frontiers
in Communication

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/fcomm.2021.689423

Advanced Second Language Learners of Mandarin Show Persistent Deficits for Lexical Tone Encoding in Picture-to-Word Form Matching

Eric Pelzl[1,2]*, *Ellen F. Lau*[1], *Taomei Guo*[3] *and Robert M. DeKeyser*[1]

[1]*University of Maryland, College Park, MD, United States,* [2]*The Pennsylvania State University, University Park, PA, United States,* [3]*Beijing Normal University, Beijing, China*

People who grow up speaking a language without lexical tones typically find it difficult to master tonal languages after childhood. Accumulating research suggests that much of the challenge for these second language (L2) speakers has to do not with identification of the tones themselves, but with the bindings between tones and lexical units. The question that remains open is how much of these lexical binding problems are problems of *encoding* (incomplete knowledge of the tone-to-word relations) vs. *retrieval* (failure to access those relations in online processing). While recent work using lexical decision tasks suggests that both may play a role, one issue is that failure on a lexical decision task may reflect a lack of learner confidence about what is *not* a word, rather than non-native representation or processing of known words. Here we provide complementary evidence using a picture-phonology matching paradigm in Mandarin in which participants decide whether or not a spoken target matches a specific image, with concurrent event-related potential (ERP) recording to provide potential insight into differences in L1 and L2 tone processing strategies. As in the lexical decision case, we find that advanced L2 learners show a clear disadvantage in accurately identifying tone mismatched targets relative to vowel mismatched targets. We explore the contribution of incomplete/uncertain lexical knowledge to this performance disadvantage by examining individual data from an explicit tone knowledge post-test. Results suggest that explicit tone word knowledge and confidence explains some but not all of the errors in picture-phonology matching. Analysis of ERPs from correct trials shows some differences in the strength of L1 and L2 responses, but does not provide clear evidence toward differences in processing that could explain the L2 disadvantage for tones. In sum, these results converge with previous evidence from lexical decision tasks in showing that advanced L2 listeners continue to have difficulties with lexical tone recognition, and in suggesting that these difficulties reflect problems both in encoding lexical tone knowledge and in retrieving that knowledge in real time.

Keywords: second language, Mandarin, lexical tone, ERPs, speech perception, Chinese, fuzzy lexical representations

OPEN ACCESS

Edited by:
Anna Chrabaszcz,
University of Pittsburgh, United States

Reviewed by:
Seth Wiener,
Carnegie Mellon University,
United States
Wanting Huang,
The University of Hong Kong, China

****Correspondence:***
Eric Pelzl
pelzlea@gmail.com

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *31 March 2021*
Accepted: *02 June 2021*
Published: *22 June 2021*

Citation:
Pelzl E, Lau EF, Guo T and DeKeyser RM (2021) Advanced Second Language Learners of Mandarin Show Persistent Deficits for Lexical Tone Encoding in Picture-to-Word Form Matching. Front. Commun. 6:689423. doi: 10.3389/fcomm.2021.689423

INTRODUCTION

People who grow up speaking a language without lexical tones typically find it difficult to master tonal languages after childhood. They may confuse or misidentify tones in speech early on (e.g., Wang et al., 1999), and they often end up with a large store of *fuzzy* second language (L2) tone word representations, that is, mental lexical representations with missing, incorrect, or uncertain tone representations (Pelzl et al., 2020). This outcome is not surprising, given that F0 (pitch) is used for many things in non-tonal languages (stress, intonation, emphasis, singing), but does not differentiate one word from another.

By far the most studied L2 tone language is Mandarin Chinese (for a review, see Pelzl, 2019). Mandarin has four citation tones, differentiated primarily by F0 height and contour (Howie, 1974; Ho, 1976). Relative to a speaker's own vocal pitch range, Mandarin Tone 1 is high and level; Tone 2 rises from mid to high; Tone 3 is low; and Tone 4 falls from high to low. Along with consonants and vowels, these tones serve to uniquely identify each syllable-sized unit (typically a morpheme or word) of spoken Mandarin.

Misidentification of Mandarin tones is common among naïve and novice learners (e.g., Wang et al., 1999; Alexander et al., 2005; Bent et al., 2006; Huang and Johnson, 2010; So and Best, 2010). For more experienced learners, tone identification and categorization abilities improve and many individuals approach native levels on categorization and identification tasks (Ling and Grüter, 2020; Pelzl, 2019; Shen and Froud, 2016; Tsukada and Han, 2019; Zou et al., 2017). Nevertheless, similarities between F0 contours among Mandarin tones can lead to confusions among some tones (e.g., in isolated syllables, Tone 3 may have a dipping contour leading it to resemble Tone 2). Such confusions can persist into intermediate and advanced levels of L2 proficiency (Lee et al., 2010; Hao, 2012; Tsukada and Han, 2019).

Given that tone identification is already a challenge, it is not surprising that using tones to differentiate words in Mandarin is also difficult for many novice learners (Wong and Perrachione, 2007; Chandrasekaran et al., 2010; Chang and Bowles, 2015). Perhaps less expected is that the same difficulties appear to persist into more advanced stages of learning, even for many learners who have achieved strong categorization or identification abilities (Han and Tsukada, 2020; Ling and Grüter, 2020; Pelzl et al., 2019; Pelzl et al., 2020). We will refer to this type of difficulty as an L2 *tone word* difficulty, that is, it is not necessarily about tones alone, but about how representations of the tone categories are bound to the lexical representations in long-term memory.

This L2 tone word difficulty may best be understood as *phonolexical* in nature. A difficult L2 phonological contrast (tones) impacts the learner's mental representations of the relevant lexical units, leading to fuzzy representations of lexical tones. While this situation is similar to documented segmental learning challenges in other L2 contexts (Díaz et al., 2012; Darcy et al., 2013; Chrabaszcz and Gor, 2014; Cook and Gor, 2015; Amengual, 2016; Cook et al., 2016; Gor and Cook, 2020; Llompart and Reinisch, 2020), L2 tone word difficulties may also differ in important ways. For instance, learning a novel L2 vowel contrast may require a learner to add new categories in their phonological vowel space—a challenge addressed by many models of L2 phonological learning (Best and Tyler, 2007; Escudero and Boersma, 2004; Flege, 1995). However, using vowels to differentiate meaningful lexical units is already a given. For tones, non-tonal language speakers must not only learn to categorize F0 contrasts on syllable-sized units, they must also learn to apply these new tone categories to the process of word recognition. This is a functional leap that may not come easily.

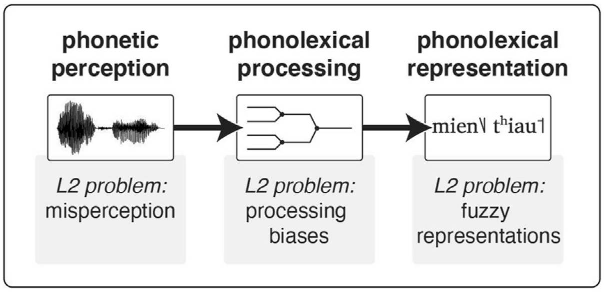

phonetic perception
phonolexical processing
phonolexical representation
mien˥˩ tʰiau˥
L2 problem: misperception
L2 problem: processing biases
L2 problem: fuzzy representations

FIGURE 1 | Word recognition "pipeline" depicting three sites of potential difficulty in L2 (tone) word recognition: phonetic perception, phonolexical processing, and phonolexical representation.

When it comes to the fuzzy L2 lexical representations that result from such difficulties, tone word representations are much like those of purely segmental representations. They will vary from low to high quality, and this is likely to be closely related to a learner's familiarity with those words (Diependaele et al., 2013; Gor and Cook, 2020). An individual word's representation may have tones that are uncertain, incorrect, or completely missing (Pelzl et al., 2020). Each of these problems has the potential to impede fluent spoken word recognition. In this study, we set out to examine how this happens in more detail, specifically asking 1) whether tone word recognition errors will persist even when we control for fuzzy L2 word knowledge, and 2) whether L2 listeners' neural responses for correctly recognized words will display early sensitivity to tones.

Three General Explanations for L2 Tone Word Processing Difficulties

To provide some theoretical context for L2 difficulties with tone word recognition, we begin with a rough sketch of three basic ways in which it might break down (for more detail, see Pelzl et al., 2019, Pelzl et al., 2020). We use the metaphor of a pipeline to capture the way these issues feed one into another (**Figure 1**).

First, an L2 listener may have difficulty in what we are calling *phonetic perception,* that is, accurately perceiving the unfamiliar sounds of the L2. Specific to tones, many beginning and novice L2 learners regularly misidentify or confuse similar tone categories when recognizing words or syllables (e.g., Chandrasekaran et al., 2010; Chang and Bowles, 2015; Wang et al., 1999; Wong and Perrachione, 2007). Our own previous research suggests many advanced learners develop excellent tone identification abilities for monosyllables (Pelzl et al., 2019), and other studies have also found strong—if not completely nativelike—L2 categorization of tones among more advanced learners (Ling and Grüter, 2020; Shen and Froud, 2016, 2019; Zou et al., 2017). While overall impressive, such results are not a claim for flawless L2 tone perception. Though not examined in detail here, tone identification results from the participants in the current study show that advanced L2 listeners may struggle to identify tones when there is a following syllable (for details, see Pelzl, 2018), showing accuracy that is notably lower than L1 participants for the same context. So then, weaknesses in tone identification remain a possible cause of persistent L2 tone word recognition difficulties. If listeners cannot faithfully perceive tones in the acoustic-phonetic signal, they will have difficulty using and encoding them. In this case, the breakdown occurs near the beginning of the word recognition pipeline (**Figure 1**). The "substance" (the perceived speech sounds) that enters the pipeline is already problematic. This could well have knock-down effects leading to fuzzy L2 lexical representations (cf. Matusevych et al., 2021).

Second, an L2 listener may have difficulty processing the perceived speech sounds as lexical cues. We are calling this *phonolexical processing*. This is the real-time process that links the perceived phonetic signal to phonolexical representations encoded in long-term memory. It roughly corresponds to the phonological step of the phonetic-phonological-lexical continuum (Wong and Perrachione, 2007; Chan and Leung, 2020), but we wish to stress the *lexical* aspect of this process, along with the phonological. For L2 learners, years of experience attending only to the important phonetic features of their L1 might interfere in real-time word recognition. Such ideas have long been part of L2 theories, under a variety of terms (e.g., *cue competition* MacWhinney and Bates, 1989; *selective perception routines* Strange, 2011; *perceptual attention* Chang, 2018). In the case of L2 tone learning, listeners might privilege segmental information over tonal information. This need not be all or nothing; contextual factors or a learner's wider knowledge about the language might impact when and how tones are used. For example, Wiener et al. (2018), Wiener et al. (2020) have shown that when speech is produced by several different speakers (as opposed to just one), L2 learners tend to rely more on their experiential knowledge of syllable + tone co-occurrence probabilities (i.e., which tones are most likely with which syllables), and less on the acoustic-phonetic signal itself. When L2 processing strategies do not give appropriate weight to tones, it can lead to errors or inefficiencies during lexical retrieval. With respect to our pipeline, in this case the function of the pipe itself is the problem. Accurately perceived spoken tone words enter, but key features get siphoned off before they reach their destination.

Finally, an L2 listener may have difficulty with *phonolexical representations*, that is, encoding the lexical units of speech in their mental lexicon. This difficulty may lead to a variety of issues for tone representations: they may be entirely missing, incorrect, only known with some degree of uncertainty, or even confidently incorrect (Pelzl et al., 2020). If a given representation is not faithful to the actual form of the spoken tone word, this has the potential to lead to a variety of problems during lexical access. If lexical activation is very strict, the appropriate lexical unit might fail to be activated due to misaligned tone representations. If lexical activation is more lenient, inappropriate competitors could become active during lexical competition (Broersma and Cutler, 2008; Broersma, 2012). This representational account puts the problem at the end of the pipeline. The perceived speech sounds enter and run through the pipe smoothly, but the destination is incorrect.

Each of these difficulties is likely to make its own contribution to the fuzziness of L2 tone word representations. In previous work (Pelzl et al., 2019), we found that advanced L2 Mandarin learners (native English speakers) displayed near-native abilities on a challenging tone identification task, suggesting excellent phonetic perception of tones. In that task, we used syllables clipped out of disyllabic words that had been produced in continuous speech. Despite their strong performance when identifying those tones in isolation, when we presented the disyllabic words themselves—many of which contained the very same spoken syllables as used in the identification task—most L2 learners performed below chance in rejecting tonal nonword competitors of real Mandarin words (e.g., nonword _fang4*zi_/fɑŋ4tsɹ/derived from real word *fang2zi* "house"). This extreme difficulty was only seen for tone nonwords, not vowel nonwords (e.g., nonword *feng2zi*/fəŋ2tsɹ/). Taken together, the excellent tone identification paired with chance performance for tone nonwords suggested to us that auditory tone perception is unlikely to be the primary source of L2 tone word recognition difficulty. A follow-up study used more clearly produced (non)words and once again found a strong difference between tone and vowel nonwords (Pelzl et al., 2020).

The Present Study

Although our previous studies provided compelling evidence of advanced L2 Mandarin learners' tone word difficulties, the use of lexical decision as the primary test may have painted a particularly dismal picture. Rejection of a nonword in a lexical decision task requires a lexical search on the part of listeners (i.e., to confirm that a nonword does *not* exist in the lexicon). So then, a person's failure on any given lexical decision trial may reflect their lack of confidence about what is *not* a word, rather than their fuzzy representation or processing of the targeted real word.

To specifically target L2 learners' knowledge and recognition of lexical tone in real words, we conducted a picture-phonology matching experiment with native Mandarin speakers and advanced L2 learners of Mandarin (cf. Desroches et al., 2009). In the picture-phonology matching task, people see an image (noodles) and then hear a word that either matches or does not match the image. The auditory targets are real words (*mian4tiao2* "noodles"), or nonwords with a mismatching vowel (*men4tiao2*) or tone (*mian1tiao2*). Unlike lexical decision, a picture-phonology matching trial requires only knowledge of the specific word targeted in a trial. If the listener can successfully bring that word to mind, their task is simply to determine whether the auditory stimulus matches it or not. Here we allowed a full 1.75 s of picture-viewing time before the onset of the word, to provide L2 listeners with plenty of time to bring the word to mind. Thus, if the listener knows the pictured word and can faithfully perceive the spoken prompt, they should be able to confidently reject the mismatching nonwords. Another motivation for the picture-phonology matching task is to place L2 word recognition into a meaningful—albeit very simple—context. While tests of isolated word recognition can be a useful tool for understanding lexical processes, most words do not occur in isolation; typically, contextual cues help listeners create expectations about what words they expect to hear. The current paradigm mirrors this real-world situation, but uses a simple picture context to avoid complications that can arise in interpreting how much of a prior sentence context L2 learners have access to (see Pelzl et al., 2019 for discussion).

Our primary question was whether, when a prior visual context was provided in this way, advanced L2 learners of Mandarin would still show the same kind of disadvantage with lexical tone information vs. vowel information that we observed in the experiments on lexical decision. If the disadvantage we observed in prior experiments was primarily due to learners' lack of confidence about what tone words they *haven't* heard, then it might disappear when the task is focused only on determining the match of a picture to a known word. However, if the disadvantage is due to phonolexical encoding or processing of tone, then it should persist under the current conditions. We were also interested in gaining more insight into how much of any tone disadvantage observed is due to phonolexical encoding vs. processing. Therefore, we conducted an offline vocabulary knowledge post-test so that we could determine whether L2 listeners persist in (incorrectly) accepting tone mismatches even when they have correct and confident knowledge of the relevant words.

More exploratorily, during the picture-phonology matching experiment we also collected concurrent electrophysiological responses in order to look for signs of differential processing of lexical tone in native and L2 listeners that might explain different behavioral performance. Because the smaller number of incorrect behavioral responses in this paradigm are insufficient for ERP analysis, we focused on examining the ERPs from trials that had a correct behavioral response. Although these represent cases in which the L2 learners, like the native listeners, succeeded in accepting real words or rejecting tone/vowel mismatches, this real-time neural data could suggest differential approaches to lexical processing that could explain the profile of errors observed in the L2 learners. In the next section, we briefly review some background on the ERP responses that might provide such clues.

The Phonological Mismatch Negativity and Late Positive Component Responses in Event-Related Potential Research

The picture-phonology matching task sets up strongly constraining lexical expectations. Prior ERP research suggests that native speakers performing such a task with words and near-neighbor nonwords are likely to show modulation in three ERP components: the phonological mismatch negativity (PMN), the N400, and a late positive component (LPC). However, because the N400 component overlaps with the other two and can be modulated by nonword status itself as well as real word expectation (cf. Newman and Connolly, 2009), we chose to focus on the PMN and LPC responses in the current study.

The phonological mismatch (or mapping) negativity typically occurs between 200 and 400 ms after stimulus onset and is hypothesized to index neural responses to unexpected/mismatching phonological content in words, relative to expected words (Connolly and Phillips, 1994; Desroches et al., 2009; Newman and Connolly, 2009; see also discussion of the "N200" in e.g., Brunellière and Soto-Faraco, 2015; Van Den Brink et al., 2001). The PMN has been consistently observed in previous ERP research of Mandarin spoken words (Zhao et al., 2011; Malins and Joanisse, 2012), although it has not always been overtly analyzed or labeled as such (Liu et al., 2006; Pelzl et al., 2019). Of particular relevance is the study by Malins and Joanisse (2012), which used a picture-word paradigm with single syllable Mandarin words. In their study, all auditory stimuli were real words, and they manipulated the relation to pictures so that either consonants, vowels, tones, or complete syllables matched/mismatched the evoked word. They found significant PMN and N400 effects for all mismatch types. An MEG study has linked the PMN to activity generated in anterior left auditory cortex (Kujala et al., 2004). Within the EEG literature, PMN peaks have appeared variably across anterior, central, and posterior electrode sites. In the present case, because our nonwords differ from real words only with respect to a tone or a vowel in the first syllable, we expect that PMN responses will be evoked in native speakers as soon as the departure from the target word becomes apparent.

Along with PMN responses, we also expect to see strong late positive components (LPCs) in native speakers. In sentence processing experiments, late positivities are often classified as P600s and are hypothesized to reflect reanalysis or repair processes when people are confronted by infelicitous syntax (Gouvea et al., 2010; Kaan and Swaab, 2003; Osterhout and Holcomb, 1992), though similar effects have been observed for lexical violation (e.g., Romero-Rivas et al., 2015; Schirmer et al., 2005) and phonological mismatches (e.g., Schmidt-Kassow and Kotz, 2009). Importantly, we observed LPCs in our previous sentence processing ERP study when L1 listeners detected tone and rhyme mismatches in nonwords (Pelzl et al., 2019). Although, not a sentence processing study, similar effects—though not analyzed—are also apparent in the later portion of waveforms for vowel and tone mismatches in Malins and Joanisse (2012, p. 2037 Figure 1). Thus, we expect to find LPCs in response to picture-phonology mismatches in the present case. These effects are often described as indexing error detection, repair, reanalysis, or reorientation processes and may be related to more general (i.e. non-linguistic) processing mechanisms (Coulson et al., 1998; Sassenhagen and Bornkessel-Schlesewsky, 2015; Sassenhagen et al., 2014; for a review, see; Leckey and Federmeier, 2019).

What differences might we expect to see in L1 and L2 ERP responses in our analysis of trials with accurate behavioral responses? Given that the LPC essentially indexes the attentional processes that lead to decisive rejections, we expected this component to align fairly well with behavioral responses across groups, such that both L1 and L2 listeners would show larger LPCs for correctly rejected tone mismatches and vowel mismatches relative to correctly accepted matching words.

However, if the L1 and L2 speakers arrive at those correct responses in different ways, we might expect to see differences across groups in the earlier PMN response. One possibility is that L2 speakers have incomplete encoding of lexical tone such that they are unable to fully retrieve it to form a prediction for the upcoming speech input. This would predict that the L2 group would show a PMN for vowel mismatches relative to matching real words, but not for tone mismatches. Another possibility is that L2 speakers use a different processing strategy across the board: they may not be able to use the picture to generate a detailed phonological prediction in the same way as native speakers do, which might manifest as an absence of PMN effects in all conditions. Such a pattern would not directly account for the tone disadvantage, but might point to differences in processing that indirectly contribute to phonolexical encoding or processing problems.

In summary, the picture-phonology matching experiment aims to create a scenario where L2 listeners are given strong odds of success in recognition of tone mismatches in a lexical context, and, by recording ERPs aims to examine L2 neural responses to the tone and vowel cues as they occur.

MATERIALS AND METHODS

Participants

We recruited 19 native English speakers, all of whom had achieved advanced levels of proficiency in spoken Mandarin Chinese (**Table 1**).[1] One was excluded due to early onset of learning (age 7), and one was removed from analyses due to excessive artifacts in EEG data. This left 17 advanced L2 participants. All participants passed two screening measures (yes/no vocabulary test and Can-do self-assessment). The measures and criteria were the same as used in Pelzl et al. (2019), Pelzl et al. (2020). Due to the difficulty of finding sufficient L2 participants, one L2 participant was accepted despite a slightly lower vocabulary score (65.7) than criterion (70). Additionally, all participants completed a tone identification task, testing their ability to identify tones produced by four different talkers that were presented either in isolated syllables, or on the first syllable of a disyllabic target (contextualized syllables). Due to space constraints, we do not present the full details of the tone identification here (see Pelzl, 2018).

Twenty-four native Chinese speakers also completed the experiment (average age = 26.1). One was excluded due to equipment failure, and three were excluded due to excessive EEG artifacts, leaving twenty L1 participants for all analyses presented below.

TABLE 1 | Background information, screening measures, and tone identification scores for L2 participants (*n* = 17).

	Mean (sd)	Range
Age at testing	25.8 (4.9)	18–38
Age of onset	17.5 (4.0)	11–25
Semesters of formal study	9.0 (5.0)	3–20
Years in immersion	3.5 (2.7)	0.7–9
Total years learning	8.3 (3.8)	3–19
Can-do self-assessment (%)	82.7 (7.6)	72.8–96.8
Vocabulary self-assessment (%)	88.2 (9.6)	65.7–100
Tone identification accuracy (%): *overall*	85.8 (7.7)	71.9–99.2
isolated syllables	89.5 (4.2)	81.2–98.4
contextualized syllables	82.1 (12.3)	57.8–100

All participants gave informed consent and were compensated for their time.

Task and Stimulus Design

In the picture-phonology matching experiment, participants saw a picture followed either by a word that matched the picture or by a nonword that mismatched the pronunciation of the word evoked by the picture.

Critical stimuli were based on a set of 96 disyllabic real words[2]. All were highly frequent imageable nouns, chosen so that a corresponding picture could be matched to each one (e.g., *mian4tiao2* 'noodles'). Words were first sought in beginning and intermediate levels of the popular L2 Mandarin textbook series *Integrated Chinese*. Additional words were chosen based on frequency in the SUBTLEX-CH corpus (Cai and Brysbaert, 2010) and the intuitions of the first author, an L2 Mandarin speaker and former Mandarin teacher.

In order to make pictures as easily identifiable as possible, photographic images were used[3] The majority of images were taken from two freely available picture databases (BOSS: Brodeur et al., 2010; Ecological SVLO: Moreno-Martínez and Montoro, 2012), with additional images culled from other free photo repositories (e.g., Wikimedia commons). A small number of difficult to find images were purchased from Adobe Stock, and two more images were created specifically for the experiment. An example image is shown in **Figure 2**. All images were placed on a white background. No attempt was made to control colors or luminosity as the neural response to the presentation of the images was not of interest. Instead we aimed to make images as recognizable as possible.

To assure that images would evoke the intended words, two rounds of picture norming were conducted. In each round, ten native Mandarin speakers generated Chinese words for 132 images. Images that were judged to perform inadequately in

[1] This experiment was part of a larger study that was the first author's dissertation (Pelzl, 2018). A brief overview of the full design is included in **Supplementary Appendix A**. Participants are the same as those described in Pelzl et al. (2020), though there were some participant exclusions in the current dataset due to excessive EEG artifacts in the picture-phonology task.

[2] There was no overlap in target words between the current study and the lexical decision stimuli reported in Pelzl et al. (2020).

[3] For the words *tian1shi3* 'angel' and *mo2gui3* 'devil', computer generated 3-D cartoon images were used, as no angels or demons were available for photos.

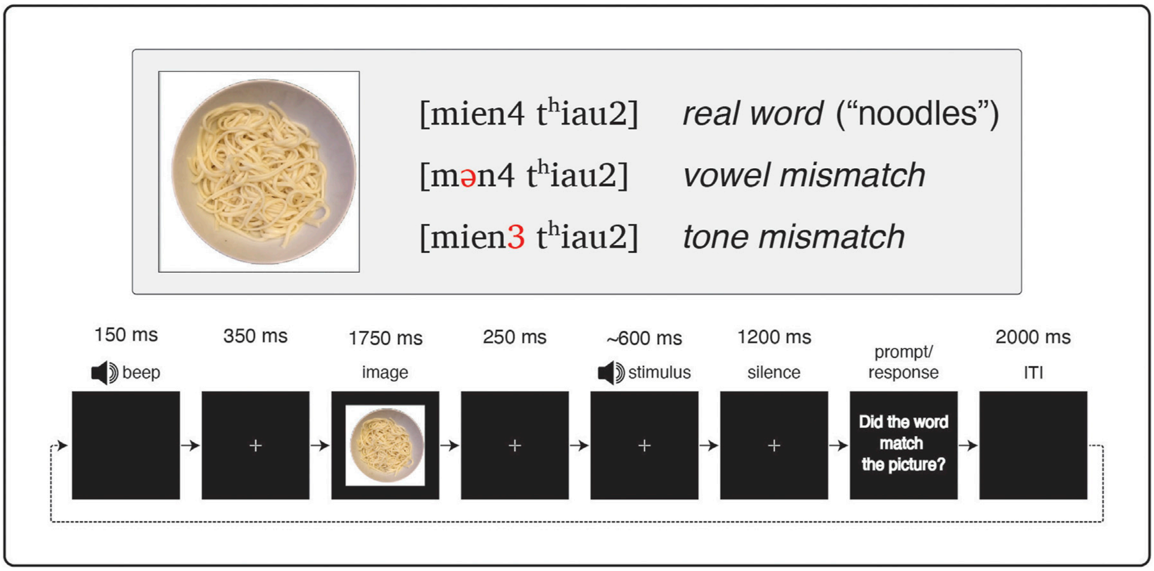

[mien4 t^hiau2] real word ("noodles")
[mən4 t^hiau2] vowel mismatch
[mien3 t^hiau2] tone mismatch
150 ms
beep
350 ms
1750 ms
image
250 ms
~600 ms
stimulus
1200 ms
silence
prompt/
response
2000 ms
ITI
Did the word match the picture?

FIGURE 2 | Example stimulus set and trial parameters for picture-phonology matching task.

the first round (less than 70% generation of the target word, or generation of problematic competitor words) were replaced and a second round of norming was conducted with a new group of ten people. The end result was a set of 96 critical images that had an average word generation rate of 86%, though a handful of items (7 total) had rather low naming rates (under 50%). Future work might try to replace either those words or images. Images for filler items were also overall highly identifiable.

The real words were further manipulated to create two types of nonwords. The first syllable of the nonwords mismatched the real word counterpart with respect to either a tone or a vowel. For example, the real word *mian4tiao2* /mien4t^hiau2/ became the vowel nonword *m**e**n4tiao2* /mən4t^hiau2/ and the tone nonword *mian**3**tiao2* /mien3t^hiau2/. All possible tone combinations and manipulations were balanced across words and nonwords. For vowel mismatches, the syllable rhyme was changed by switching, adding, or deleting a single vowel sound (monophthong), though in a handful of cases, changes affected multiple vowel sounds (diphthongs). As much as possible, repetition of first syllables was avoided across stimuli, though some exceptions had to be made to accommodate the limited availability of imageable words that were likely to be known by L2 participant (all stimuli are available in the **Supplementary Appendix B**).

These procedures resulted in a total of 96 critical real word/vowel nonword/tone nonword triplets. An additional 16 real words with accompanying images were selected as fillers. Given all the constraints noted above, it was not possible to limit selection of fillers to words with a balanced occurrence of tones, and many filler items had neutral tones on the second syllable.

Three lists were constructed to balance images and words across participants. For each list, four unique pseudo-random presentation orders were prepared, with conditions balanced so that no more than three trials in a row would require the same response type (yes or no).

We also designed an offline vocabulary post-test. For each L2 participant, the test included all real word counterparts for vowel and tone mismatched nonwords encountered during the picture-phonology matching task (64 words total; words that occurred in the 'real word' condition were not tested). Each item provided Chinese characters and toneless Pinyin. Participants supplied tones (numbers 1-4 for each syllable), an English definition, and a confidence rating from 0–3 for both the tones and the definition of each item. Participants were informed that the 0–3 scale had the following meaning: *0 = I don't recognize this word; 1 = I recognize this word, but am very uncertain of the tones/meaning; 2 = I recognize this word, but am a bit uncertain of the tones/meaning*; *3 = I recognize this word, and am certain of the tones/meaning*. This scale remained visible as a reference through the duration of the test. For any tones or definitions they did not know, participants were instructed to leave the answer blank and supply "0" for confidence.

Procedures

Thirty-six participants (24 L1 and 11 L2) were tested in the lab at Beijing Normal University (BNU). Seven additional L2 participants were tested under conditions as similar as possible in the lab at the University of Maryland (UMD). Each participant was seated in front of a computer monitor and fit with an EEG cap. Auditory stimuli were presented using a single high-quality audio monitor (JBL LSR305) placed centrally above the computer monitor.

This experiment was conducted as part of a larger study (see Pelzl, 2018), it followed a lexical decision task (reported in Pelzl et al., 2020), and was itself followed by a picture-word matching task that examined N400 responses for clear lexical violations (details in Pelzl, 2018). For the picture-phonology matching experiment, participants began by completing eight practice items with stimuli not included in the experiment, and then completed 112 picture-phonology matching trials. Stimuli were presented in seven blocks of 16 trials, with self-paced breaks between each block. Trial parameters are illustrated in **Figure 2**. The beginning of each trial was signaled with a 'beep', followed by a fixation cross. After 350 ms, a picture was displayed. Then, after 1.75 s the image was replaced by a fixation cross. Still 250 ms later the auditory stimulus was presented, followed by 1.2 s of silence at which point the fixation cross was replaced by a question prompt: "Did the word match the picture?" (or equivalent in Chinese for L1 participants). After the participant's response, there was a 2 s pause before the next trial began. The entire picture-phonology matching experiment lasted approximately 15 min.

The long display time for the images (1.75 s) was determined after piloting and with the logic that, for this experiment we wanted to maximize the opportunity for L2 learners to recognize images and their associated words. This design allows (but does not compel) participants to utilize explicit knowledge of tones in retrieving target items. The design serves as a proof-of-concept for this approach, testing L2 ability to utilize tone cues under near optimal circumstances.

After completion of the ERP experiments, participants completed the offline vocabulary test.

Vocabulary Posttest

The offline vocabulary posttest produced four data points for each mismatching nonword trial that an L2 participant encountered: an accuracy score for the tones and definition they supplied, a confidence rating for the tones, and a confidence rating for the definitions. Accuracy was scored correct (1) or incorrect (0). For example, if the real word target was 面条 *miantiao*, the only correct response would be "42" (*mian4tiao2* "noodles"). Any deviations from these two tones would be marked incorrect. Definitions were scored similarly using a list of acceptable definitions generated prior to scoring. Confidence ratings were recorded as a number from 0 to 3. One participant's vocabulary test data was lost due to a coding error, leaving a total of 1,024 trials (64 per participant) for the analysis.

Electroencephalogram Recording

Raw electroencephalogram (EEG) was recorded continuously at a sampling rate of 1,000 Hz using a Neuroscan SynAmps data acquisition system and an electrode cap (BNU: Quik-CapEEG; UMD: Electrocap International) mounted with 29 AgCl electrodes at the following sites: *midline*: Fz, FCz, Cz, CPz, Pz, Oz; *lateral*: FP1, F3/4, F7/8 FC3/4, FT7/8, C3/4, T7/8, CP3/4, TP7/8, P4/5, P7/8, and O1/2 (UMD*:* had FP2, but *no* Oz). Recordings were referenced online to the right mastoid and re-referenced offline to averaged left and right mastoids. The electro-oculogram (EOG) was recorded at four electrode sites: vertical EOG was recorded from electrodes placed above and below the left eye; horizontal EOG was recorded from electrodes situated at the outer canthus of each eye. Electrode impedances were kept below 5kΩ. The EEG and EOG recordings were amplified and digitized online at 1 kHz with a bandpass filter of 0.1–100 Hz.

EEG Data Processing

Consistent with the approach used in the related study reported in Pelzl et al. (2020), data from fifteen central electrodes (F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CP3, CPz, CP4, P3, Pz, P4) were chosen for final analysis. To reduce some mild non-normality in the data, any trial with an absolute value greater than 50 μV was removed prior to final data analysis. Finally, only trials that elicited correct behavioral responses (correct acceptance or correct rejection) were retained for final analysis.

Data from one L1 participant was excluded due to equipment failure. Data from three additional L1 participants and one L2 participant were excluded due to having greater than 40% artifacts on experimental trials (a second L2 participant's data was borderline at 41.67% trials rejected due to artifacts, but was retained due to the difficulty of obtaining advanced L2 data). After excluding these participants, artifact rejection affected 10.55% of experimental trials (L1 8.31%; L2 13.18%). A single average amplitude was obtained for each trial for each electrode for each subject in an early PMN window (200–400 ms) and a later LPC window (400–600 ms). These windows were chosen on the basis of previous research and by visual inspection of grand average waveforms. We recognize the reliance on visual inspection for window selection as a potential limitation, and future work should improve on it in line with advice in Luck and Gaspelin (2017).

After exclusion of incorrect trials, the final PMN dataset contained 42,613 data points (80.0% out of total possible 53,290 data points: L1 = 88.1%; L2 = 70.4%) and the LPC dataset contained 42,610 data points (80.0% out of total possible 53,290 data points. L1 = 88.1%; L2 = 70.4%).

RESULTS

Behavioral Results and Analysis

Reliability for picture-phonology matching data was high (List A: $\alpha = 0.91$; List B: $\alpha = 0.93$; List C: $\alpha = 0.94$). Descriptive results are shown in **Table 2**. Overall, L1 listeners were more accurate than L2 listeners. Whereas L1 listeners were least accurate in judging vowel mismatches, L2 listeners were least accurate in judging tone mismatches.

To further investigate response patterns, we also computed d-prime (d') for each participant, contrasting vowel mismatches with matching real words, and tone mismatches with real words. Laplace smoothing was used to correct for infinite values (Jurafsky and Martin, 2009; Barrios et al., 2016). As with accuracy, d' results suggest overall higher sensitivity to mismatches for L1 listeners, with better scores for tone mismatches compared to vowel mismatches (vowel $d' = 3.49$, sd = 0.49; tone $d' = 3.91$, sd = 0.41). In contrast, L2 had less sensitivity overall, with vowel mismatches detected more readily

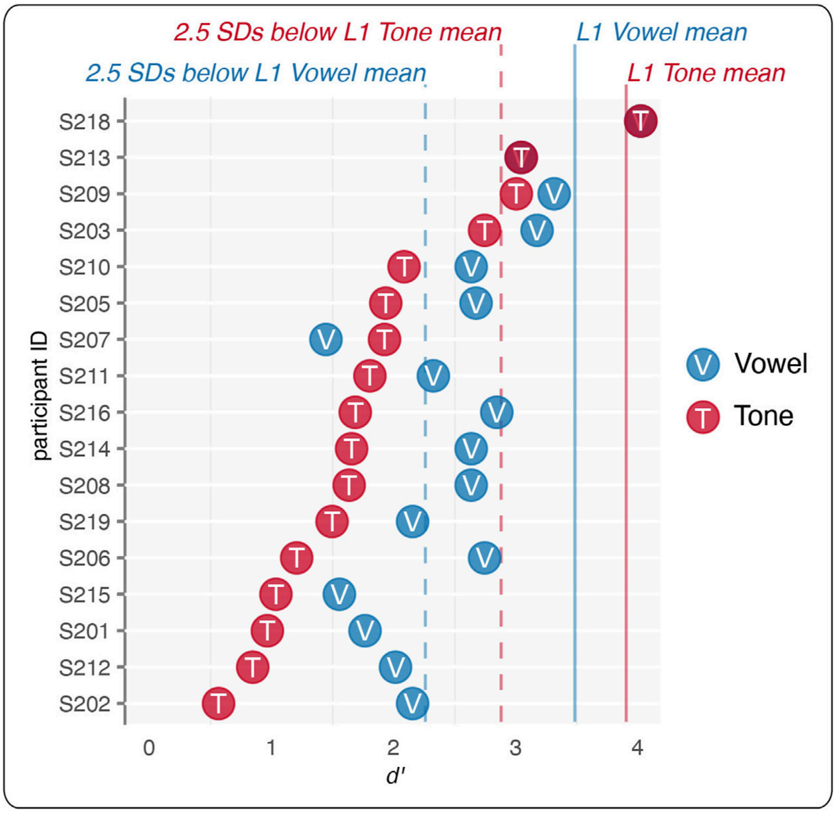


FIGURE 3 | Individual L2 participants' (n = 17) *d′* results for vowel (V) and tone (T) mismatch conditions in the picture-phonology matching task.

than tone mismatches (vowel *d′* = 2.54, sd = 0.66; tone *d′* = 1.87, sd = 0.91).

Figure 3 depicts individual d′ results for each L2 participant. All but three L2 participants had tone d′ values more than 2.5 standard deviations below the L1 mean for tone mismatches, while vowel mismatches were more mixed. More importantly, in all but three cases (S218, S213, and S207), individual L2 participants had lower d′ values for tone mismatches than for vowel mismatches.

All statistical analyses reported below were conducted in *R* (version 4.0.3, R Core Team, 2020). Mixed-effects models were fit using the *lme4* package (version 1.1.21, Bates et al., 2015b). Effects coding was applied using the *mixed* function in *afex* (Singmann et al., 2017). Rather than examining general model outcomes that were not of importance for our research questions (e.g., whether there is a main effect of group or condition), we focus on the specific outcomes of interest (Schad et al., 2020), which we specified using the *multcomp* (Hothorn et al., 2008) and *emmeans* (Lenth, 2020) packages (full model results are reported in the **Supplementary Appendix C**).

Accuracy results were submitted to a mixed-effect logistic regression (using the *bobyqa* optimizer) with crossed random effects for subjects and items. The dependent variable was accuracy (1, 0). Fixed effects included the factors *condition* (real word, tone mismatch, vowel mismatch), and *group* (L1, L2), and their interaction. The maximal random effects model was fit first (Barr et al., 2013; Bates et al., 2015a). Model convergence difficulties were addressed by suppressing correlations in random effects (using "expand_re = TRUE" in the *mixed* function). The best fitting model was determined by model comparison conducted through likelihood ratio tests, building from the maximal model (which was rejected due to convergence issues) to progressively less complex models. The final model included by-subject random intercepts and slopes for the effect of condition, and by-item random intercepts and slopes for condition and group and their interaction: (*glmer* model formula)*:* accuracy ~ condition * group + (condition | subject) + (condition * group || item).

TABLE 2 | Descriptive accuracy results for picture-phonology matching task.

Group	Condition	Mean acc. % (sd)
L1 (*n* = 20)	Real	97.5 (15.6)
	Vowel	92.7 (26.1)
	tone	98.1 (13.6)
L2 (*n* = 17)	Real	87.5 (33.1)
	Vowel	88.0 (32.5)
	tone	69.1 (46.2)

The critical comparison was whether the L2 group displays a difference between vowel and tone accuracy. To complete the picture, we also examined how this difference compares to the same contrast in the L1 group. Critical comparisons are

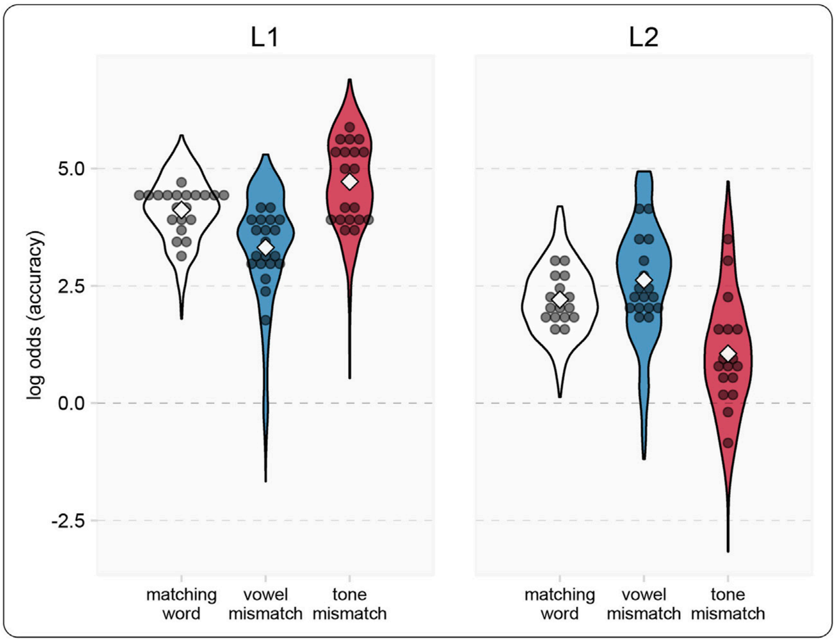


FIGURE 4 | Violin plots of model estimated log odds of a correct response in the picture-phonology matching task. Width of the plot indicates distribution of model estimated responses. White diamonds indicate group means. Gray circles indicate individual participant mean scores. The zero line indicates chance.

TABLE 3 | Critical comparisons for vowel and tone accuracy in the picture-phonology matching task.

Comparison	b	SE	95% CI[a]	Z	p (>/z/)[b]
L1 vowel vs tone	−1.36	0.51	[−2.55, −0.18]	-2.69	0.007
L2 vowel vs tone	1.73	0.36	[0.88, 2.58]	4.75	<0.001
L1 vs L2: Vowel vs tone	−3.09	0.54	[−4.35, −1.83]	-5.72	<0.001

[a] *Asymptotic confidence intervals.*
[b] *Adjusted using Bonferroni-Holm method.*

summarized in **Table 3**, and model results are depicted in **Figure 4**. L2 listeners were significantly more accurate in rejection of vowel mismatches than of tone mismatches. They were about two and a half times more likely to incorrectly accept tone mismatches than vowel mismatches (30.9/12 = 2.6). There was also a statistically significant difference in accuracy between mismatch conditions for the L1 group, with L1 more accurate for tone than vowel mismatches. Compared to L1, the accuracy difference between mismatch conditions for L2 was larger and in the opposite direction.

In summary, whereas L1 listeners had more difficulty detecting vowel mismatches than tone mismatches, L2 listeners had more difficulty detecting tone mismatches than vowel mismatches.

Table 4 displays descriptive results for the offline vocabulary test, along with related accuracy for those items in the picture-phonology matching task. We find that, overall, L2 learners were quite confident of the definitions they provided (mostly ratings of high or mid confidence), and that higher confidence appears to relate strongly to the accuracy of those definitions. In other words, learners know which words they know and which they do not. Learners' confidence ratings for their explicit tone knowledge indicate less certainty for tones than for definitions. Although overall accuracy for tones is lower than for definitions, it still does appear to track with confidence ratings. That is, L2 learners generally know which tones they know and which they do not. However, even for the tones they know most confidently, they are still inaccurate for more than one in ten of those tones (*mean* = 86% when counting nonword conditions together). Whereas accurate knowledge of definitions always appears to impact performance in the picture-phonology matching task, accurate knowledge of tones appears to relate only to tone nonword items. This makes sense, as tone knowledge is largely irrelevant for vowel nonword items.

Descriptively, then, we find that L2 offline knowledge suggests some difficulties in accurate encoding of tones in explicit lexical representations, and that this appears to impact accuracy for correct rejection of tone mismatches.

As in Pelzl et al. (2020), we conducted an exploratory "Best case Scenario" analysis using only the subset of trials that targeted nonwords for which an L2 participant had indicated correct and confident knowledge (confidence rating = 3) of both tones and definitions for the real word counterparts. This comprised 256 tone nonword and 255 vowel nonword trials (511 total, 47% of total mismatch trial data). Mean accuracy for vowel nonwords was 93% (sd = 26%); mean accuracy for tone nonwords was 80% (sd = 40%). The accuracy results were submitted to a generalized linear mixed effects model following procedures outlined for previous analyses. The model included the fixed effect of nonword condition. The maximal model was fit, and included

TABLE 4 | Results of offline vocabulary test requiring L2 participants ($n = 16$) to supply tones, definitions, and confidence ratings for the real word counterparts of critical mismatching nonwords. Tone accuracy indicates whether supplied tones were correct. Picture-phonology (pic-pho) accuracy indicates whether the related nonwords were correctly rejected in the matching task.

	condition	**conf. rating**	**k (items)**	**definition acc. %**	**pic-pho acc. %**
confidence ratings and accuracy of L2 supplied definitions	vowel nonword	3 (high)	465	98	90
		2 (mid)	27	81	78
		1 (low)	7	43	57
	tone nonword	3 (high)	470	99	70
		2 (mid)	23	83	61
		1 (low)	5	4	40
	condition	conf. rating	k (items)	tone acc. %	pic-pho acc. %
confidence ratings and accuracy of L2 supplied tones	vowel nonword	3 (high)	300	87	92
		2 (mid)	170	52	84
		1 (low)	29	28	83
	tone nonword	3 (high)	309	84	77
		2 (mid)	163	50	57
		1 (low)	35	37	51

random intercepts for subjects and items, and random slopes for the by-subject and by-item effects of condition. There was a significant difference in accuracy for tone and vowel nonwords ($b = -7.71$, $SE = 3.06$, 95%, $z = -2.51$, $p = 0.012$).

In summary, after accounting for offline L2 word knowledge and subjective confidence of that knowledge, L2 learners still showed a more limited ability to reject tone mismatches than vowel mismatches. At the same time, we should not ignore the observable improvement that occurred when results were limited to items known correctly and with certainty. Accuracy for vowel mismatches rose from 88 to 93%, and for tone mismatches the increase was even greater, from 69 to 80%, indicating that—at least among this group of learners—eliminating fuzzy (incorrect and uncertain) lexical representations appears to partially account for performance deficits for both tones and vowels.

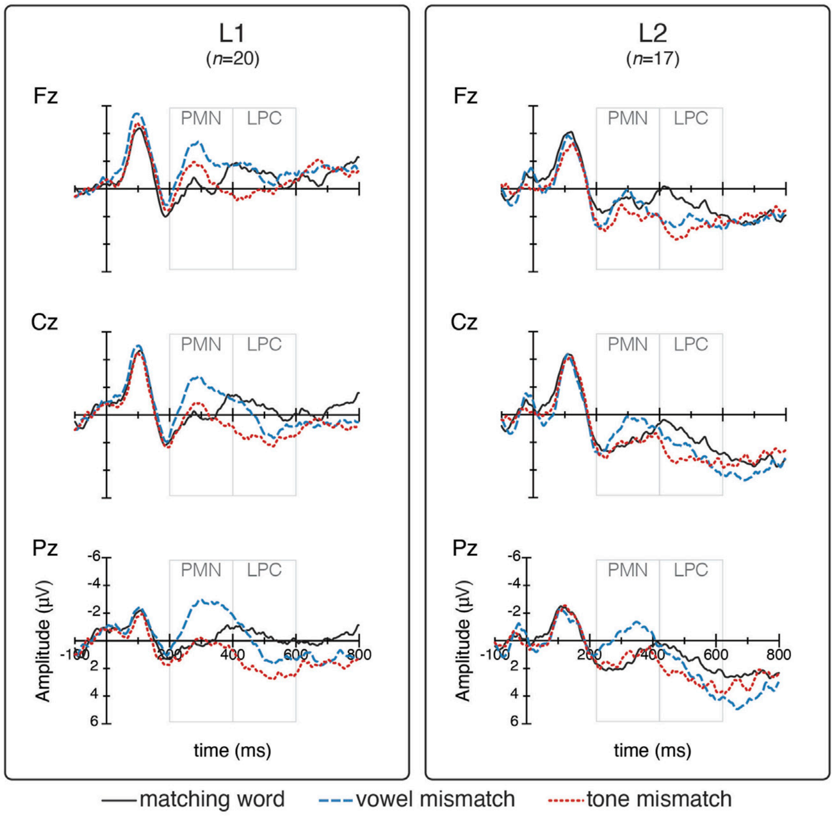
ERP Results and Analyses for Phonological Mismatch Negativity and Late Positive Component Windows

Mean amplitudes for ERP responses in the time windows for the PMN (200–400 ms) and LPC (400–600 ms) are displayed in **Table 5**. Grand average ERP waveforms are depicted in **Figure 5**. The L1 group appears to have strong negativities for vowel mismatches in the PMN window; though L2 responses are more positive overall, over centro-posterior electrodes the same pattern holds, with vowel mismatches showing the most negative amplitude among condition. In the LPC window, responses for real words are most negative (least positive), followed by vowel mismatches, with tone mismatch responses being the most positive. While there are differences in absolute amplitudes between groups, over centro-posterior electrodes the overall ordering of responses (real, vowel, tone) is similar within L1 and L2 groups.

Average amplitudes for correct trials in the PMN and LPC windows were submitted to linear mixed-effects regression model with crossed random effects for subjects and items. Fixed effects were *condition* (match, mismatch) and *group* (L1, L2) and their interaction. Convergence difficulties were addressed by specifying uncorrelated random effects. Effects coding was used, and *p*-values were obtained using Satterthwaite's method. The maximal model that successfully converged was fit first and was then compared to less complex models to test random effects. The final maximal models for both data sets were parallel, and included random slopes for subjects and items, with electrodes nested under subjects. The models also included by-item random.

Though our primary interest in this study is in L2 sensitivity to vowels and tones, in order to evaluate L2 responses, we need to compare them to an L1 baseline. To this end, we report critical comparisons for three relevant contrasts (matching word vs. vowel mismatch, matching word vs. tone mismatch, vowel mismatch vs. tone mismatch) within and between L1 and L2 groups. Results for the PMN window are shown in **Table 6**, and depicted in violin plots in **Figure 6**. For both the L1 and L2 group, responses to vowel mismatches were significantly more negative than both matching word and tone mismatch responses. Despite the similar overall pattern of their responses, there were interactions between group and condition. For the L1 group, the magnitude of differences for the matching word vs. vowel mismatch and vowel vs. tone mismatch were significantly larger than the same contrasts for L2 participants. In other words, though L1 vowel mismatch responses were stronger overall, the same pattern of responses applied for both groups, with *neither* group showing strong PMN deflections for tone mismatches.

Results for the LPC window are shown in **Table 7**, and depicted in violin plots in **Figure 7**. For both the L1 and L2 group, responses to tone mismatches were significantly more positive than responses to matching words. For the L1, but not the L2 group, tone mismatch responses were more positive than vowel mismatch responses. Vowel mismatch responses did not differ significantly from matching word responses. Interactions between groups and conditions There was a significant interaction between group and condition for the contrast of matching words vs vowel mismatches with the L2 difference being larger than the L1 difference. There was



only trials with correct behavioral responses included

FIGURE 5 | Grand average waveforms for L1 and L2 participants. Time windows highlighted for PMN (200–400) and LPC (400–600). (Waveforms for all 15 electrodes are available in the **Supplementary Appendix D**).

TABLE 5 | Mean amplitude (in μV) and standard error (SE) of PMN and LPC responses (correct trials only).

		PMN		*LPC*	
group	**condition**	**mean amp**	**(SE)**	**mean amp**	**(SE)**
L1	Real	0.24	(0.10)	−0.37	(0.11)
L1	Vowel	−1.66	(0.10)	−0.13	(0.12)
L1	Tone	0.20	(0.10)	1.39	(0.11)
L2	Real	1.35	(0.11)	0.66	(0.12)
L2	Vowel	0.20	(0.11)	1.52	(0.13)
L2	tone	1.69	(0.13)	2.53	(0.15)

also a significant interaction between groups and vowel vs. tone mismatches, the size of the difference being larger for L1 than for L2 responses. Confidence intervals for all comparisons suggest some imprecision and so results should be interpreted with appropriate caution.

DISCUSSION

In order to test advanced L2 Mandarin learners' sensitivity to lexical tones, we conducted a picture-phonology matching task with L1 and L2 speakers of Mandarin. Key results can be summarized as follows. 1) L2 participants were less accurate at rejecting tone mismatches than vowel mismatches—the opposite pattern from L1 participants who were more accurate in all conditions overall, but less accurate for vowel than tone mismatches. 2) After limiting the analysis to trials for words L2 participants knew correctly and confidently, their accuracy for both tone and vowel mismatch trials increased, but tone mismatch trials still remained significantly less accurate than vowel mismatch trials. For ERP results, which targeted only trials with correct behavioral responses, 3) in the early PMN window, both L1 and L2 listeners displayed significantly more negative responses to vowel mismatches, than to either matching words or tone mismatches. Though there were differences in the magnitude of effects between L1 and L2, the overall patterning of responses was similar. 4) In the later LPC window, both groups displayed strong positive responses following tone mismatches, with some differences in the magnitude of responses to vowel mismatches relative to tone mismatches and real words. Below, we discuss these results in more detail, while also connecting them to broader discussions of L2 tone word learning and fuzzy L2 lexical representations.

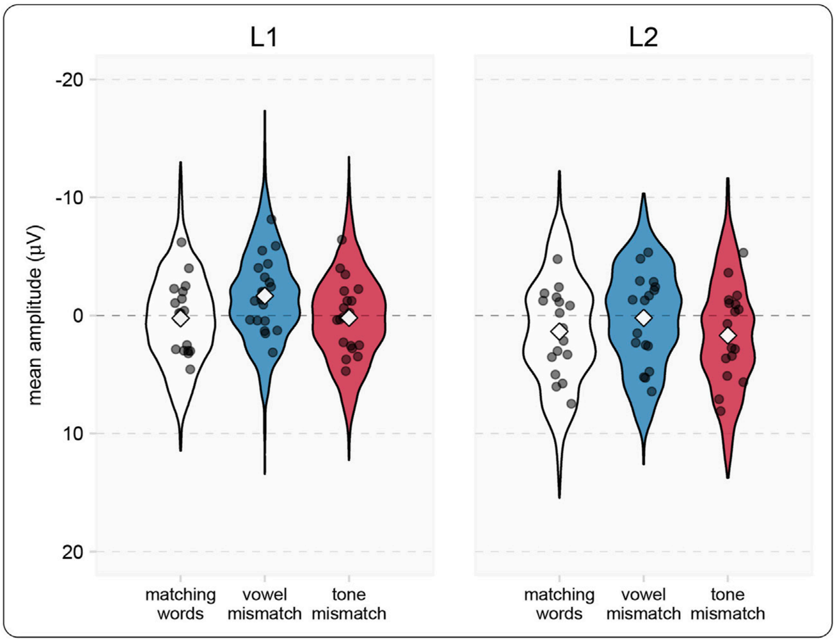


FIGURE 6 | Model estimates for PMN (200–400 ms) amplitude in the picture-phonology matching task (correct trials only). White diamonds indicated model estimated group means for each condition, with shaded areas representing the distribution of estimated responses. Each gray dot indicates an individual participant's mean amplitude in the condition.

TABLE 6 | ERP results and analyses for PMN window (200–400 ms).

Comparison	b	SE	95% CI[a]	Z	p (>/z/)[b]
L1 match vs vowel	2.10	0.38	[1.08, 3.12]	5.52	<0.001
L1 match vs tone	0.05	0.34	[−0.86, 0.96]	0.14	1.000
L1 vowel vs tone	−2.05	0.37	[−3.05, −1.04]	−5.48	<0.001
L2 match vs vowel	1.17	0.39	[0.13, 2.21]	3.03	0.010
L2 match vs tone	−0.14	0.36	[−1.09, 0.82]	−0.39	1.000
L2 vowel vs tone	1.31	0.39	[0.27, 2.35]	3.38	0.004
L1 vs L2: Match vs vowel	0.92	0.20	[0.40, 1.45]	4.74	<0.001
L1 vs L2: Match vs tone	0.19	0.21	[−0.37, 0.74]	0.90	1.000
L1 vs L2: Vowel vs tone	−0.74	0.21	[−1.30, −0.18]	-3.56	0.002

[a]Asymptotic confidence intervals.
[b]Adjusted using Bonferroni-Holm method.

Tones Are Difficult (Again)

Our results echo those seen in previous studies, indicating that—for nontonal L1 speakers—mastery of tone words is a major L2 learning challenge (Han and Tsukada, 2020; Ling and Grüter, 2020; Pelzl et al., 2019; Pelzl et al., 2020). Given the nature of the picture-phonology matching task, the present results are perhaps the clearest indication yet of how difficult L2 tone word recognition is. As noted above, the picture-phonology matching task was less demanding than previously used lexical decision tasks. As long as a person knew the pictured word and its tones, they could directly judge whether the target matched or not. There was no need to search their mental lexicon to verify the *absence* of a nonword. Nevertheless, we found that the L2 group made errors on 31% of tone mismatch trials overall, compared to 12% of vowel mismatch trials. When we limited consideration to correctly and confidently known words, they still made errors on 20% of tone mismatch trials. In other words, for these L2 participants, explicit knowledge of tone words only accounted for, at most, one-third of their tone errors.

Three General Accounts of Tone Difficulties

As outlined in our introduction (**Figure 1**), there are three broad accounts that could uniquely or jointly explain these outcomes, positing perception, processing, or representation as the locus of L2 tone word breakdowns. Present results do not allow us to determine the relative contribution of these accounts to lexical tone learning difficulties. At the same time, they do suggest directions for future study.

First, though the present experiment did not directly test auditory perception, the overall accuracy for tones after limiting analysis to correctly and confidently known words (mean = 80%) bears a striking resemblance to the same L2 participants' overall accuracy for tone identification in disyllabic contexts (*mean* = 82%, see **Table 1**). In other words, it may well be the case that, once explicit knowledge of tones has been established, L2 listeners' remaining tonal difficulties are due primarily to difficulties perceiving tones faithfully in *multisyllabic strings.* Since our first study (Pelzl et al., 2019), the particular difficulty of disyllabic, as opposed to monosyllabic tone words, has remained an open question. Studies with naïve, novice, and intermediate proficiency L2 participants have reported this pattern in tone category identification tasks (Broselow et al., 1987; Hao, 2018; Sun, 1998; see also; Chan and Leung, 2020).

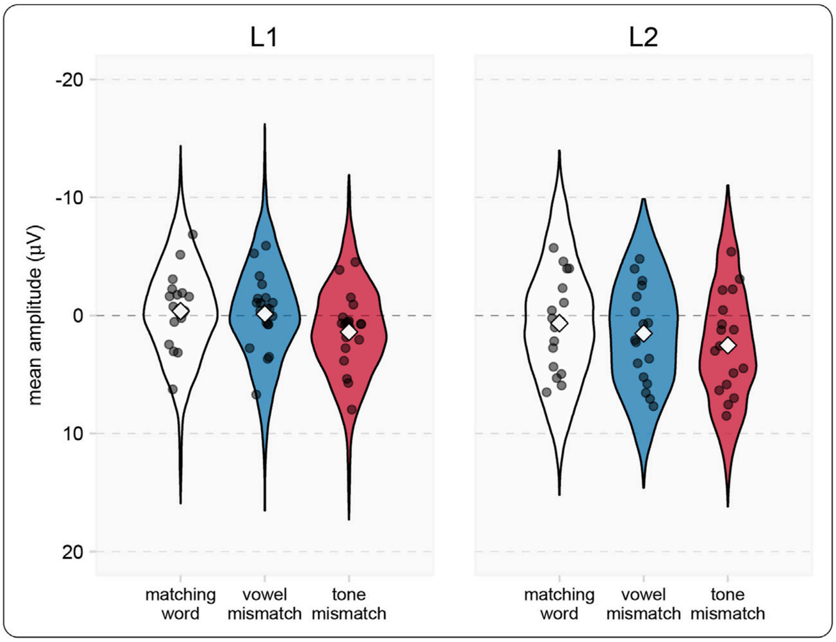


FIGURE 7 | Model estimates for LPC (400–600 ms) amplitude in the picture-phonology matching task (correct trials only). White diamonds indicated model estimated group means for each condition, with shaded areas representing the distribution of estimated responses. Gray dots indicate individual participants mean amplitude in the condition.

TABLE 7 | ERP results and analyses for LPC *window (400–600)*.

Comparison	b	SE	95% CI[a]	Z	p (>/z/)[b]
L1 match vs vowel	0.01	0.44	[−1.17, 1.19]	0.03	1.000
L1 match vs tone	−1.71	0.36	[−2.68, −0.74]	−4.71	<0.001
L1 vowel vs tone	−1.72	0.41	[−2.83, −0.61]	−4.16	<0.001
L2 match vs vowel	−0.95	0.45	[−2.15, 0.25]	−2.13	0.133
L2 match vs tone	−1.57	0.38	[−2.59, −0.54]	−4.09	<0.001
L2 vowel vs tone	0.61	0.43	[−0.54, 1.77]	1.43	0.462
L1vs L2: Match vs vowel	0.96	0.22	[0.38, 1.55]	4.42	<0.001
L1vs L2: Match vs tone	−0.15	0.23	[−0.76, 0.47]	−0.63	1.000
L1vs L2: Vowel vs tone	−1.11	0.23	[−1.73, −0.49]	−4.78	<0.001

[a] *Asymptotic confidence intervals.*
[b] *Adjusted using Bonferroni-Holm method.*

In a tone word training study with naive learners, Chang and Bowles (2015) found disyllabic words to be much more challenging than monosyllabic words. That longer strings of syllables are more difficult is not surprising in and of itself, but the exact cause of the difficulty remains unclear. In Pelzl et al. (2019), our tone identification task used monosyllables clipped from context, thus preserving the coarticulation of the tones, but removing the potentially useful contextual cues provided by neighboring syllables—and L2 participants performed with near-native accuracy on all but Tone 2. In contrast, the tone identification, lexical decision and picture-phonology matching tasks used with the participants in the present study (and in Pelzl et al., 2020) were produced more slowly and clearly, and all maintained the contextual cues, nevertheless, most L2 learners showed some difficulties. Future research will need to examine additional factors that may be impacting multisyllable tone perception, such as memory constraints (e.g., *the phonological loop* Baddeley, 1968), L1 prosodic biases that might operate across multiple syllables (Braun et al., 2014; Braun and Johnson, 2011; Schaefer and Darcy, 2014; So and Best, 2010; So and Best 2014), and potential ordering effects in the perception of co-articulated tones (Xu, 1994; Xu, 1997).

Second, the phonolexical processing account can also naturally explain the incorrect responses on trials where participants reported correct and confident knowledge in the offline task. Despite having explicit knowledge of the pictured words, L2 listeners may have occasionally allowed their native processing biases to take over, ignoring tonal cues as they accessed words.

We have argued elsewhere (Pelzl et al., 2020) that tones are often redundant with other available cues. Most Mandarin words are longer than a single syllable, making the likelihood of a plausible minimal tone pair competitor low. Perhaps more importantly, the broader context will usually guide interpretation of what is heard. SLA scholars have long noted the difficulties associated with redundant cues in an L2 (e.g., VanPatten, 1996; DeKeyser, 2005). Insofar as tones are redundant with other available cues, recent discussion of the phenomena of *unlearning* and *blocking* may provide insights to the source of L2 failures to learn them (see, especially, Nixon, 2020; but also Ellis, 2006; MacWhinney and Bates, 1989). First, through long experience with a non-tonal L1, L2 Mandarin listeners have unlearned tone cues—that is, they have learned through negative evidence that F0 height and shape on vowels/syllables is not informative for speech comprehension, and thus have down-weighted such cues. When confronted with new F0 cues in the L2, they need to re-weight these cues appropriately, but because tones typically co-occur with other disambiguating cues, there is little opportunity for prediction error to guide this re-weighting process. This leaves primarily statistical learning mechanisms to guide the development of L2 tone processing. Indeed, statistical learning mechanisms have been shown in L2 tone learning for *highly frequent* syllable + tone co-occurrence probabilities (Wiener et al., 2018).

It has also been proposed that as vocabulary size increases, minimal pairs might push learners toward more sensitivity for difficult L2 contrasts (cf. discussion of L2LP model in Wiener et al., 2019; see also Bundgaard-Nielsen et al., 2011; Llompart and Reinisch, 2020). While not denying that minimal pairs *could* play a role in improved L2 outcomes, our own work so far has given us a rather pessimistic view of the strength of minimal pairs in typical L2 tone learning. Though it is not difficult to find tonal minimal pairs in Mandarin if one goes looking for them, for L2 learners these pairs accrue very gradually over time, and it is likely that many other developing L2 abilities will allow learners to further capitalize on contextual cues to the detriment of tones. Thus, it may be that, for most L2 learners, only the most frequent tone words will ever be processed phonolexically.

Returning to present results, with respect to the representational account, if the explicit knowledge of words directly captures how those words are encoded in phonolexical representations, the representational account cannot explain persistent L2 difficulties for correctly and confidently known words in the present data (note, this would be true for vowel mismatches as well). Still, as we will consider in more detail below, the representational account cannot be fully rejected as a contributor to the current pattern of results, as it could be that explicit knowledge of tone words was not the main source L2 listeners drew upon when judging whether a tone word matched a picture.

ERPs

Results from our PMN analysis suggest that, for *correctly* judged trials, both L1 and L2 listeners had the strongest (most negative) response to vowel mismatches. This suggests that L2 listeners are able to generate phonological expectations based on context, at least when there is plenty of time available to do so after the context appears, as in the current study. At the same time, PMN results failed to show significant differences between tone mismatches and real words, suggesting that mismatching vowels may affect ERPs earlier than tones for *all* listeners. This pattern of results is consistent with several previous studies which provided contextual cues for lexical expectation in phrases or sentences and found reduced N400s for tone mismatches relative to rhyme mismatches (Hu et al., 2012; Pelzl et al., 2021; Zou et al., 2020), though such differences do not always appear (Brown-Schmidt and Canseco-Gonzalez, 2004; Schirmer et al., 2005; Pelzl et al., 2019). On the other hand, as we expected for these correctly judged trials, in the later LPC time window both groups displayed strong positive deflections for tone mismatches relative to the matched word condition. In fact, at the LPC the tone mismatch response was significantly larger than the vowel mismatch response for the L1 speakers.

The seemingly similar delayed response to tone mismatches relative to vowel mismatches across groups might be tied to the nature of tone contrasts, especially as they occur in contextualized syllables. For many trials, it may be that in order to identify the F0 contour as it unfolds over time, more of the syllable needs to be available than in the case of vowels; sometimes listeners may even need the contextual information of the following syllable in order to make the identity of the tone unambiguously wrong (cf. J. Huang and Holt, 2009). This extended perceptual analysis for tones could be too late to impact the early phonological perception computations that may be driving the phonological mismatch negativity. In contrast, mismatching vowels reveal themselves almost immediately, which could drive a stronger negativity across the PMN and the N400 time-windows.

It is also worth noting, however, that visual examination of anterior electrodes indicates a numerical trend toward a PMN for tone mismatches in the L1 group that is not visible in the L2 group (**Figure 5**; see also **Supplementary Appendix D** for waveforms of all 15 tested electrodes). Therefore, it could be that we did not have enough power to reliably detect differences between real word and tone mismatch responses. Perhaps the nesting of electrodes in our models (rather than testing electrode locations as a fixed effect) washed out effects that were more prominent at some sites than others. That is, had we targeted only frontal electrodes, an L1 PMN for tones would have been observed. Perhaps the large LPC observed for tones in the L1 group actually began early enough to wash out PMN and N400 effects at posterior electrodes. In contrast, significant L2 PMNs for tones seem less much weaker, regardless of electrode site. If this were the case, it would constitute some evidence in favor of a different processing timeline for L1 and L2 listeners. However, further targeted investigation with new datasets would be needed to draw any such conclusions[4].

Regardless of how we understand the group differences in the PMN window, the later LPCs indicate that for both groups, correctly rejected trials ultimately lead to the same process of repair or reanalysis. The subtle differences between the patterns observed at the LPC, however, are intriguing. As in several prior ERP studies (Pelzl, 2018; Pelzl et al., 2021), in the L1 group we observed a slightly larger LPC for tone mismatches than vowel mismatches. In the L2 group there was no such tendency. Although we had no predictions for group differences for correct trials in this later time window, it is tempting to speculate that the slightly larger LPC for tone

[4] Another explanation for between group differences might be the greater number of correctly judged trials available for analysis. Since L1 responded correctly more often, the magnitude of PMN responses was greater. However, this doesn't fit with the response patterns of the L1 group itself, as they responded incorrectly to more vowel trials than tone trials.

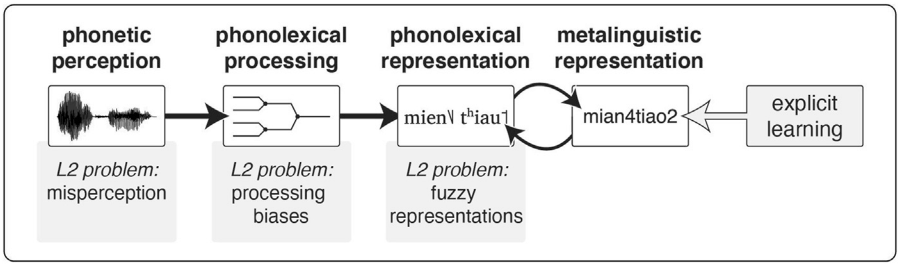

phonetic perception
phonolexical processing
phonolexical representation
metalinguistic representation
mien˥˩ tʰiau˧˥
mian4tiao2
explicit learning
L2 problem: misperception
L2 problem: processing biases
L2 problem: fuzzy representations

FIGURE 8 | Expanded problem space for L2 tone word recognition. Explicit learning of tones in relation to words may result in separate metalinguistic representations that interact with, but do not necessarily directly reflect the information encoded in phonolexical representations.

mismatches in L1 is a reflection of stronger sensitivity to them—perhaps increased attempts by L1 listeners to reanalyze the input or consider alternative explanations for the unexpected tone. This may be an interesting avenue for future investigation.

Fuzzy Tone Word Representations and Metalinguistic Tone Word Knowledge

Present results once again demonstrate that L2 tone words fit well under the umbrella of fuzzy L2 lexical representations. As with some other L2 instances of fuzzy lexical representations, the fuzziness can be directly linked to the difficulty of novel L2 speech sounds (Broersma and Cutler, 2008; Broersma and Cutler, 2011; Broersma, 2012; Díaz et al., 2012). However, L2 tone word difficulties may also be somewhat unique. Rather than competing with existing L1 phonological contrasts, tones may exist outside the native language phonological space, requiring learners to use F0 cues in a new way. For this reason, some of the fuzzy lexical effects found in L2 tone studies may be qualitatively different from those documented in other L2 contexts. In particular, there is a possibility for metalinguistic tone knowledge to play a very strong role in L2 tone word recognition.

As in other areas of L2 learning, the contrast between implicit and explicit knowledge might be a key for understanding L2 tone word outcomes (DeKeyser, 2003; Suzuki and DeKeyser, 2017). Whereas L2 learners spend great effort to establish metalinguistic tone word representations (encoded in writing via Pinyin romanization), these metalinguistic representations may be a separate form of knowledge that is not automatically drawn upon during word recognition. This is depicted in **Figure 8**. While implicit (fuzzy) tone word lexical representations guide L2 word recognition in the earliest, automatic stages, the metalinguistic tone word representations can serve to identify words (with effort) in later stages. While most often the implicit and metalinguistic representations will be aligned, occasionally, it may happen that despite correct explicit word knowledge, L2 speakers might still have weakly developed implicit phonolexical representations. As these fuzzy representations take the lead during word recognition, they can lead to occasional behavioral errors, even in tasks that allow learners to draw heavily on their explicit knowledge.

LIMITATIONS

While results of the present study are consistent with previous work showing weaknesses in advanced L2 tone perception, we acknowledge some clear limitations. First, the sample of participants, especially L2 participants, was relatively small. Advanced L2 Mandarin learners are difficult to find, but this practical consideration does not affect the statistical facts: it certainly could be the case that we had insufficient power to detect smaller differences between groups and/or conditions, especially for ERP outcomes. Though difficult, it is worth striving to improve in this regard in future work (Brysbaert, 2020).

Second, present results may have been impacted by an ordering effect. As part of a larger set of experiments, the picture-phonology matching task always followed a lexical decision task (see **Supplementary Appendix A**). No stimuli were shared between the lexical decision and picture-phonology matching experiments, but it is possible that L2 participants were more aware of tones in the picture-phonology experiment as they had already experienced the lexical decision task. We did not consider this a problem for the current study, where we aimed to give L2 learners the best chance possible to succeed at the task.

CONCLUSION

This study provides converging evidence of weaknesses in tone word recognition by advanced L2 learners. Learners have clear difficulty in encoding tones in explicit long-term memory, and "Best Case Scenario" results suggest that, even when they do succeed in encoding tones, they do not always succeed at utilizing tones during online Mandarin word recognition. ERP results suggested L1 listeners use early sensitivity to phonological cues to successfully reject mismatching vowels, but there was no clear evidence of other ERP effects in either the L1 or L2 group.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/3aue9/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board at the University of Maryland, College Park. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EP, EL, and RD contributed to conception and design of the study. EP and TG conducted the research in Beijing. EP conducted the research at Maryland. EP processed the data, performed the statistical analysis, and wrote the first draft of the manuscript. EL contributed sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was supported in part by NSF-IGERT grant 0801465 and Ling-DDRI grant 1728851 to pay for data collection. Open access publication fee provided through the NSF SBE Postdoctoral Research Fellowship under Grant No. 2004279. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.689423/full#supplementary-material

REFERENCES

Alexander, J. A., Wong, P. C., and Bradlow, A. R. (2005). Lexical Tone Perception in Musicians and Non-musicians. *Interspeech*, 397–400. Available at: http://groups.linguistics.northwestern.edu/speech_comm_group/publications/2005/Alexander-Wong-Bradlow-2005.pdf (Accessed May 13, 2015).

Amengual, M. (2016). The Perception of Language-specific Phonetic Categories Does Not Guarantee Accurate Phonological Representations in the Lexicon of Early Bilinguals. *Appl. Psycholinguistics* 37 (05), 1221–1251. doi:10.1017/S0142716415000557

Baddeley, A. D. (1968). How Does Acoustic Similarity Influence Short-Term Memory? *Q. J. Exp. Psychol.* 20 (3), 249–264. doi:10.1080/14640746808400159

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal. *J. Mem. Lang.* 68 (3), 255–278. doi:10.1016/j.jml.2012.11.001

Barrios, S. L., Namyst, A. M., Lau, E. F., Feldman, N. H., and Idsardi, W. J. (2016). Establishing New Mappings between Familiar Phones: Neural and Behavioral Evidence for Early Automatic Processing of Nonnative Contrasts. *Front. Psychol.* 7, 1–16. doi:10.3389/fpsyg.2016.00995

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious Mixed Models. ArXiv:1506.04967 [Stat]. Available at: http://arxiv.org/abs/1506.04967 (Accessed May 4, 2018). doi:10.5821/palimpsesto.14.4719

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting Linear Mixed-Effects Models Usinglme4. *J. Stat. Soft.* 67 (1), 1–48. doi:10.18637/jss.v067.i01

Bent, T., Bradlow, A. R., and Wright, B. A. (2006). The Influence of Linguistic Experience on the Cognitive Processing of Pitch in Speech and Nonspeech Sounds. *J. Exp. Psychol. Hum. Perception Perform.* 32 (1), 97–103. doi:10.1037/0096-1523.32.1.97

Best, C. T., and Tyler, M. D. (2007). "Nonnative and Second-Language Speech Perception," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*. Editors O.-S. Bohn and M. J. Munro (Amsterdam: John Benjamins), 13–34. doi:10.1075/lllt.17.07bes

Braun, B., Galts, T., and Kabak, B. (2014). Lexical Encoding of L2 Tones: The Role of L1 Stress, Pitch Accent and Intonation. *Second Lang. Res.* 30 (3), 323–350. doi:10.1177/0267658313510926

Braun, B., and Johnson, E. K. (2011). Question or Tone 2? How Language Experience and Linguistic Function Guide Pitch Processing. *J. Phonetics* 39, 585–594. doi:10.1016/j.wocn.2011.06.002

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., and Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLoS ONE* 5 (5), e10773. doi:10.1371/journal.pone.0010773

Broersma, M., and Cutler, A. (2011). Competition Dynamics of Second-Language Listening. *Q. J. Exp. Psychol.* 64 (1), 74–95. doi:10.1080/17470218.2010.499174

Broersma, M., and Cutler, A. (2008). Phantom Word Activation in L2. *System* 36, 22–34. doi:10.1016/j.system.2007.11.003

Broersma, M. (2012). Increased Lexical Activation and Reduced Competition in Second-Language Listening. *Lang. Cogn. Process.* 27 (7–8), 1205–1224. doi:10.1080/01690965.2012.660170

Broselow, E., Hurtig, R. R., and Ringen, C. (1987). "The Perception of Second Language Prosody," in *Interlanguage Phonology: The Acquisition of a Second Language Sound System*. Editors G. Ioup and S. H. Weinberger (New York, NY: Newbury House Publishers), 350–364.

Brown-Schmidt, S., and Canseco-Gonzalez, E. (2004). Who Do You Love, Your Mother or Your Horse? An Event-Related Brain Potential Analysis of Tone Processing in Mandarin Chinese. *J. Psycholinguist Res.* 33 (2), 103–135. doi:10.1023/b:jopr.0000017223.98667.10

Brunellière, A., and Soto-Faraco, S. (2015). The Interplay between Semantic and Phonological Constraints during Spoken-word Comprehension. *Psychophysiol.* 52 (1), 46–58. doi:10.1111/psyp.12285

Brysbaert, M. (2020). Power Considerations in Bilingualism Research: Time to Step up Our Game. *Bilingualism* 1, 1–6. doi:10.1017/S1366728920000437

Bundgaard-Nielsen, R. L., Best, C. T., and Tyler, M. D. (2011). Vocabulary Size Matters: The Assimilation of Second-Language Australian English Vowels to First-Language Japanese Vowel Categories. *Appl. Psycholinguistics* 32 (1), 51–67. doi:10.1017/S0142716410000287

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS One* 5 (6), e10729. doi:10.1371/journal.pone.0010729

Chan, R. K., and Leung, J. H. (2020). Why Are Lexical Tones Difficult to Learn? *Stud. Second Lang. Acquis.* 42 (1), 33–59. doi:10.1017/S0272263119000482

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. M. (2010). Individual Variability in Cue-Weighting and Lexical Tone Learning. *J. Acoust. Soc. America* 128 (1), 456–465. doi:10.1121/1.3445785

Chang, C. B., and Bowles, A. R. (2015). Context Effects on Second-Language Learning of Tonal Contrasts. *J. Acoust. Soc. America* 138 (6), 3703–3716. doi:10.1121/1.4937612

Chang, C. B. (2018). Perceptual Attention as the Locus of Transfer to Nonnative Speech Perception. *J. Phonetics* 68, 85–102. doi:10.1016/j.wocn.2018.03.003

Chrabaszcz, A., and Gor, K. (2014). Context Effects in the Processing of Phonolexical Ambiguity in L2. *Lang. Learn.* 64 (3), 415–455. doi:10.1111/lang.12063

Connolly, J. F., and Phillips, N. A. (1994). Event-related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *J. Cogn. Neurosci.* 6 (3), 256–266. doi:10.1162/jocn.1994.6.3.256

Cook, S. V., and Gor, K. (2015). Lexical Access in L2. *Ml* 10 (2), 247–270. doi:10.1075/ml.10.2.04coo

Cook, S. V., Pandža, N. B., Lancaster, A. K., and Gor, K. (2016). Fuzzy Nonnative Phonolexical Representations Lead to Fuzzy Form-To-Meaning Mappings. *Front. Psychol.* 7, 1–17. doi:10.3389/fpsyg.2016.01345

Coulson, S., King, J. W., and Kutas, M. (1998). ERPs and Domain Specificity: Beating a Straw Horse. *Lang. Cogn. Process.* 13 (6), 653–672. doi:10.1080/016909698386410

Darcy, I., Daidone, D., and Kojima, C. (2013). Asymmetric Lexical Access and Fuzzy Lexical Representations in Second Language Learners. *Ml* 8 (3), 372–420. doi:10.1075/ml.8.3.06dar

DeKeyser, R. M. (2003). "Implicit and Explicit Learning," in *The Handbook of Second Language Acquisition*. Editors C. J. Doughty and M. H. Long (Malden, MA: Blackwell Publishing), 313–348.

DeKeyser, R. M. (2005). What Makes Learning Second-Language Grammar Difficult? A Review of Issues. *Lang. Learn.* 55 (S1), 1–25. doi:10.1111/j.0023-8333.2005.00294.x

Desroches, A. S., Newman, R. L., and Joanisse, M. F. (2009). Investigating the Time Course of Spoken Word Recognition: Electrophysiological Evidence for the Influences of Phonological Similarity. *J. Cogn. Neurosci.* 21 (10), 1893–1906. doi:10.1162/jocn.2008.21142

Díaz, B., Mitterer, H., Broersma, M., and Sebastián-Gallés, N. (2012). Individual Differences in Late Bilinguals' L2 Phonological Processes: From Acoustic-Phonetic Analysis to Lexical Access. *Learn. Individual Differences* 22 (6), 680–689. doi:10.1016/j.lindif.2012.05.005

Diependaele, K., Lemhöfer, K., and Brysbaert, M. (2013). The Word Frequency Effect in First- and Second-Language Word Recognition: A Lexical Entrenchment Account. *Q. J. Exp. Psychol.* 66 (5), 843–863. doi:10.1080/17470218.2012.720994

Ellis, N. C. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Appl. Linguistics* 27 (2), 164–194. doi:10.1093/applin/aml015

Escudero, P., and Boersma, P. (2004). Bridging the gap between L2 Speech Perception Research and Phonological Theory. *Stud. Sec. Lang. Acq.* 26 (04). doi:10.1017/S0272263104040021

Flege, J. E. (1995). "Second Language Speech Learning: Theory, Findings, and Problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Editor W. Strange (Timonium, MD: York), 233–277.

Gor, K., and Cook, S. V. (2020). A Mare in a Pub? Nonnative Facilitation in Phonological Priming. *Second Lang. Res.* 36 (1), 123–140. doi:10.1177/0267658318769962

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The Linguistic Processes Underlying the P600. *Lang. Cogn. Process.* 25 (2), 149–188. doi:10.1080/01690960902965951

Han, J.-I., and Tsukada, K. (2020). Lexical Representation of Mandarin Tones by Non-tonal Second-Language Learners. *J. Acoust. Soc. America* 148 (1), EL46–EL50. doi:10.1121/10.0001586

Hao, Y.-C. (2018). Contextual Effect in Second Language Perception and Production of Mandarin Tones. *Speech Commun.* 97, 32–42. doi:10.1016/j.specom.2017.12.015

Hao, Y.-C. (2012). Second Language Acquisition of Mandarin Chinese Tones by Tonal and Non-tonal Language Speakers. *J. Phonetics* 40 (2), 269–279.

Ho, A. T. (1976). The Acoustic Variation of Mandarin Tones. *Phonetica* 33, 353–367. doi:10.1159/000259792

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biom. J.* 50 (3), 346–363. doi:10.1002/bimj.200810425

Howie, J. M. (1974). On the Domain of Tone in Mandarin. *Phonetica* 30 (3), 129–148. doi:10.1159/000259484

Hu, J., Gao, S., Ma, W., and Yao, D. (2012). Dissociation of Tone and Vowel Processing in Mandarin Idioms. *Psychophysiol.* 49 (9), 1179–1190. doi:10.1111/j.1469-8986.2012.01406.x

Huang, J., and Holt, L. L. (2009). General Perceptual Contributions to Lexical Tone Normalization. *J. Acoust. Soc. America* 125 (6), 3983–3994. doi:10.1121/1.3125342

Huang, T., and Johnson, K. (2011). Language Specificity in Speech Perception: Perception of Mandarin Tones by Native and Nonnative Listeners. *Phonetica* 67 (4), 243–267. doi:10.1159/000327392

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kaan, E., and Swaab, T. Y. (2003). Repair, Revision, and Complexity in Syntactic Analysis: An Electrophysiological Differentiation. *J. Cogn. Neurosci.* 15 (1), 98–110. doi:10.1162/089892903321107855

Kujala, A., Alho, K., Service, E., Ilmoniemi, R. J., and Connolly, J. F. (2004). Activation in the Anterior Left Auditory Cortex Associated with Phonological Analysis of Speech Input: Localization of the Phonological Mismatch Negativity Response with MEG. *Cogn. Brain Res.* 21 (1), 106–113. doi:10.1016/j.cogbrainres.2004.05.011

Leckey, M., and Federmeier, K. D. (2019). The P3b and P600(s): Positive Contributions to Language Comprehension. *Psychophysiology* 57, e13351. doi:10.1111/psyp.13351

Lee, C.-Y., Tao, L., and Bond, Z. S. (2010). Identification of Multi-Speaker Mandarin Tones in Noise by Native and Non-native Listeners. *Speech Commun.* 52 (11–12), 900–910. doi:10.1016/j.specom.2010.01.004

Lenth, R. (2020). Emmeans: Estimated Marginal Means, Aka Least-Squares Means. (R package version 1.5.3.) [Computer software]. Available at: https://CRAN.R-project.org/package=emmeans.

Ling, W., and Grüter, T. (2020). From Sounds to Words: The Relation between Phonological and Lexical Processing of Tone in L2 Mandarin. *Second Lang. Res.*, 026765832094154, doi:10.1177/0267658320941546

Liu, Y., Shu, H., and Wei, J. (2006). Spoken Word Recognition in Context: Evidence from Chinese ERP Analyses. *Brain Lang.* 96 (1), 37–48. doi:10.1016/j.bandl.2005.08.007

Llompart, M., and Reinisch, E. (2020). The Phonological Form of Lexical Items Modulates the Encoding of Challenging Second-Language Sound Contrasts. *J. Exp. Psychol. Learn. Mem. Cogn.* 46 (8), 1590–1610. doi:10.1037/xlm0000832

Luck, S. J., and Gaspelin, N. (2017). How to Get Statistically Significant Effects in Any ERP experiment (And Why You Shouldn't). *Psychophysiol.* 54 (1), 146–157. doi:10.1111/psyp.12639

MacWhinney, B., and Bates, E. (1989). *The Cross-Linguistic Study of Sentence Processing* (New York, NY: Cambridge University Press).

Malins, J. G., and Joanisse, M. F. (2012). Setting the Tone: An ERP Investigation of the Influences of Phonological Similarity on Spoken Word Recognition in Mandarin Chinese. *Neuropsychologia* 50 (8), 2032–2043. doi:10.1016/j.neuropsychologia.2012.05.002

Matusevych, Y., Kamper, H., Schatz, T., Feldman, N. H., and Goldwater, S. (2021). A Phonetic Model of Non-native Spoken Word Processing. ArXiv:2101.11332 [Cs]. Available at: http://arxiv.org/abs/2101.11332 (Accessed March 12, 2021).

Moreno-Martínez, F. J., and Montoro, P. R. (2012). An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables. *PLoS ONE* 7 (5), e37527. doi:10.1371/journal.pone.0037527

Newman, R. L., and Connolly, J. F. (2009). Electrophysiological Markers of Pre-lexical Speech Processing: Evidence for Bottom-Up and Top-Down Effects on Spoken Word Processing. *Biol. Psychol.* 80 (1), 114–121. doi:10.1016/j.biopsycho.2008.04.008

Nixon, J. S. (2020). Of Mice and Men: Speech Sound Acquisition as Discriminative Learning from Prediction Error, Not Just Statistical Tracking. *Cognition* 197, 104081. doi:10.1016/j.cognition.2019.104081

Osterhout, L., and Holcomb, P. J. (1992). Event-related Brain Potentials Elicited by Syntactic Anomaly. *J. Mem. Lang.* 31, 785–806. doi:10.1016/0749-596x(92)90039-z

Pelzl, E., Lau, E. F., Guo, T., and DeKeyser, R. (2019). Advanced Second Language Learners' Perception of Lexical Tone Contrasts. *Stud. Second Lang. Acquis* 41 (1), 59–86. doi:10.1017/S0272263117000444

Pelzl, E., Lau, E. F., Guo, T., and DeKeyser, R. (2020). Even in the Best-Case Scenario L2 Learners Have Persistent Difficulty Perceiving and Utilizing Tones in Mandarin. *Stud. Second Lang. Acquis*, 1–29. doi:10.1017/S027226312000039X

Pelzl, E., Lau, E. F., Jackson, S. R., Guo, T., and Gor, K. (2021). Behavioral and Neural Responses to Tone Errors in Foreign-Accented Mandarin. *Lang. Learn.* 71, 414–452. doi:10.1111/lang.12438

Pelzl, E. (2018). *Second Language Lexical Representation and Processing of Mandarin Chinese Tones*. College Park: University of Maryland.

Pelzl, E. (2019). What Makes Second Language Perception of Mandarin Tones Hard? *Csl* 54 (1), 51–78. doi:10.1075/csl.18009.pel

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Romero-Rivas, C., Martin, C. D., and Costa, A. (2015). Processing Changes when Listening to Foreign-Accented Speech. *Front. Hum. Neurosci.* 9, 1–15. doi:10.3389/fnhum.2015.00167

Sassenhagen, J., and Bornkessel-Schlesewsky, I. (2015). The P600 as a Correlate of Ventral Attention Network Reorientation. *Cortex* 66, A3–A20. doi:10.1016/j.cortex.2014.12.019

Sassenhagen, J., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2014). The P600-As-P3 Hypothesis Revisited: Single-Trial Analyses Reveal that the Late EEG Positivity Following Linguistically Deviant Material Is Reaction Time Aligned. *Brain Lang.* 137, 29–39. doi:10.1016/j.bandl.2014.07.010

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to Capitalize on A Priori Contrasts in Linear (Mixed) Models: A Tutorial. *J. Mem. Lang.* 110, 104038. doi:10.1016/j.jml.2019.104038

Schaefer, V., and Darcy, I. (2014). Lexical Function of Pitch in the First Language Shapes Cross-Linguistic Perception of Thai Tones. *Lab. Phonology* 5 (4), 489–522. doi:10.1515/lp-2014-0016

Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., and Chen, H.-C. (2005). Brain Responses to Segmentally and Tonally Induced Semantic Violations in Cantonese. *J. Cogn. Neurosci.* 17 (1), 1–12. doi:10.1162/0898929052880057

Schmidt-Kassow, M., and Kotz, S. A. (2009). Event-related Brain Potentials Suggest a Late Interaction of Meter and Syntax in the P600. *J. Cogn. Neurosci.* 21 (9), 1693–1708. doi:10.1162/jocn.2008.21153

Shen, G., and Froud, K. (2016). Categorical Perception of Lexical Tones by English Learners of Mandarin Chinese. *J. Acoust. Soc. America* 140 (6), 4396–4403. doi:10.1121/1.4971765

Shen, G., and Froud, K. (2019). Electrophysiological Correlates of Categorical Perception of Lexical Tones by English Learners of Mandarin Chinese: An ERP Study. *Bilingualism* 22 (2), 253–265. doi:10.1017/S136672891800038X

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (2017). Afex: Analysis of Factorial Experiments. R package version 0.17-8. Available at: http://cran.r-project.org/package=afex [Computer software].

So, C. K., and Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Lang. Speech* 53 (2), 273–293. doi:10.1177/0023830909357156

So, C. K., and Best, C. T. (2014). Phonetic Influences on English and French Listeners' Assimilation of Mandarin Tones to Native Prosodic Categories. *Stud. Second Lang. Acquis* 36 (02), 195–221. doi:10.1017/S0272263114000047

Strange, W. (2011). Automatic Selective Perception (ASP) of First and Second Language Speech: A Working Model. *J. Phonetics* 39 (4), 456–466. doi:10.1016/j.wocn.2010.09.001

Sun, S. H. (1998). *The Development of a Lexical Tone Phonology in American Adult Learners of Standard Mandarin Chinese*. Honolulu, HI: Second Language Teaching & Curriculum Center.

Suzuki, Y., and DeKeyser, R. (2017). The Interface of Explicit and Implicit Knowledge in a Second Language: Insights From Individual Differences in Cognitive Aptitudes. *Lang. Learn.* 67 (4), 747–790. doi:10.1111/lang.12241

Tsukada, K., and Han, J.-I. (2019). The Perception of Mandarin Lexical Tones by Native Korean Speakers Differing in Their Experience with Mandarin. *Second Lang. Res.* 35 (3), 305–318. doi:10.1177/0267658318775155

Van Den Brink, D., Brown, C. M., and Hagoort, P. (2001). Electrophysiological Evidence for Early Contextual Influences during Spoken-word Recognition: N200 versus N400 Effects. *J. Cogn. Neurosci.* 13 (7), 967–985. doi:10.1162/089892901753165872

VanPatten, B. (1996). *Input Processing and Grammar Instruction in Second Language Acquisition*. New York, NY: Ablex Publishing.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American Listeners to Perceive Mandarin Tones. *J. Acoust. Soc. America* 106, 3649–3658. doi:10.1121/1.428217

Wiener, S., Ito, K., and Speer, S. R. (2018). Early L2 Spoken Word Recognition Combines Input-Based and Knowledge-Based Processing. *Lang. Speech* 61 (4), 632–656. doi:10.1177/0023830918761762

Wiener, S., Ito, K., and Speer, S. R. (2020). Effects of Multitalker Input and Instructional Method on the Dimension-Based Statistical Learning of Syllable-Tone Combinations. *Stud. Second Lang. Acquis* 43, 155–180. doi:10.1017/S0272263120000418

Wiener, S., Lee, C. Y., and Tao, L. (2019). Statistical Regularities Affect the Perception of Second Language Speech: Evidence From Adult Classroom Learners of Mandarin Chinese. *Lang. Learn.* 69 (3), 527–558. doi:10.1111/lang.12342

Wong, P. C. M., and Perrachione, T. K. (2007). Learning Pitch Patterns in Lexical Identification by Native English-speaking Adults. *Appl. Psycholinguistics* 28 (04), 565–585. doi:10.1017/s0142716407070312

Xu, Y. (1997). Contextual Tonal Variations in Mandarin. *J. Phonetics* 25, 61–83. doi:10.1006/jpho.1996.0034

Xu, Y. (1994). Production and Perception of Coarticulated Tones. *J. Acoust. Soc. America* 95 (4), 2240–2253. doi:10.1121/1.408684

Zhao, J., Guo, J., Zhou, F., and Shu, H. (2011). Time Course of Chinese Monosyllabic Spoken Word Recognition: Evidence from ERP Analyses. *Neuropsychologia* 49 (7), 1761–1770. doi:10.1016/j.neuropsychologia.2011.02.054

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* 20 (5), 1017–1029. doi:10.1017/S1366728916000791

Zou, Y., Lui, M., and Tsang, Y.-K. (2020). The Roles of Lexical Tone and Rime during Mandarin Sentence Comprehension: An Event-Related Potential Study. *Neuropsychologia* 147, 107578. doi:10.1016/j.neuropsychologia.2020.107578

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.