Original Article

second
language
research

Second Language Research
1–25
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0267658320941546
journals.sagepub.com/home/slr
SAGE

From sounds to words: The relation between phonological and lexical processing of tone in L2 Mandarin

Wenyi Ling and Theres Grüter
University of Hawai'i at Mānoa, USA

Abstract

Successful listening in a second language (L2) involves learning to identify the relevant acoustic–phonetic dimensions that differentiate between words in the L2, and then use these cues to access lexical representations during real-time comprehension. This is a particularly challenging goal to achieve when the relevant acoustic–phonetic dimensions in the L2 differ from those in the L1, as is the case for the L2 acquisition of Mandarin, a tonal language, by speakers of non-tonal languages like English. Previous work shows tone in L2 is perceived less categorically (Shen and Froud, 2019) and weighted less in word recognition (Pelzl et al., 2019) than in L1. However, little is known about the link between categorical perception of tone and use of tone in real time L2 word recognition at the level of the individual learner. This study presents evidence from 30 native and 29 L1-English speakers of Mandarin who completed a real-time spoken word recognition and a tone identification task. Results show that L2 learners differed from native speakers in both the extent to which they perceived tone categorically as well as in their ability to use tonal cues to distinguish between words in real-time comprehension. Critically, learners who reliably distinguished between words differing by tone alone in the word recognition task also showed more categorical perception of tone on the identification task. Moreover, within this group, performance on the two tasks was strongly correlated. This provides the first direct evidence showing that the ability to perceive tone categorically is related to the weighting of tonal cues during spoken word recognition, thus contributing to a better understanding of the link between phonemic and lexical processing, which has been argued to be a key component in the L2 acquisition of tone (Wong and Perrachione, 2007).

Keywords

categorical perception, eye-tracking, L2 processing, lexical tone, Mandarin Chinese, spoken word recognition

Corresponding author:
Wenyi Ling, Department of Second Language Studies, University of Hawai'i at Mānoa, 1890 East-West Road, Moore Hall, Room 570, Honolulu, HI 96822, USA.
Email: Wenyi9@hawaii.edu

I Introduction

For most listeners, recognizing words in speech is an effortless process. Yet the automaticity of this process, especially in a native language (L1), belies the number and complexity of processing steps involved. These often become more salient in a second language (L2), where processing is less automatized. As in L1, word recognition in L2 involves at least two critical steps: (1) perceiving phonological categories from acoustic input, and (2) using perceived phonemes to activate word candidates in the mental lexicon (Cutler, 2012). Infants are born with sensitivity to various sound contrasts, and soon tune in to those that are meaningful in the language(s) they are exposed to (Werker and Tees, 1984; Yeung et al., 2013). This leads to increasing sensitivity to language-specific contrasts and decreasing sensitivity to contrasts irrelevant in the L1(s). By adulthood, L1 phonological categories are deeply entrenched, allowing L1 listeners to assign acoustic input to phonological categories rapidly and without effort. For L2 learners, the challenge is not only to learn potentially new phonological categories, but also to automatically activate and use this knowledge to identify phonological contrasts in highly variable utterance realizations in fluent speech. Laboratory experiments provide ample evidence that L2 learners have great difficulty with phonological contrasts that do not exist in their L1 (Dupoux et al., 2008; Strange and Dittmann, 1984; Pelzl et al., 2019; Qin et al., 2017), and that they perceive L2 phonemic variants less categorically than native listeners (Ling et al., 2016; Shen and Froud, 2019).

These difficulties at the phonological level in L2 speech perception lead to further challenges at the level of lexical access during L2 word recognition. As a result of not being able to efficiently utilize all relevant phonological cues, it is likely that L2 listeners will activate a greater number of word candidates, leading to more extensive competition for selection compared to L1 processing (Cutler, 2012; Weber and Cutler, 2004). Evidence from recent studies shows that L2 learners were more likely to activate pseudo-homophones, partially overlapped non-words and cross-language vocabulary (Broersma and Cutler, 2008, 2011; Dijkstra et al., 2000; Marian and Spivey, 2003; Weber and Cutler, 2004). Such difficulties in L2 lexical processing have been attributed to L2 listeners' inability to access or utilize relevant phonological cues, which in turn may be the result of inability or difficulty to identify non-native phonemes in the L2 acoustic signal (Broersma, 2012; Broersma and Cutler, 2008, 2011; Qin, 2017).

While there is abundant work demonstrating L2 listeners' difficulties in speech perception and lexical access separately, only few studies have attempted to investigate the relation between the two directly. The goal of the present study is to examine this relation in the context of a linguistic property well-known to be challenging in L2 acquisition, namely lexical tone. Over half of the world's languages are tonal languages, where pitch as a phonological category is part of lexical representations alongside segmental information (Yip, 2002).[1] Mandarin, one of the most widely spoken and studied tonal languages and the target language of the present study, has four lexical tones, which are often described numerically according to fundamental frequency (F0) in a system known as 'Chao tone letters', dividing the natural pitch range of the normal speaking voice into five levels, with 1 as the lowest and 5 as the highest (Chao, 1930; T1 = 55, T2 = 35, T3 = 214, T4 = 51). In Mandarin, tones can distinguish between two words consisting of the same segments. For example, /ma/ with high level pitch (T1) means ‘mother’, but with rising pitch (T2) means ‘hemp’. The important linguistic status of pitch in tonal languages is captured by its representation on a separate tier from segmental and prosodic material in underlying mental representations, although in the surface realization of words, tone can only be realized in association with segments (Yip, 2002).

Lexical tone presents a novel phonological contrast to L2 learners from non-tonal L1 (e.g. English) backgrounds, thus raising challenges at the levels of both speech perception and lexical access. Previous research, discussed in more detail below, has investigated the challenges presented by lexical tone in L2 speech perception and word recognition separately, but little remains known about what Wong and Perrachione (2007: 565) have referred to as the ‘phonetic–phonological–lexical continuity in adult nonnative word learning’. In the present study, we examine to what extent L1 and L2 speakers of Mandarin perceive tone categorically, as well as to what extent they use tonal cues along with segmental cues for lexical access during real-time listening. This allows us to then examine to what extent the two are related. To this end, we report findings from a visual world eye-tracking experiment designed to investigate the role of tonal cues in Mandarin word recognition, as well as from a tone identification task, assessing listeners’ ability to perceive Mandarin tone categorically. Based on the findings from these two tasks, we address the following two research questions:

- Research question 1: How do L1 and L2 listeners weigh tonal cues relative to segmental cues in Mandarin spoken word recognition?
- Research question 2: How does L2 learners’ use of tonal cues in spoken word recognition relate to their ability to perceive tone categorically?

II Background

1 Processing of tone in L1

While at the phonological level Mandarin has four tone categories, described by discrete labels as illustrated above, the acoustic realization of tone in actual speech varies substantially with different talkers, contexts and speech rate (Jongman et al., 2006). Thus, a critical part of processing words in tonal languages involves identifying discrete tone categories from highly variable acoustic realizations. To study the categorical perception of tone, experimental research uses synthetic stimuli from tone-pair continua created by manipulating pitch (and potentially other suprasegmental parameters, such as duration). Research going back to Wang (1976), who conducted identification and discrimination tasks with synthetic tokens varying along the T1–T2 continuum, has found that L1 Mandarin listeners perceive tone more categorically than naive L1 English listeners with no experience of tonal languages. Subsequent studies including continua of all six possible tone pairs in Mandarin have confirmed these findings (Hallé et al., 2004; Ling and Schafer, 2016; Peng et al., 2010; Shen and Froud, 2016; Xi et al., 2010; Xu et al., 2006).

At the level of lexical processing in tonal languages, most previous work has focused on the independent contributions of tones versus segments in the process of lexical access. Early studies reached a general agreement that tones may be a weaker cue than segments. By using a homophone judgment task of written characters in Mandarin, Taft and Chen (1992) found that native listeners took longer and were less accurate when a pair of words differed only by tone than only by vowel quality. Cutler and Chen (1997) examined how native speakers of Cantonese (a tonal language similar to Mandarin, but with six tones) process lexical tones in spoken words. Results from a lexical decision task showed that listeners were most likely to accept a non-word as a word when the only difference between the two was in tones. A speeded same-different judgment task with the same participants showed slower and less accurate responses when the only difference between two words was in tones. The researchers concluded that lexical tone might be a weaker cue due to the comparatively late availability of the tonal cue compared to segmental cue. Similar results were obtained by Ye and Connine (1999) in a simple tone-vowel detection task. However, they also found that tone might play a more important role when presented in context. As pointed out by Liu and Samuel (2007), the tasks employed in many early studies are mostly sensitive to sub-lexical processing, where listeners could respond without accessing lexical meaning.

In tasks requiring listeners to access lexical representations, findings have indicated more similar roles for tonal and segmental cues. Malins and Joanisse (2010) conducted a real-time spoken word recognition experiment using the visual word paradigm, in which native Mandarin speakers listened to words and were asked to select the corresponding picture from a set of four, including the target (e.g. *chuang2* ‘bed’), a competitor overlapping with the target in different phonological domains (segmental: *chuang1* ‘window’; cohort: *chuan2* ‘ship’; rhyme: *huang2* ‘yellow’; tonal: *niu2* ‘cow’), and two phonologically unrelated distractors. Of critical interest were the segmental and cohort conditions, where the divergence of competitors from targets was argued to occur at comparable time points (see appendix in Malins and Joanisse, 2010). Eye gaze patterns did not differ between the two conditions, leading the authors to conclude that ‘tonal and segmental information are accessed concurrently and play comparable roles’ in word recognition by Mandarin native speakers (Malins and Joanisse, 2010: 407). A follow-up study (Malins and Joanisse, 2012) using an event-related potential (ERP) paradigm supported the conclusion that tonal and segmental cues were both accessed as soon as they were available, while also indicating potentially different underlying processing mechanisms for tonal versus segmental information.

So far, we have considered two lines of research on the role of tone relative to segments, each leading to somewhat different conclusions. In experiments encouraging sub-lexical processing and not requiring access to lexical meanings, tone appears to be a weaker cue than segments (Cutler and Chen, 1997; Ye and Connine, 1999). In spoken word recognition tasks, which require full lexical access, tones and segments appear to be on a more equal footing (Malins and Joanisse, 2010, 2012). A third line of research, focusing specifically on the interaction of tonal and segmental cues, provides evidence for the dynamic relationship between different cues, and draws attention to the limitations of isolating individual cues in experimental contexts, emphasizing the importance of processing tonal cues along with segmental cues as one entity in successful word recognition. In a form priming task, Sereno and Lee (2015) found weak priming effects when the prime and target matched only in segmental content, but much larger effects when they matched in both segmental and tonal content (see also Lee, 2007), indicating that the syllable as a unit, rather than its individual component parts, may play a critical role in word processing. Using the Garner speeded classification paradigm, Tong et al. (2008) investigated the interaction between tones and segments from a different angle. In this task, participants have to classify stimuli according to a specific dimension (e.g. consonant, tone, or vowel) while ignoring variation on other dimensions. Results from native Mandarin speakers showed that variation in vowel quality interfered more in the classification of stimuli by tone than the reverse, which led the authors to conclude that the processing of tone was not independent, but integrated into the processing of vowels. Further evidence comes from another speeded classification experiment conducted by Lin and Francis (2014). Using stimuli constituting legitimate words in both English and Mandarin, they found that L1 Mandarin listeners showed symmetric interference between consonants and tone, regardless of whether stimuli were presented in an English or a Mandarin context, while L1 English listeners did not show any interference from the non-target dimension in either direction. They suggested that native Mandarin listeners process tonal and segmental cues in a more integrated manner regardless of the ambient language, while English listeners with no experience of tonal languages processed them more separately.

These findings are consistent with ‘selective perception routines (SPRs)’ as proposed in Strange’s (2011) Automatic Selective Perception (ASP) model: L1 Mandarin listeners have developed the routine of perceiving and processing tonal cues along with segmental cues for lexical access through life-long experience with their tonal L1. This L1 SPR constitutes a highly over-learned pattern so that even when instruction and stimuli are in English, native Mandarin listeners are unable to inhibit the automatic processing of tone. On the other hand, English listeners unaccustomed to processing pitch at a lexical level might treat pitch as intonation and process it separately from segmental content, since in English, suprasegmental features are rarely distinctive at a lexical level. Even in some rare cases, when English words only differ from each other in stress, English listeners do not appear to use such suprasegmental information during lexical access (Cutler, 1986). English listeners are thus more likely to process suprasegmental features such as pitch at a post-lexical level, without immediate impact on lexical access (Lin & Francis, 2014). Framed within the ASP model, the task for L2 learners then is to inhibit entrenched L1 SPRs, and learn new SPRs to optimize use of the most reliable linguistic cues in the L2. A key goal of our study is to gain a better understanding of the extent to which English learners of Mandarin are able to do so in the context of using tonal and segmental cues during lexical processing.

2 Processing of tone in L2

While lexical tone presents a novel phonological dimension for L2 learners with non-tonal L1s, previous work provides ample evidence that listeners with non-tonal language backgrounds are sensitive to pitch contrasts (Hallé et al., 2004), and that accuracy of tone identification can be improved significantly even after short-term training (Wang, 2013; Wang et al., 1999; Wong and Perrachione, 2007). At the same time, L2 learners, even those with intermediate to advanced proficiency, show persistent difficulty with the processing of tone at a lexical level (Pelzl et al., 2019; Qin, 2017). This gap between perception and lexical processing of tones has begun to be addressed in a recent study by Pelzl et al. (2019). They found that although English-speaking learners of Mandarin were as successful as native Mandarin speakers at identifying tones in isolated syllables in a tone identification task, the same learners were less likely to correctly reject non-words in a lexical decision task than native speakers when the non-words and words differed only by tone, indicating a 'disconnect between L2 abilities to categorize tones as phonetic objects and abilities to utilize those categories as lexical cues' (Pelzl et al., 2019: 69). A third task using an ERP paradigm provided further neurophysiological evidence of learners' difficulty in using tone as a cue in lexical processing. These findings present important evidence for a dissociation between phonological and lexical processing ability related to tone in L2, based on group-level analyses comparing L2 with L1 speakers on different tasks. Remaining unexplored is the relation between these abilities at the level of the individual learner, a question that we aim to address in the present study.

Many factors can contribute to this disconnect between phonological and lexical processing of tones by L2 learners. At a phonological level, although L2 learners were able to achieve native-like accuracy in identifying the four standard Mandarin tones in isolated syllables (Pelzl et al., 2019), L2 listeners may have more difficulty handling the lack of invariance of tone realization in actual speech, where they have to assign a specific token quickly into a phonological tone category. In a study by Ling et al. (2016), L2 learners of Mandarin (L1 English) with various proficiency, along with native Mandarin listeners and native English listeners with no experience of Mandarin, participated in an identification task designed to assess categorical perception of tone. In each trial, listeners heard a token from one of the six synthesized continua of all possible Mandarin tone pairs (e.g. T1–T2, T1–T3) with manipulated pitch, and they were instructed to decide which tone they heard in each pair. Results showed that L2 learners patterned between the other two groups on the steepness of identification slopes, suggesting L2 learners perceive tone less categorically than native speakers, yet more categorically than listeners with no experience of a tonal language. Moreover, the steepness of identification slopes, i.e. the degree of categorical perception, correlated significantly with learners' proficiency in Mandarin. These findings are consistent with those of Shen and Froud (2016), who reported results from advanced learners of Mandarin on a similar identification task with T1–T4 and T2–T3 pairs showing that advanced learners' performance was similar to that of native speakers. However, in a follow-up ERP study (Shen and Froud, 2019), native Mandarin speakers, but not L2 learners, showed electrophysiological evidence of categorical perception of tones. L2 learners also showed larger P300 responses, suggesting more attentional focus in the processing of tones. As Shen and Froud (2019: 263) pointed out, '[i]f phonetic categorization by adult learners is less efficient and requires more attention compared to native Chinese speakers, higher-level processes such as lexical access may also be more challenging for learners.' More specifically, we might expect L2 learners to rely on tonal cues less than native listeners at a lexical level. This is consistent with Pelzl et al.'s (2019) finding that L2 learners were more likely to reject segmentally mismatched than tonally mismatched non-words, while native listeners rejected them at equal rates, indicating that, compared to native listeners, L2 learners allocated less weight to tones relative to segments.

In addition to the bottom-up difficulty of drawing phonological information from acoustic input, the top-down influence of highly over-learned L1 automatic selective perception routines (SPRs; Strange, 2011) may also make L2 learners less likely than native speakers to use tonal cues in word recognition. In English, suprasegmental features are rarely lexically distinctive. Although some English words differ from each other only in stress, English listeners do not appear to use such suprasegmental information during lexical access, presumably because such cases are exceedingly rare (Cutler, 1986). In Mandarin, on the other hand, tone is a key component of lexical representations and is indispensable in word processing.

The question that arises is whether long-term exposure and learning experience lead to increased use of tonal cues in L2. Zou et al. (2017) addressed this question using an ABX classification task with beginner and advanced Dutch learners of Mandarin. Results showed that, like the naive Dutch control group, beginner learners with 8–20 months learning experience were less accurate at classifying stimuli based on tones alone than Mandarin native speakers. The advanced learners with 3–14 years learning experience showed significantly higher accuracy than both the Dutch control group and the beginner learners, and did not differ significantly from the native Mandarin speakers. These findings indicate that the development of tone acquisition in L2 involves increasing cue strength on tonal cues along with improving proficiency. However, this study only investigated processing at a phonological level, in a task that did not involve lexical access. We thus still do not know whether advanced L2 learners will use tonal and segmental cues to similar extents as native listeners in word recognition. It remains possible that during a more resource-demanding, higher-level process such as lexical access, attention to the comparatively less familiar tonal cue will remain reduced even in more advanced learners. We explore this possibility here in a visual-world eye-tracking experiment, inspired by Malins and Joanisse (2010), in which we investigate how L1 and L2 speakers of Mandarin weigh tonal cues along with segmental cues in real-time word recognition.

As mentioned earlier, despite the wealth of research on tone in speech perception and lexical processing, very little is known about the relation between the two, especially in L2. As far as we are aware, only one study has directly examined this ‘phonetic–phonological–lexical continuity’ (Wong and Perrachione, 2007: 565) in L2 tone learning. In their study, English speakers with no experience of tone languages participated in multisession training during which they were exposed to English pseudo-words with pitch patterns resembling Mandarin tones. A tone identification task and a spoken word recognition task were conducted to measure learners’ ability of perceiving pitch patterns and word learning success. Results showed a significant correlation between participants’ performance in the two tasks, supporting a continuity of phonetic–phonological and lexical abilities in adult L2 word learning. The generalizability of these findings to the acquisition of tone in L2 Mandarin, however, is somewhat limited given that stimuli consisted of English pseudo-words with Mandarin tones, and participants were naive English listeners with no learning experience of a tonal language. Inspired by Wong and Perrachione’s (2007) work, we aim to shed further light on the phonetic–phonological–lexical continuity in L2 tone learning by exploring the relation between performance on a categorical perception and a real-time word recognition task among English-speaking L2 learners of Mandarin.

III Methods

1 Participants

A total of 30 native and 34 L2 speakers of Mandarin participated in this study. Data from 5 L2 participants was excluded due to exposure to Chinese in childhood (4) or professional music experience[2] (1), leaving data from 29 L2 speakers for analysis. Native speakers (21 female, mean age: 25.6 years, range: 20–36) were recruited from among the international student community at the University of Hawai‘i at Mānoa (*n* = 20), as well as at Peking University (*n* = 10). All native speakers reported being born in Mainland China and self-identified as native speakers of Mandarin.

L2 learners (11 female, mean age = 24.4 years, range: 19–41) were recruited at the University of Hawai‘i at Mānoa (*n* = 3), as well as at Peking University, the University of Hong Kong and Chinese University of Hong Kong (*n* = 26). All L2 learners self-identified as native speakers of English and started to learn Mandarin after age 12 (mean age of onset: 20.0 years, *SD* = 4.5). To ensure basic familiarity with the Mandarin vocabulary used in the experimental materials, only participants who were taking or had taken 3rd-year Chinese classes (or above) in the USA, or intermediate/advanced classes in China, were admitted to the study. L2 proficiency was assessed through self-ratings of speaking (*M* = 2.8, *SD* = 1.1), listening (*M* = 2.9, *SD* =1.1) and reading (*M* = 3.0, *SD* = 1.0) skills on a 5-point scale, as well as through a listening proficiency test adapted from the Hanyu Shuiping Kaoshi (HSK or Chinese Standard Exam) level 4 (Confucius Institute in Atlanta, 2017). The HSK level 4 test is a boundary test to differentiate intermediate to advanced proficiency. One L2 learner did not complete the listening proficiency test. The remaining 28 scored from 40% to 100% (*M* = 76.9%, *SD* = 17.5%). Information on music experience and language experience was collected by questionnaire. The study protocol was approved by the Institutional Review Board at University of Hawai‘i at Mānoa, and participants were compensated with extra course credit or a small amount of money.

2 Spoken word recognition task

a Materials. Linguistic stimuli in the visual-world eye-tracking experiment consisted of 12 sets of 5 monosyllabic words. All words were easily imageable common nouns, and were composed of a consonant onset and a rhyme (for a complete list of stimuli, see Appendix 1). Each set consists of (1) a target (e.g. *gou3* ‘dog’); (2) a segmental competitor (SC: *gou1* ‘hook’), which completely matches the target in segmental content but differs in tone; (3) a rhyme competitor (RC: *shou3* ‘hand’), which matches the target in segmental and tonal content of the rhyme but differs in onset; (4) a vowel competitor (VC: *dou4* ‘bean’), which matches the target in segmental but not tonal content of the rhyme and also differs in onset; and (5) a distractor (*qiu2* ‘ball’), which does not share either segmental or tonal content with the target.[3] The SC and VC competitors were included to allow for the critical comparison between competitors that differ from the target in tone only (SC) versus competitors that differ in both tone and segmental (onset) content (VC). Rhyme competitors (RC), which share tone and vowel but not onset segmental content with the target, were included to assess potential late co-activation due to overlapping rhymes (Allopenna et al., 1998).

In order to examine potential differences in word frequency between stimulus types, we used word frequency indices (log10W) from the SUBTLEX-CH corpus (Cai and Brysbaert, 2010). A one-way ANOVA showed no significant differences between targets, the three types of competitors, and distractors ($F(4,55) = 1.58, p = .19$). Since indeces of frequency in an L1 corpus may not be fully reflective of frequency experienced by L2 learners, we also used HSK vocabulary level as an index of word difficulty for L2 learners: level 1 ‘easiest’ to level 6 ‘most difficult’). Words not listed in the HSK vocabulary were given a value of 7 (5 words). L1 word frequency (log10W) and HSK level index correlated moderately ($\tau = -0.50, p < 0.001$), indicating some consistency between the two values. A one-way ANOVA with HSK level as the dependent variable showed no significant differences between stimulus types ($F(4,55) = 1.24, p = 0.30$) either. Differences in word frequency were thus unlikely to greatly influence looks to targets, competitors, and distractors in the present study.

Visual scenes contained three areas of interest (AOIs): the target, one of the three competitors (SC, RC, or VC), and a distractor (Figure 1). Participants saw each target in three conditions (SC, RC, or VC), for a total of 36 experimental trials. The location of different AOIs was rotated across conditions. The order of items was pseudo-randomized, and interspersed with 54 filler trials. Fillers were constructed from the same 60 words used in the experimental trials, but words constituting competitors or distractors in experimental trials acted as targets in filler trials. Fillers were created to approximately balance the occurrence of each image as a named vs. unnamed referent. All participants were presented with the same 90 trials in two blocks, with 18 experimental trials and 27 fillers in each block. Block order was counterbalanced across participants. Three initial practice trials similar to filler trials familiarized participants with the task.

The auditory stimuli were produced by an adult female native speaker of Mandarin at a slow speed in a sound-proof booth at 44.1 kHz and recorded in Praat (Boersma and Weenink, 2016). Each noun was spoken preceded by the carrier phrase *qing3xuan3 . . .* (‘please choose . . .’) three times in citation form with full realization of tones. One token of each noun was selected according to intensity and sound quality. Nouns were then extracted and concatenated with the same token of the carrier phrase for all items, with noun onset 2,042 ms after onset of the carrier phrase. Average duration of target nouns was 658 ms (range: 429–913). Two native speakers of Mandarin confirmed the naturalness of the concatenated sentences.

b Procedure. Prior to the visual-world eye-tracking experiment, participants completed a self-paced vocabulary familiarization task followed by a naming test, in order to ensure understanding of all vocabulary and word-image associations. In the familiarization task, participants were presented with all 60 words (12 sets * 5 words) in random order in PsychoPy (Peirce, 2007). Each word was presented auditorily once together with its corresponding image, Chinese characters and English gloss. Participants were instructed to take their time to familiarize themselves with each word and image before pressing the space bar to move on. They were told they would be tested in a naming task afterwards. In order to make sure participants paid attention and were familiar with the

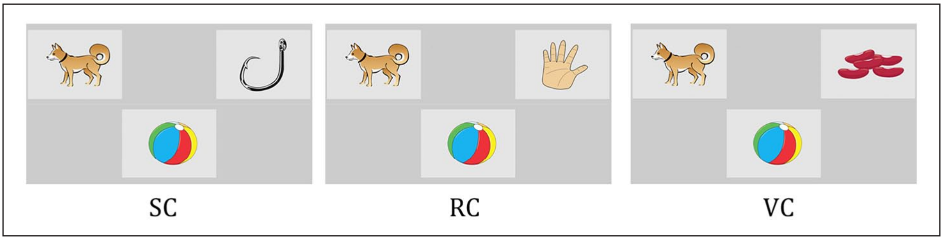

Figure 1. Examples of visual scenes in the three conditions.
Note. Location of areas of interest (AOIs) was rotated in the actual materials. SC = segmental competitor. RC = rhyme competitor. VC = vowel competitor.

word-image associations, 10 words were selected and used as test trials in the naming task. To ensure L2 learners were familiar with the tested words, we asked three Chinese instructors to check the stimuli and select the 10 words their students might have most difficulty with. For the 10 test items in the naming task, all 5 words not listed in HSK level 1–6 were included in addition to another 5 words selected as difficult by instructors. In each test trial, participants saw an image and after 500 ms they heard a beep and were required to name the picture. Participants needed to produce 9/10 words with correct pronunciation of segments to pass the naming task; otherwise they were asked to repeat the familiarization task and naming test until they met criterion. All native listeners and 11 L2 learners passed the naming task the first time. 15 L2 learners repeated the familiarization task once and 3 L2 learners repeated it twice. No feedback was provided during the naming task.

After passing the naming task, participants proceeded to the main part of the spoken word recognition experiment. The experiment was conducted on an SMI RED250 eye-tracker sampling at 250 Hz (for participants tested in Hawai‘i), or a mobile REDn Scientific eye-tracker sampling at 60 Hz (for participants tested in China). Participants were instructed to click on one of the three images in the scene after listening to the auditory instruction *qing3xun3* (‘Please choose’) + NOUN, preceded by 1,500 ms preview of the visual scene. Mouse-click and eye fixation were recorded through SMI ExperimentSuite software. Fixation data were binned into 20 ms samples. Preliminary analyses showed no differences in the structure of the data from the two trackers, thus all data was combined for further analysis.

3 Identification task

a Materials. One syllable (/pi/) was used to create critical stimuli, and another syllable (/kwo/) was used for creating practice stimuli. Both syllables occur with all four tones in Mandarin. To avoid effects of co-articulation and tone sandhi, the stimuli were constructed from isolated Mandarin words spoken by a female native speaker of Mandarin (the same speaker as in the other task) three times in a row with slow speed and a clear voice. The stimuli were recorded at 44.1 kHz in Praat. Tokens were selected based on both intensity and quality. Four tokens (4 tones) of /pi/ were selected and used to construct tone pair continua for each of the six possible tone pairs in Mandarin (e.g. T1–T2). Six 9-step tone continua were generated by equalizing duration within the continuum, and linearly changing the pitch and intensity between the end points with the PSOLA method (Moulines and Laroche, 1995) in Praat (Boersma and Weenink, 2016). Six 4-step tone continua for /kwo/ were synthesized in the same manner and used for the practice trials. Two native Mandarin speakers confirmed the naturalness of all stimuli.

b Procedure. Stimuli were blocked by tone pair to reduce listeners' memory load and avoid uninformative answers (e.g. the middle step of the T2–T4 continuum has a flat pitch, which is similar to T1). Each block started with information about the upcoming tone pair and two exposure trials with tone labeled endpoints. Blocks were counterbalanced across participants. All stimuli were presented three times in random order within each block, resulting in a total of 162 trials (6 tone pairs * 9 steps * 3 presentations). An initial practice block was used to familiarize participants with the task by using the T1–T2 continuum with the syllable /kwo/. Participants were asked to classify each token by pressing the appropriate tone-number key on the keyboard (e.g. *1* or *2* in block T1–T2), followed by the space bar to initiate the next trial at their own pace. Participants were encouraged to guess if unsure.

4 General procedure

Before coming to the lab, all participants completed a web-based questionnaire to collect information on basic demographics, language background, music experience, and self-ratings of speaking, listening and reading ability in both Mandarin and English. In the lab session, all participants completed the visual-world eye-tracking experiment followed by the identification task. L2 learners additionally completed the listening proficiency test at the end.

IV Results

1 Data trimming

Data from a total of 2,124 trials (59 participants, 36 experimental items) on the eye-tracking experiment was first inspected for valid mouse-click responses. Trials with no mouse-click (L1: 1, L2: 1) and trials in which the participant clicked on an image before noun onset (L1: 4) were excluded, as were trials in which the timing of the click exceeded 3 *SD*s of the group's average reaction time (RT) (L1: 16, L2: 17). The remaining data were inspected for track loss. Trials with more than 16% (= $M+3SD$) missing sample points were excluded (L1: 25, L2: 4). In all, a total of 3.2% of the data (68/2124 trials; L1: 4.3%, L2: 2.1%) was discarded.

2 Spoken word recognition task

a Mouse-click data. Participants' accuracy in selecting the named target is illustrated in Figure 2 (left). While the L1 group showed similar accuracy rates across conditions,

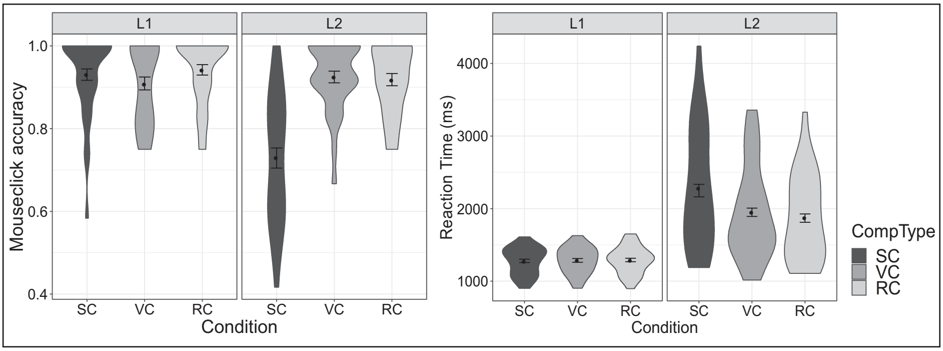

Figure 2. Accuracy (left) and reaction time (right) by group and condition.
Notes. The dot represents the mean in each condition and group. Error bars indicate one standard error by participant in each condition and group. SC = segmental competitor. RC = rhyme competitor. VC = vowel competitor.

accuracy in the L2 group was substantially lower in the SC condition than in the other two. For statistical analysis, accuracy data were submitted to a generalized linear mixed effect model. This and all subsequent statistical analyses were conducted in R (version, 3.6.0, R Core Team, 2019), using the lme4 package (version 1.1-21, Bates et al., 2015). Fixed effects included Group (L1, L2; contrast-coded and centered), Condition (SC, RC, VC; simple coded with VC as reference level) and their interactions. Maximal random effect structures justified by the design were attempted, and reduced if convergence problems arose (Barr et al., 2013). Model comparisons were carried out using the *anova()* function to identify the best-fitting model.

Table 1 presents the output of the best-fitting model. The significant negative estimate for group ($b = -0.72, p = .01$) indicates lower accuracy overall in the L2 than in the L1 group. The significant negative estimate for SC ($b = -1.33, p = .03$) indicates overall lower performance in the SC than in the (reference-level) VC condition. In other words, participants were less accurate in selecting the named target when there was a competitor that differed only in tone versus a competitor differing in both tone and segmental content. Importantly, this effect interacted with Group ($b = -2.04, p < .001$), prompting follow-up analyses within each group separately. Within the L1 group, there were no significant differences in accuracy across conditions (all $p > .4$). In the L2 group, on the other hand, accuracy in the SC condition was significantly lower compared to the VC ($b = -2.24, p < .001$) and RC ($b = -3.10, p = .002$) conditions, with no significant differences between the latter two ($b = 0.86, p = .49$).

Analyses of reaction time (RT) in trials with correct mouse-clicks (Figure 2 right) echoed the results from the analysis of accuracy. RTs in the L2 group (M_{L2} = 2,036, SD_{L2} = 716) were substantially longer overall than in the L1 group (M_{L1} = 1,285, SD_{L1} = 184), indicating generally greater difficulty in recognizing Mandarin words among L2 learners. Statistical analysis was conducted using mixed-effect models of the inverse-gaussian family due to the skewed distribution of the RT data (Lo and Andrews,

Table 1. Results of generalized linear mixed-effects model for accuracy for Formula (glmer): accuracy ~ group * condition + (1 | participant) + (1 + condition | item).

	b	SE	z	p
Intercept	3.04	0.23	13.07	< .001
Group	−0.72	0.28	−2.59	.01
RC	0.18	0.89	0.20	.84
SC	−1.33	0.61	−2.18	.03
Group × RC	−0.68	0.45	−1.51	.13
Group × SC	−2.04	0.40	−5.09	< .001

2015). Otherwise the same modeling strategies were followed as in the analysis of accuracy. Results from the best-fitting model *RT ~ group * condition + (1 s| participant) + (1 + condition | item*) confirmed that the L2 group took longer than the L1 group in making correct choices ($b = 700.69, p < .001$), and that participants took longer in the SC than in the VC condition ($b = 111.47, p = .02$). The interaction between Group and Condition (SC vs. VC) was significant ($b = 181.19$, $p < .001$), prompting follow-up analyses within each group. In the L1 group, there were no differences in RT by condition (SC vs. VC: $b = 20.78, p = .29$; RC vs. VC: $b = 2.44, p = .90$; SC vs. RC: $b = -18.35, p = .37$). The L2 group, by contrast, took substantially longer to make correct choices in the SC condition compared to the VC ($b = 194.13, p < .001$) and the RC ($b = 216.17$, $p < .001$) condition, with no difference between the latter two ($b = -21.93, p = .46$).

In sum, the L2 group achieved accuracy comparable to the L1 group in the RC and VC conditions, and within the L2 group, learners were equally fast on correct target selections in these two conditions. In the SC condition, on the other hand, where tone was the only cue distinguishing the target from the competitor, L2 participants were significantly less accurate than L1 participants, and took longer on correct selections than in the other two conditions. L2 participants also showed substantially more variability on both accuracy and RT in the SC condition than L1 participants (Figure 2). It is possible that this variability stems from the inclusion of L2 participants who were unable to distinguish words by tone alone, and were thus simply guessing in the SC condition. In order to identify such participants, we examined the probability of a participant guessing in the SC condition based on a binomial distribution. Assuming that the critical choice was between the target and the competitor (even though there was a third, phonologically unrelated distractor in the scene), chance was assumed to be at .5. Adopting an alpha level of .05, the binomial distribution indicates that correct responses on at least 9 out of 12 items represents performance significantly above chance. All participants in the L1 group met this criterion, as did 15 out of the 29 L2 learners. We will refer to this subgroup as the 'L2-above-chance learners'. The remaining 14 L2 participants were at chance ('L2-at-chance learners'). Proficiency measured on the listening task was higher in the L2-above-chance ($M = 0.86$, $SD = 0.13$) than the L2-at-chance ($M = 0.67$, $SD = 0.17$) subgroup ($b = -0.19, p = .002$).

In order to examine whether L1–L2 differences in the SC condition persist when comparing only L2 learners with statistically significant sensitivity to tones (the above-chance-subgroup) with L1 speakers, we reran the analysis of accuracy reported above with Group treated as a 3-level rather than a 2-level factor (L1, L2-above-chance, L2-at-chance; simple coded with L1 as reference level). Results showed no significant difference in overall accuracy between the L2-above-chance and the L1 group ($b = -0.005$, $p = .99$), while the L2-below-chance group performed significantly below both ($bs > |1.20|, ps < .001$). Interactions between Group and Condition (SC–VC) remained significant for both the L2-at-chance vs. L1 ($b = -2.17, p < .001$) and the L2-above-chance vs. L1 ($b = -1.80, p = .001$) comparisons, but were non-significant for the L2-at-chance vs. L2-above-chance comparison ($b = -.37$, $p = .5$). Within-group analyses showed no significant differences between the RC and VC conditions in either L2 subgroup. In the L2-at-chance group, accuracy in SC was significantly worse than in the VC condition ($b = -2.06, p < .001$); in the L2-above-chance group, this difference was only marginally significant ($b = -1.58, p = .053$).

We thus find the pattern of results from the initial comparison between the L1 and L2 groups repeated in the comparison between the L1 and the L2-at-chance subgroup. This is unsurprising given that this L2 subgroup was defined by chance performance when the recognition of the target critically required reliance on tone. More importantly, we also find the pattern largely repeated, though somewhat weaker, in the comparison between the L1 and the L2-above-chance group. Notably, the interaction between group and the SC–VC comparison remained significant, and follow-up analysis within the L2-above-chance group still showed a marginal trend towards lower accuracy in the SC than the VC condition. Analogous analyses of RT on correct responses further showed that, unlike in the L1 group (see above), RTs in the SC vs. the VC condition were longer in both the L2-at-chance ($b = 258.35, p = .002$) as well as the L2-above-chance ($b = 179.05, p < .001$) subgroups. These findings suggest that even for L2 learners with demonstrated above-chance ability to recognize target nouns by tone alone, performance does not fully mirror that of L1 speakers.

b Eye-movement data. In order to further explore these differences between the L2-above-chance and the L1 groups, we investigated the time course of participants' looks to targets and competitors in the visual scene as they were listening to the noun in real time. Figure 3 illustrates L1 and L2-above-chance participants' looking patterns in the SC, VC and RC conditions on trials in which they selected the correct target. Visual inspection of fixation patterns in the L1 group shows little evidence of competition in any condition, with looks to competitors decreasing sharply, along with looks to phonologically unrelated distractors, about 200 ms after the onset of the noun. In the L2-above-chance group, looks to the target in the VC and RC conditions increase on a similar timescale as in the L1 group, but asymptote at a lower level. At the same time, looks to competitors remain more persistent, a pattern that appears particularly evident in the SC condition.

Statistical analysis was conducted to address our research question on L1 and L2 listeners' relative weighting of tonal and segmental cues. Specifically, our goal was to assess whether competition from a competitor differing only in tone would be stronger

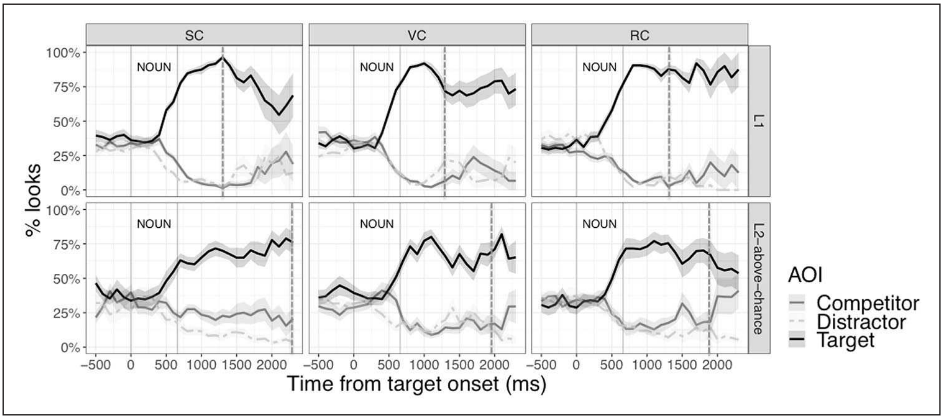

Figure 3. Proportion of looks to Target, Competitor and Distractor by condition and group in trials with correct mouse-click responses.

Notes. Zero on the *x*-axis along with the first vertical gray line represents the acoustic onset of the target noun. The second vertical line represents mean target noun offset. The dashed line represents average mouse-click reaction time by group and condition. Ribbons show one standard error.

than from a competitor differing in both tone and segmental content, and whether this effect would be more pronounced in the L2 than in the L1 group. To this end, we compared the proportion of looks to the competitor (versus the target) in the SC versus the VC and RC conditions in both groups. The large difference in RT between the two groups (see above), however, raised the difficult question of the appropriate time period within which to analyse these looking patterns. We decided to honor the variability in the timing of participants' decisions, as captured already by RT, and focus on a participant-driven time window, extending from 200 ms after noun onset (taking into consideration the time needed to execute a ballistic eye movement; Matin et al., 1993) until mouse-click, i.e. until the participant selected the (correct) target in a given trial. Within this period, which varied by trial, we calculated the proportion of frames with fixations to the competitor out of fixations to target and competitor combined. This measure captures the proportion of time the participant spent looking at the competitor before making a final decision.

A linear mixed-effect model with Group (L1, L2-above-chance; contrast-coded and centered) and Condition (simple-coded, VC as reference) as fixed effects was fitted to these data. Given the highly non-normal distribution of the outcome measure at the trial level, models at the trial level including both random effects for participants and items proved to be a poor fit. We therefore decided to aggregate data over participants and over items, and fit two separate models to each (Barr, 2008). Table 2 presents the output from the best-fitting models, which showed similar patterns in the by-participant and by-item aggregations. The main effect of Group was significant ($b_1 = 0.09, p_1 < .001; b_2 = 0.11, p_2 < .001$), indicating that the L2 learners were overall more likely than native speakers to look at competitors. An overall trend for more looks to competitors in the SC (vs. VC) condition also emerged ($b_1 = 0.04, p_1 = .02; b_2 = 0.03, p_2 = .10$). This trend did not interact with Group, yet in light of our research question, we decided to explore its nature

Table 2. Results of linear mixed-effect model for proportion of looks to competitor, aggregated over participants and over items.

	b	SE	t	p
*Formula (lmer): PropCompetitor ~ group * condition + (1 | participant):*				
Intercept	0.20	0.01	21.82	< .001
Group	0.09	0.02	4.51	< .001
RC	−0.01	0.02	−0.96	.34
SC	0.04	0.02	2.39	.02
Group × RC	0.03	0.03	0.93	.35
Group × SC	0.04	0.03	1.34	.18
*Formula (lmer): PropCompetitor ~ group * condition + (1 | item):*				
Intercept	0.21	0.02	11.44	< .001
Group	0.11	0.02	6.45	< .001
RC	−0.01	0.02	−0.44	.66
SC	0.03	0.02	1.65	.10
Group × RC	0.03	0.04	0.81	.42
Group × SC	0.03	0.04	0.73	.47

further through models fit to the data from each group separately. In the L1 group, no differences between SC vs. VC condition ($b_1 = -0.02, p_1 = .16; b_2 = -0.03, p_2 = .35$) or RC vs. VC condition ($b_1 = 0.02, p_1 = .21; b_2 = 0.02, p_2 = .48$) emerged. Within the L2 above-chance group, a significant difference was found between the SC and VC condition in the by-participant ($b_1 = 0.06, p_1 = .04$) but not in the by-item data ($b_2 = 0.05, p_2 = .18$); no significant differences were observed between the RC and VC conditions ($b_1 = 0.005, p_1 = .86; b_2 = 0.007, p_2 = .83$). In sum, even in trials with correct mouse-click, L2 listeners with the ability to discriminate words by tone showed more consideration of competitors overall, and tended to look at competitors more when tone was the only differing cue between targets and competitors than when they differed in both tone and segmental content; native speakers, by contrast, did not show any differences between conditions.

3 Identification task

Following standard procedures in categorical perception studies, we calculated the proportion of participants' Sound-A vs. Sound-B responses (e.g. Sound A = T1, Sound B = T2, in T1–T2 pair) for each tone pair and step. The results, collapsed over the six different tone pairs, are illustrated in Figure 4. At both endpoints, participants in all groups were highly accurate at identifying tone on tokens with natural pitch and intensity, indicating no general difficulty in perceiving standard tones on isolated syllables. The slope of identification curves serves as a measure of the degree of categorical perception, with

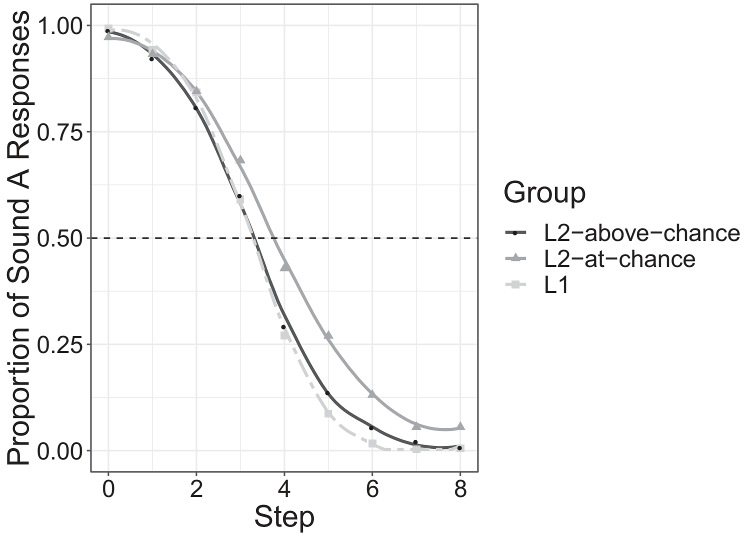

Figure 4. Identification curve averaged across participants and tone pairs by group.

steeper slopes indicating more categorical perception (Xu et al., 2006). Visual inspection of Figure 4 shows that the L2-at-chance group had the shallowest curve and the L1 group had the steepest curve, while the L2-above-chance group patterned between the two. Slope values for each participant and tone pair were submitted to a linear mixed effect model with Group as a fixed effect and participants and tone pairs as random effects. Result showed that the slope for the L2-above-chance group ($M = -1.82$, $SD = 0.57$) was significantly steeper than for the L2-at-chance group ($M = -1.49$, $SD = 0.60$; $b = 0.34$, $p = .002$), but also significantly shallower than for the L1 group ($M = -2.57$, $SD = 0.61$; $b = -0.75$, $p < .001$).

Finally, in order to investigate whether listeners' ability to perceive tone categorically was related to their use of tonal cues in spoken word recognition, we conducted correlation tests between participants' identification slopes and their proportion of looks to the competitor in the SC condition in the visual world task. Recall that the latter was calculated over trials in which participants correctly selected the target. Since the number of such trials was low and variable in the L2-at-chance group, which was likely guessing in the SC condition, we confine this analysis to the L2-above-chance and the L1 groups. As illustrated in Figure 5, whereas no significant relation was found in the L1 group ($\tau = 0.17$, $p = .19$), a strong positive correlation emerged in the L2-above-chance group ($\tau = 0.63$, $p = .001$), showing that learners with steeper identification slopes were less likely to look at competitors. These findings suggest that difficulty with using tonal cues in L2 spoken word recognition is related to the ability to perceive tone categorically at a phonological level.

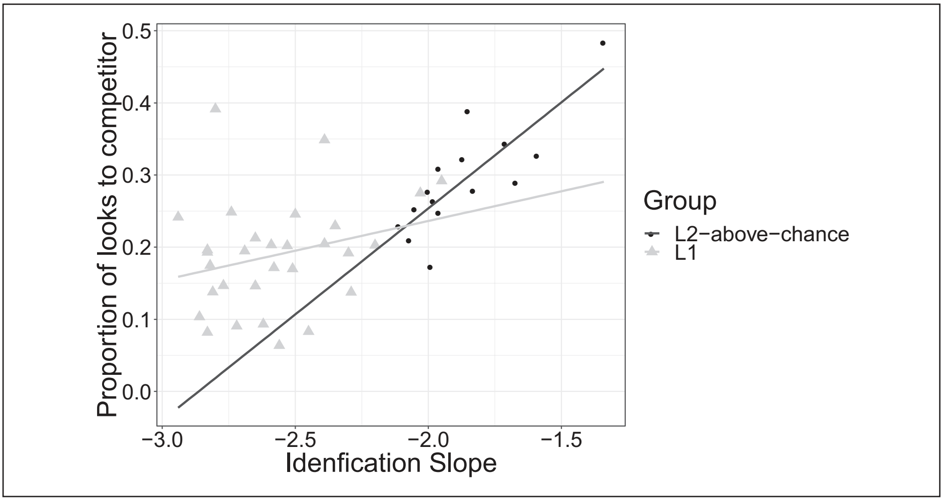

Figure 5. Scatterplot of identification slope (*x*-axis) and proportion of looks to competitor (*y*-axis) by group.
Notes. More negative slope values indicate steeper slopes and are indicative of more categorical perception.

V Discussion

The goal of this study was to investigate how L2 learners of Mandarin make use of tone during real-time word recognition (RQ1), and whether their use of tonal cues in word recognition relates to their ability to perceive tone categorically on isolated syllables (RQ2). To address RQ1, we used the visual world paradigm to assess L1 and L2 listeners' ability to use tone as a distinguishing cue in real-time word recognition. For this purpose, our analyses focused on the comparison between the SC condition, where target (e.g. *gou3* 'dog') and competitor (SC: *gou1* 'hook') overlapped completely in segmental content but differed by tone, and the VC condition, where target and competitor (VC: *dou4* 'bean') differed in segmental content as well. A third condition in which target and competitor (RC: *shou3* 'hand') differed in onset but not rhyme was included to examine potential co-activation of rhyme competitors as found in previous work on English (Allopenna et al., 1998). Notably, Malins and Joanisse (2010) reported no significant rhyme effects for L1 Mandarin speakers in their study. We obtained the same outcome in this study for both L1 and L2 speakers: no significant differences emerged between the RC and VC conditions for accuracy, RT, or looks to competitors, in either group. As recent work by Teruya and Kapatsinski (2019) has demonstrated, however, rhyme effects appear to be confined to disyllabic words, where the overlap between target and competitor is more extensive (e.g. *speaker–beaker* in Allopenna et al., 1998). In both Malins and Joanisse's (2010) and our study, all stimuli were monosyllabic. The absence of rhyme competition is thus consistent with Teruya and Kapatsinski's (2019) observation, and will need further investigation in Mandarin in studies including disyllabic words. We will therefore confine the remainder of our discussion to the comparison between the SC and VC conditions most directly relevant to RQ1.

The overall comparison between the L1 and L2 groups showed that the latter were slower and less accurate in spoken word recognition. However, these main effects of Group interacted significantly with Condition. With regard to accuracy, the L2 group differed from the L1 group only in the SC condition, indicating that L2 learners were able to recognize words based on segmental differences as well native speakers (albeit still more slowly), yet they had considerably more difficulty in using tone alone to distinguish between words. This is consistent with Pelzl et al.'s (2019) observation that L2 learners were less accurate than native speakers at rejecting tonally, but not segmentally, mismatching non-words in a lexical decision task. Our findings add further evidence from a more ecologically valid listening task with natural stimuli showing that L2 learners allocate less weight to tonal cues than native listeners during language processing.

Further analyses within the L2 group indicated that even though we had admitted to the study only learners who were at least at the 3rd-year-Chinese level or equivalent by standards of US college-level instruction, almost half of these learners (14/29) were not significantly above chance at distinguishing words by tone alone. This is striking testimony to the observation by the Foreign Service Institute (US Department of State, 2016) that Mandarin is 'exceptionally difficult for native English speakers', with the acquisition of lexical tones known to be one of the most challenging aspects for adult L2 learners of Mandarin (Wang et al., 2006). When comparing only the L2-above-chance (n = 15) subgroup with the L1 group, the main effect of Group on accuracy disappeared; however, the interaction with Condition remained significant. Follow-up analyses showed a remaining marginal trend in the L2-above-chance subgroup towards lower performance in the SC compared to the VC condition. In order to further explore these remaining differences between native speakers and more advanced L2 learners who are able to differentiate words by tone alone above chance levels, we compared looking patterns to targets and competitors during real-time listening in the L1 and the L2-above-chance groups.

Overall, the L2 learners spent proportionally more time looking at the competitor before clicking on the (correct) target than native speakers, indicating greater uncertainty in all conditions. A main effect of Condition also emerged, indicating more consideration of competitors in the SC than in the VC condition; interestingly, this effect was not qualified by an interaction with Group, suggesting greater competition when words differed only by tone in both groups. Further exploratory inspection of this effect within each group, however, showed no difference between the SC and the VC condition for the L1 group. The effect thus appears to be driven predominantly by the performance of the L2 learners, and it is possible that the reduced number of participants in this L2 subgroup did not afford sufficient power to detect an interaction.

Returning to RQ1, our findings from real-time word recognition provide strong evidence that distinguishing words by tone alone remains difficult even for advanced learners of Mandarin. We have shown that even learners who are able to accomplish this task with above-chance accuracy take substantially more time to do so than native speakers, and show more uncertainty in the process, as indicated by proportionally more looks to competitors minimally differing by tone only. This persistent between-group difference is consistent with Strange's (2011) Automatic Selective Perception (ASP) model: While the learners in the L2-above-chance group have clearly acquired enough knowledge of tone to distinguish between words in most cases – as indicated by their above-chance accuracy in the SC condition – they still appear to rely predominantly on their L1 selective perception routines (SPRs), with focus predominantly on segmental contrasts, during L2 lexical processing. In other words, they do not (yet) appear to have developed sufficiently automatized selective perception routines that allocate tone the weight it has in L1 processing.

Whether native-like SPRs are ever attained in L2 development is a critical question that we must leave for future research with long-term immersed and highly proficient L2 learners to further explore. The present study does, however, allow us to consider the issue of L2 development to some extent, namely by looking at the relationship between learners' use of tonal cues in word recognition and their ability to perceive tone categorically: our second research question. In order to allow us to address RQ2, a 9-step tone identification task was included, modeled after standard procedures in the categorical perception literature (e.g. Hallé et al., 2004). As expected, native listeners showed steeper identification slopes, indicating more categorical perception, than L2 learners. When we divided the L2 group into above-chance and at-chance subgroups as defined by performance on the word recognition task, we found significantly steeper slopes in the L2-above-chance than in the L2-at-chance group, although the identification slopes of both were significantly shallower than those in the L1 group. The L2-above-chance group also performed significantly better than the L2-at-chance group on the independent listening comprehension task. This is consistent with a previous study by Ling et al. (2016), which showed a correlation between proficiency and the degree of categorical perception of tone among L2 learners of Mandarin. These findings indicate that increasing language experience can lead to more categorical perception of tone, even among learners whose L1 does not instantiate this phonological contrast. At the same time, our finding that the more advanced L2 subgroup still differed significantly from the L1 group contrasts with the results obtained by Shen and Froud (2016), who found no significant differences between advanced learners and native speakers. However, several factors could have contributed to this inconsistency. In particular, Shen and Froud (2016) included only two tone pairs (T1–T4 and T2–T3), and the number of participants in their study was more limited (10 L1 and 10 L2 speakers), thus the statistical power to detect any between-group effects may have been more limited. Importantly, the two studies converge in the observation that L2 listeners tend to perceive tone more categorically with increasing learning experience.

Returning to RQ2, our critical analysis consisted of assessing potential correlations between participants' performance on the identification and the word recognition tasks. More specifically, we assessed correlations between participants' slope parameters on the identification task with their proportional looking to segmental competitors in the visual world experiment. Only trials with correct final selections in the visual-world task were included. No significant correlations were obtained within the L1 group, potentially due to limited variance on both tasks: as expected in performance relying on highly automatized routines. For the L2 group, we confined our analyses to the L2-above-chance subgroup, i.e. the learners who demonstrated that they had the ability to reliably distinguish words by tone alone. Within this group, we found a strong and highly significant correlation between performance on the two tasks: Learners who perceived tone more categorically were also less likely to look at segmental competitors. This observation constitutes the first evidence that we are aware of that the ability to perceive tone categorically and the use of tonal cues in lexical processing are directly related at the level of individual learners in the L2 acquisition and processing of Mandarin. As such, these findings provide support for Wong and Perrachione's (2007) claims about the continuity between phonetic, phonological and lexical skills in L2 tone learning. We caution, however, that our findings show correlation, not causation. Future work, including longitudinal data and training studies, will be needed to identify the causal direction of these effects. Yet the strong contingency between the two that we have observed here provides a promising starting point for such further investigations, leading to potentially important insights for L2 training and curriculum design in the future.

VI Conclusions

Drawing on evidence from categorical perception and real-time spoken word recognition, we found that English-speaking L2 learners of Mandarin, even those with considerable L2 experience, differed from native Mandarin speakers in both the extent to which they perceived tone categorically as well as in their ability to use tonal cues to distinguish between words in real-time listening comprehension. At the same time, we observed substantial variability among L2 learners' performance on both tasks, with at least some of this variability due to more experienced learners showing patterns of performance more similar to those observed in the L1 group. Critically, the present study provides the first direct evidence showing that the ability to perceive tone categorically is related to the weighting of tonal cues during spoken word recognition among adult L2 learners with demonstrated ability of distinguishing words by tone alone. This finding supports Wong and Perrachione's (2007) conclusions, which were based on a laboratory-based artificial-language training study with naive participants, regarding the importance of the continuity between phonetic, phonological and lexical abilities in the learning of tone, and extends them to the acquisition of lexical tone in Mandarin by L2 learners with language experience gained outside the laboratory. A better understanding of the links between learners' developing abilities across linguistic domains, often studied in isolation from each other in different research subfields, is essential not only for theoretical models of L2 development and processing, but also for more applied purposes such as curriculum design and instruction in Mandarin-as-a-foreign-language contexts, where the acquisition of tone remains a topic of central concern.

Acknowledgements

We thank Amy Schafer for insightful suggestions, Shiyi Lu for sharing lab space at Peking University, and Yongyi Wu and Elaine Lau for helping recruit participants in China. We are also grateful for valuable feedback from audiences at the Second Language Research Forum 2018 (SLRF 2018), Boston University Conference on Language Development 2018 (BUCLD 2018), and the Language Acquisition Reading Group at the University of Hawai‘i, as well as from two anonymous *Second Language Research* reviewers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (BCS-1824082), Ruth Crymes Memorial Grant, and Elizabeth Carr Holmes Scholarship.

ORCID iDs

Wenyi Ling https://orcid.org/0000-0003-1889-5706
Theres Grüter https://orcid.org/0000-0001-6354-9787

Notes

1. In some African and American languages, tones can be used to convey morphological, syntactic, semantic and pragmatic information. Here we focus on lexical tones only.
2. Previous studies found that musicians were better than non-musicians at identifying, discriminating, and imitating lexical tones (e.g. Bidelman et al., 2013).
3. Due to the limited number of natural words forming such sets of five, we were unable to fully control the tone pairs in the stimuli. However, the number of different tone types were approximately balanced across targets, the tree types of competitors and distractors.

References

Allopenna PD, Magnuson JS, and Tanenhaus MK (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38: 419–39.

Barr DJ (2008) Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language* 59: 457–74.

Barr DJ, Levy R, Scheepers C, and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–78.

Bates D, Machler M, Bolker B, and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Bidelman GM, Hutka S, and Moreno S (2013) Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLoS One* 8: e60676.

Boersma P and Weenink D (2016) *Praat: Doing phonetics by computer: Version 6.0.16* [computer software]. Available at: https://www.fon.hum.uva.nl/praat (accessed June 2020).

Broersma M (2012) Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes* 27: 1205–24.

Broersma M and Cutler A (2008) Phantom word activation in L2. *System: An International Journal of Educational Technology and Applied Linguistics* 36: 22–34.

Broersma M and Cutler A (2011) Competition dynamics of second-language listening. *Quarterly Journal of Experimental Psychology* 64: 74–95.

Cai Q and Brysbaert M (2010) SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One* 5: e10729.

Chao YR (1930) A system of tone letters. *Le Maître phonétique* 45: 24–27.

Confucius Institute in Atlanta (2017) *HSK vocabulary and sample tests*. Atlanta, GA: Confusicus Institute in Atlanta. Available at: http://confucius.emory.edu/hsk_and_resources/hsk/hsk_samples.html (accessed June 2020).

Cutler A (1986) Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech* 29: 201–20.

Cutler A (2012) *Native listening: Language experience and the recognition of spoken words*. Cambridge: MIT Press.

Cutler A and Chen H (1997) Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics* 59: 165–79.

Dijkstra T, Timmermans M, and Schriefers H (2000) On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language* 42: 445–64.

Dupoux E, Sebastián-Gallés N, Navarrete E, and Peperkamp S (2008) Persistent stress 'deafness': The case of French learners of Spanish. *Cognition* 106: 682–706.

Hallé PA, Chang YC, and Best CT (2004) Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics* 32: 395–421.

Jongman A, Wang Y, Moore C, and Sereno JA (2006) Perception and production of Mandarin Chinese tones. In: Bates E, Tan LH, and Tseng O (eds) *Handbook of Chinese Psycholinguistics: Volume 1: Chinese*. Cambridge: Cambridge University Press, pp. 209–17.

Lee CY (2007) Does horse activate mother? Processing lexical tone in form priming. *Language and Speech* 50: 101–23.

Lin M and Francis AL (2014) Effects of language experience and expectations on attention to consonants and tones in English and Mandarin Chinese. *The Journal of the Acoustical Society of America* 136: 2827–38.

Ling W and Schafer AJ (2016) Tone pair similarity and the perception of Mandarin tones by Mandarin and English listeners. *Proceedings of the 5th International Tonal Aspects of Language (TAL)*. Available at: https://isca-speech.org/archive/TAL_2016/pdfs/20-Ling-Schafer.pdf (accessed June 2020).

Ling W, Schafer AJ, and Grüter T (2016) Identification and discrimination of tone by L2 learners of Mandarin. Unpublished oral presentation at the Second Language Research Forum, New York, NY, USA.

Liu S and Samuel AG (2007) The role of Mandarin lexical tones in lexical access under different contextual conditions. *Language and Cognitive Processes* 22: 566–94.

Lo S and Andrews S (2015) To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology* 6: 1171.

Malins JG and Joanisse MF (2010) The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language* 62: 407–20.

Malins JG and Joanisse MF (2012) Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia* 50: 2032–43.

Marian V and Spivey M (2003) Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition* 6: 97–115.

Matin E, Shao KC, and Boff KR (1993) Saccadic overhead: Information-processing time with and without saccades. *Perception and Psychophysics* 53: 372–80.

Moulines E and Laroche J (1995) Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication* 16: 175–205.

Peirce JW (2007) PsychoPy-psychophysics software in Python. *Journal of neuroscience methods* 162: 8–13.

Pelzl E, Lau EF, Guo T, and DeKeyser R (2019) Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition* 41: 59–86.

Peng G, Zheng HY, Gong T, et al. (2010) The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics* 38: 616–24.

Qin Z (2017) How native Chinese listeners and second-language Chinese learners process tones in word recognition: An eye-tracking study. Unpublished PhD thesis, University of Kansas, KS, USA.

Qin Z, Chien Y, and Tremblay A (2017) Processing of word-level stress by Mandarin-speaking second language learners of English. *Applied Psycholinguistics* 38: 541–70.

R Core Team (2019) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (accessed June 2020).

Sereno JA and Lee H (2015) The contribution of segmental and tonal information in Mandarin spoken word processing. *Language and Speech* 58: 131–51.

Shen G and Froud K (2016) Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America* 140: 4396–96.

Shen G and Froud K (2019) Electrophysiological correlates of categorical perception of lexical tones by English learners of Mandarin Chinese: An ERP study. *Bilingualism: Language and Cognition* 22: 253–65.

Strange W (2011) Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics* 39: 456–66.

Strange W and Dittmann S (1984) Effects of discrimination training on the perception of /r–l/ by Japanese adults learning English. *Perception and Psychophysics* 36: 131–45.

Taft M and Chen H (1992) Judging homophony in Chinese: The influence of tones. In: Chen H Tzeng OJL (eds) *Language processing in Chinese*. Oxford: North-Holland, pp. 151–72.

Teruya H and Kapatsinski V (2019) Deciding to look: Revisiting the linking hypothesis for spoken word recognition in the visual world. *Language, Cognition and Neuroscience* 34: 861–88.

Tong Y, Francis AL, and Gandour JT (2008) Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes* 23: 689–708.

US Department of State, Foreign Service Institute, School of Language Studies (2019) *Languages*. Washington, DC: Author. Available at: https://www.state.gov/key-topics-foreign-service-institute/foreign-language-training (accessed June 2020).

Wang WSY (1976) Language change. *Annals of the New York Academy of Sciences* 280: 61–72.

Wang X (2013) Perception of Mandarin tones: The effect of L1 background and training. *The Modern Language Journal* 97: 144–60.

Wang Y, Jongman A, and Sereno JA (2006) L2 acquisition and processing of Mandarin tone. In: Bates E, Tan LH, and Tseng O (eds) *Handbook of Chinese Psycholinguistics: Volume 1: Chinese*. Cambridge: Cambridge University Press, pp. 250–56.

Wang Y, Spence MM, Jongman A, and Sereno JA (1999) Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America* 106: 3649–58.

Weber A and Cutler A (2004) Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language* 50: 1–25.

Werker JF and Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7: 49–63.

Wong PC and Perrachione TK (2007) Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics* 28: 565–85.

Xi J, Zhang L, Shu H, Zhang Y, and Li P (2010) Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience* 170: 223–31.

Xu Y, Gandour JT, and Francis AL (2006) Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* 120: 1063–74.

Ye Y and Connine CM (1999) Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes Special Issue: Processing East Asian Languages* 14: 609–30.

Yeung HH, Chen KH, and Werker JF (2013) When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language* 68: 123–39.

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

Zou T, Chen Y, and Caspers J (2017) The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of Mandarin tones. *Bilingualism: Language and Cognition* 20: 1017–29.

Appendix 1. Experimental stimuli and English gloss in parentheses.

Stimuli set	Target	Segmental competitor	Rhyme competitor	Vowel competitor	Distractor
1	Cha1 (fork) [2.33; 5]	Cha2 (tea) [3.10; 1]	Sha1 (sand) [2.87; 5]	Fa4 (hair) [3.80; 3]	Bi3 (pen) [3.52; 3]
2	Cheng2 (orange) [1.99; 6]	Cheng4 (scale) [1.64; 6]	Sheng2 (rope) [2.56; 5]	Deng4 (stool) [1.79; 7]	Guo1 (pot) [2.54; 5]
3	Chi3 (ruler) [2.79; 5]	Chi4 (wing) [1.74; 5]	Zhi3 (finger) [3.76; 5]	Shi1 (lion) [2.22; 4]	Mao4 (hat) [2.66; 3]
4	Dao3 (island) [3.47; 5]	Dao1 (knife) [3.44; 4]	Cao3 (grass) [2.87; 3]	Pao4 (cannon) [2.43; 5]	Jian4 (arrow) [3.02; 7]
5	Di2 (flute) [2.07; 7]	Di4 (floor) [4.26; 3]	Bi2 (nose) [2.59; 3]	Ji1 (chicken) [3.47; 2]	Gua1 (melon) [2.48; 2]
6	Fang2 (house) [3.38; 2]	Fang1 (square) [3.23; 5]	Tang2 (candy) [3.18; 3]	Tang1 (soup) [3.00; 4]	Huo3 (fire) [3.55; 4]
7	Gou3 (dog) [4.07; 1]	Gou1 (hook) [2.43; 6]	Shou3 (hand) [4.18; 3]	Dou4 (bean) [2.61; 5]	Qiu2 (ball) [3.85; 2]
8	Gu3 (bone) [2.89; 5]	Gu1 (mushroom) [1.08; 7]	Shu3 (mouse) [2.74; 5]	Ku4 (pants) [2.76; 3]	Xiang4 (elephant) [3.70; 5]
9	Jing1 (whale) [1.95; 7]	Jing3 (well) [2.68; 6]	Bing1 (ice) [3.20; 6]	Ting2 (pavilion) [1.94; 5]	He2 (river) [3.06; 3]
10	Shao2 (spoon) [1.97; 5]	Shao4 (whistle) [1.91; 6]	Tao2 (peach) [2.50; 5]	Bao1 (bag) [3.57; 3]	Xie2 (shoes) [3.37; 3]
11	Tu4 (rabbit) [2.60; 5]	Tu2 (picture) [3.06; 3]	Shu4 (tree) [3.33; 3]	Shu1 (book) [3.85; 1]	Nao3 (brain) [3.26; 5]
12	Zhu1 (pig) [3.30; 4]	Zhu2 (bamboo) [1.90; 6]	Shu1 (comb) [2.11; 5]	Gu3 (drum) [2.61; 4]	Niao3 (bird) [3.34; 3]

Note. The first value in square brackets denotes word frequency (log10W) of each item from the SUBTLEX-CH corpus and the second value denotes Hanyu Shuiping Kaoshi (HSK or Chinese Standard Exam) level.