LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Behavioral and Neural Responses to Tone Errors in Foreign-Accented Mandarin

Eric Pelzl,[a,b] Ellen F. Lau,[a] Scott R. Jackson,[a] Taomei Guo,[c] and Kira Gor[a]

[a]University of Maryland, College Park, [b]The Pennsylvania State University, and [c]Beijing Normal University

Abstract: Previous event-related potentials (ERP) research has investigated how foreign accent modulates listeners' neural responses to lexical-semantic and morphosyntactic errors. We extended this line of research to consider whether pronunciation errors in Mandarin Chinese are processed differently when a foreign-accented speaker makes them relative to when a native-accented speaker makes them (a conceptual replication using the materials from Pelzl et al., 2019). We evaluated behavioral judgments, the N400, and late positive component while native speakers listened to native and foreign-accented sentences containing tone and rhyme pronunciation errors. We observed effects that suggested that the participants were prone to detect errors in foreign-accented speech more often in sentences with no critical word deviation but also were less likely to reject critical tone errors produced by the foreign-accented speaker. ERP results showed a main effect of accent on late positive components that suggested a difference in degree for sensitivity to foreign-accented compared to native-accented pronunciation errors rather than a completely different response pattern. We found no effect of accent on N400s, with statistically significant differences between tone and rhyme errors regardless of speaker accent.

Thanks to Junjie Wu, Di Lu, Yongben Fu, Shuhua Li, and Chunyan Kang for help running participants in Beijing, to Man Li for help proofreading initial rounds of Chinese stimuli, to Anna Chrabaszcz for helpful advice and MATLAB scripts, and to Brendan Cone for help editing many auditory sentence files. Thanks also to Kara Morgan-Short and three anonymous reviewers for their constructive comments and questions. This research was supported in part by NSF-IGERT grant 0801465, NSF-EAPSI grant 1514936, and NSF grant 1749407. Supplementary materials, stimuli, data, and R code can be found at https://osf.io/vysx7/

Correspondence concerning this article should be addressed to Eric Pelzl at Moore Building 111, Penn State University, University Park, PA 16802, USA. E-mail: ezp218@psu.edu.

The handling editor for this article was Kara Morgan-Short.

DOI: 10.1111/lang.12438

Keywords accented speech; pronunciation; lexical tones; Mandarin, event-related potentials

Introduction

Understanding how native listeners process foreign-accented speech has important practical applications and can provide unique insights into basic speech comprehension mechanisms. Using event-related potentials (ERPs) to measure neural responses to errors in spoken sentences, a number of studies have shown that foreign-accented speech can modulate typical responses to lexical and semantic errors (Hanulíková, van Alphen, van Goch, & Weber, 2012; Romero-Rivas, Martin, & Costa, 2015). These effects have been attributed to differences in listeners' expectations about the linguistic properties of foreign-accented speech or to listeners' increased difficulty processing that speech.

We extended this line of research, using ERPs to examine whether native Chinese listeners respond differently to pronunciation errors made by a native or by a foreign-accented Mandarin speaker. Because Mandarin is a tonal language, both segmental (consonants, vowels) and suprasegmental (tones) pronunciation errors are possible (Chen, 1999; Wan & Jaeger, 1998). Adult English speakers who learn Mandarin as a second language (L2) are known to have difficulty mastering production of tones (e.g., Chen, Wee, Tong, Ma, & Li, 2016). With this in mind, we manipulated highly constraining sentences so that sometimes the critical expected word was mispronounced by either a single tone or a single rhyme (vowel and syllable-final/n/or/ŋ/), resulting in a nonword. While their electroencephalogram (EEG) was recorded, Chinese listeners heard sentences spoken by both a native Mandarin speaker and an American L2 speaker of Mandarin and judged whether the sentences sounded correct.

Our ERP analyses focused on the N400 and the late positive component (LPC). The N400 is a negative-going deflection that peaks around 400 milliseconds after stimulus onset. N400s can be observed in response to each word in a sentence as the sentence unfolds over time and will reliably grow more negative in response to more unexpected words (the N400 effect). In the sentence processing literature, the N400 is typically used as an index of difficulty in access to or integration of lexical targets (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980, 1984; Lau, Phillips, & Poeppel, 2008). LPCs are positive-going deflections that occur late—following N400s in sentence processing studies—though with large variation in onset and duration. They can be reliably elicited by a wide-range of mismatches in sentence comprehension, including syntactic violations (Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994), lexical mismatches (e.g., Romero-Rivas et al., 2015; Schirmer, Tang, Penney, Gunter, & Chen, 2005), and phonological mismatches (e.g., stress mismatches in Schmidt-Kassow & Kotz, 2009). We have noted below relevant functional interpretations of the LPC in context, with the caveat that functional accounts of LPC effects are actively debated (see Leckey & Federmeier, 2019).

We compared N400s and LPCs in listeners' responses to pronunciation errors in Mandarin to determine (a) whether the presence of a foreign accent impacts how listeners process pronunciation errors and (b) whether listeners process tone and rhyme errors in the same way.

Background Literature

Previous Event-Related Potential Studies on Accented Speech Processing

A growing number of ERP studies have examined how a speaker's accent can impact listeners' processing of lexical-semantic and syntactic errors, finding notable differences depending on the nature of the error and of the listeners' previous experience with accented speech (Caffarra & Martin, 2019; Grey & van Hell, 2017; Hanulíková et al., 2012; Romero-Rivas et al., 2015; Xu, Abdel Rahman, & Sommer, 2019).

In general, effects of accent tend to be attributed to two main causes (cf. Lev-Ari, 2015). First, listeners may hold different expectations about the linguistic properties of foreign-accented speech. Some syntactic errors or pronunciation patterns that would be odd for a native speaker may be considered normal for an accented L2 speaker (Brunellière & Soto-Faraco, 2013; Caffarra & Martin, 2019; Grey & van Hell, 2017; Hanulíková et al., 2012). Second, the accented speech itself may be less intelligible and thus more difficult to process, requiring more effort and time on the part of listeners who are unfamiliar with the novel pronunciation patterns (Goslin, Duffy, & Floccia, 2012; Romero-Rivas et al., 2015).

Hanulíková et al. (2012) observed LPCs (also called P600s) when a native Dutch speaker committed grammatical gender violations in Dutch, but did not observe LPCs when a Turkish-accented Dutch speaker made the same violations. They suggested that one way to interpret these results is that they show Dutch listeners have learned from broader experience to expect frequent grammatical gender errors from Turkish-accented speakers. This leads to reduced attempts to repair those errors as indexed by the late positivities.

The same pattern of LPCs occurring in response to native errors, but not foreign-accented ones, has since been found for a variety of languages and linguistic features (Caffarra & Martin, 2019; Romero-Rivas et al., 2015; Xu et al., 2019). Although providing general support for the role of listener expectations in foreign-accented speech processing, these studies have introduced additional complexities. For example, Grey and van Hell (2017) found that gender mismatched pronouns in English sentences elicited frontal negativities (the Nref) for native English speech but not for Chinese-accented English. They also conducted an exploratory subset analysis considering listeners' subjective familiarity with Chinese-accented speakers. Listeners who were able to identify the foreign accent as Asian showed early negativities in response to pronoun errors, but those who could not identify the accent showed no response. Caffarra and Martin (2019) manipulated the typicality of grammatical errors. Whereas L2 speakers of Spanish often make grammatical gender errors, they rarely make number errors. Caffarra and Martin found that both types of errors elicited late positive responses for native accented speech, but only the atypical number violation errors elicited such responses in foreign-accented speech. Follow-up correlational analysis found that the size of late positive responses for typical errors (but not atypical errors) in foreign-accented Spanish was related to listeners' familiarity with that type of accented speech. Speakers who were more familiar with English-accented Spanish speakers showed weaker late positive responses to typical errors (i.e., grammatical gender violations) than those who reported little experience with this type of accent.

Although listener expectations can offer a compelling explanation for ERP results, it may be more fitting to attribute observed patterns to the role of processing difficulty in some cases. Most studies have used only one or two accented speakers from the same language background, but Romero-Rivas et al. (2015) presented Spanish listeners with accented speech produced by a variety of differently accented L2 speakers (e.g., French, Greek, Italian, Japanese). They found a general increase in N400 responses to (correct) words in foreign-accented speech relative to native speech, though this effect decreased from the first to second half of the experiment. In the second half of the experiment, they also began presenting semantically misfitting words on some trials (e.g., "Coming to Barcelona we always cross a **piano* in the highway.") and found stronger N400 effects for foreign-accented speech than for native Spanish speech. They interpreted these outcomes as an indication of the relative difficulty listeners had in accessing words in foreign-accented speech.

Among accent studies, however, Romero-Rivas et al.'s (2015) findings represent just one of many different patterns of N400 results. Hanulíková et al. (2012) found no N400 differences for lexical-semantic violations between accent conditions (see similar results for world knowledge effects in Foucart, Costa, Morís-Fernández, & Hartsuiker, 2020). Grey and van Hell (2017) found N400 effects only for native-accented English semantic errors, with delayed negativities for the same errors in Chinese-accented English. Xu et al. (2019) similarly reported finding N400 effects only for native-accented German speech errors (i.e., slips of the tongue), but no significant N400 differences for such errors in Chinese-accented German. Finally, Goslin et al. (2012) found weaker N400 effects for (correct) foreign-accented words compared to native and regionally accented words. These studies differed in many ways (e.g., task, number of talkers, nature of lexical targets, proportion of errors, accent of speakers, and target languages). Varied outcomes should not then be surprising, though they do make any generalization across studies rather difficult.

The studies that we have reviewed examined responses to various lexical and syntactic errors in foreign-accented speech. Many of these errors were typical of the L2 speaker group under investigation. However, no studies have specifically investigated pronunciation errors—arguably one of the most common error types in L2 speech.

Accented Pronunciation and Pronunciation Errors

An obvious property of foreign-accented pronunciation is that it differs from native pronunciation; but not all differences have the same origins within a speaker or equal consequences for listeners (for a broad discussion, see Derwing & Munro, 2015). This led us to draw a distinction between accented pronunciation and pronunciation errors.

We conceptualized foreign accent as a shifted pattern of pronunciation in the dimensional space of phonetic features (e.g., F0, duration), with some or all of the features of the target phonological system realized in novel ways (Kleinschmidt & Jaeger, 2015). This may include both phonemic and subphonemic patterns. This accented shift is systematic, rather than random; a given sound will consistently be realized in a particular form (or range of forms). For example, a Spanish-accented speaker may systematically produce English/i/(as in *sheep*) as closer to the vowel/ɪ/(as in *ship*) (Wade, Jongman, & Sereno, 2007). This may cause some difficulty for listeners, but it is not a random speech error on the part of the speaker. This means that, given enough exposure, a listener could learn and adapt to the patterns of this type of foreign-accented speech.

Unlike accent-shifted pronunciation patterns, pronunciation errors often occur without an obvious pattern except that of being infelicitous. Furthermore, although they may sometimes appear to be shifted versions of the target sound, pronunciation errors may also include categorical errors, where one phoneme is swapped for another (e.g., *sheep* produced sounding like *shape*). Superficially, this type of speech error resembles slips of the tongue

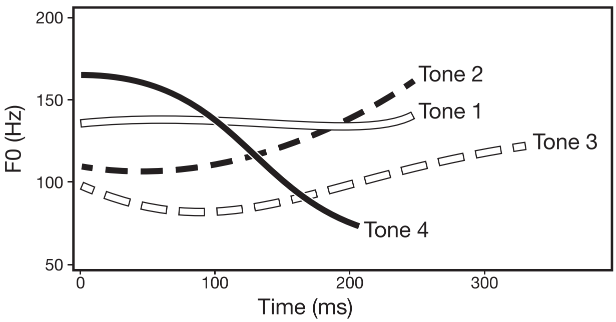

Figure 1 Example of F0 contours for the four citation tones of Mandarin.

that are typical even in native speech (J.-Y. Chen, 1999; Garnham, Shillcock, Brown, Mill, & Cutler, 1981). However, unlike slips of the tongue, the underlying cause of L2 pronunciation errors is incomplete or uncertain knowledge of the phonological composition of a L2 word. Because there is little pattern to such errors, listeners may not be able to adapt to them so easily.

Tone Errors in Foreign-Accented Mandarin

The distinction between accented pronunciation and pronunciation errors is useful for understanding tone production in foreign-accented Mandarin. Mandarin has four contrastive lexical tones (Figure 1): Tone 1 is a high level tone, indicated in Pinyin romanization with a level diacritic (¯); Tone 2 is a rising tone (´); Tone 3 is a low or low-dipping tone (ˇ); and Tone 4 is a falling tone (`). Although swapping tones can lead to different words (e.g., *mā* "mother" vs. *mǎ* "horse"), this is not always the case. The majority of words in Mandarin are disyllabic (Duanmu, 2007), so that tone errors will often result in a nonword. For example, when intending to say *mǎyǐ* "ant", with a low tone on the first syllable, if the speaker instead produced *māyǐ*, with a high tone, this would be a nonword.

As we have already mentioned, native English speakers who learn Mandarin often struggle to master control of lexical tones (Miracle, 1989; Shen, 1989; Sun, 1998; Wang, Jongman, & Sereno, 2003). Just as accents affect other phonological features, there may be shifted patterns of tone realizations in a L2 Mandarin speaker's production (Hao, 2018; Wang et al., 2003; Yang, 2016; Zhang, 2016). But when it comes to pronunciation errors, perhaps unlike most segmental errors, multidirectional category substitutions are common for tones in foreign-accented Mandarin because L2 speakers often swap any one tone for any other (N. F. Chen et al., 2016). This can occur due to interference from native language prosody (White, 1981; Yang, 2016), but it can also occur due to misremembered, inaccurately encoded, or completely forgotten tones for specific words (Pelzl, 2018). In other words, in addition to any accented features that may characterize L2 Mandarin speech, frequent pronunciation errors—especially of tones—are an additional characteristic of that speech. If tone errors are frequent enough, a reasonable response from listeners would be to ignore them. However, it is uncertain if listeners can completely ignore pronunciation errors (cf. Pelzl, Carlson, Guo, Jackson, & van Hell, 2020). So then, an alternative possibility is that frequent tone errors will simply make speech comprehension less fluent and more effortful.

Lexical Tones in Mandarin Speech Processing

Although there have been no previous ERP studies of accented speech in tonal languages, several studies have examined Mandarin and Cantonese listeners' responses to tonal and segmental errors in native speech. Brown-Schmidt and Canseco-Gonzalez (2004) presented listeners with constraining sentences in which they manipulated sentence-final monosyllabic words so that they sometimes mismatched listener expectations. There were three manipulations: segments changed while tones stayed the same (e.g., expected *tāng* "soup" vs. unexpected *shū* "book"); tones changed while segments stayed the same (e.g., *tāng* vs. *táng* "candy"); or both tone and segments changed (e.g., *tāng* vs. *dǒng* "understand"). Brown-Schmidt and Canseco-Gonzalez found N400 effects for segmental, tonal, and combined violations, with no differences among them. Schirmer et al. (2005) obtained the same pattern of results with a similar design using Cantonese sentences and listeners. These studies suggested that native Chinese listeners are equally sensitive to all types of phonological violations; critically for the current study, this included tones (for similar results with isolated words, see Malins & Joanisse, 2012; Zhao, Guo, Zhou, & Shu, 2011).

These studies used monosyllabic critical words and tested responses to vowel or tone changes that produced other meaningful words or morphemes. In other words, the critical violations were simultaneously phonological and lexical in nature. Pelzl, Lau, Guo, and DeKeyser (2019) also examined tone manipulations but used disyllabic critical words. Pronunciation errors—particularly tone errors—often function differently in disyllabic words because the error will less likely result in another existing word, whereas monosyllabic errors often do so. Pelzl et al. (2019) focused on N400 and LPC responses to tone and rhyme mismatches in native listeners and L2 Mandarin listeners. As with the other Chinese ERP studies, Pelzl et al. (2019) also found clear N400 responses for tone and rhyme mismatches in native Chinese listeners (but not in L2 listeners). However, there were also apparent differences between word and nonword responses. Compared to the expected word condition, the N400 for completely mismatching words remained more negative well beyond the 300–500 milliseconds N400 window that they tested, whereas responses in the tone and rhyme condition (which resulted in nonwords) became strongly positive in the LPC window (450–800 milliseconds) and differed significantly from both the expected word and the semantic mismatch conditions (which both resulted in real words). This finding suggested that nonword violations elicited different late processes from those elicited by real word violations. This result contrasted somewhat with that of Schirmer et al. (2005), who also found strong LPCs after tonal, segmental, or complete (both tone and segmental) violations but no differences between violation types. As we have already suggested, one reason for the differences between studies may be that the violations used by Schirmer et al. (2005) all resulted in unexpected real words, but the pronunciation violations in Pelzl et al. (2019) study resulted in nonwords.

In summary, previous ERP studies have suggested that Chinese listeners process segmental and tonal cues in largely the same manner during word recognition, but most of this research used violations that were both phonological and lexical. The present study used the design and many of the stimuli of Pelzl et al. (2019) to investigate how rhyme and tone pronunciation errors—resulting in nonwords—are processed in foreign-accented Mandarin.

The Present Study

In the present study, we considered the potential roles of listener expectations and processing difficulties in the case of foreign-accented Mandarin. We aimed to answer two questions. First, we asked: To what extent do native Mandarin listeners process pronunciation errors differently for native and foreign-accented speech? We hypothesized that native Mandarin listeners might expect more pronunciation errors, of all kinds, in foreign-accented speech relative to the speech of native speakers. This might manifest as smaller LPCs—and perhaps also smaller N400s—for both rhyme and tone pronunciation errors in foreign accented speech compared to that of native speakers. Alternatively, if accented speech increases difficulty in retrieving words, N400s might increase for foreign-accented pronunciation errors compared to the same errors in native speech. Second, we asked: To what extent do native Mandarin listeners process pronunciation errors differently for tones and rhymes? Given the particular difficulty L2 speakers have with lexical tone, we hypothesized that listeners would have heightened expectations for tone errors compared to rhyme errors. This might manifest as a greater reduction in N400 and/or LPC amplitude for tone errors compared to rhyme errors in foreign-accented speech.

To answer these questions, we conducted an ERP study with native Mandarin Chinese listeners in Beijing, China. The listeners heard sentences produced by either a native speaker of Mandarin (native accent) or by a L2 speaker (foreign accent). Sentences were highly constraining and pushed the listeners to expect a specific critical word. We compared ERPs for critical words that were expected with ERPs for three types of mismatching words: words that had a rhyme mismatch, a tone mismatch, or a complete word mismatch.

Method

Participants

We recruited 42 native speakers of Mandarin Chinese from Beijing Normal University and surrounding communities. All participants were from Mandarin speaking regions of China, were right-handed, and had no reported history of language or hearing impairment. Additionally, they were highly educated (either currently undergraduate students or already graduated). After we had processed the data, we removed eight participants from later analysis due to excessive EEG artifacts, leaving a final sample of 34 participants (17 male, 17 female; $M_{age} = 23$ years, $SD = 2.7$; range 18–30). All participants gave informed consent and were compensated for their time. The University of Maryland's Institutional Review Board had approved all procedures prior to data collection.

Event-Related Potential Experiment: Design and Materials

To investigate the research questions, we designed an auditory ERP sentence experiment with a behavioral sentence judgment task accompanying sentence trials. Materials for this study are freely available at https://osf.io/vysx7/.

Auditory materials consisted of 240 constraining sentences, each occurring with four critical word conditions (expected word, tone mismatch, rhyme mismatch, and word mismatch) and two accent conditions (native, foreign). Figure 2 illustrates the experimental paradigm. The expected word condition (e.g., *kèrén* "guest") provided a reduced N400 baseline for highly predictable words. The word mismatch condition (e.g., *zázhì* "magazine") should have induced the typical increased N400 effect relative to expected words. The tone mismatch condition tested N400 effects for mispronounced tones (e.g., a rising tone on ***ké*** *rén* [a nonword] vs. the expected falling tone on *kèrén*). The rhyme mismatch tested N400 effects for mispronounced rhymes (e.g., ***kù*** *rén* [a nonword] vs. the expected *kèrén*). The rhyme manipulation always affected vowels, but, in some cases, also affected word-final consonants (see Appendix S1

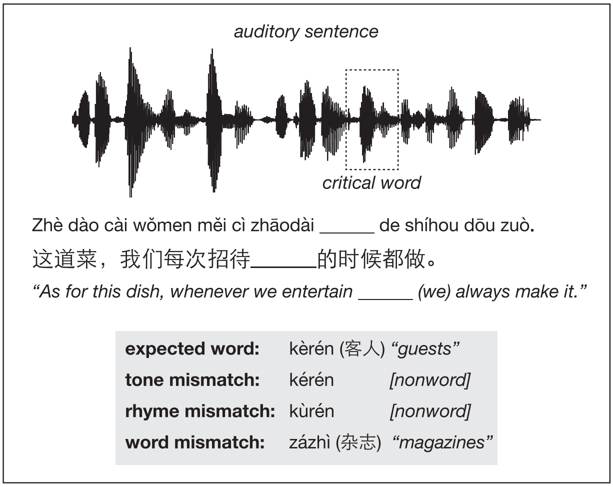

Figure 2 Example of stimulus sentence and critical words (full list of stimuli available in Appendix 5 in the online Supporting Information).

in the online Supporting Information for more details and Appendix S5 for full list of stimuli). Because these last two conditions used nonwords, N400 responses might have differed from those in the word mismatch condition (regardless of accent condition) due to the lack of lexical meaning and the correspondence of the second syllable with that of the expected word, which might have eased recovery of that word.

Each participant heard 60 sentences in each critical word condition, half produced by a male native Mandarin speaker from Beijing and half produced by a male, advanced L2 speaker from the United States.[1]

ERP Experiment: Procedures

We conducted the experiment in a quiet room in the EEG lab at Beijing Normal University. The participants sat in front of a computer with two audio speakers (Edifier R1600TIII) placed to the left and right of the display computer monitor. The experiment was run using MATLAB (Version 2013a; Mathworks, 2013) and the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Instructions were delivered on the computer monitor in Chinese and we verified comprehension of these instructions orally in Chinese. After participants had been fitted with the EEG cap, we showed them the output of the EEG on the monitor and familiarized them with the impacts of blinks and other movements on the EEG signal.

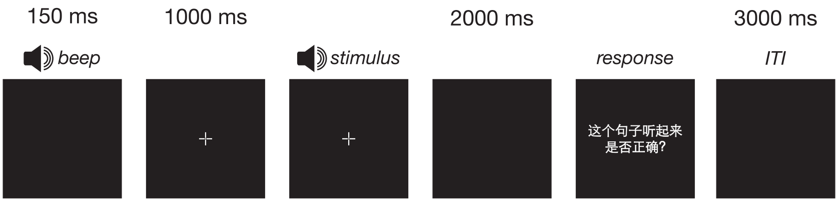

Figure 3 Trial structure and timing parameters. ITI = inter-trial interval.

Prior to beginning the experiment, we had familiarized the participants with the accompanying behavioral sentence judgment task. The experimenter explained the three types of errors that might occur in sentences. The participants then completed eight practice sentences not used in the experiment itself, half spoken by a native speaker, half spoken by a nonnative speaker. After each sentence, the participants were required to answer the question: *Zhege juzi tingqilai shifou zhengque?* "Did this sentence sound correct or not?" They answered by pressing J for yes and F for no. To illustrate what was intended by the Chinese word *zhengque* "correct," we provided feedback after each practice sentence, both in written form presented on the monitor as well as by oral feedback from a native-speaking Chinese lab assistant. Feedback was intended to make clear that only three types of error should be construed as not correct, namely: a semantically misfitting word (*bu he yujing* "does not fit context"); a word with its rhyme mispronounced (*fayin bu dui* "pronunciation is incorrect"); or a word with its tone mispronounced (*yudiao/shengdiao bu dui* "tone is incorrect"). Feedback drew attention to the specific word on which an error occurred and to what type of error had occurred.

Figure 3 illustrates the timing parameters. We asked participants to delay their responses until the prompt that occurred 2,000 milliseconds after the end of the auditory stimulus. After every 15 trials, participants took a self-paced break. The entire experiment included 240 sentences and lasted approximately one hour, not including preparation of the electrode cap.

Data Processing

We recorded raw EEG continuously at a 1,000 Hz sampling rate using a Neuroscan SynAmps data acquisition system and an electrode cap mounted with 29 silver chloride electrodes at the following sites: midline: Fz, FCz, Cz, CPz, Pz, and 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. 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 5 kΩ. The EEG and EOG recordings were amplified and digitized online at 1 kHz with a bandpass filter of 0.1–100 Hz.

For each trial for each participant, we extracted epochs time-locked to the onset of the critical word from −100:1,000 milliseconds, using preprocessing routines from the EEGLAB (Version 10.2.5.8b; Delorme & Makeig, 2004) and ERPLAB (Version 3.0.2.1; Lopez-Calderon & Luck, 2014) toolboxes to reject trials with ocular or muscular artifacts. Muscle potential, sweat, and alpha wave artifacts were identified using the peak-to-peak artifact rejection routine (for this, a 200-millisecond window was moved across the data in 100-millisecond increments and any epoch where the peak-to-peak voltage exceeded 100 microvolts [μV] was rejected). We used the step function artifact rejection routine to identify eye-blink (40 μV threshold) and eye-movement (25 μV threshold) artifacts, followed by visual confirmation of the identified artifacts by the experimenters. We removed eight participants from further analysis because they had greater than 40% of their trials rejected due to artifacts. We exported trial-level data to R (Version 3.6.1; R Core Team, 2019) for further analysis. Prior to model fitting, extreme amplitude values above |50| were excluded. For the remaining data, we retained 62,882 (86%) of 73,440 trials in the N400 window (total trials: native accent: expected word = 7,953; tone mismatch = 7,829; rhyme mismatch = 7,813; word mismatch = 7,868; foreign accent: expected word = 7,928; tone mismatch = 7,789; rhyme mismatch = 7,772; word mismatch = 7,930). In the LPC window, we retained 62,356 (85%) of 73,440 trials (total trials: native accent: expected word = 7,862; tone mismatch = 7,760; rhyme mismatch = 7,733; word mismatch = 7,815; foreign accent: expected word = 7,877; tone mismatch = 7,728; rhyme mismatch = 7,697; word mismatch = 7,884).

Behavioral Sentence Judgment Task: Analysis

We fitted mixed-effects logistic regression models with crossed random effects for subjects and items using the *mixed*() function from the afex package (Version 0.25-1; Singmann, Bolker, Westfall, & Aust, 2017), and the lme4 package (Version 1.1-21; Bates, Mächler, Bolker, & Walker, 2015) in R, with the setting *family* = *binomial*. The dependent variable was response type (binary: yes = 1, no = 0). Fixed effects included critical word (expected word, tone mismatch, rhyme mismatch, word mismatch), speaker accent

Table 1 Mean percentage of sentences judged acceptable (yes response) for the sentence judgment task

	Expected word	Tone mismatch	Rhyme mismatch	Word mismatch
Accent	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Native	91 (29)	14 (35)	7 (26)	5 (22)
Foreign	58 (49)	24 (42)	11 (32)	3 (17)

(native, foreign), and their interaction. We assumed that the participants would vary in both their responses to mismatch types, their responses to accented speech, and the way these variables interacted. Because our critical words occurred in sentences, we also expected by-item (sentence) variability given that responses to sentences would vary by critical word and accent. Based on this reasoning, we aimed to model by-subject and by-item random intercepts, and by-subject random slopes for the interaction of word and accent, and by-item random slopes for the effect of word and accent. We applied effects coding using the *mixed*() function in afex, and *p* values were obtained using the likelihood ratio test ("LRT") method. We have reported model results as chi-square tests from a mixed-effects ANOVA, which allows for a convenient summary of main effects and interactions in mixed-effects models. Models were run with the bobyqa optimizer in lme4 (Bates et al., 2015). Our goal was to retain the best fitting model with the maximal random effects structure described above (Barr, Levy, Scheepers, & Tily, 2013). The maximal model failed to converge, so we ran models with suppressed correlations for random effects and settled for a slightly less than maximal model, using lmer formula:

$$score \sim word * accent + (word + accent \parallel subj) + (word * accent \parallel item). \quad (1)$$

In order to examine possible adaptation effects, we also tested models including variables for experiment trial and experiment half. However, likelihood ratio tests indicated that these more complex models failed to improve model fit, so we did not consider them further.

Results

Table 1 shows the mean percent of sentences judged acceptable (yes responses) in the sentence judgment task. Descriptively, the participants tended to accept expected word sentences from the native accented speaker while rejecting

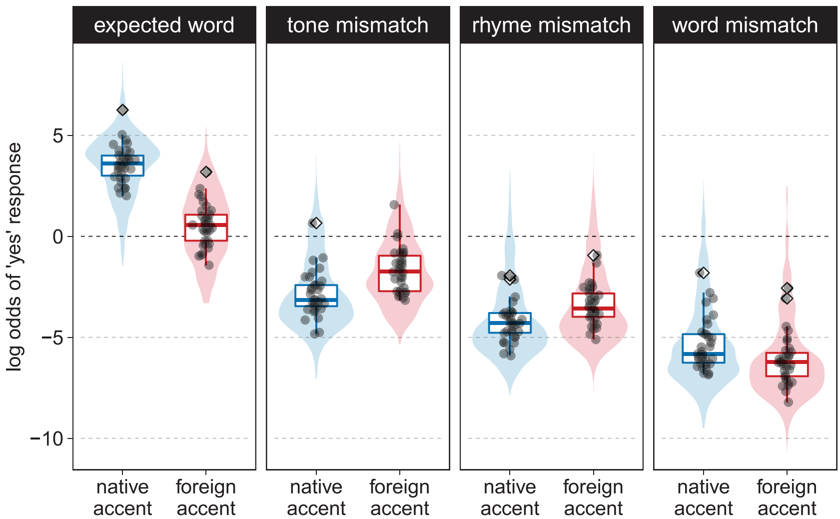

Figure 4 Model estimated log odds of accepting a sentence (yes response) for the sentence judgment task. Shaded areas behind the boxplots indicate the distribution of responses. Gray dots indicate the mean of a single participant's responses; white diamonds indicate that a participant's mean is an outlier. [Color figure can be viewed at wileyonlinelibrary.com]

sentences that contained mismatch words. The trends were generally similar for the sentences from the foreign-accented speaker, except that the participants also often rejected his sentences when there was no mismatch manipulation (i.e., for expected word sentences), suggesting that they might have been sensitive to other aspects of the sentence besides the critical word. There was large variation across the participants, especially for foreign-accented speech.

Figure 4 depicts model estimates of sentence acceptance rates with boxplots and violin plots. We submitted the results to a mixed-model ANOVA that indicated a statistically significant main effect of word, $\chi^2(3) = 199.20$, $p < .001$, and of accent, $\chi^2(1) = 4.30$, $p = .038$, and a significant word-by-accent interaction, $\chi^2(3) = 202.30$, $p < .001$.

To address our research questions, we conducted post hoc comparisons using the multcomp package (Hothorn, Bretz, & Westfall, 2008) and used the Bonferroni-Holm procedure to correct for multiple comparisons (see Table 2). To test whether accent affected the participants' processing of pronunciation errors, we compared the difference in the size of model estimates for

Table 2 Post hoc comparisons for behavioral sentence judgment task model

Comparison	*b*	*SE*	95% CI[a]	*z*	*p* (>/*z*/)[b]
Accent effects					
Native: Exp – Foreign: Exp	3.18	0.30	[2.36, 4.00]	10.75	< .001
Native: Tone – Foreign: Tone	−1.31	0.29	[−2.11, −0.50]	−4.51	< .001
Native: Rhyme – Foreign: Rhyme	−0.79	0.34	[−1.72, 0.15]	−2.34	.058
Native: Word – Foreign: Word	0.72	0.46	[−0.56, 2.00]	1.56	.239
Interactions of accent effects					
(Nat: Exp × Word) – (For: Exp × Word)	2.46	0.53	[1.00, 3.92]	4.68	< .001
(Nat: Tone × Word) – (For: Tone × Word)	−2.03	0.52	[−3.47, −0.58]	−3.89	< .001
(Nat: Rhyme × Word) – (For: Rhyme × Word)	−1.50	0.55	[−3.03, 0.02]	−2.74	.025
(Nat: Tone × Rhyme) – (For: Tone × Rhyme)	0.52	0.36	[−0.49, 1.53]	1.43	.239
Tone and rhyme effects across accents					
Tone – Rhyme	−3.41	0.62	[−5.13, −1.70]	−5.51	< .001

Note. Nat = native accent; For = foreign accent; Exp = expected word; Tone = tone mismatch; Rhyme = rhyme mismatch; Word = word mismatch.
[a] Asymptotic confidence intervals.
[b] Adjusted using Bonferroni-Holm method.

judgments of critical word conditions between accents. We called this difference the accent effect.

We found a large and statistically significant accent effect for the expected word condition, with listeners less likely to accept sentences produced by the L2 speaker. This result might seem surprising, but it was reasonable in the context of the task. Though our experimental manipulation was tied to the critical words, the participants' responses were based on entire sentences. Thus, a participant could reasonably have rejected expected word sentences due to perceived infelicities in pronunciation anywhere in the sentence. In this light, the results suggested that some participants had difficulty accepting any sentences produced with a foreign accent.

To answer our first research question concerning whether Mandarin listeners respond differently to pronunciation errors in different accents, we examined the accent effect for mismatch conditions. We found a statistically significant accent effect for sentences containing tone mismatches, with the participants more likely to accept such sentences when produced by the foreign-accented speaker.[2] At the same time, we failed to find significant accent effects for sentences with word or rhyme mismatches—though rhymes trended toward an effect.

To better understand the relationships between word and accent conditions, we examined the interactions between accent effects and conditions (e.g., a difference in size of accent effects for rhymes vs. tones). Typically, we would have used the expected word condition as a baseline for these comparisons, but, because of the large accent effect for expected word sentences, this would have been comparing apples and oranges. In contrast, the accent effect for word mismatch sentences was very small, making it a suitable choice as a baseline for testing interactions. As Table 2 shows, the interaction effect indicated that the accent effect for expected word sentences was different from the effect for word mismatch sentences, such that foreign-accented sentences with expected words were more likely to be rejected. The interaction effect for tone and word mismatches also differed, such that foreign-accented sentences with tone mismatches were more likely to be judged acceptable. The interaction effect for rhyme and word mismatches also differed, again indicating that rhyme mismatches were more likely to be judged acceptable in the foreign accent. In summary, these results further supported a tendency for the participants to judge foreign-accented speech as infelicitous more often than native speech in sentences with expected words, but also to be more accepting (or less sensitive) when they judged sentences with foreign-accented tone and foreign-accented rhyme mismatches.

To answer our second research question about possible differences in accent effects for tone and rhyme mismatches, we conducted two additional comparisons. First, we compared the interaction of accent effects for rhyme and tone mismatches. This interaction failed to reach statistical significance. In order to understand the underlying trend, we further compared tone and rhyme mismatch effects across accents. There was a significant difference in log odds, with the participants more likely to accept sentences containing tone mismatches than sentences with rhyme mismatches. This suggested that, behaviorally, the participants' judgments were less sensitive for tones compared to rhymes—regardless of speaker accent.

Given the nature of the sentence judgment task, these results should be treated with some caution because participant responses cannot be unambiguously tied to the critical word manipulations. Additionally, the results were highly variable across participants. When we looked at individual participants' mean raw accuracy (of responding *correct*) in the expected word condition, they ranged from accepting as few as 21% of accented sentences up to as many as 95%.

Event-Related Potential Analysis

To analyze results from the ERP experiment, we fitted mixed-effects linear regression models with crossed random effects for subjects and items using the afex and lme4 packages in *R*. We ran separate analyses for the 300–500 milliseconds (N400) and 600–900 milliseconds (LPC) windows over nine posterior electrodes. ERPs were time-locked to the onset of critical words. Analyses included all critical trials, regardless of the accuracy of behavioral judgments because judgments occurred after the end of a trial and were not necessarily tied to the critical words themselves. The dependent variable was mean amplitude, calculated as a mean for each subject and each item at each electrode averaged across the 300–500 milliseconds and 600–900 milliseconds windows. We selected these windows to capture typical N400 and LPC effects and to follow windows used in Pelzl et al. (2019), though we used longer epochs and thus a slightly adjusted LPC window in the present study (600–900 vs. 450–800; full model output is available in Appendix S3 in the online Supporting Information; for interested readers, a parallel analysis using repeated measures ANOVA is available in Appendix S4 in the online Supporting Information).

The spatial dimension of ERP data introduces challenges for mixed-effects models (see discussion in Winsler, Midgley, Grainger, & Holcomb, 2018). To constrain variability across electrodes, as seen in other recent ERP studies that

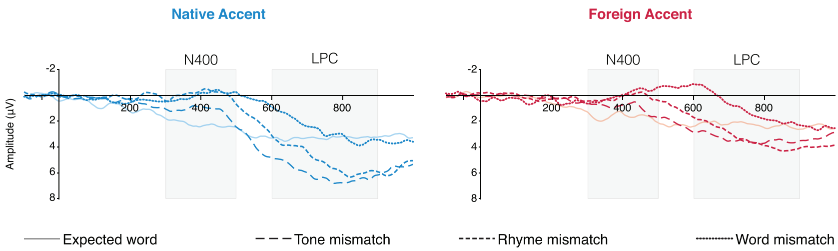

Figure 5 Grand mean waveforms averaged across posterior electrodes (low-pass filter of 20Hz applied for visualization only), highlighting the N400 (300–500 milliseconds) and late positive component (LPC; 600–900 milliseconds) windows (waveforms for individual electrodes available in Appendix S2 in the online Supporting Information). [Color figure can be viewed at wileyonlinelibrary.com]

have used mixed-effects models, we restricted our analysis to a smaller set of electrodes where N400 and LPC effects are typically observed (for a similar approach, see Kuperberg, Brothers, & Wlotko, 2020). We chose nine central posterior electrodes: CP3, CPz, CP4, P3, Pz, P4, O1, Oz, and O2.[2]

For both analyses (N400, LPC), fixed effects included critical word (effect coded: expected word, tone mismatch, rhyme mismatch, word mismatch), accent (effect coded: native or foreign), and their interaction. We fitted the maximal random effects structure including by-subject and by-item random intercepts and slopes for both of the fixed effects and their interaction. Results below are reported from *F* tests for mixed-model ANOVAs computed on linear mixed-effects models using the *mixed*() function from afex (Singmann et al., 2017). The lmer/mixed formula for the final models was:

$$mean.amplitude \sim word * accent + (word * accent \mid subj) + (word * accent \mid item). \quad (2)$$

We evaluated possible effects of adaptation by also testing models with the additional variables trial (1–240) and half (first, second), but, because these failed to improve model fit, we did not pursue them further.

N400 Analysis

Figure 5 shows the grand mean waveforms averaged across all nine posterior electrodes. Table 3 provides the mean amplitudes for word and accent conditions in the N400 window. Mean results suggested small differences between accent conditions and more negative responses for rhyme and word

Table 3 Mean amplitude of response in microvolts for the N400 window (300–500 milliseconds)

	Expected word	Tone mismatch	Rhyme mismatch	Word mismatch
Accent	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
Native	2.22 (0.11)	1.18 (0.11)	−0.23 (0.11)	−0.12 (0.10)
Foreign	1.58 (0.11)	1.11 (0.11)	0.06 (0.11)	0.06 (0.10)

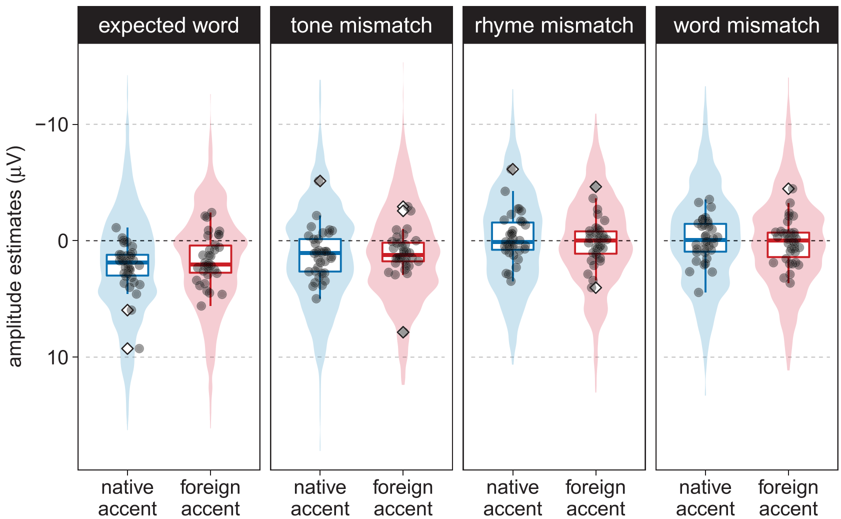

Figure 6 Boxplots depicting the model estimates (amplitude) for the N400 (300–500 milliseconds) window over posterior electrodes. Shaded areas behind the boxplots indicate the distribution of estimates. Gray dots indicate the mean amplitude of each participant's responses; white diamonds indicate that a participant's mean is an outlier. [Color figure can be viewed at wileyonlinelibrary.com]

mismatches relative to the expected word condition. Mixed-model ANOVA results showed a significant main effect for word, $F(3, 93.22) = 11.26$, $p < .001$. The main effect of accent, $F(1, 103.78) = 0.07$, $p = .800$, and the word-by-accent interaction failed to reach significance, $F(3, 81.13) = 0.29$, $p = .831$. Figure 6 depicts the model estimates.

To answer our research questions, we conducted post hoc comparisons that are reported in Table 4. In all comparisons, b values can be interpreted as amplitude differences (in μV). In order to capture overall trends, especially with respect to N400 effects, we tested effects of the mismatch conditions across

Table 4 Post hoc comparisons for N400 model

Comparison	*b*	*SE*	95% CI[a]	*z*	*p* (>/*z*/)[b]
Across accent effects					
Tone mismatch – Expected word	−1.58	0.76	[−3.66, 0.49]	−2.09	.185
Rhyme mismatch – Expected word	−3.83	0.78	[−5.96, −1.69]	−4.90	< .001
Word mismatch – Expected word	−3.79	0.77	[−5.90, −1.67]	−4.90	< .001
Rhyme mismatch – Tone mismatch	−2.25	0.76	[−4.33, −0.17]	−2.95	.019
Interactions of accent effects					
(Nat: Tone × Exp) – (For: Tone × Exp)	−0.25	0.90	[−2.71, 2.22]	−0.27	1.000
(Nat: Rhyme × Exp) – (For: Rhyme × Exp)	−0.68	0.87	[−3.05, 1.69]	−0.79	1.000
(Nat: Word × Exp) – (For: Word × Exp)	−0.61	0.87	[−2.99, 1.76]	−0.70	1.000
(Nat: Tone × Rhyme) – (For: Tone × Rhyme)	0.44	0.79	[−1.74, 2.61]	0.55	1.000

Note. Nat = native accent; For = foreign accent; Exp = expected word; Tone = tone mismatch; Rhyme = rhyme mismatch; Word = word mismatch.
[a]Asymptotic confidence intervals.
[b]Adjusted using Bonferroni-Holm method.

Table 5 Mean amplitude of response in microvolts for the late positive component window (600–900 milliseconds)

	Expected word	Tone mismatch	Rhyme mismatch	Word mismatch
Accent	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
Native	2.96 (0.13)	5.77 (0.13)	4.51 (0.13)	2.44 (0.12)
Foreign	2.05 (0.12)	3.43 (0.12)	3.00 (0.12)	1.10 (0.12)

accents. There were statistically significant N400 effects in response to rhyme and word mismatches, with amplitudes more negative than those of expected word responses. Relative to other conditions, the 95% confidence intervals were narrow, with upper bounds falling approximately 1.7 μV below zero. The difference between responses to expected words and tone mismatches failed to reach statistical significance, but the difference between rhyme and tone mismatches was statistically significant. In other words, we found relatively strong N400 effects ($>$3.5 μV) for word and rhyme mismatches but failed to find N400 effects for tone mismatches, regardless of accent.

Post hoc comparisons of the interactions of accent effects indicated no significant interaction effects. So, with respect to our first question about responses to foreign-accented pronunciation errors, we failed to find evidence of differences between accent conditions. If this null result reflected a true lack of meaningful differences between accent conditions, this would suggest foreign accent does not strongly impact N400 responses. However, these results should be treated as very uncertain because the wide confidence intervals suggested a large degree of imprecision.

For our second research question, we did find evidence of differences in responses to tone and rhyme mismatches but no accent effect. In other words, regardless of accent, the participants showed N400 effects for rhyme errors, but no—or at least very small—responses to tone errors. Again, in light of the wide confidence intervals, these results should be treated as uncertain.

Late Positive Component Analysis

Table 5 shows mean amplitudes for word and accent conditions in the LPC window. Mean results suggested more positive amplitudes for native accented speech across critical word conditions and more positive amplitudes for tone and rhyme mismatches within both accent conditions. Across all conditions, the amplitude of LPCs for foreign-accented speech was about 1–1.5 μV less positive than those for native speech. A mixed-model ANOVA showed

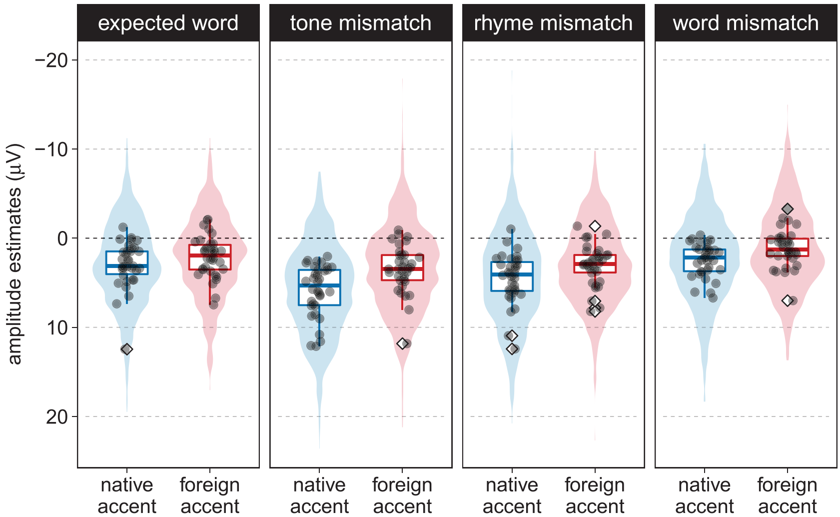

Figure 7 Boxplots depicting the model estimates (amplitude) for the late positive component (600–900 milliseconds) window over posterior electrodes. Shaded areas behind the boxplots indicate the distribution of estimates. Gray dots indicate the mean amplitude of a single participant's responses; white diamonds indicate that a participant's mean is an outlier. [Color figure can be viewed at wileyonlinelibrary.com]

significant main effects for word, $F(3, 96.85) = 17.34$, $p < .001$, and accent, $F(1, 109.69) = 30.77$, $p < .001$. The word-by-accent interaction was not significant, $F(3, 74.17) = 1.57$, $p = .203$. Figure 7 depicts the model estimates.

Table 6 depicts post hoc comparisons. In order to capture overall LPC trends, we tested effects of the mismatch conditions across accents. There were statistically significant LPC effects in response to tone and rhyme mismatches, with amplitudes more positive than those of expected word responses. Wide confidence intervals suggested a large degree of uncertainty, with plausible values (under the given model's assumptions) ranging from fairly weak effects (<1 μV) to very large effects (>6.5 μV). The difference between responses to expected words and word mismatches failed to reach statistical significance, and there was also no statistically significant difference between tone and rhyme mismatch LPCs.

Though there was a main effect of accent on LPC responses, post hoc comparisons of the interactions of accent effects indicated no significant differences. With respect to our first research question, we failed to find evidence of an accent effect on the participants' responses to pronunciation errors. As

Table 6 Post hoc comparisons for late positive component model

Comparison	b	SE	95% CI[a]	z	p (>/z/)[b]
Across accent effects					
Tone mismatch – Expected word	4.15	0.93	[1.60, 6.69]	4.46	<.001
Rhyme mismatch – Expected word	2.62	0.85	[0.30, 4.94]	3.09	.014
Word mismatch – Expected word	−1.54	0.79	[−3.71, 0.63]	−1.95	.259
Rhyme mismatch – Tone mismatch	−1.52	0.97	[−4.19, 1.14]	−1.57	.469
Interactions of accent effects					
(Nat: Tone × Exp) – (For: Tone × Exp)	1.83	0.87	[−0.55, 4.22]	2.10	.214
(Nat: Rhyme × Exp) – (For: Rhyme × Exp)	0.79	0.94	[−1.77, 3.35]	0.85	.845
(Nat: Word × Exp) – (For: Word × Exp)	0.67	0.96	[−1.95, 3.29]	0.70	.845
(Nat: Tone × Rhyme) – (For: Tone × Rhyme)	1.04	0.97	[−3.68, 1.60]	−1.08	.845

Note. Nat = native accent; For = foreign accent; Exp = expected word; Tone = tone mismatch; Rhyme = rhyme mismatch; Word = word mismatch.
[a] Asymptotic confidence intervals.
[b] Adjusted using Bonferroni-Holm method.

with the N400 results, wide confidence intervals suggested a large degree of imprecision and uncertainty.

Post hoc comparisons also targeted the critical comparison of LPC responses to tone and rhyme mismatches. We found no significant differences either across or between accents, though descriptively, tone mismatch responses were more strongly positive than rhyme mismatch responses, and the magnitude of both mismatch responses was larger for native accented speech.

General Discussion

We conducted an ERP study with native listeners of Chinese who judged the accuracy of sentences produced by either a native speaker of Mandarin or a L2 Mandarin speaker with a foreign (American English) accent, and we compared N400 and LPC responses to critical words that were expected, mispronounced by either a tone or a rhyme (in both cases thus becoming a nonword), or were contextually inappropriate. We wanted to know whether listeners' expectations for more pronunciation errors in foreign-accented speech result in reduced ERP responses to such errors (relative to native-accented speech) and whether this effect is larger for tone than rhyme errors.

In summary, our findings showed that:

1. Behavioral judgments of sentences indicated less accuracy (i.e., responding *correct* when there was in fact an incorrect pronunciation) for tone mismatches when produced by the foreign-accented speaker compared to when produced by the native-accented speaker. There was also less accuracy to tone mismatches (relative to rhyme mismatches), regardless of accent. In addition, there was a large accent effect for the expected word condition, with the participants less likely to accept sentences produced by the L2 speaker.
2. There were statistically significant N400 effects for rhyme and word mismatches but not for tone mismatches; there was also a significant difference between tone and rhyme mismatch N400s. There were no significant interactions with accent conditions.
3. LPCs for foreign-accented speech displayed weaker positive deflections overall, and there were significant LPCs for tone and rhyme mismatches in both accent conditions, but there were no significant interactions with accent conditions.
4. Both behavioral and ERP data suggested a high degree of variability in the participants' responses.

Below, we discuss these results in more detail, tying together behavioral and ERP results.

The Role of Foreign Accent in Mandarin Speech Processing

Our results provide some evidence that foreign-accent may shape how Mandarin listeners process pronunciation errors. The behavioral judgments suggested that many participants accepted more sentences with critical tone mispronunciations in foreign-accented speech than in native speech. What might explain this difference?

One possible explanation is that upon identifying a speaker as nonnative, the participants used their knowledge of the world to adapt their expectations. Knowing that tone errors are typical in foreign-accented Mandarin, listeners may choose to apply less stringent criteria for error when judging a L2 speaker's tones.

A second possible explanation is that, after hearing enough tone errors, the participants might have begun to change the way that they used acoustic-phonetic cues (i.e., F0 or other features of tones) in the speech of a foreign-accented speaker. Once they noticed that tones were often unhelpful or even misleading when trying to recognize words, the participants might have begun to rely on tones less and use other available cues (e.g., segments, context) more. Idemaru and Holt (2011) showed that listeners down-weight some types of redundant acoustic-phonetic features if those features stop being informative (i.e., useful for comprehension). This down-weighting might be a way to ignore pronunciation errors; however, it is unclear if it can occur for phonological features (Pelzl et al., 2020). Many studies have shown listener adaptation to accented features of L2 speech (e.g., Baese-Berk, Bradlow, & Wright, 2013; Bradlow & Bent, 2008; Reinisch & Holt, 2014; Xie, Theodore, & Myers, 2017), but these accented pronunciation features were still informative. That is, by adapting, listeners become more efficient at comprehending L2 speech. In the case of pronunciation errors, however, listeners might achieve smoother comprehension if they ignore the error entirely, rather than trying to interpret it. Thus, it may be that the participants accepted more tone mismatch sentences from the L2 speaker because they began to down-weight the speaker's tones.

Finally, a third explanation for our behavioral results is that phonetic information in foreign-accented speech is more difficult for listeners to process, leading to decreased accuracy in the judgment of pronunciation errors. Relative to rhyme errors (or at least those in our stimuli), tone errors might have been subtler, leading to even lower accuracy in judgments of tones compared to rhymes. This trend has been observed previously with native speech

(Cutler & H.-C. Chen, 1997; Pelzl et al., 2019), and in the present study is supported by the participants' lower rejection rate of tone mismatches for both speakers. Perhaps, then, foreign-accent simply makes detection of tone errors even more difficult.

As we noted earlier, we recommend caution in interpreting the present sentence judgment results, given that responses could not be unambiguously tied to the mismatch manipulations (despite the training that the participants were given prior to the experiment). Future work should use less ambiguous measures to establish whether the behavioral patterns observed here are reliable.

If our behavioral results suggest effects for foreign-accented speech, our ERP results are less clear. N400s showed no evidence of an effect for accent. LPCs showed a main effect of accent, but no interactions with mismatch conditions. In other words, for LPCs there appears to be a difference of degree in response to foreign-accented speech errors compared to native speech errors, regardless of the type of error. Superficially, this might seem inconsistent with previous studies that found significantly reduced LPCs to foreign-accented speech errors (Caffarra & Martin, 2019; Hanulíková et al., 2012; Romero-Rivas et al., 2015), but, despite a failure to reach statistical significance, our trends were clearly in the same direction as those studies, suggesting somewhat attenuated LPCs for foreign-accented speech—both for tone and rhyme mismatches.

Recent work on LPC/P600 effects has highlighted that, in many cases, these late responses are largely indistinguishable from the P3b response (e.g., Leckey & Federmeier, 2019). Importantly, the amplitude of the P3b is sensitive to salience, such that stimuli that are more salient elicit larger P3bs. In this case, weaker responses for foreign-accented pronunciation errors might indicate less salience for these deviations relative to native speech. This line of interpretation accords well with the reasoning presented by Hanulíková et al. (2012), who argued that a lack of N400 differences for lexical-semantic violations indicated that listeners had little difficulty comprehending the speaker's message; instead, it was largely their different expectations for usage of grammatical gender that reduced LPCs for foreign-accented speech. These ideas also fit well with models of speech perception such as the ideal adapter (Kleinschmidt & Jaeger, 2015) or noisy channel (Gibson et al., 2017; Gibson, Bergen, & Piantadosi, 2013), which make explicit the role of listeners' high-level expectations in driving responses.

Although our data comport with this line of thinking, we cannot rule out alternative accounts that would attribute trends in behavioral and neural responses to difficulties perceiving the accented speech itself. In the present study, it could also have been the case that the participants in fact detected more frequent pronunciation errors in the foreign-accented speech than in the native speech—that is, they may have detected errors in addition to the critical mismatching words, and done so even in the expected word sentences. Increasing the frequency of task-relevant events is known to attenuate the P3 (Leckey & Federmeier, 2019; Squires, Wickens, Squires, & Donchin, 1976). Thus, unintended pronunciation deviations by the foreign-accented speaker may have impacted LPC responses by increasing the overall frequency of errors in the foreign-accented stimuli. Finally, there is no reason pronunciation errors would necessarily be expected to function in the same way that syntactic violations have functioned in previous studies.

Future work along these lines should seek to more effectively tease apart alternative accounts, exercising more control on the properties of foreign-accented speech. The use of judgment tasks or the avoidance of them should also be carefully considered, as such tasks no doubt impact ERP responses, even when decisions are delayed after each trial, as they were in the current study. If the LPCs of interest are, in fact, P3b responses, then the wealth of relevant literature can help guide design choices.

Potential Differences Between Rhymes and Tone Mismatches

In addition to the accent-specific behavioral effect for tone mismatches, we also found a more general behavioral effect for tone mismatches. That is, regardless of a speaker's accent, when attempting to judge sentences for pronunciation deviations, the participants were more consistent in correctly rejecting rhyme than tone mismatches. These behavioral differences were mirrored by N400 effects in ERPs. In both accent conditions, tone N400 effects were weaker than rhyme N400 effects and failed to differ from expected word N400s. Though we did not set out to test this difference in native speech, the finding is somewhat novel and requires discussion.

First, our results are consistent with behavioral outcomes from Pelzl et al. (2019), who found less accuracy for tone mismatches relative to rhymes both in a disyllabic lexical decision task and in a sentence judgment task that used a subset of the stimuli used here (with only the native Mandarin speaker). However, consistent with other ERP studies reviewed above (Brown-Schmidt & Canseco-Gonzalez, 2004; Schirmer et al., 2005), Pelzl et al. (2019) failed to find statistically significant ERP differences between tone and rhyme mismatches, unlike the current study.

Given the close relationship between this study and that of Pelzl et al. (2019), the difference in ERP results was unexpected. There were three main differences between the studies. First, errors were overall more frequent in this study (75% overall) compared to the errors found in Pelzl et al. (2019; 50% overall). However, this did not seem to impact behavioral judgments because Pelzl et al. (2019) reported 16% acceptance rates for tonal mismatches and 5% for rhymes compared to 14% tone and 7% rhyme acceptance for native-accented mismatches in the present study. Second, the number of stimuli in the present study (240 sentences) was greater than that in Pelzl et al. (2019; 180 sentences). This meant that we had a more relaxed criterion for cloze probabilities of critical words in the present study. However, it is unclear how this would have led to differing results. Moreover, any differences should have been addressed by the counter-balancing that we implemented in the current study across lists, so that any such effects would have been similar for tone and rhyme mismatch conditions. Finally, the present study had both a native and a foreign-accented speaker (in contrast to Pelzl et al., 2019, which only had a native speaker). Perhaps encountering the accented speaker in the stimuli affected the expectation for tone errors globally—and may have in fact increased it if there were incidental (i.e., not experimentally manipulated) tone errors in sentences. This could have reduced sensitivity for all tone mismatches regardless of accent. Clearly, these results will need to be replicated and further work will be needed to determine the cause or causes of N400 differences between different studies.

Individual Differences

A clear feature of the present data was large variability among the participants, especially for the foreign-accented speaker in the expected word condition of the behavioral task. Figure 8 illustrates visible trends in this variability, showing that about a third of participants (solid black lines) tended to reject most sentences spoken by the foreign-accented speaker even if the critical word was the expected word. At the other extreme, a handful of participants (dashed lines) appeared to accept almost every expected word and most tone mismatch sentences from the foreign speaker. Patterns in the ERP data were not so distinctive—though variability was still clearly a feature of these data as well.

Previous ERP studies (Caffarra & Martin, 2019; Grey & van Hell, 2017) found that people's (lack of) experience with foreign-accented speakers was connected to their ERP responses to foreign-accented speech errors. We might then speculate that individual participant differences in our behavioral and ERP data related to their previous experience with L2 Mandarin speakers (i.e., listeners with more experience and thus stronger expectations for L2

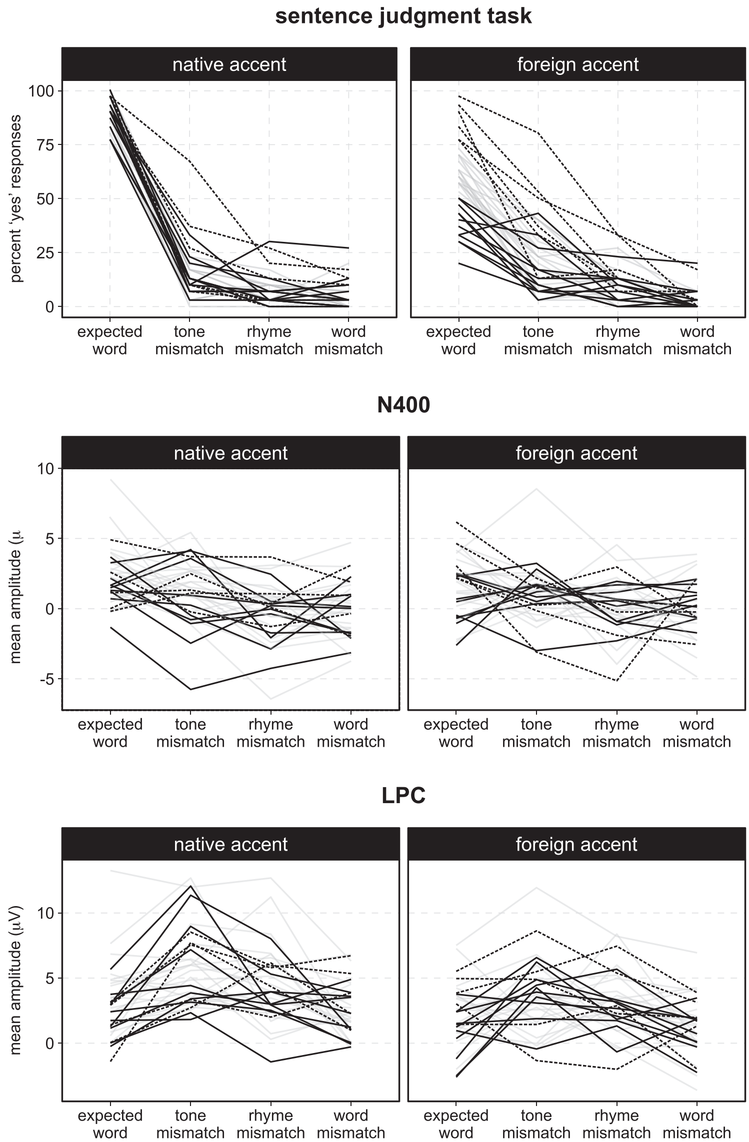

Figure 8 Individual participant patterns across measures. Each line indicates a single participant's mean (percent or amplitude) across critical word conditions. Line type is associated with performance in the expected word sentences for the foreign-accented speaker. Solid black lines indicate < 50% yes responses; dashed black lines indicate > 75% yes responses; gray lines indicate more moderate judgments. LPC = late positive component.

tone errors might have reduced N400/LPC responses to those errors). Unfortunately, we did not gather this type of background information from our participants.

In addition to participants' experience with foreign-accented Mandarin speakers, their experience with native but regionally accented Mandarin speakers could also have led to individual differences in task performance. China's native language environment is highly diverse, with at least seven major Chinese language varieties (e.g., Mandarin, Cantonese, or Shanghainese), and almost countless local dialects of each (Norman, 1988). Although Mandarin is the common language, people's other Chinese languages can have strong influences on their Mandarin pronunciation. Individual differences in experience with these many flavors of Mandarin might influence sensitivity to tone (and rhyme) deviations from the standard pronunciation. Regional differences of this sort have previously been shown to impact outcomes in lexical tasks (Wiener & Ito, 2016). Future work might explore these issues further, perhaps contrasting listeners' processing and judgments of regional and foreign-accented Mandarin speech.

Finally, we cannot rule out a strategic component to participants' behaviors. Even though they were told specifically what types of speech errors should lead to rejections, some participants might have settled on differing strategies to carry out the judgment task. For example, some may have focused more heavily on individual words, but others might have tried to take in the sentences more holistically. This could certainly have influenced outcomes, especially at the behavioral level.

Although we cannot provide a satisfactory account of the variability in our data, such variability is not necessarily surprising. Work by Tanner and Van Hell (2014) and Tanner (2019) has illustrated how simple measures of central tendencies at the group level can often misrepresent individual ERP patterns. We also acknowledge that variability is sometimes an indication of noisy data. Given these realities, the presentation of estimates of variability (standard errors, confidence intervals), as provided in the current study, should be a standard feature of ERP studies in order to help researchers determine how closely results reflect reliable group trends and to allow for comparison of variability across studies.

Limitations and Future directions

First, this study used just a single L2 Mandarin speaker. It is certain this speaker's Mandarin accent was not representative of all L2 speaker accents, or degrees of accent. A small post hoc accentedness rating task suggested that he had a moderate to strong accent (see Appendix S1 in the online Supporting Information for details). Future work should follow the example of other studies by establishing accentedness a priori (cf. Grey & van Hell, 2017) and perhaps by using a variety of L2 speakers (Romero-Rivas et al., 2015).

Although the sample size of the current study compared well with other accented speech and Chinese tone ERP studies, this does not mean it was sufficient to find relevant effects. Future studies should strive to substantially increase participant numbers (Brysbaert & Stevens, 2018) and also to reduce the complexity of the experimental designs when possible (Luck & Gaspelin, 2017).

Our participants (young, urban, university educated) cannot be considered representative of the Chinese population more broadly. Future work might make efforts to expand the population of Chinese people recruited to join research (see Andringa & Godfroid, 2019 for similar arguments).

Finally, as we noted in our review, Pelzl et al. (2019) found visible differences between word and nonword N400 effects, with word mismatch N400s continuing much longer than those of nonwords. These trends were replicated here, suggesting there may be distinctive N400 effects for words compared to nonwords. The current study was not designed to pinpoint the reasons for such differences. The extended N400 for word mismatches, compared to both tone and rhyme nonwords, may be due to the disambiguating information provided by the following syllable in the tone and rhyme conditions, which could allow recovery of the originally expected lexical target. Alternatively, the quickly attenuating N400 response for nonwords might indicate a failure to access any word at all. Finally, it could be that the key difference is indexed not in the N400, but in the strong LPC that follows for nonwords. Future research might attempt to distinguish between pronunciation errors that lead to words compared to nonwords, while controlling possible confounds such as word length (monosyllable vs. disyllable). By taking advantage of the distinctive properties of monosyllabic words (few nonword neighbors) and disyllabic words (many nonword neighbors), researchers may find ways to disentangle effects that are otherwise confounded. Pursuing this line of work in the context of foreign-accented speech processing might be optimal because speech errors will be ecologically valid and could provide a more realistic assessment of the importance of accurate tones (or segments) than is typically accomplished by artificial manipulations of native speech.

Conclusion

The present study set out to explore the relationship between L2 accent and pronunciation errors in the context of Mandarin Chinese sentence comprehension. Results suggest that native Mandarin listeners tend to be less accurate in judging tonal errors in foreign-accented speech compared to the same errors in native-accented speech. The cause of this difference remains unclear. ERP results were inconclusive. We observed a main effect of accent on LPCs, and a trend for reduced LPCs in response to pronunciation errors in foreign-accented speech compared to native-accented speech. The results also suggest that tone mismatches may be processed differently from rhyme mismatches, regardless of a speaker's accent. Finally, we found evidence of variability across listeners. Such variability may be a valuable object of study in future work because it could provide additional insight into the factors that influence foreign-accented speech processing in Mandarin.

Final revised version accepted 23 June 2020

Notes

1 There was one extreme outlier who accepted 80% of tone mismatches in the foreign accent and 67% in the native accent. To be certain results were not due solely to this participant, we removed this participant and reran the analysis. There was no substantive difference in outcomes, except that the interaction of accent effects of rhyme vs. word (i.e., native-speaker rhyme vs. native-speaker word compared to foreign-speaker rhyme vs. foreign-speaker word) was not significant when this participant's data were excluded, $b = -1.39$, $SE = .592$, $z = -2.36$, $p = .07$.

2 We have reported an additional analysis using anterior electrodes in Appendix S4 in the online Supporting Information. We abandoned an earlier attempt to model electrodes nested under subjects due to model fitting difficulties.

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://osf.io/vysx7/. All proprietary materials have been precisely identified in the manuscript.

References

Andringa, S., & Godfroid, A. (2019). Call for participation. *Language Learning*, *69*, 5–10. https://doi.org/10.1111/lang.12338

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*, EL174–EL180. https://doi.org/10.1121/1.4789864

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 443–446.

Brown-Schmidt, S., & 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. *Journal of Psycholinguistic Research*, *33*, 103–135.

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, *125*, 82–93. https://doi.org/10.1016/j.bandl.2013.01.007

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, article 9, 1–20. https://doi.org/10.5334/joc.10

Caffarra, S., & Martin, C. D. (2019). Not all errors are the same: ERP sensitivity to error typicality in foreign accented speech perception. *Cortex*, *116*, 308–320. https://doi.org/10.1016/j.cortex.2018.03.007

Chen, J.-Y. (1999). The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue. *Applied Psycholinguistics*, *20*, 289–301. https://doi.org/10.1017/S0142716499002064

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, *84*, 46–56. https://doi.org/10.1016/j.specom.2016.07.005

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics*, *59*, 165–179.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins.

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). New York, NY: Oxford University Press.

Foucart, A., Costa, A., Morís-Fernández, L., & Hartsuiker, R. J. (2020). Foreignness or processing fluency? On understanding the negative bias toward foreign-accented speakers. *Language Learning*, 1–43. https://doi.org/10.1111/lang.12413

Garnham, A., Shillcock, R., Brown, G. D. A., Mill, A. I. D., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous speech. *Linguistics*, *19*, 805–818. https://doi.org/10.1515/ling.1981.19.7-8.805

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, *28*, 703–712. https://doi.org/10.1177/0956797617690277

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, *122*, 92–102. https://doi.org/10.1016/j.bandl.2012.04.017

Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, *42*, 93–108. https://doi.org/10.1016/j.jneuroling.2016.12.001

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*, 878–887.

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

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1939–1956. https://doi.org/10.1037/a0025641

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148–203. https://doi.org/10.1037/a0038695

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*, 12–35. https://doi.org/10.1162/jocn_a_01465

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*, 463–470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potential reflect semantic incongruity. *Science*, *207*, 203–205.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*, 920–933. https://doi.org/10.1038/nrn2532
Leckey, M., & Federmeier, K. D. (2019). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology*, *57*, 1–15. https://doi.org/10.1111/psyp.13351
Lev-Ari, S. (2015). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5*, 1–12. https://doi.org/10.3389/fpsyg.2014.01546
Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 1–14. https://doi.org/10.3389/fnhum.2014.00213
Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't): How to get significant effects. *Psychophysiology*, *54*, 146–157. https://doi.org/10.1111/psyp.12639
Malins, J. G., & 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*, 2032–2043. https://doi.org/10.1016/j.neuropsychologia.2012.05.002
MATLAB ([Computer software]. (2013). Natick, MA: The MathWorks Inc.
Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, *24*(3), 49–65.
Norman, J. (1988). *Chinese*. New York, NY: Cambridge University Press.
Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elcited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.
Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 786–803.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.
Pelzl, E. (2018). *Second language lexical representation and processing of Mandarin Chinese tones* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.
Pelzl, E., Carlson, M. T., Guo, T., Jackson, C. N., & van Hell, J. G. (2020). Tuning out tone errors? Native listeners do not down-weight tones when hearing unsystematic

tone errors in foreign-accented Mandarin. *Bilingualism: Language and Cognition*, 1–8. https://doi.org/10.1017/S1366728920000280

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, *41*, 59–86. https://doi.org/10.1017/S0272263117000444

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 539–555. https://doi.org/10.1037/a0034409

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, *9*, 1–15. https://doi.org/10.3389/fnhum.2015.00167

Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., & Chen, H.-C. (2005). Brain responses to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience*, *17*, 1–12.

Schmidt-Kassow, M., & Kotz, S. A. (2009). Event-related brain potentials suggest a late interaction of meter and syntax in the P600. *Journal of Cognitive Neuroscience*, *21*, 1693–1708.

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

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). Afex: Analysis of factorial experiments [Computer software]. Retrieved from http://cran.r-project.org/package=afex

Squires, K., Wickens, C., Squires, N., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, *193*, 1142–1146. https://doi.org/10.1126/science.959831

Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, *111*, 210–237. https://doi.org/10.1016/j.cortex.2018.10.011

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301. https://doi.org/10.1016/j.neuropsychologia.2014.02.002

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*, 122–144. https://doi.org/10.1159/000107913

Wan, I.-P., & Jaeger, J. (1998). Speech errors and the representation of tone in Mandarin Chinese. *Phonology*, *15*, 417–461. https://doi.org/10.1017/S0952675799003668
Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*, 1033–1042. https://doi.org/10.1121/1.1531176
White, C. M. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of the Chinese Language Teachers Association*, *16*, 27–56.
Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, *56*, 38–51. https://doi.org/10.1016/j.wocn.2016.02.001
Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience*, *33*, 1063–1082. https://doi.org/10.1080/23273798.2018.1455985
Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 206–217. https://doi.org/10.1037/xhp0000285
Xu, J., Abdel Rahman, R., & Sommer, W. (2019). Perceived language competence modulates criteria for speech error processing: Evidence from event-related potentials. *Language, Cognition and Neuroscience*, *35*, 752–765. https://doi.org/10.1080/23273798.2018.1562558
Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. Philadelphia, PA: John Benjamins.
Zhang, H. (2016). Dissimilation in the second language acquisition of Mandarin Chinese tones. *Second Language Research*, *32*, 427–451.
Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, *49*, 1761–1770. https://doi.org/10.1016/j.neuropsychologia.2011.02.054

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Additional Details for Stimuli Design and Creation.
Appendix S2. Waveforms for all Critical Electrodes.
Appendix S3. Full Mixed-Effects Model Output.
Appendix S4. Additional Analyses.
Appendix S5. List of Stimuli.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Mandarin Listeners May Treat Tone Errors Differently Depending on the Accent of the Speaker Who Makes Them

What This Research Was About and Why It Is Important

Understanding how and why listeners respond to foreign accents the way they do could be helpful for second language learners and those who talk with second language speakers. Researchers have previously investigated foreign-accented speech in several Indo-European languages. We examined Mandarin Chinese in order to see how accent interacts with lexical tones. Tones are pitch changes on syllables that affect word meanings. Because tones are not common in Indo-European languages, many adults who speak those languages find tones very difficult to produce accurately when learning Mandarin. For this reason, we wanted to know how native Mandarin speakers would respond to tone errors made by a foreign-accented speaker. Chinese listeners heard sentences spoken by either a native or foreign-accented speaker, and judged whether the sentences sounded correct. We measured behavioral judgments and brain activity to see how listeners responded to different types of speech errors. We found that they judged tone errors somewhat differently depending on who made them.

What the Researchers Did

- Brain activity ("brain waves") of 34 native Chinese speakers was recorded while they listened to 240 sentences, half spoken by a native Mandarin speaker and half spoken by a foreign-accented speaker.
- By design, many of the sentences contained one of three types of error: mispronounced tones, mispronounced vowels, or words that did not make sense in context.
- After hearing each sentence, listeners judged whether it had errors (such as pronunciation or word errors).

What the Researchers Found

- Native Chinese listeners were more likely to accept sentences with pronunciation errors (especially tones) as "correct" when the error was made by the foreign-accented speaker than when it was made by a native speaker.
- They tended to judge foreign-accented sentences as "not correct" even when there was no intentional pronunciation error in the sentence—though this varied widely across listeners.

- Listeners' brain waves did not show strong differences in how they responded to the two speakers, but they did suggest that listeners might always be less sensitive to incorrectly pronounced tones compared to vowels.

Things to Consider

- While the judgments of listeners were clearly influenced by the speaker's accent, the best interpretation of the results isn't clear. Listeners may have expected tone errors from the foreign-accented speaker and so tended to ignore them, or they may have had difficulty hearing and then judging the foreign-accented speaker's errors.
- Some listeners had very different responses to others. More research is needed to understand how "typical" Chinese listeners respond to tone errors in foreign-accented speech, and what leads to differences among people. For example, a person may be less bothered by errors if they have many foreign-accented friends.
- In the future, research like this might help learners and teachers of tonal languages decide how important it is to master tones. For now, it is too early to make strong teaching recommendations.

Materials and data: Materials are publicly available at https://osf.io/vysx7/.
How to cite this summary: Pelzl, E., Lau, E. F., Jackson, S. R., Guo, T., & Gor, K. (2020). Mandarin listeners may treat tone errors differently depending on the accent of the speaker who makes them. *OASIS Summary* of Pelzl et al. (2021) in *Language Learning*. https://oasis-database.org