Mandarin Tone Acquisition as a Multimodal Learning Problem:

Tone 3 Diacritic Manipulation

Adam A. Bramlett

A Scholarly Paper
Submitted in Partial Fulfillment of
Masters of Arts in Second Language Studies

Department of Second Language Studies
University of Hawai‘i at Mānoa

First Reader: Theres Grüter, Ph.D.
Second Reader: William O’Grady, Ph.D.

Author’s Note
I have no conflicts of interest to report.
Acknowledgements

Throughout the writing of this master’s Scholarly Paper, I have received a great deal of support. This study would not have been possible without the generous support of the Elizabeth Carr-Holmes Scholarship, which allowed me to provide compensation to participants in the tone learning identification task.

I would first like to thank my advisor, Theres Grüter, Ph.D., whose expertise was invaluable in formulating the research questions and methodologies used in this study. I would also like to thank William O’Grady, Ph.D., whose experience and depth of knowledge helped me build the study’s design and understand its implications.

I would like to thank the many Second Language Studies, Linguistics, and East Asian Languages Literatures departments’ faculty members and fellow graduate students for constantly challenging me and teaching me about the tools that were necessary to make this study possible.

Lastly, I would like to thank my wife, Ding Wang-Bramlett, for her love and continual support that allowed me to pursue my dreams by continuing to strive to become a better researcher and friend every day. Leonidas and you are the reason I am inspired do this work.
Abstract

Through a tone knowledge survey and tone identification training task, I explore the extent to which visual representation of tone affects tone identification and tone learning. I use T3 as a case study in understanding the extent to which visual representation affects identification. Recent work on tone learning suggests there may be a link between different multimodal cueing mechanisms and efficient learning of tones. McGinnis (1997) finds that Pinyin diacritics are one of the most effective representation systems for learning Mandarin tones. However, Pinyin is also historically used as a descriptive representation of native speaker phonological knowledge. The dual purpose of the visual representation systems of tone creates a tension between a descriptive model of tone and an L2 learning aid. T3 shows extreme variation from a full T3 that falls and rises to a low T3. Zhang (2018) claims that Mandarin learners are not taught about T3 low features. The first experiment in this study found that current Mandarin learners disproportionately lack explicit knowledge of T3 low features but not other T3 features or any other tones’ features. Next, a tone identification training task was conducted to see if changing the visualization of T3 to low T3 would affect the learning of T3 and other tones. Findings show that T3 and T2 identification improved in the condition where T3 is represented visually as a low tone rather than full tone. Together, these findings suggest that many Mandarin learners do not know about T3 low features and that learning T3 with a low-tone visualization can positively affect the identification of T3 and T2. Using multimodal means to improve the representation of T3 may help reduce confusion in early stages and remedy long term acquisition problems of T3 and T2.

Keywords: non-native speech perception, multimodal learning, Mandarin, tone 3, HVPT
Introduction

Tone acquisition continues to receive an increasing amount of scholarly attention in both first language (L1) and second language (L2) acquisition research (e.g., Liu et al. 2011; Wiener, 2018; Zhang, 2018). Research focusing on tone acquisition for learners without lexical tone experience sits at an interesting intersection between theories on learning and non-native speech perception. Learning lexical tone is a perceptual learning task, which requires an investigation of more basic processes (perception, attention, and salience) (Liu et al., 2011; Cintrón-Valentín & Ellis, 2016). Mandarin tone 3 (T3) provides a potentially fruitful opportunity for investigating the basic processes of non-native speech perception and learning. T3 is notorious for both its variation in forms and difficulties to acquire in L2 learning (e.g., Zhang, 2018; Duanmu, 2007). In phonological research, much debate surrounds T3 concerning its basic form. The most common representations are referred to as low T3 (21) and full T3 (214)/falling-rising (further explanation of tone number systems will be discussed, in the Role of Multimodal Representations in Tone Learning, below). There has been a long-standing debate in the phonology literature on the representation of T3. However, the conversation has only just begun in non-native speech perception.

In this study, I aim to understand the relationship between visual representation of tone and tone learning in Mandarin. It is therefore necessary to introduce Mandarin tones and their respective acquisition, the relationship between visual representations and multimodal learning, and the current teaching of Mandarin tones. For this reason, this paper starts with an introduction to Mandarin Tones and Acquisition, explaining the concept of tone and each Mandarin tone, thereafter. I continue by examining the findings in Mandarin tone acquisition research with particular focus on T3. Next, the Role of Multimodal Representations in Tone Learning explores
the intimate and historical relationships between visual tone representations and multimodal learning. Lastly, the Congruence Between Pinyin and Mandarin Tones is explained with particular focus on T3. I will then introduce two experiments and present the research questions and methods used for this study, which are designed to investigate: 1) experienced L2 Mandarin learners' explicit knowledge of Mandarin tone dimensions (experiment 1), and 2) alternative visual representations of Mandarin tones in a multimodal (audio-visual) experiment, which tests the effectiveness of two variations of diacritics for learning Mandarin tones (experiment 2). I will end each experiment with a discussion of their respect results. Lastly, I will provide a general discussion of all results and end with concluding remarks.

**Mandarin Tones and Acquisition**

All languages use pitch variation to some extent (Gussenhoven, 2004). In tonal languages, pitch variation is implemented to change a word’s meaning. The lexical tone systems of such languages are complex and have extensive study from multiple fields. As such, there are various views on the nature of lexical tone. Broadly speaking, lexical tone can be defined by the contrastive pitch variation on words (Hyman, 2001). In this view, Cantonese and Swedish are both tonal languages; Cantonese, in this view, is considered a densely tonal language where Swedish is considered a more sparsely tonal language. Zhu (2010, p. 273) offers a more exclusive definition of lexical tone or pitch variation, which requires a phonologicalized pitch category to have three or more members and different lexical meanings (as cited in Zhu & Wang, 2015). Under both definitions, Mandarin is a tonal language. In Mandarin, the tones are associated with syllables; each syllable is a tone bearing unit (TBU) (Gussenhoven, 2004). There are twelve possible syllable structures in Mandarin and four possible tones (e.g., Lin & Wang,
1992; Duanmu, 2007). In Mandarin, every syllable is capable of bearing tone, but not required (see Duanmu, 2007, for further discussion of stress and tone).

Mandarin tones can be classified with two principal perceptual features, contour (pitch direction of change) and pitch height (Gandour, 1983). Tone 1 (T1) is a high flat tone, tone 2 (T2) is a rising contour tone (low to high), tone 4 (T4) is a falling contour tone (high to low), and T3 is a falling-rising contour tone (T3 will be further discussed below). Temporal features and amplitude have both been found to be secondary cues for tone identification (Kong & Zeng 2006). However, pitch height (F0, the acoustic correlate of pitch height) remains the primary cue for distinguishing tone in Mandarin (Liang, 1963; Lin, 1988; Whalen & Xu, 1992).

Mandarin Tone 3 has inspired research from phonology to psychology. Most T3 research focuses on its variation in forms (e.g., Duanmu, 2007) or late acquisition for L1 and L2 learners (e.g., Chao, 1951; Tsung, 1987). This paper sits at the intersection of these research areas; I will, therefore, summarize insights from both.

Research spans a continuum between support for the variant forms of T3, T3 as 214 (falling-rising) and 21 (low tone). A full discussion of tone representation (e.g., 214, 21) will be provided in the next section, *Role of Multimodal Representations in Tone Learning*. Research which proposes that the basic or underlying representation of T3 is 214 generally views the 21 variation of T3 as a reduction of the full form (Zhang, 2018). In other research, 21 is regarded as the basic form, while the 214 variation is seen as one of the variant allophones of T3. In recent years, T3 is starting to be viewed as a low tone phonologically (Duanmu, 2007; Xu, 2015) and perceptually (Shen et al., 2013). Zhang (2018) proposes a system which has a greater focus on tone height (register) rather than on pitch contour. The problem of T3 phonological representation has been noted by many people (e.g., Yip, 2002; Duanmu, 2007; Zhang, 2018).
However, Zhang (2018) is the first to link the debate for low T3 to T3 tone learning. Regardless, both forms exist in spoken Mandarin. The 21 variation is more common in speech at regular speed, but 214 is not completely absent (Duanmu, 2007). The 214 variation is more common for slowed speech or in citation form; however, 21 is not unacceptable in slowed speech or citation form.

In terms of acquisition, lexical tone learning for learners without a tonal language background is challenging; however, many studies have found effective results with respect to short term focused tone learning for L2 learners with and without prior tone learning experience (Bowles et al., 2016; Chandrasekaran et al., 2010; Chang & Bowles, 2015; Godfroid et al., 2017; Liu et al., 2011; Perrachione et al., 2011; Wang et al., 1999; Wong & Perrachione, 2007). An early high variable phonetic training (HVPT) study found an increase in correct tone identification starting at 67% and improving to 90% after a short training period (Wang et al., 1999). While acquisition of each Mandarin tone poses its own difficulties, T1 is generally agreed to be acquired first in L1 and L2 acquisition (Yip, 2002; Zhang, 2018). T2 and T3 are consistently found to be the most difficult to acquire in perceptual studies (Duanmu, 2007; Yip, 2002; Zhang, 2018). T3 is the most difficult to distinguish for non-native speakers (Kiriloff, 1969; Wang et al., 1999) as it is often confused for T2 and T4. Further, for both adult L2 learners and children, T2 and T3 are the most difficult to fully acquire (Leather, 1983; Li & Thompson, 1977; Shen & Lin, 1991; Sun, 1998; Wiener, 2017; Zhang, 2018). Although, T3 is known to be acquired later than other tones, Tsung (1987) convincingly argues that low T3 was found to be acquired first in several early longitudinal production studies of L1 acquisition (Li & Thompson, 1977; Chao, 1957).
To summarize, the underlying phonological representations of Mandarin tones are still being debated. In particular, T3 has two primary forms, T3 low (21) and T3 falling-rising (214). In the L2 acquisition of Mandarin tone, T3 is often acquired later and with more difficulty. In the following section, I discuss the role of tone representations in tone learning, and the respective insight that multimodal learning research can offer.

The Role of Multimodal Representations in Tone Learning

Lexical tone is a word level feature. However, its orthographic representation is made difficult by confusion on how it should be made salient in written form. Like many other tone languages, Mandarin tone is described primarily as a series of F0 patterns. The F0 patterns of Mandarin tones have been conceptualized with various representation systems. Chao (1930) was the first to propose a tonal representation system. His system, the five-point scale (FPS), uses successive numbers to describe tone on each syllable, “1” being the lowest pitch and “5” being the highest pitch. Change in number in FPS implies contour. Chao (1930) also introduced Gwoyeu Romatzyh (GR), which aims to emphasize the internal importance (word-level relationship between tone and syllable) of tone by using spelling alternations for different tones. Like the current study, GR spelling alternations were created with the intention to make the tonal system of Mandarin accessible to speakers of non-tonal languages (Chao, 1948). Hanyu Pinyin (Pinyin) was first developed from Chao’s FPS system by a group of linguists in the 1950s and modified to its current standard in 1995 (Zhou, 1995). Pinyin is the predominant romanization system used today. See table 1 for examples of the FPS system, Pinyin, and GR.
Table 1

*FPS, Pinyin, and GR Example*

<table>
<thead>
<tr>
<th>Name</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPS</td>
<td>ma55</td>
<td>ma35</td>
<td>ma21(4)</td>
<td>ma51</td>
</tr>
<tr>
<td>Pinyin</td>
<td>mā</td>
<td>má</td>
<td>mă</td>
<td>mā</td>
</tr>
<tr>
<td>GR</td>
<td>mha</td>
<td>ma</td>
<td>maa</td>
<td>mah</td>
</tr>
</tbody>
</table>

These systems of tonal representation, in essence, aim to use visual means in order to make tone a salient feature of the romanization system. That is, they use multimodal means to represent phonological or phonetic categories to assist the learning of tones. These systems were developed for two primary reasons: 1) to descriptively represent the categories of tone, e.g., representing T3 as mā, and 2) to aid the learning of tones by L2 learners (in the case of GR). However, it is not necessarily true that the descriptive representation of a tone and the best solution for L2 learners would be the same. In other words, it is possible that the underlying representation of tone for native speakers may be most accurately represented by one visual form, while the most effective visual representation for tone learners could be a different visual form. More recent work on tone is gaining the necessary tools to understand the effectiveness of tone representation systems and their sub-parts for L2 learning, as I outline below.

Recent work has continued this tradition by testing new systems which utilize other visual means such as color, shape, and a combination of cueing mechanisms. McGinnis (1997) examined the relationship between romanization systems and learning efficacy. His study examined Mandarin learning through two romanizations of Mandarin, GR and Pinyin. Comparisons between Pinyin and GR showed that students learning with the same content material and teaching methods have faster learning and longer retention with better long-term accuracy when using Pinyin. It is notable that Pinyin bears better learning results for tones of
both non-tonal L1 and pitch-accent L1 learners (English and Japanese, respectively). The consistent diacritics-markings above the syllable to represent tone categories-in Pinyin are helpful for recognizing tone as a category, while GR represents tone with various spelling alternations depending on the vowels of the respective word (see table 1).

In Liu et al. (2011), the researchers examined the subparts of Pinyin in an identification task attempting to understand the relationship between Mandarin tone learning and the subparts of the Pinyin system (Pinyin-spelling romanization and tone diacritics). The results favored the combination of diacritics and Pinyin-spellings over Pinyin-spellings with FPS marked numbers and diacritics without Pinyin-spellings. That is, participants trained using the regular Pinyin system showed better identification than participants training using a system which only uses diacritics or a system which uses numbers to replace the diacritics. It is suggested that this may reflect a modality principle which utilizes both visual and auditory information of the pitch contour and register (pitch height) (Mayer & Moreno, 2002). The visual representation input helps confirm the auditory input and allows for a more robust category to be formed. Liu et al. (2011) argue that the correspondence between the pitch contours and the pitch height of both visual and auditory stimuli created an ease in processing, where the numeric indicators were more burdensome; the lack of attention to the direction of pitch change by non-tonal L2 learners of Mandarin (Gandour, 1983) is remedied by making salient the visual representation of tones, which is supported by the intuitive nature of relating pitch height variation to height in physical space (Rusconi et al., 2006). Godfroid et al. (2017) continued the investigation of multimodal learning for Mandarin tone through an online training session which explored color, shape, symbol, and number. The researchers used several different cueing mechanisms to examine the learning of tones through an identification task. While color variation did show improvement,
symbol and number showed more robust results, better identification, across participants and cueing mechanisms.

While some form of Pinyin has been used to visually represent tone for Mandarin since the 1950s, Sumby and Pollack (1954) were the first to explore speech perception as a multimodal cognitive task. Likewise, Casasanto et al. (2003) were the first to provide evidence for a link between spatial movement and pitch movement in the general cognitive domain; their experiment showed identical auditory stimuli with higher visual spatial placement to be perceived as higher pitch. Similarly, Mitchel and Weiss (2011) found cross-modal input of non-linguistic tone to show decreased learning in treatments which have less cross-modal (visual and auditory) coherence. In regard to lexical tone specifically, several more recent experiments have tested facial and gestural congruency in tone learning with improved learning in conditions that provide congruency across audio-visual input (e.g., Hannah et al., 2017; Baills et al., 2019).

Stated simply, visuospatial information has been found to be consistently tied to pitch information across general perception studies and language-specific learning studies for tone.

In summary, the dual purpose of visual representation systems for tone creates a tension between a descriptive model of tone and an L2 learning aid. T3’s basic representation for native speakers is still being debated. However, the low T3 (21) is the more common form of T3 in connected speech (Duanmu, 2007). While the respective frequencies of T3 low and full T3 in and out of the classroom are currently unknown, the question remains open as to whether T3 low would be a better representation of T3 for L2 learners. Results of multimodal research support highly congruent visual and auditory information in tone learning. Variation of Mandarin T3 provides an opportunity for exploring congruency between visual and auditory input. The next
section explores the learning of tones using Pinyin to better understand Mandarin tone learning with particular emphasis on the congruency between tones and their visual representations.

**Congruence Between Pinyin and Mandarin Tones**

As previously mentioned, Mandarin tone can be described with two principal perceptual features: pitch height and pitch direction of change (Gandor, 1983). Here, I describe the visual coherence between Pinyin tone representation and Mandarin tones. Mandarin T1 is a high flat tone represented in "mā", which is congruent across pitch levelness, however incongruent in representing the height of T1. In Mandarin T2 and T4, mà and mà, are congruently represented in Pinyin for both pitch height and pitch contour. Due to its variation in form, T3 is more complicated. Mandarin T3 is represented in "mǎ", which accurately represents the falling and rising of full T3, however misrepresents the onset, or starting point, of the tone contour. For low T3, pitch height is implicit (lowness is part of the full T3 diacritic) within the falling-rising contour of Pinyin, however the contour is incongruent.

At the time of writing this study, major textbooks that use Pinyin utilize “v”, (e.g., mǎ) as the diacritic for T3. When compared to “—”, “/”, and “\” (T1, T2, and T4), “v” (T3) visually implies the most contour. However, in several studies, the lowness of T3 has been found to be the only perceptually salient cue for T3 recognition in context, out of context, and in varying sentence positions (Duanmu, 2007; Xu, 1997; Xu, 2015). Furthermore, the falling-rising T3 form (which does occur in citation form and slowed speech) is less common than low T3 in connected speech (Shi & Li, 1997).

Furthermore, Suo (2019) found success in the teaching of low T3 over full T3. She found that the use of low T3 (21) showed improved results in both listening and production performance than the use of 214. In her study, four Mandarin classrooms were separated into two
conditions and taught with one of two T3 variations for the duration of 9.5 weeks. In the first condition, students were taught using full T3. In the second condition, students were taught using low T3. For the second condition, teachers did not change the Pinyin T3 to low form. Teachers simply refrained from using the full variation of T3 even when producing T3 independently. Results showed that there was a statistically significant difference between both students’ oral reading and listening scores between the two conditions. The listening task is a forced-choice task, which requires students to decide which tones that they heard by selecting a disyllabic word. The oral reading task is a production task, which was rated by three trained native Mandarin speakers. A delayed post-test was conducted 18 weeks after the instruction and effects remained significant between conditions. Results in this study consistently showed that teaching T3 as low tone improved students' learning of Mandarin tones.

Although auditory input in the Mandarin classroom may tend to exemplify the full form of T3 (frequent slowed speech and citation form speech), no studies have directly examined the use of full and low T3 in and out of the classroom. This visual cue difference may be causing long-term acquisition problems as the salience of the contour for T3 is the most salient part of T3 rather than the lowness in pitch of T3. The misalignment between auditory and visual input (Mandarin classroom aural input and explicit instruction) may lead to perceptual incongruence. The difficulty of T3 across L2 Mandarin learners and the variation in forms of T3 calls for an experiment considering the effect of congruence between pitch height and the visual representation of T3 itself. Although it has been claimed that low T3 is better for L2 learners (Suo, 2019; Zhang, 2018), no studies have explored current learners’ explicit knowledge of low T3 or the effects of using low T3 as a visual representation of T3.
Experiment 1: Tonal Feature Survey

In previous literature, teaching T3 as a low tone has shown improved learning of both perception and production (Suo, 2019). Although it has been claimed that Mandarin learners are unaware of T3 lowness (Zhang, 2018), no studies have examined Mandarin learners’ explicit knowledge of low T3. As previously stated, the principal perceptual features of Mandarin tones include contour and pitch height (Gandour, 1983). Testing learner knowledge of these dimensions of tone may help uncover if and where students have confusion. In a basic sense, one of the most straightforward ways of classifying tones is by contour and register. Further, T1 and T3 both have principal heights that are salient, high and low, respectively. For T2, T3, and T4, direction of contour is also important. However, these features are not equally represented in the Mandarin Pinyin system. For example, T1 and T3 diacritics are incongruent with actual tonal features for tone height. That is, the T3 diacritic does not show lowness and T1 diacritic does not show highness (lowness of T3 is only implicit within the contour itself). However, this same contrast is not true for contour, which is represented fairly well besides an exaggeration of slope (for T2, T3, and T4). One question that can be asked is if participants will show equal explicit knowledge of T3 for questions pertaining to contour and height. This leads to the following research question:

To what extent do current Mandarin L2 learners exhibit confusion as a result of incongruence between tonal features and the Pinyin diacritic of T3?

One way to think about this question is by comparing T3 to other tones. Is accuracy for T3 explicit knowledge equal to T1, T2, and T4? Here we can predict that participants’ explicit knowledge of T1 and T3 height will be similar. This is because both tones are incongruent between their Pinyin diacritics and actual pitch contours. Further, participants' knowledge of T3
contour will be equally accurate to that of T2 and T4 contour. This is because T2, T3, and T4 contours are all congruent with their respective tonal diacritics. A separate way of thinking about this problem is by looking at knowledge of T3 height compared to T3 contour. That is, is knowledge about T3 equal across dimensions of tone (height and contour)? We can predict that participants will show unequal knowledge of T3 for height and contour, with height having lower accuracy than contour. Said in another way, participants' accuracy for T3 will be lower for height but higher for contour.

Participants

97 learners of Mandarin participated in the tonal feature survey. Only participants that were studying Mandarin at the point of responding to the survey are included. Inclusion criteria required participants to be learning Mandarin in a US University setting. The mean length of studying Mandarin for participants is 2.2 years with a range of 1-4 years.

Materials

This tone knowledge survey is composed of 12 multiple choice questions. Among the 12 questions, three questions were designed for each tone, which relate to the dimensions of tone type, direction, and height. As previously mentioned, Mandarin tones can be categorized by pitch height and pitch contour alone. However, for this questionnaire, pitch contour is broken down into two sub-categories, type and direction. Hence, each participant produces 12 data points (3 questions per tone times 4 tones). The questions were designed to explore participants' explicit knowledge of Mandarin tone features. The tone feature questions consisted of identical questions for each tone, asking about tone type (level, contour; 1a), direction (falling, rising, falling–rising, rising–falling, this tone does not have a contour; 1b), and primary height (high, low, neither; 1c). Each tone has a corresponding correct answer by type, direction, and height: tone 1 (level, this
tone does not have a contour, high), tone 2 (contour, rising, neither), tone 3 (contour, falling-rising, low), tone 4 (contour, falling, neither). An example for each category of tone question is found below, presented in (1), with 3rd tone being used in all examples.

(1) a. Tone type question:

Is the 3rd tone of mandarin a flat tone or a contour tone? e.g., mā

A) flat
B) contour

b. Tone direction question:

What type of contour tone is 3rd tone? e.g., mā

A) rising
B) falling
C) rising-falling
D) falling-rising
E) this tone isn’t a contour tone

c. Tone height question:

What pitch height is most important for the 3rd tone? e.g., mā

A) high
B) low
C) this does not have a most important pitch height

Procedure and Analysis Approach

Participation in this survey was completely voluntary. Data were collected using a Google Forms questionnaire, which was distributed to university-level Mandarin instructors. The Mandarin instructors then distributed the survey to their students. The data is automatically recorded in Google Forms and was converted to CSV format for processing in R (R Core Team, 2021). The survey took participants approximately eight minutes.
Each correct answer was coded as 1 while incorrect answers were coded as 0. The final dataset is organized by five variables: participant ID, tone (T1, T2, T3, T4), question category (type, direction, height), score (1, 0), and length of study (1-4 years). All data for T1 from contour direction is removed as T1 does not have a contour direction. All data for T2 and T4 are removed from the height questions as T2 and T4 do not have a predominant pitch height. Next, the dataset was further subset by category of question to do statistical analysis. Four logistic regressions were performed to test the differences between correct and incorrect answers for tone in questions relating to type, direction, and height, separately. The first 3 logistic regressions were performed to test the hypothesis that T3 accuracy is the same as the other tones for questions relating to type, direction, and height. The last logistic regression was conducted to test the hypothesis that T3 accuracy is different between the dimensions of height and both type and direction.

**Results**

Results of this survey show a general confusion for T3 questions in the question category of height. Amongst the 97 participants, 54 incorrectly (55.7%) answered the height question for T3, while only 12 out of 97 (12.4%) participants answered the same question incorrectly for T1. “High” is the correct answer for T1 and “low” is the correct answer for T3. Amongst incorrect answers, participants answered T1 as “low” and NA (“this tone does not have a most important pitch height”) four and eight times, respectively. For T3 incorrect answers, participants answered “high” and NA eight and 46 times, respectively. Figure 1 shows the distribution of answers. Results of the binary logistic regression, glm (formula = correct/incorrect ~ tone, family = binomial), indicated that there was a significant association between tone and explicit knowledge of height ($B = 1.9577$, $SE = 0.3084$, $Z = 6.349$, $p < .0001$).
Figure 1

**Number of answers for each tone**

![Chart showing number of answers for each tone]

However, for the category of tone type, all four tones showed similar amounts of correct and incorrect answers. The number of correct answers across participants for each tone are as follows: T1 95 (97.9%), T2 94 (96.9%), T3 92 (94.8%), T1 91 (93.8%). Results are displayed below in figure 2. Results of the binary logistic regression indicated that there was no significant association between tone and explicit knowledge of direction.

Figure 2

**Number of answers for each tone**

![Chart showing number of answers across tones and response levels]
Similarly, for the category of tone direction, T2, T3, and T4 also showed similar amounts of correct and incorrect answers. The number of correct answers across participants for each tone are as follows: T2 95 (97.9%), T3 96 (98.9%), T1 94 (96.9%). For T2 incorrect answers, two participants responded falling and none responded falling-rising, rising-falling, or NA (“this tone is not a contour tone”). For T3 incorrect answers, one participant answered falling and no participants answered rising, rising-falling, or NA. Lastly, for T4, 3 participants answered rising-falling and none answered rising, falling-rising, or NA. Results are displayed below in figure 3. Results of the binary logistic regression indicated that there was no significant association between tone and explicit knowledge of type.

Figure 3

Number of Answers for Each Tone

Lastly, when comparing T3 correct answers across question category (height, type, direction). Results of the binary logistic regression, glm (formula = correct/incorrect ~ question category, family = binomial), indicated that there was a significant association between T3 question category and explicit knowledge ($B = 4.564, SE = 1.005, Z = 4.54, p < .0001$). This
difference can be seen under T3 for figure 4. Further, figure 4 combines the above information by showing the percent correct for each tone and category.

**Figure 4**

*Number of Correct Answers by Tone across Category*

![Bar chart showing percent correct for each tone and category.]

**Discussion**

Results of the survey showed that T3 does show some confusion that could be due to incongruence between its Pinyin diacritic and tonal features. In particular, T3 shows confusion in the dimension of height. T3 performed lower in height when compared to T1 in height. Also, accuracy on T3 height was significantly lower than T3 contour (direction and type). However, incongruence cannot be the only explanation for incorrect answers. This is because T1 questions performed statistically better than T3 questions in the dimension height even though both T3 and T1 are incongruent in the dimension of height.

I will start with an interpretation of the fourth logistic regression. It was hypothesized that participants would perform worse on T3 height questions than T3 contour questions. This was confirmed by the logistic regression. One explanation for why participants are more likely to get T3 height question incorrect is that the Pinyin diacritic of T3 does not make T3 lowness information salient. In this regard, this data supports the hypothesis that incongruence between
Pinyin diacritics and actual tonal features is causing students to get the question incorrect. However, a secondary explanation is that participants are getting the answer to contour questions correct because of congruence. Participants are explicitly aware that the Pinyin diacritics are informative, which is true for the contour questions, but not the height question of T3.

However, the issue is complicated by the results of the within question analysis (comparing within question). While the hypothesis that correctness on T3 in the contour dimensions is similar to T2 and T4 can be accepted, we must reject the null hypothesis that correctness of T1 and T3 are the same in the dimension of height. As mentioned, both T1 and T3 tonal features are incongruent in the dimension of height compared with their respective Pinyin diacritics.

Here, I will go through three further possibilities to explain why participants were more likely to answer the T3 height question incorrectly in comparison to the T1 height question. The first explanation is that while T3 and T1 diacritics are both incongruent with their actual pitch contours, the ways that they are incongruent are different. T1 is incongruent in only height. To make the T1 diacritic completely congruent with T1 tonal features, it would only need to be raised to the onset height of T4 (or in FPS 55). However, T3 is more complicated due to its variation in forms. For the 214 variation, the Pinyin diacritic would only need to be lowered with respect to the onset position, which may increase the salience of T3 being low. For the 21 variation, it would need to appear as a low falling tone. Either way, the difference between T1 and T3 height incongruence may not be directly comparable. The second explanation is that the T3 incongruence for Pinyin is not taught while the T1 incongruence is taught. Pinyin is one of many resources that L2 Mandarin learners use to improve their Mandarin knowledge. T1’s pitch contour is incongruent with its Pinyin counterpart, but T1 is taught in Mandarin classes as a
high-level tone, while T3 low is not taught. This leads to the third possible explanation, T3 has more variation and only one of its forms is taught (Zhang, 2018). The actual phonological representation is still being debated. However, in the university setting, T3 is consistently taught as 214 without explanation of the 21 variation. It may be that it is best to represent T3 as a low form, instead of a full T3, for beginning learners to make the lowness salient as Tsung (1987) and Zhang (2018) suggest.

This survey was done to explore the explicit knowledge of Mandarin learners. I designed the questions to examine participants’ metalinguistic knowledge of the features for each Mandarin tone. Because this knowledge is metalinguistic knowledge, it would not be expected that L1 speakers of Mandarin would demonstrate a similar understanding of tonal features. However, this survey still provides insight into the features that current Mandarin learners have been taught or can infer from their experience with Mandarin tones and Pinyin.

In summary, this tonal survey confirms that current L2 Mandarin learners, as a group, lack explicit knowledge pertaining to T3 lowness. That is, whether it is from tonal incongruence or not being taught, L2 learners of Mandarin disproportionately display correct knowledge of tone features for T1, T2, and T4, but not T3. Specifically, nearly all learners show accurate knowledge of T3 falling-rising features but are unaware of the low features. Many participants show confusion about incongruence between Pinyin diacritics and the tonal features of T3. However, the source of this confusion is still unclear. While the Pinyin diacritic of T3 may be misleading for the low variation of T3, T1 does not suffer from the same problem. The next question that remains unanswered is if naïve learners of Mandarin would perform better on T3 learning if the tonal diacritic were manipulated to visually represent the 21 variation of T3.
Experiment 2: Tone Identification Task

Experiment 1 confirms that tested participants that are currently Mandarin learners lack explicit knowledge pertaining to T3 lowness, confirming the claim of Zhang (2018), that the lowness of T3 is not taught. One can now ask if changing the T3 diacritic would affect the perceptual congruence and learning of T3. As McGinnis (1997) indicates, Pinyin diacritics are one of the most effective representation systems for Mandarin tones in L2 learning as they represent F0 movement, which provides a visual link between pitch movement and tone category. Phonologically, there is debate between the actual representation of full T3 and low T3, while scholars generally agree upon tones 1, 2, and 4. If the full representation of T3 (e.g., mā) is the most accurate phonological representation of T3 for native speakers, it is still not necessarily true that L2 learners learn most effectively with this abstracted form in tone acquisition. In the case that the most accurate underlying representation of T3 is low tone, the “v” visual representation may lead L2 learners' attention to be guided to contour instead of the salient features of T3 (lowness). Regardless, the most effective visual representation in L2 tone learning does not necessarily match native speaker underlying representations. The accuracy of visual representation of a tone is crucially important for the L2 learner, as it acts as a multimodal confirmation of the auditory stimuli. The learning studies mentioned above have found that visual representation of tone can impact the efficiency and efficacy of learning lexical tones of Mandarin (e.g., Liu et al., 2011; Hannah et al., 2016; Baills et al., 2018). Mandarin tone learning has been studied extensively. Suo (2019) shows that orally teaching T3 as a low tone has better learning results across classrooms. However, no studies have considered changing the shape of tone diacritics for L2 tone learning to explore the effects of congruence between pitch height and visual height. While earlier studies have claimed that teaching T3 as a low tone improves
Multimodal Learning of Mandarin Tone

perceptual categorization of T3 (Suo, 2019) and claimed that T3 is primarily a low tone (Zhang, 2018), no study has developed an alternative solution for T3 diacritic representation, nor experimentally tested the effects of multimodal training to confirm the hypothesis of improved learning of T3 as a low tone. The following research questions can then be asked.

1) To what extent does changing the third tone diacritic to a low tone affect identification of T3?

2) To what extent does improved identification of T3, if there is any, improve categorization of T1, T2, and T4?

3) To what extent does changing the third tone diacritic affect learning of Mandarin tones throughout training for overall identification and rate of learning?

These research questions lead to the following hypotheses. The first hypothesis, in regard to RQ1, in this experiment is that manipulating the T3 diacritic to a low tone will result in better identification of T3 with less confusion for T2 and T4. The second hypothesis, for RQ2, is that improved identification of T3 will improve the identification of other tones. Specifically, previous studies have claimed that T3 is confused for and with T2 and T4 (Kiriloff, 1969; Wang et al., 1999). Here I attempt to understand if improved learning of T3 will improve the identification of T2 and T4. It is unclear from previous work if the confusion of T3 is bidirectional. In analysis, I will analyze the distribution of wrong answers for T2 and T4 to see if there are bidirectional confusions for T3 or simply a unidirectional confusion (e.g., T3 is incorrectly selected for T2, but T2 is not incorrectly selected for T3). For both identification of T3 and other tones, improved identification will be examined through accuracy, total number of correct responses over the total amount of trials.
Lastly, for RQ 3, it is predicted that participants that use the low T3 diacritic will have better overall identification of Mandarin tones throughout training. Overall, in this case, means collapsing over tone. Here, I ask if there will be an overall advantage to learning Mandarin tones with one T3 diacritic versus another. That is, I am examining the effects of changing the T3 diacritic on identification over training for amount correct for all tones, not just for any particular tone (as RQ 1 and RQ 2 do). In McGinnis (1997), participants learning Mandarin tones with Pinyin showed an improved starting point (identified more accurately from the start), with even better identification after a semester of using Pinyin over GR. Like his study, I predict that changing T3 will improve the identification of T3 and other Mandarin tones from the very beginning with better identification throughout the training. It is unclear whether rates of learning will change between groups in this short training session. However, because of perceptual ease of processing the low T3, it is therefore predicted that participants using low T3 will also learn at faster rates on average. Here, the prediction is that participants using low T3 will on average show better tone identification overall throughout training and learn at faster rates.

**Participants**

25 consenting adults with access to a computer running Google Chrome and owning a pair of headphones participated in the experiment for either course credit through the University of Hawaii at Manoa or an Amazon gift card valued at 5 USD. None of the participants had experience learning a tonal language. The experiment was done remotely by each participant in their own time. Headphone use was required and tested before the experiment began. Instructions were sent to the participants after they contacted the researcher with interest in participating. Instructions were written in a simple and clear way. No participants were confused
by the procedure throughout testing. All participants reported normal or corrected hearing and eyesight. The mean age of participants was 32 years of age.

Language background information was collected through a questionnaire to ensure that participants did not speak a tonal language that they were unaware is tonal. Language experience was self-reported: participants were instructed to list all known and studied languages in order of dominance, from strongest to weakest language. One participant was excluded in data analysis for experience with a tonal language. The remaining participants were randomly assigned into 1 of 2 conditions, namely system 1 and system 2, which will be described below in materials.

Questionnaire data revealed no statistically meaningful differences between participants in the two conditions with respect to language background, age, musical training, or vocal training (see the appendix for detailed information).

Materials

The materials consisted of 32 target real Mandarin words recorded using the Shure MV88 application and recorder in soundproof booths by four phonetically trained Mandarin native speakers who are all Mandarin teachers in a university setting. This study used high variable phonetic training (HVPT) for its training identification task, which has been found to improve categorization of tones in prior tone learning research (e.g., Wiener, 2020). Four speakers were used to create more variability in productions of tones to increase category learning of tones. When speaker input is highly variable, learners must generalize over speaker-specific and word-specific acoustic similarities to correctly categorize tone with specific phonetic patterns (see Thomson, 2012, for a full review of HVPT and Wiener, 2020, for use of HVPT in a production study of Mandarin tones). All sounds were recorded in the carrier sentence “我说 x 这个词 (I say the word X)”. The complete sentences were presented in a random order so that speakers
would not get used to a specific ordering of tones or syllables so that speakers were encouraged
to pronounce the words more naturally as they occur. Speakers were asked to say each sentence
naturally as if speaking with a friend. The 32 real words are made up of four contrasting initials
and two contrasting finals each recorded for all four tones (4 initials \times 2 finals \times 4 tones = 32
unique sounds by each speaker). The carrier sentence was designed to avoid sandhi tone issues.
While research about sandhi tone is meaningful and interesting, it is not the focus of this study.
The eight syllables used for Mandarin consisted of ba, bi, la, li, ma, mi, da, and di. Words were
then manually extracted using Praat by two Mandarin speakers with training in phonetic
transcription from complete sentences (Boersma & Weenink, 2021). The experiment was
designed using Python and carried out on Psychopy’s Pavlovia online platform (Peirce et al.,
2019; van Rossum, 1995).

Procedure

The experiment consisted of an identification training task. Each of the 128 individual
tokens constitute a single trial. The 128 tokens were split up into four 32 trial-blocks. Blocks
were organized by syllable, tone, and speaker. That is, each block had eight words from each
speaker with two of each tone but only one recording of each possible syllable. Each of the four
blocks have an equal amount of each tone and syllable. Each block is made up of 32 unique
recordings that are only presented once throughout the entire experiment. Each block contains
eight of each tone with a total of 32 of each Mandarin tone throughout the experiment. Token
order was randomly presented within each block.

Each participant was assigned to one of two groups which I will refer to as system 1 and
system 2, each group has 12 participants. System 1 marks tones like the standard Pinyin system,
with the exception of T1 being raised to match the FPS system (55). System 2 is the same as
system 1 except for the visual representation of T3. In system 2, T3 is represented by a low falling tone or 21 in the FPS system. The two systems have identical words and block design. However, the visual representations are different. For tones 1, 2, and 4 the tones are represented the same. However, T3 is represented differently between the two systems. See figure 5 for the visual representations for the two systems.

**Figure 5**

*Example of Visual Representation of the Two Systems Taken from Experimental Trials*

<table>
<thead>
<tr>
<th>System 1</th>
<th>System 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Visual Representation" /></td>
<td><img src="image2.png" alt="Visual Representation" /></td>
</tr>
</tbody>
</table>

The order of the blocks was counterbalanced to reduce the chances of block-specific problems being confounded with improvement in identification. On each trial, participants were asked to identify which visualization best matches the sound that is played using the “Q”, “P”, “Z”, “M” keys (see figure 5 for the keys next to each tone for the two systems). These keys are chosen to best match the position on screen. Each of the tones is consistently labeled with the same key. That is, T1 corresponds to “Q”, T2 corresponds to “P”, T3 corresponds to “Z”, and T4 corresponds to “M” (two examples for the two groups can be found in figure 5). After each trial, the participant was given individualized feedback with variable text depending on the correctness of their response. Feedback told them if their response was correct or incorrect, replayed the
exact same sound file and showed the correct answer, as seen in figure 6. The participants were instructed to identify the tone they heard by listening to the word played and by using feedback from previous trials.

**Figure 6**

*Examples of Feedback for T1 Correct Answer (left) and T1 Incorrect Answer (right) on Trials*

Click “space” to continue.

---

**Results**

To test the hypotheses formed from research questions 1, 2, and 3 stated above, a two-way ANOVA and a linear regression were used to better understand participants' identification of tones, with particular emphasis on T3. The hypotheses of research questions 1 and 2 were explored with a two-way ANOVA that uses accuracy as the dependent measure. Research question 3, which aims to examine overall learning through identification and learning rate, was explored using a multiple linear regression. Answer selection by tone and block is shown in figure 7. Statistical test results will be provided below.
Figure 7

Raw Selection of Tone Responses across Tone Trials and Blocks

Note. Blocks are organized from top to bottom (Block 1-Block 4) and tone trials of each tone are organized from left to right (T1, T2, T3, and T4).

For the two-way ANOVA, tone (four levels: T1, T2, T3, and T4; within-subject) and system (two levels: system 1 and system 2; between-subject) constituted the independent variables (IVs) and mean response accuracy by participant constituted the dependent variable (DV). Accuracy scores were calculated by dividing each participant's correct answers for a given tone by the total number of trials for that tone, which was always 32. There was a significant effect of tone ($F(3, 376) = 5.69, p < 0.001$, partial $\eta^2 = .043$) and system ($F(1, 376) = 28.05, p < 0.001$, partial $\eta^2 = .069$) with a significant interaction effect between tone and system ($F(3, 376) = 2.68, p < 0.05$, partial $\eta^2 = 0.020$). In order to explore the nature of the interaction, pairwise comparisons between the two systems were conducted for each of the four tones, with alpha levels adjusted for multiple comparisons ($.05/4 = .0125$). The output of these comparisons,
summarized in Table 2, indicated significant differences between the two systems for both T2
and T3, but not for T1 and T4. A visualization of accuracy data is provided in figure 8.

Table 2

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Difference</th>
<th>Lower</th>
<th>Upper</th>
<th>P adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1: System 2-T1: System 1</td>
<td>0.1120</td>
<td>-0.0432</td>
<td>0.2671</td>
<td>0.3540</td>
</tr>
<tr>
<td>T2: System 2-T2: System 1</td>
<td>0.1797</td>
<td>0.0246</td>
<td>0.3348</td>
<td>0.0109*</td>
</tr>
<tr>
<td>T3: System 2-T3: System 1</td>
<td>0.2188</td>
<td>0.0636</td>
<td>0.3738</td>
<td>0.0006**</td>
</tr>
<tr>
<td>T4: System 2-T4: System 1</td>
<td>0.0286</td>
<td>-0.1265</td>
<td>0.1837</td>
<td>0.9993</td>
</tr>
</tbody>
</table>

*Note.* *p < 0.0125, **p < 0.0025, ***p < 0.00025

**Figure 8**

*Mean Response Accuracy by Participant and Tone*

*Note.* The left side of each plot is a box plot that shows the first and third quantiles and median. The right side is a violin plot, which shows the distribution of data. A point plot is added to show actual data points of participants.
T3 and T2 showed better accuracy in system 2; However, this does not tell us anything about the incorrect answers for each tones’ trials. From figure 9, for T3 trials, it is clear that T4 is the primary distractor for system 1 participants but not for system 2 participants. For system 1, T4 and T3 were selected 11.42 and 11.50 average times for T3 trials per participant. However, for system 2 participants, T4 and T3 were selected an average of 5.75 and 18.5 times per participant. This means that the T4 diacritic is selected just as often as T3 for T3 trials in system 1, but not system 2. For T2 trials, T3 seems to be the primary distractor item for participants in system 1 but not system 2. Participants in system 1 on average selected T3 and T2 9.58 and 13.75 times, respectively. For comparison, system 2 participants selected T3 4.33 times, where they selected T2 an average of 19.50 times. One of the contributing reasons for accuracy improving seems to be less confusion for other tones. However, the confusion is tone specific and not distributed equally across tone answers. When comparing system 1 and system 2, T4 is selected more often for T3 trials. Similarly, for T2 trials, T3 is often chosen in T2’s place. However, this confusion for both T3 and T2 is unidirectional. In system 1, T4 is chosen for T3 trials and T3 is chosen for T2 trials, but the opposite is not true. T2 is not equally confused for T3 trials and T3 is not equally chosen on T4 trials.
To answer RQ3, a multiple linear regression was calculated to predict correctness based on system and trial number. For this analysis, I collapsed percent correct by trial over tones. Results of the multiple linear regression, \( \text{lm (formula = percent correct \sim system*trial number)} \), indicated that there was a significant effect for system and trial number (\( F(2, 252) = 26.3, p < .001, R^2 = .23 \)). The individual predictors were examined further and indicated that system (\( t = 4.42, p < .001 \)) and trial number (\( t = 3.185, p < .01 \)) were significant predictors in the model. However, no significant interaction was found. Results are visualized in figure 10. The dotted line is showing chance performance that would be expected if participants were choosing completely randomly. When examining the participants as groups, both systems perform above chance from the very beginning. The significant effect of system confirms that learning T3 as a low tone improves overall tonal identification throughout the training session. The main effect of trial
number shows that both groups improved identification over the training session. However, because there was no interaction between system and trial number, we must accept the null hypothesis of rate of learning being the same between groups.

**Figure 10**

*Percent Correct Answers by Trial across Systems*

![Figure 10](image)

*Note.* The grey dotted line is chance performance.

**Discussion**

Results of this experiment show that changing the diacritic of T3 affects not only identification of T3 itself but also improves identification of T2. Both groups performed better in the later trials than in earlier trials, which supports the efficacy of short-term tone training found in earlier studies (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Chang & Bowles, 2015). Although participants using system 2 on average outperformed participants using system 1, rates of learning between systems were unaffected.

Results of this experiment showed an overall improved identification of T3 for participants that were trained using system 2. This supports the first hypothesis that system 2
participants correctly identified T3 sounds by choosing T3 more often on T3 trials. Further, T2 identification is improved for system 2 participants which means that the T2 visualization was chosen more often on T2 trials. This appears to be because of less confusion for T3. This study has reproduced the findings of previous studies for T2 confusion with T3 (e.g., Leather, 1983; Li & Thompson, 1977; Shen & Lin, 1991). However, this confusion is unidirectional. T3 is commonly selected for T2 trials; however, T2 is selected least often for T3 trials. Further, this confusion is remedied by making the visual contours more visually congruent with phonetic features. While system 1 participants showed confusion for T3 on T2 trials, system 2 participants did not confuse T3 for T2 any more often than T1 or T4. It is notable that both groups were unlikely to choose T2 for T3 trials. This is meaningful because previous literature has claimed that T3 is often confused for T2 and T4 (Kiriloff, 1969; Wang et al., 1999). While this is true for participants in system 1 for choosing T4 on T3 trials, it is not true for either T2 or T4 with system 2.

RQ2 was answered with the same two-way ANOVA as RQ1. In the same way that T3 was analyzed, I analyzed the accuracy of T1, T2, and T4. For accuracy between systems, no difference was found for T1 or T4. However, T2 accuracy was shown to be statistically different between the two systems. This shows that improving identification of one tone can affect the identification of another tone. This can be confirmed by examining T2 trials’ incorrect answers, the accuracy improvement for T2 is mostly due to participants choosing T3 less often in system 2. As stated above, this means that system 1 participants reproduced the confusion of T3 and T2 found in several previous studies (e.g., Shen & Lin, 1991), while system 2 participants were better able to distinguish between the two tonal categories. This particular improvement is interesting for two reasons. 1) Many studies have found that T2 and T3 confusion is long term
even for very advanced L2 learners (e.g., Zhang, 2018). Improvement in any way for naïve beginners may lessen the burden of this problem in the long-run for Mandarin L2 learners. 2) Within the regular Pinyin system or system 1, T3 looks equally similar to T2 and T4. However, in the system 2, T3 is much less visually similar to T2 but remains similar to T4. The contour direction within system 2 for T3 and T4 is the same. Having no improvement for T4 accuracy may be because of a persisting similarity between the visual contours. Likewise, the improvement for T2 and T3 accuracy could be due to the visual distinctiveness of the two contours, which allows for the more salient featural differences of both T2 and T3 to be noticed rather than their similarities.

Lastly, to answer RQ3, I performed a linear regression to see if system and trial number predict the proportion of correctness across participants. Like early studies, this study shows that short term focused learning can improve tone learning (e.g., Chang & Bowles, 2015). Further, the analysis revealed a difference between both system and trial number. The results indicate that participants from both groups improved over the training session. With the addition that system 2 participants performed better than system 1 participants throughout training. From this perspective, we can say that system 2 participants learned the tones more effectively than system 1 participants by having better identification upon starting perception and improving to a higher overall identification of Mandarin tones. In this way, we can say that system 2 participants on average learned the categories of Mandarin tones better than participants in system 1.

While there was a difference between tone identification over the training session, there was no meaningful difference between rates of learning. For this study, I interpret this the same way that Liu et al. (2011) interpret their results of better scores without faster rates. The rates of learning are the same between groups. Meaning that participants in system 1 and system 2 both
Multimodal Learning of Mandarin Tone

successfully improved over the training. However, the advantages of system 2 are purely perceptual ease in processing. For L2 learners, processing T3 tones as a low tone forms a more perceptually immediate link between visual and auditory information. Helping them by bootstrapping tone learning on top of a previously acquired link between visio-spatial information and pitch information. This means that system 2 participants are not better at tone learning. Rather, system 2 participants are able to take advantage of knowledge that they already have to identify tones more correctly, specifically by confusing T2 and T3 less often from the start. Previous literature has repeatedly found that T2 and T3 show the most confusion in acquisition for both L1 children and L2 adults (Leather, 1983; Li & Thompson, 1977; Shen & Lin, 1991; Sun, 1998; Wiener, 2017; Zhang, 2018). This study confirms that, at least for L2 adult learners, this process can be made less problematic by making the primary perceptual features of T2 and T3 more salient rather than their similarities.

**General Discussion**

This study attempts to understand how multimodal input affects the process of learning tone categories in a tonal language. This study starts by exploring Mandarin learners’ explicit knowledge of Mandarin tones and finds that the majority of participants did not have explicit knowledge pertaining to T3 lowness. Next, this study looked to examine the effects of changing the visual representation of tones by manipulating the T3 diacritic to a low tone. Specifically, this study has found evidence for improved identification of Mandarin tonal categories when manipulating the tonal diacritic of Mandarin T3. This study makes no attempt to claim that T3 should be represented as a low tone descriptively, I only claim that the use of low T3 visualization improves the learning of Mandarin tones for adults when compared to full T3.
In more general terms of tone learning outside of Mandarin, this study, like many other multimodal tone learning studies, has shown that some ways of representing tone are more or less useful for learning tones (Liu et al., 2011; McGinnis, 1997). The particular implications of this would be language specific. However, what is clear is that visual representation of a tone provides crucial information to the learner that allows them to learn tonal categories more or less effectively (Godfroid et al., 2017). One question that remains unanswered in the literature is if visual representations would be effective for variation in tone. This is particularly important for the case of T3. While this study has found that Mandarin T3 is learned more effectively by low T3 than full T3, it is important to remember that this study was done for in-context T3s (T3s that were spoken in context). That is, every sound was extracted from a carrier sentence. The actual T3 pronunciations have not been analyzed for lowness or fullness. It is unclear if representing a tone closer to the actual pitch contour would be equally helpful in a manner that is consistent with this study. Meaning that it may not necessarily be better to have variation in the representation of a tone visualization. However, because this study uses in-context T3 for trials, it is not unreasonable to assume that most of the T3 sound files were not full T3s (Duanmu, 2007). In this respect, this study reconfirms the link between visio-spatial information and pitch information found in studies of general perception (Casasanto et al., 2003). Like other recent studies looking at multimodal learning of tones (e.g., Baills et al., 2019; Hannah et al., 2017), this study has found that the manipulation of visual information improves the learning of tone by expanding this relationship to a more abstract visual means of tonal diacritics.

For the case of T3, earlier studies have consistently found high confusion between T2 and T3 (e.g., Sun, 1998). This study confirms this confusion between T2 and T3, but only for participants that learn T3 using the full “v” visual representation. In contrast, participants who
learn T3 using the low T3 do not confuse T3 with T2 any more often than any other tone. While, as stated, it may be best to represent T3 descriptively as a full T3, the findings of this study show that while it is important to descriptively understand the features of a language, it is equally important to understand the starting points of learners. The tension between using Pinyin as a descriptive model of Mandarin tone and an L2 learning aid has created an unnecessarily difficult task for the L2 learner. The findings from earlier studies for persistent problems with acquiring T3 in adult L2 acquisition may be partially resolved by changing the visual representation of T3. However, changes like this should only be done if future studies can continue to replicate the results of this study in more longitudinal experiments that carefully examine the ways that students are learning T1, T2, and T4 alongside T3. While the low T3 of system 2 showed improved learning, it is possible and likely that another representation could be even better (e.g., a low dot for T3 and a raised line for T1). It remains unclear which visual representation would be best for L2 learners. Specifically, it is still unclear if the low T3 visualization of system 2 would help or hinder the learning of the full T3 variation. However, for tone learning, it is clear that visual representation matters.

**Conclusion**

This study has two primary findings: 1) that many current Mandarin L2 learners lack explicit knowledge of the low features of T3 and 2) that manipulating the visualization of a tone to be more phonetically representative of F0 height can improve the learning of tones by acting as a multimodal confirmation of auditory information. Moreover, the T3 tone can be learned with the “v” diacritic or the 21 low tone diacritic presented in this study. However, the T3 “v” representation causes learners to have a more difficult time with making an immediate connection between auditory information and visual information. In a practical situation,
teaching T3 as a full tone is not detrimental to learning. However, teaching T3 as a low tone does improve the ease of learning T3 and affects the overall accuracy of tonal identification. System 2 participants performed better throughout the identification training task overall because of T3 and T2 identification improvements; but the rates of learning between systems remained the same. When we are considering tone learning, it is essential to examine the differences between descriptive representation and L2 learning aid. While the description of a tone is crucially important for understanding the nature of tones in a particular language, we must not let these debates hinder the learning and potential successes of L2 learners.
References


Appendix

Language background results
A two-sample t-test was run to compare the difference in the amount of language spoken/learned by participants across the two groups. No significant difference was found ($t(22) = -0.92, p = 0.37$).

Musical training results:
A Mann-Whitney U test was run to compare the differences in the participants’ musical training across groups. No significant difference was found ($W=65, p = 0.97$).

Age results:
A Mann-Whitney U test was run to compare the differences in the participants’ musical training across groups. No significant difference was found ($W=69.5, p = 0.85$).

Vocal Training results:
A Mann-Whitney U test was run to compare the differences in the participants’ vocal training across groups. No significant difference was found ($W=85, p = 0.27$).