakers in overall pitch contour. For Tone 4, analysis of Figure 3d shows similar deviations from the native norm in the pretest (deviation score: -0.38) and posttest (deviation score: -0.28), specifically that the pitch height was noticeably higher at the end of the contour, and that the overall shapes did not match the native norm.

## DISCUSSION

Responding to the first research question, namely what impact visual feedback training has on lexical tone production for words in isolation, the results suggest improvement for two of the four tones examined. Based on the qualitative analysis of the figures and deviation scores, there was possible improvement from pretest to posttest when compared to the native speakers for Tone 1 and Tone 2, but not for Tone 3 and Tone 4. This suggests partial support for the original predictions, as it was expected that Tone 1 would show a high level of improvement, and there were suggested improvements in two out of four tones. However, unexpectedly the deviation scores from Tone 3 and Tone 4 did not suggest any improvement from pretest to posttest, although Tone 3 was expected to show less improvement based on its reported difficulty (e.g., Tao & Guo, 2008). Using a visual feedback paradigm to train tone at the word level may give students the opportunity to see their deviations from the native speakers' productions, which is the goal of the visual feedback method.

Two specific findings should be noted. First, these results, coupled with previous findings in the literature (Chen, 2022; Chun et al., 2015; Wang, 2008), suggest that visual feedback may be an effective method of training native speakers of non-tonal languages to perceive and produce L2 lexical tone. Second, it should be noted that, as originally hypothesized, the suggested improvement resulting from visual feedback varied across the individual tones. Specifically, the predicted improvement was only found in Tones 1 and 2. Given the relatively small data set in this pilot study, it remains to be seen whether improvement in productions of Tone 3 and 4 lag behind Tones 1 and 2, or if visual feedback is not effective for certain tones.

As previous research has shown a degree of generalizability of visual feedback training from words in isolation to words in connected speech at the segmental level (e.g., Offerman & Olson, 2016), the second research question addressed whether similar generalizability may be found for lexical tone. Results in the current study showed little support for generalizability of gains from the word to the phrasal level, with analysis of the figures and deviation scores only suggesting improvement for Tone 3. It is worth noting that generalizability for lexical tone may be fundamentally different than generalizability for segmental features, as embedding words in connected speech provides greater opportunity for cross-linguistic transfer of prosodic features from L1 intonation. In short, it is plausible that at the phrasal level, L1 English intonation negatively impacted the participants' tone productions.

Overall, qualitative analysis of the time series graphs and comparison of deviation scores for tones at the word level show that visual feedback may be a useful method training L2 lexical tone. However, the analyses of tones at the phrasal level may indicate that visual feedback training is not necessarily generalizable from word to phrasal levels for all L2 features. Further investigation into the generalizability of other suprasegmental features is warranted. From a pedagogical perspective, the results at the phrasal level show the potential limitations of visual feedback training. Future studies should explore additional training methods that might be used in conjunction with visual feedback for training at the phrasal level.

## ACKNOWLEDGMENTS

We would like to thank the Chinese Language Program at Purdue University for allowing us to run the current study in the second semester Chinese course.

## ABOUT THE AUTHORS

Alexis Zhou is a PhD student in the Department of Linguistics at Purdue University. Her research interests include second language acquisition, second language speech perception/production, and Chinese linguistics.
Purdue University
Department of Linguistics
100 N. University St.
West Lafayette, IN 47907
765-494-7161
atews@purdue.edu

Dr. Daniel J. Olson is Associate Professor of Spanish and Linguistics at Purdue University. His research focuses on phonetics and psycholinguistics in bilingual populations, particularly the acquisition of second language phonetics.
Purdue University
School of Languages and Cultures
640 Oval Dr.
West Lafayette, IN 47907
765-494-3828
danielolson@purdue.edu

## REFERENCES

Auer, E. T., Bernstein, L. E., & Tucker, P. E. (2000). Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory & Cognition, 28*(5), 789–797.

Boersma, P., & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.23.

Chao, Y. R. (1948). *Mandarin primer*. Harvard: Harvard University Press.

Chen, G.T. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics 2*(2), 159–171.

Chen, M. (2022). Computer-aided feedback on the pronunciation of Mandarin Chinese tones: Using Praat to promote multimedia foreign language learning. *Computer Assisted Language Learning, Advance online access,* 1–26.

Chen, Q. (1997). Toward a sequential approach for tonal error analysis. *Journal of the Chinese Language Teachers Association, 32,* 21–39.

Chun, D. M. (2007). Come ride the wave: But where is it taking us? *CALICO Journal*, *24*(2), 239–252.

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

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

Dummit, N. (2008). *Chinese through tone & color*. New York: Hippocrene Books.

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279.

Jiang, Y., & Chun, D. M. (2021). Web-based intonation training helps improve ESL and EFL Chinese students' oral speech. *Computer Assisted Language Learning, Advance online access*, 1–29.

MATLAB. (2022). *Version 2022a*. Natick, Massachusetts: The MathWorks Inc.

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*(3), 347–353.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, *59*, 45–60.

Offerman, H. M., & Olson, D. J. (In press). Speech visualization for pronunciation instruction: Exploring instructor support in L2 learner attitudes towards visual feedback. In S. McCrocklin (Ed.). *Technological resources for second language pronunciation learning and teaching: Research-based approaches* (pp. 1–33). Lexington Books.

Rose, P. (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communication, 6*, 343–351.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R.W. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Singh, L., & Fu, C. S. L. (2016). A new view of language development: The acquisition of lexical tone. *Child Development*, *87*(3), 834–854.

Tao, L., & L. Guo (2008). Learning Chinese tones: a developmental account. *Journal of the Chinese Language Teachers Association*, *43*(2), 17–46.

Wang, X. (2008). Training for learning Mandarin tones. In F. Zhang & B. Bakers (Eds.), *Handbook of research on computer-enhanced language acquisition and learning* (pp. 259–273). 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. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043.

Yip, M. (2002). *Tone.* Cambridge: Cambridge University Press.

Zhang, H. (2017). The effect of theoretical assumptions on pedagogical methods: a case study of second language Chinese tones. *International Journal of Applied Linguistics*, *27*(2), 363–382.