Linking Production of Mandarin Tonal Contrasts with Musicality in Adult Learners

Chen Gao[1] (cgao1@gradcenter.cuny.edu)
C. Donnan Gravelle[2,3] (cgravelle@gradcenter.cuny.edu)
Shan Jiang[4] (sjiang@gradcenter.cuny.edu)
Patricia J. Brooks[1,2,3,4] (patricia.brooks@csi.cuny.edu)

[1]PhD Program in Educational Psychology, CUNY Graduate Center, New York, NY USA
[2]PhD Program in Psychology, CUNY Graduate Center, New York, NY USA
[3]Department of Psychology, College of Staten Island, CUNY, Staten Island, NY USA
[4]PhD Program in Speech-Language-Hearing Sciences, CUNY Graduate Center, New York, NY USA

Abstract

Mandarin Chinese is a tonal language where variations in voice pitch distinguish word meaning. Acquiring tonal contrasts presents challenges for adult second-language learners. Undergraduates (N = 59) completed a computer-assisted language learning protocol, where they engaged in listening and repeating Chinese disyllabic nouns and matching them with corresponding pictures. Tone production accuracy was assessed at pretest/posttest using complexity invariant distance (CID), a quantitative metric of the distance between time series (learner vs. native-speaker productions). Word-level analyses found lower CID scores at posttest, indicating improvements in tone production after three blocks of word-picture matching. Fine-grained syllable-level analyses showed lower CID scores for first vs. second syllables, suggesting a primacy advantage. Accuracy on the Music Ear Test predicted lower CID scores, linking musicality with aptitude in learning tonal contrasts. No effects of nonverbal intelligence or language background were found. CID offers a robust method of assessing tone production accuracy for future studies.

Keywords: tone production and perception; L2 acquisition; Mandarin Chinese; musicality; complexity invariant distance measurement; individual differences

Introduction

Mandarin Chinese, a tonal language, uses pitch variations to distinguish the meanings of words that are otherwise phonetically identical (Bluhme & Burr, 1971). These pitch distinctions are fundamental to communication, making accurate perception and production of tones crucial for language proficiency. Mastering tonal contrasts presents a unique and significant challenge for adult second-language (L2) learners, especially those whose native languages are non-tonal (e.g., English). Unlike non-tonal languages, where meaning is conveyed primarily through segmental sounds, Mandarin's reliance on pitch adds a layer of complexity for learners to navigate (Kiriloff, 1969; Shen, 1989).

The difficulties that adult L2 learners face in acquiring Mandarin tones are well-documented (Kiriloff, 1969; Pelzl, 2019). Perceiving and producing the five tones of Mandarin (high-level, rising, low-dipping, falling, and neutral) requires reliance on fine-tuned auditory discrimination and control over vocal pitch modulation. This complexity is compounded by the fact that tonal contrasts are not only segmental but extend over the syllable, potentially increasing cognitive load (Shen, 1989). While targeted training in tone identification can improve performance, adult L2 learners generally fail to achieve native-like accuracy in tone perception and production (Pelzl, 2019).

Disyllabic words comprise about 40% of word tokens and 60% of word types in the Mandarin Chinese Conversational Corpus (Liu et al., 2016). Yet, most studies examining L2 tone perception and production have focused on monosyllables, which are markedly easier to process than disyllables (Chang & Bowles, 2015; Hao, 2018). Mandarin learners often struggle to identify tones in disyllabic words, where tonal register and contour shift based on the adjacent tones (Xu, 1997). Since coarticulation causes Mandarin tones in disyllabic words to differ considerably from monosyllables, studying tone acquisition using disyllabic words provides a better representation of the challenges associated with learning the language.

Moreover, syllable position within disyllabic words may also influence tone recognition and production (Chang & Bowles, 2015). Liu et al. (2011) found that college students learning Mandarin as a new language exhibited higher accuracy in recognizing tones in second syllables compared to first syllables of disyllabic words. Similarly, Hao (2018) observed better tone perception for second syllables among intermediate learners. Yu (2021) found that native Mandarin speakers preferred to assign stress to final syllables of words, reflecting Mandarin's prosodic characteristics where final syllables often carry a greater prominence. While tone recognition appears to benefit from the prosodic prominence of final syllables, how syllable position affects tone production is less straightforward. Hao (2012, 2018) found no overall effect of position but noted that certain tones were produced more accurately in word-final position.

Non-linguistic Influences

Previous studies of individual differences in L2 learning report positive associations between measures of musicality and accuracy in perceiving tonal contrasts in unfamiliar languages like Mandarin (Christiner et al., 2022; Delogu et

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

al., 2006; Han et al., 2019), Thai (Götz et al., 2023) and Norwegian (Kempe et al., 2012, 2015). This suggests that individuals with musical training may have a higher aptitude for learning tonal contrasts. The association between musicality and tone perception also raises the possibility of leveraging music (i.e., melodic intonation) to improve learning of tonal languages (Howe & Baumgartner, 2024).

In addition to musicality, nonverbal intelligence, assessed via visual-spatial pattern completion tasks, has been linked with individual differences in adult language learning outcomes (Brooks & Kempe, 2013; Rose et al., 2023) and is considered a general indicator of aptitude (Grigorenko et al., 2000). Consequently, we included measures of nonverbal intelligence and musicality in our study.

Complexity Invariant Distance (CID) Metric

Accuracy of tone production is difficult to measure. One often-used strategy is to have native speakers rate participants' tone productions (He et al., 2016; Li & DeKeyser, 2017; Weiner et al., 2020). This approach relies on subjective perception, and introduces a great deal of information loss, as responses may be scored as "correct" or "incorrect" without allowing for graded judgment. Additionally, this approach lacks scalability and requires the number of trials to be low or the data processing timeline extended, for scoring to be manageable.

Another approach is to compare the pitch curves (i.e., fundamental frequency, f0) of the learner's word productions with those of a native speaker and compute the distance between the two curves as a deviation score (Wang et al., 2003). While this approach provides a more objective measure of production accuracy, the choice of a distance metric for computing deviance scores is non-trivial. As a non-linear time series, the choice of distance metric must take into consideration the complexity of the curves. To compute an objective measurement of tone production accuracy, the current study employs complexity invariant distance (CID) as a metric (Batista et al., 2014). CID provides a more nuanced comparison of learners' tonal productions with those of native speakers that adjusts for the complexity of pitch contours (Zhou & Olson, 2023).

Current Study

This study builds on prior work on individual differences in Mandarin tone acquisition (Christiner et al., 2022; Delogu et al., 2006; Han et al., 2019), with a focus on musicality and nonverbal intelligence as indicators of learning aptitude. Using a computer-assisted language learning (CALL) protocol, we assessed accuracy in vocabulary comprehension and tone production after one hour of practice in listening and repeating Mandarin disyllabic nouns and matching them with corresponding pictures. Following Lee and Kalyuga (2011), we also varied the presence/absence of captioning during training blocks to test whether captions might enhance learning. This manipulation had no effect on learning outcomes and is not discussed at length in this report due to space constraints.

Using CID as an objective, quantitative metric, we examined whether accuracy in tone production improved from pretest to posttest and whether it varied by syllable position. Based on Hao (2018) and Liu et al. (2011), we hypothesized that learners would show higher accuracy on second syllables than initial syllables of disyllabic words. We also predicted that adult learners with higher musical aptitude and nonverbal intelligence would have more success in learning Mandarin tonal contrasts.

Method

Participants

College students (N = 59, 39 women, 20 men, M_{age} = 20.9, SD_{age} = 3.1, *range* 18–33) were recruited via the SONA research participation system at a minority-serving public college in the Northeastern United States. Race/ethnicity was reported as follows: 39.0% White, 33.3% Black, 25.4% Hispanic/Latinx, and 5.1% Asian (not mutually exclusive).

After obtaining informed consent, we administered a language background questionnaire (adapted from Rose et al., 2023) on Qualtrics to determine eligibility. Participants were asked to provide information about their native language(s) and other languages studied at school, spoken at home, or used abroad. We excluded potential participants with prior knowledge of Chinese. On average, participants reported knowledge of 2.6 languages, including English (*SD* = 0.9, *range* 1–5). Three participants reported knowledge of Tagalog, a syllable-timed language with a pitch accent. No other participants reported prior experience with a tonal language. The performance of the three Tagalog speakers was comparable to that of other participants. Language background will not be discussed further here as preliminary analyses showed no relation to learning outcomes.

Materials

Chinese Vocabulary. Stimuli were 36 disyllabic Mandarin words comprising 18 pairs of segmental homophones, i.e., words with identical phonemes but differing in tonal patterns (e.g., yan3jing1 [眼睛, eyes] vs. yan3jing4 [眼镜, eyeglasses]), and corresponding pictures; see Table 1 for examples. Segmental homophones varied in tone on the first syllable (7 pairs), second syllable (3 pairs), or both (8 pairs). One additional word pair (ma1 [妈, mother], ma3 [马, horse]) was used to explain task procedures. Words were recorded by native Chinese speakers (two women, one man) in a sound booth and normalized for volume.

Table 1: Examples of Mandarin disyllabic homophones.

狮子 shīzi [lion]		柿子 shìzi [persimmon]	
熊猫 xióngmāo [panda]		胸毛 xiōngmáo [chest hair]	

Culture Fair Test. We administered a computerized adaptation of the Culture Fair Test, Scale 4, Form A (Cattell, 1973) to assess participants' nonverbal intelligence (Rose et al., 2023). The assessment comprised two timed tests presented via Qualtrics. Test 1 (Series, 13 items, 3 minutes), asked participants to identify the correct geometric pattern from six options to complete a series of four images. Test 2 (Classification, 14 items, 4 minutes), asked participants to select which two of five patterns were similar and differed from the other three. Scores were calculated as the number of items correct (M = 12.2, SD = 2.9, *range* 5–20).

Musical Ear Test. The Musical Ear Test (Wallentin et al., 2010) was used to assess participants' musical ability. We administered the test on the Gorilla research platform. The test presents 52 pairs of melodic phrases played with sampled piano sounds. Participants were asked to indicate whether the two melodic phrases were identical or not via button press. Scores were calculated as the percentage of items correct (M = 64.2%, SD = 11.1%, *range* 40.0–87.5%).

Procedure

Each participant was tested individually in a single 2-hour session conducted remotely and recorded on Zoom. Participants were instructed to wear headphones throughout the session. At the start of the session, the participant shared their screen with the camera hidden. The experimenter remained present to monitor attention and answer general questions. Tasks were presented in a fixed order (see Figure 1), with written instructions provided on the screen.

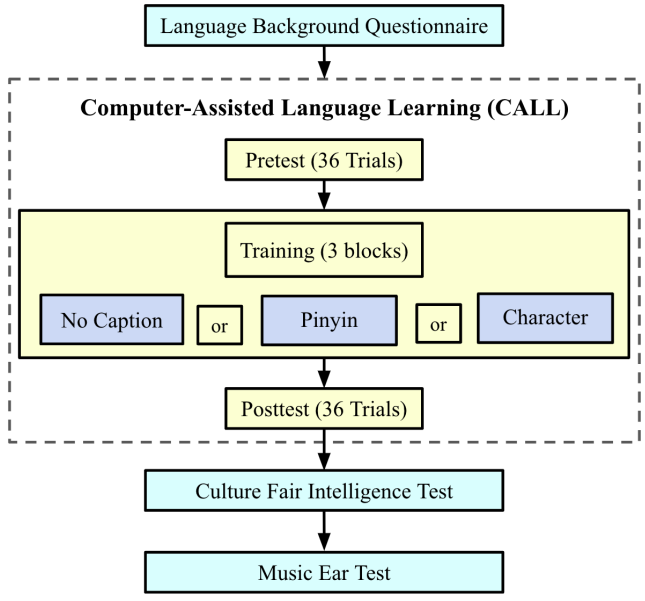
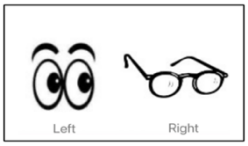
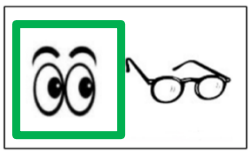
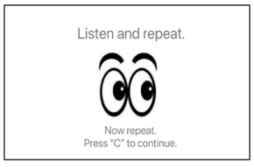


Figure 1: Order of tasks

Computer-Assisted Language Learning (CALL). The CALL protocol was programmed in PsychoPy and delivered via Pavlovia (Peirce et al., 2019) on Google Chrome. The protocol was fully self-paced and organized into three parts: pretest (one block of 36 trials), training (three blocks of 36 trials), and posttest (one block of 72 trials). Trials proceeded as the participant entered responses, with short breaks between blocks. In an effort to support vocabulary learning, comprehension feedback was provided after each response. The experimenter did not provide any feedback besides general encouragement between blocks.

Pretest/Posttest. The pretest and posttest were structured identically and utilized a word comprehension task combined with listen-and-repeat practice. Chinese nouns were presented one at a time with a corresponding pair of pictures representing the two possible meanings of the homophonic word pair. In each trial, the participant was asked to match the word they heard with the corresponding picture, using a two-option forced-choice task with the depictions of the segmental homophones side by side. After making their choice, the participant received feedback with the Mandarin noun repeated aloud as the correct picture was shown. The participant was then instructed to repeat the Mandarin word aloud to assess accuracy in tone production and repeated it again after hearing the word a second time. As summarized in Table 2, this procedure allowed us to determine whether vocabulary comprehension and/or listen-and-repeat accuracy of Mandarin tones improved from pretest to posttest. For the pretest (36 trials), all recorded stimuli were drawn from the same native speaker to minimize variability. For the posttest (72 trials), half of the trials were identical to the pretest, and the other used recordings drawn from two other native speakers.

Table 2: Pretest and posttest format.

Tasks and Instructions	Picture Presented
Vocabulary Comprehension *Instructions:* Listen to the recording and use the right or left arrow key to select the picture matching what you hear. *Female voice:* 眼睛, yǎnjīng *Feedback:* After response, Mandarin word (眼睛, yǎnjīng) is replayed as correct picture [eyes] is shown.	 
Listen-and-Repeat *Instructions:* Listen and repeat the word out loud twice, when you see the words "Now Repeat" on the screen. Press "C" to advance to the next trial.	

Training. Participants were randomly assigned to one of three learning conditions: no captions (n = 20), pinyin [Romanized text] captions (n = 19), and character captions (n = 20). In each of the three training blocks, participants were given the task of matching Mandarin spoken words with the

corresponding pictures. The format and instructions for the training blocks were identical to the pretest/posttest, with the exception that there were no instructions to repeat the words (i.e., comprehension practice alone). Also, in two of the learning conditions, the pictures were presented with captions. In each block, the recorded stimuli were drawn from two different speakers, (i.e., two female speakers, male speaker and female1, male speaker and female2). Half the words were spoken by one speaker and the other half by a different speaker, with the words presented in random order. In the two captioning conditions (Chinese characters and pinyin text), the captions appeared above the corresponding pictures throughout the training blocks.

Complexity Invariant Distance Calculations

Complexity Invariant Distance (CID) is a metric used to compare two time series while taking into account the complexity of the data. It calculates the Euclidean distance between the time series and adds a correction factor for the complexity of the series so that a complex series is viewed as "closer" to another complex series (Batista et al., 2014). Without this correction, Euclidean distance is biased in favor of simpler time series. CID is particularly useful in evaluating accuracy in tone production because it accounts for the inherent complexity of pitch contours characterized by variation in duration as well as the pitch curve (i.e., fundamental frequency, f0). This metric provides a single deviation score indicating how closely the learner reproduced the native speaker model.

Audio files were split into individual words (word-level analysis) and first and second syllables (syllable-level analysis) using PRAAT (Boersma, 2001). Pitch curves for the participant and the native speaker productions were extracted at both word and syllable levels using the Parselmouth library in Python (Jadoul et al., 2018). Following Zhou and Olson (2023), we normalized and resampled the pitch curves at 20 points per word (word-level analysis) or syllable (syllable-level analysis). This allowed for point-by-point comparisons in pitch curves. We used linear interpolation for sampling because preliminary analyses found that linear interpolation was more reliable than a Fourier series. We then compared each normalized word or syllable production with the native speaker model and calculated CID as the difference between curves. Python code for the CID calculations is available in Supplemental Materials (Gao et al., 2025).

Figure 2 provides a word-level example comparing a 20-year-old female participant's production of 眼睛 yǎnjīng [eyes] with the native speaker (model). In this example, the CID score was 17.8 at pretest and 10.7 at posttest, indicating higher listen-and-repeat accuracy for 眼睛 yǎnjīng [eyes] at posttest (i.e., less distance between the learner and the native speaker). Figure 3 provides a syllable-level example, i.e., a 22-year-old male participant producing the two syllables of 眼睛 yǎnjīng [eyes] on a single trial. Here, the CID for the first syllable (1.9) was markedly lower than for the second syllable (22.3) indicating higher accuracy in tone production for the initial syllable.

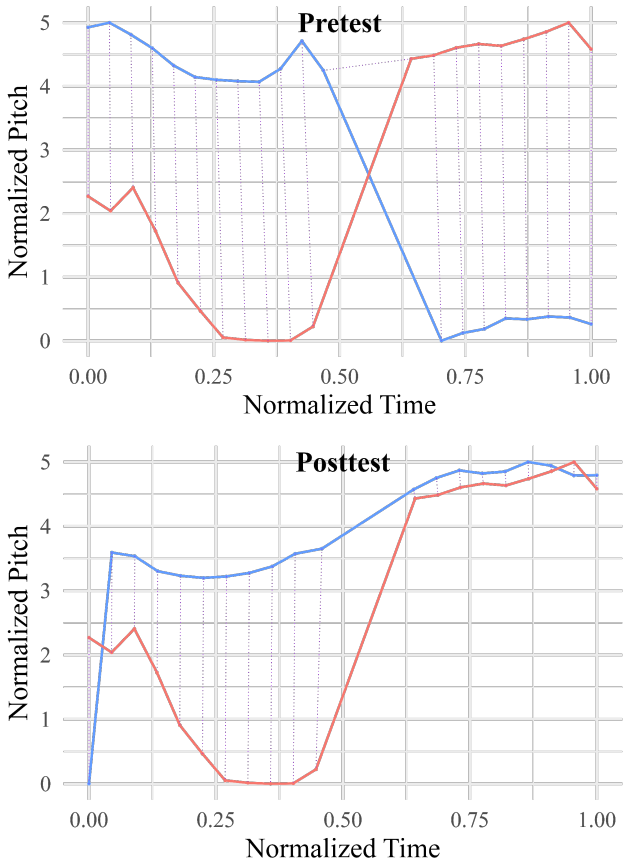


Figure 2: Word-level example comparing the pitch curves of 眼睛 yǎnjīng [eyes] at pretest and posttest for the participant (blue line) vs. the native speaker model (red line). Dotted purple lines indicate pointwise distance.

Results

Vocabulary Comprehension Accuracy

We found no improvement in participants' ability to associate Mandarin homophones with their corresponding pictures (Pretest: M = 65.2%, SD = 9.3; posttest: M = 63.8%, SD = 11.1). This null result underscores the difficulties participants faced in mapping the homophones onto their referents despite multiple blocks of trials with feedback. Moreover, we found no interpretable differences in performance across captioning conditions, suggesting insufficient exposure to benefit from captions. Participants in the Chinese character condition were more accurate at pretest (M = 66.8%, SD = 8.6) than those in the no captions condition (M = 63.7%, SD = 10.0), i.e., before seeing any captions in the training blocks. This difference was maintained at posttest. Regarding individual differences, we found no effects of musicality or nonverbal intelligence in predicting vocabulary comprehension. Given that these findings failed to demonstrate benefits of captioning conditions or aptitude, we can only conclude that one session of training was simply insufficient to detect factors associated with individual differences in learning tonal contrasts.

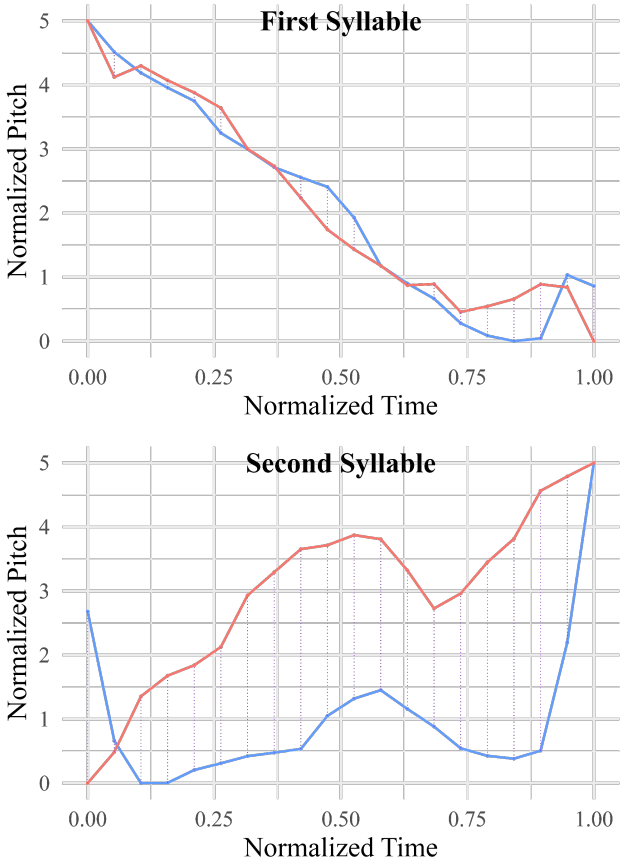


Figure 3: Syllable-level example comparing first and second syllable pitch curves of 眼睛 yǎnjīng [eyes] for the participant (blue line) vs. the native speaker model (red line). Dotted purple lines indicate pointwise distance.

Listen-and-Repeat Accuracy

For this analysis, we examined tone production for matched pretest and posttest trials (i.e., the 36 stimulus words spoken by the same female speaker) using the first repetition on each trial. To analyze production accuracy at the word and syllable level, we conducted hierarchical linear regression models using CID scores as the dependent variable. Though preliminary analyses of CID scores at pretest/posttest indicated no effects of captioning conditions on tone pronunciation accuracy, we retained learning condition as a factor in the models. Thus, both regression models included dummy variables for pretest/posttest and learning condition (i.e., comparing pinyin and character caption conditions with the no captions [reference] condition), and included musical ability (Musical Ear Test) and nonverbal ability (Culture Fair Test) as predictor variables. For the syllable-level analysis, syllable position was included as an additional predictor. Random intercepts were included for items and participants. Note that smaller CID scores indicate a smaller distance between pitch curves; thus, negative regression coefficients indicate an association between the variable and higher accuracy in tone production.

Word level analysis. We found that tone production accuracy measured at the word level increased from pretest to posttest as indicated by lower CID scores (Pretest: M = 13.8, SD = 6.4, *range* 1.6–51.6; posttest: M = 13.3, SD = 5.3, *range* 2.7–38.9; b = –0.49, SE = 0.14, p < .001). Regarding learning conditions: CID scores for the pinyin captions group (b = –0.68, SE = 0.44, p = .132) and the character captions groups (b = –0.82, SE = 0.44, p = .065) did not differ from the no captions group. Neither musical ability (b = –2.57, SE = 1.64, p = .123) nor nonverbal intelligence (b = 0.00, SE = 0.06, p = .961) were significant predictors of CID scores.

Syllable level analysis. In contrast to the word-level analysis, tone production accuracy measured at the syllable level did not differ significantly from pretest to posttest (b = –0.32, SE = 0.19, p = .092; pretest: M = 14.8, SD = 9.3, *range* = 0.5–66.9; posttest: M = 14.5, SD = 9.2, *range* = 0.7–92.0). However, we found a significant effect of syllable position (b = 3.09, SE = 0.19, p < .001; first syllable: M = 13.1, SD = 8.8, *range* = 0.5–66.9; second syllable: M = 16.2, SD = 9.43, *range* = 0.7–92.0), with higher accuracy for the first syllable than the second syllable. Moreover, musical ability was associated with performance, such that participants with higher Musical Ear Test scores had significantly lower CID scores at the syllable level (b = –5.22, SE = 2.21, p = .022). In contrast, nonverbal intelligence was unrelated to tone production accuracy (b = –3.00, SE = 2.34, p = .206). As in the word-level analysis, there was no effect of learning condition. CID scores for pinyin captions (b = –0.48, SE = 0.60, p = .427) and character captions groups (b = –0.27, SE = 0.59, p = .653) did not differ from the no captions group.

To further examine the effect of syllable position, we examined which tones presented the greatest difficulty in first vs. second syllable; see Table 3. For first syllables, differences in CID scores were minimal. For second syllables, learners did especially poorly on T0 (neutral tone), an unstressed syllable that receives its tone from the preceding syllable. They also did poorly on T3, which is realized as a full dipping tone only in final position.

Table 3: CID scores for Mandarin tones by syllable position.

Tone	First Syllable *M* (*SD*)	Second Syllable *M (SD)*
T0	N/A	18.9 (10.2)
T1	14.2 (6.2)	14.9 (6.9)
T2	13.0 (8.0)	11.7 (9.5)
T3	12.2 (10.0)	18.2 (8.5)
T4	12.6 (11.1)	15.4 (10.2)

Discussion

This study investigated Mandarin tone acquisition in adult L2 learners, focusing on their ability to learn new disyllabic words (vocabulary comprehension) and reproduce the tonal contours accurately (listen-and-repeat). Despite multiple

blocks of trials with feedback, participants showed no improvement in matching Chinese homophones with their corresponding pictures, underscoring the difficulty of using tonal contrasts as an informative cue to word meaning. Distinguishing the tonal contours of disyllabic homophones and mapping them on to their references is enormously difficult for adult L2 learners, even with corrective feedback provided trial-by-trial throughout training.

Although we found no evidence of improvement in vocabulary comprehension, modest gains were observed in learners' ability to replicate pitch contours when asked to listen-and-repeat the Mandarin words. Our study is the first large-scale implementation of CID as a measure of Mandarin tone production accuracy. Here we applied the measure to tone productions of 59 participants (72 trials each), whereas Zhou and Olson (2023) reported data from just 5 participants. We analyzed CID scores at both the word and syllable levels and found somewhat different results due to differences in the precision of the pitch curve alignment. Analyzing disyllabic words as a single time series conflates two separate tones, potentially masking localized L2 production errors. Prior research (Hao, 2018) indicates that tone accuracy varies by syllable position due to contextual and prosodic influences—effects that are blurred in word-level analyses. Syllable-level CID offers finer resolution, revealing position-specific deviations and clearer insight into learner difficulties. For preprocessing, we resampled all recordings to 20 pitch points per word/syllable. Thus, the word-level data had half the resolution of the syllable-level data, potentially introducing noise during interpolation. Additionally, we normalized pitch over the entire word, which might have added noise.

Word-level analysis showed significant pretest-to-posttest improvements in tone production accuracy, with relatively low variability in CID scores. However, at the syllable level, this improvement was no longer significant. Instead, we found participants repeated first-syllable tones with higher accuracy than second-syllable tones—contrary to our hypothesis and prior work suggesting that second-syllable tones are more recognizable (Hao, 2018; Liu et al., 2011). This may be because specific tones (e.g., T3) resemble canonical forms in word-final position (Wu et al., 2023), making them easier to identify at the ends of words.

Another factor at play may be the stage of L2 learning. Both Hao (2018) and Liu et al. (2011) tested learners who had much more exposure to Mandarin (i.e., a semester or more of instruction) than one hour of practice, as in our study. Adult native speakers of English tend to perceive first syllables as stressed, and thus more prominent, due to the trochaic bias (i.e., most disyllabic English nouns have first-syllable stress). English-speakers' bias to assign stress to first syllables may extend to Mandarin (Yu, 2021). In Yu's study, English speakers judged Mandarin nouns to have first-syllable stress, whereas Mandarin native speakers preferred to assign stress to final syllables. Due to the trochaic bias, our participants may have perceived first syllable tones as more acoustically prominent than second syllable tones, which may have enhanced their ability to reproduce them. Whether this primacy bias favoring first syllables persists over subsequent exposure to Mandarin as a new language is an important topic for future research.

As is expected in studies of adult L2 learning, there was considerable variability in participants' tone production accuracy, possibly reflecting language learning aptitude. In the fine-grained syllable-level analysis, learners who scored higher on the Musical Ear Test (Wallentin et al., 2010) were more accurate in reproducing pitch contours of Mandarin tones. The effect of musicality is in keeping with prior work (Christiner et al., 2022; Delogu et al., 2006; Han et al., 2019). Each of these studies found L2 learners with musical training and/or aptitude to be more accurate in perceiving and/or producing Mandarin tones than learners who lacked musical training. One interpretation is that same or similar auditory mechanisms are involved in processing music and non-native speech sounds, e.g., tracking pitch or melody, leading to enhanced sensitivity to tonal contours and contrasts in individuals with musical training or aptitude (Götz et al., 2023; Kempe et al., 2015).

In contrast to the effect of musicality, there was no effect of nonverbal ability (visual-spatial pattern completion) on learning outcomes. This contrasts with a previous study where learners with higher nonverbal ability were more successful in distinguishing Norwegian pitch contours than their counterparts with lower nonverbal ability (Kempe et al., 2012). More work is needed to determine how nonverbal ability may be leveraged to support language learning. Moreover, in contrast to previous research (Wayland & Guion, 2004), we found no evidence that participants' language background influenced learning outcomes (see also Hao, 2012). However, only three participants in our study had prior knowledge of a tonal language (Tagalog). Hence, additional research is needed to draw firm conclusions about possible transfer effects across languages.

Conclusion

Our findings underscore the complexity and difficulty of Mandarin tone acquisition for adult L2 learners (Pelzl, 2019). Learners' difficulties may be exacerbated in the context of learning disyllabic words, which constitute the majority of Mandarin word types (Liu et al., 2016). The modest gains we observed in tone production in the word-level analysis indicate adult L2 learners' capacity to register tonal contours from limited exposure. However, our participants were largely unsuccessful in mapping the contrasting word forms with their referents. To enhance learning, future studies might use perceptual training to increase tone sensitivity prior to introducing the word-learning task and include monosyllabic words as examples of canonical tones (Chang & Bowles, 2015). Use of CID as a quantitative measure increases the feasibility of conducting future studies of Mandarin tone production and should be compared with subjective coding (native speaker intuitions). The findings provide additional evidence that musicality may enhance the learning of tonal languages, which may inform educational strategies and curriculum design.

Acknowledgements

The research was supported by a CUNY Graduate Center Doctoral Student Research Award to the first author. Supplemental materials include a list of stimuli, de-identified data files, and analysis scripts.

References

Batista, G.E.A.P.A., Keogh, E.J., Tataw, O.M., & De Souza, V. M. A. (2014). CID: An efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery, 28*(3), 634–669. https://doi.org/10.1007/s10618-013-0312-3

Bluhme, H., & Burr, R. (1972). An audio-visual display of pitch for teaching Chinese tones. *Studies in Linguistics, 22*, 51-57.

Boersma, P. & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International, 5*, 341-345.

Brooks, P. J., Kempe, V. (2013). Individual differences in adult foreign language learning: The mediating effect of metalinguistic awareness. *Memory & Cognition, 41*, 281–296. https://doi.org/10.3758/s13421-012-0262-9

Cattell, R. B., & Cattell, A. K. S. (1973). *Measuring intelligence with the culture fair tests.* Institute for Personality and Ability Testing.

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America*, *138*(6), 3703–3716. https://doi.org/10.1121/1.4937612

Christiner, M., Renner, J., Groß, C., Seither-Preisler, A., Benner, J., & Schneider, P. (2022). Singing Mandarin? What short-term memory capacity, basic auditory skills, and musical and singing abilities reveal about learning Mandarin. *Frontiers in Psychology*, *13*, 895063. https://doi.org/10.3389/fpsyg.2022.895063

Delogu, F., Lampis, G., & Olivetti Belardinelli, M. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing, 7*(3), 203–207. https://doi.org/10.1007/s10339-006-0146-7

Gao, C., Gravelle, C. D., Jiang, S., & Brooks, P. J. (2025). Supplemental Material: Linking production of Mandarin tonal contrasts with musicality in adult learners. https://doi.org/10.17605/OSF.IO/N2GEH

Götz, A., Liu, L., Nash, B., & Burnham, D. (2023). Does musicality assist foreign language learning? Perception and production of Thai vowels, consonants and lexical tones by musicians and non-musicians. *Brain Sciences*, *13*(5), 810. https://doi.org/10.3390/brainsci13050810

Grigornko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The CANAL-F theory and test. *The Modern Language Journal, 84*(3), 390–405. https://doi.org/10.1111/0026-7902.00076

Han, Y. Q. , Goudbeek, M., Mos, M., Swerts, M. (2019). Mandarin tone identification by tone-naïve musicians and non-musicians in auditory-visual and auditory-only conditions. *Frontiers in Communication, 4*, 70. https://doi.org/10.3389/fcomm.2019.00070

Hao, Y. C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication. 97*, 32-42. https://doi.org/10.1016/j.specom.2017.12.015

He, Y., Wang, Q., Wayland, R. (2016). Effects of different teaching methods on the production of Mandarin tone 3 by English-speaking learners. *Chinese as a Second Language, 51*, 252-265. https://dx.doi.org/10.1075/csl.51.3.02he

Howe, J. H. & Baumgartner, E. S. (2024). Howe, J. H., & Baumgartner, E. S. (2024). Enhancing tonal-language learning through music: A review of experimental methods and melodic intonation therapy influences. *Review of Education, 12*(2), e3480. https://doi.org/10.1002/rev3.3480

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Kempe, V., Bublitz, D., & Brooks, P. J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *The British Journal of Psychology, 106*(2), 349-366. https://doi.org/10.1111/bjop.12092

Kempe, V., Thoresen, J. C., Kirk, N. W., Schaeffler, F., & Brooks, P. J. (2012). Individual differences in the discrimination of novel speech sounds: Effects of sex, temporal processing, musical and cognitive abilities. *PloS one, 7*(11), e48623. https://doi.org/10.1371/journal.pone.0048623

Kiriloff, C. (1969). On the auditory perception of tones in Mandarin. *Phonetica, 20*(2-4), 63-67. https://doi.org/10.1159/000259274

Lee, C. H., & Kalyuga, S. (2011). Effectiveness of on-screen pinyin in learning Chinese: An expertise reversal for multimedia redundancy effect. *Computers in Human Behavior, 27*(1), 11–15. https://doi.org/10.1016/j.chb.2010.05.005

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition, 39*, 593-620. https://dx.doi.org/10.1017/S0272263116000358

Liu, Y., Tseng, S., & Jang, J. R. (2016). Deriving disyllabic word variants from a Chinese conversational speech corpus. *The Journal of the Acoustical Society of America, 140*(1), 308-321. https://doi.org/10.1121/1.4954745

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning, 61*(4), 1119–1141. https://doi.org/10.1111/j.1467-9922.2011.00673.x

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pelzl, E. (2019). What makes second language perception of Mandarin tones hard?: A non-technical review of evidence from psycholinguistic research. *Chinese as a Second Language, 54*(1), 51-78. https://doi.org/10.1075/csl.18009.pel

Rose, M. C., Brooks, P. J., Lodhi, A. K., & Cortez, A. (2023). Benefits of testing and production for learning Turkish as a new language. *Language Learning, 74*(2), 365–401. https://doi.org/10.1111/lang.12602

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

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences, 20*(3), 188–196. https://doi.org/10.1016/j.lindif.2010.02.004

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. https://doi.org/10.1121/1.1531176

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning, 54*(4), 681–712. https://doi.org/10.1111/j.1467-9922.2004.00283.x

Wiener, S., Chan, Marjorie K. M., Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *The Modern Language Journal, 104*, 152-168. https://dx.doi.org/10.1111/modl.12619

Wu, Y., Adda-Decker, M., & Lamel, L. (2023). Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position. *Speech Communication, 146*, 45-52. https://doi.org/10.1016/j.specom.2022.11.001

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61–83. https://doi.org/10.1006/jpho.1996.0034

Yu, V. Y. (2021). Effects of syllable position, fundamental frequency, duration and amplitude on word stress in mandarin chinese. *Journal of Psycholinguistic Research*, *50*(2), 293–312. https://doi.org/10.1007/s10936-020-09731-6

Zhou, A., & Olson, D. (2023). Quantitative methods for analyzing second language lexical tone production. *Languages, 8*(3), 209. https://doi.org/10.3390/languages8030209