



# Reassessing the electrophysiological evidence for categorical perception of Mandarin lexical tone: ERP evidence from native and naïve non-native Mandarin listeners

Yang Agnes Gao<sup>1,2</sup> · Joseph C. Toscano<sup>2</sup> · Chilin Shih<sup>1,3,4</sup> · Darren Tanner<sup>1,4,5</sup>

Published online: 30 October 2018  
© The Psychonomic Society, Inc. 2018

## Abstract

Some studies have argued that native speakers of tonal languages have been shown to perceive lexical tone continua in a more categorical manner than speakers of non-tonal languages. Among these, Zhang and colleagues (NeuroReport 23 (1): 35-9) conducted an event-related potential (ERP) study using an oddball paradigm showing that native Mandarin speakers exhibit different sensitivity to deviant tones that cross category boundaries compared to deviants that belong to the same category as the standard. Other recent ERP findings examining consonant voicing categories question whether perception is truly categorical. The current study investigated these discrepant findings by replicating and extending the Zhang et al. study. Native Mandarin speakers and naïve English speakers performed an auditory oddball detection test while ERPs were recorded. Naïve English speakers were included to test for language experience effects. We found that Mandarin speakers and English speakers demonstrated qualitatively similar responses, in that both groups showed a larger N2 to the across-category deviant and a larger P3 to the within-category deviant. The N2/P3 pattern also did not differ in scalp topography for the within- versus across-category deviants, as was reported by Zhang et al. Cross-language differences surfaced in behavioral results, where Mandarin speakers showed better discrimination for the across-category deviant, but English speakers showed better discrimination for within-category deviants, though all results were near-ceiling. Our results therefore support models suggesting that listeners remain sensitive to gradient acoustic differences in speech even when they have learned phonological categories along an acoustic dimension.

**Keywords** Categorical perception · Mandarin · Lexical tones · ERP · N2 · P3

During speech perception, listeners must map continuous acoustic cues onto discrete linguistic representations. In general, this can be described as a process of categorization, in that individual speech sound tokens must be assigned to functionally equivalent classes (Holt & Lotto, 2010). Early models

of speech perception proposed that this process discards fine-grained acoustic differences, leading to *categorical perception* of speech sounds (Liberman, Harris, Hoffman, & Griffith, 1957), resulting in better discrimination of stimuli across category boundaries than equivalently separated stimuli within the same category (i.e., separated by the same distance in acoustic space). Electrophysiological data have also been used to argue for categorical perception, where listeners' mismatch negativity (MMN) event-related potential (ERP) responses have been shown, in some studies, to be larger for changes that cross a phonemic boundary than for changes within the same phoneme category (Dehaene-Lambertz, 1997).<sup>1</sup>

✉ Yang Agnes Gao  
ygao2@villanova.edu

<sup>1</sup> Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>2</sup> Department of Psychological and Brain Sciences, Villanova University, 800 E Lancaster Ave, Villanova, PA 19085, USA

<sup>3</sup> Department of East Asian Languages and Cultures, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>4</sup> Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>5</sup> Neuroscience Program, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA

<sup>1</sup> It is important to note that *categorical perception* and *category-based responses* are distinct concepts. Categorical perception states that perceptual information is processed relative to a category structure (in an extreme case, it is an “all-or-nothing” response where within-category distinctions are not perceived). However, behavioral responses can still be based on categories (either graded or all-or-nothing) even if perceptual representations are non-categorical.

In addition to segmental contrasts, a number of studies have investigated whether lexical tone<sup>2</sup> is categorically perceived. More than 70% of the world's languages are tone languages (Yip, 2002), in which words are distinguished by tonal (f0) features. For example, Mandarin uses contrasts in pitch height and/or direction to distinguish lexical items (Gandour, 1978). In Mandarin, there are four lexical tones, which use different pitch contours to discriminate lexical items: /mā/ (level f0 contour with a slight drop at the end of the utterance) means *mother*, /má/ (rising f0) means *hemp*, /mǎ/ (falling-rising f0) means *horse*, and /mà/ (falling f0) means *scold*. Previous studies have argued that lexical tone is perceived categorically, based on both behavioral and ERP data (Xu, Gandour, & Francis, 2006; Xi, Zhang, Shu, Zhang, & Li, 2010; Zhang, Xi, Wu, Shu, & Li, 2012). However, some of these studies are problematic methodologically (e.g., low N's; inappropriate use of high-pass filters in ERP data), and they have often not investigated non-native listeners' responses to tone stimuli. This is a critical comparison, as these listeners can distinguish the sounds, but do not have categories corresponding to phonological differences.

More importantly, however, researchers now agree that listeners are sensitive to within-category phonetic differences, based on experiments measuring reaction times (Pisoni & Tash, 1974), category goodness ratings (Massaro & Cohen, 1983; Miller, 1994), eye-movements in the visual-world paradigm (McMurray, Tanenhaus, & Aslin, 2002), MEG responses (Frye et al., 2007) and ERP responses (Toscano, McMurray, Dennhardt, & Luck, 2010). These results argue against the strongest form of categorical perception, demonstrating that listeners are sensitive to graded differences along acoustic continua. Thus, while listeners' phonological judgments in identification and discrimination tasks appear categorical, their underlying perceptual processes are not based on discrete phonological category representations.

It remains unclear, however, how listeners' knowledge of phonological category structure affects perceptual and categorization processes. The current study aims to address this by investigating perception of lexical tones and comparing brain responses of native Mandarin listeners, who have phonological knowledge of tones, and naïve English listeners, who do not. This allows us to study how pre-existing category information influences listeners' sensitivity to within-category acoustic differences. In the following sections, we briefly review previous studies that have provided evidence for and against categorical perception of both segmental contrasts and tones using behavioral and electrophysiological measures. We then present the results of an ERP experiment designed to measure Mandarin-speaking and English-speaking

participants' categorization of sounds varying along Mandarin tone continua.

## Perception of lexical tone

Behavioral evidence for categorical perception is based on two criteria (Liberman et al., 1957; Repp, Healy, & Crowder, 1979). First, listeners' identification functions for sounds varying along acoustic continua should show abrupt changes and steep slopes between phoneme categories, such that listeners consistently categorize sounds as one or the other. Second, peaks in discrimination accuracy must correspond to the location of category boundaries in the identification task, operationally defined as the point along a continuum at which the identification function switches from one category to the other. These two tasks have been used to study listeners' perception of speech sounds varying along a wide range of acoustic continua and phonological feature dimensions (Liberman, Delattre, & Cooper, 1952; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

A number of studies examining lexical tone perception in adult listeners have used the categorical perception framework to understand how native speakers of tone languages categorize stimuli varying along tone continua (Xu et al., 2006; Xi et al., 2010; Zhang et al., 2012). One way that this has been studied is by editing f0 contours in spoken syllables in a graded manner, with the ends of the continua representing prototypical contours for tones. In the categorical perception framework, equivalent f0 distances between two tones would be treated differently, depending on whether the two tones are heard as members of the same category or as members of different categories. That is, two members of the same category would be less discriminable than two tokens from different categories, even with an equivalent acoustic difference between them. Xi et al. (2010) used a categorical perception task to examine listeners' perception of tone stimuli varying along an 11-step continuum from tones 2 to 4, embedded in the syllable /ba/. Identification data showed typical categorization functions, with a steep slope at the category boundary, and discrimination performance showed a peak at the category boundary, consistent with categorical perception.

Listeners' responses in these tasks are also influenced by their language background. Xu et al. (2006) conducted a cross-language study that compared native Mandarin and English speakers' perception of a level to rising Mandarin tone continuum (tone 1 to tone 2). Results showed evidence consistent with categorical perception only for Mandarin listeners, who showed a sharp identification slope and greater discrimination peak at the category boundary. English speakers, however, showed no clear peak at the boundary, and, indeed, were equally good at discriminating between-

<sup>2</sup> Lexical tone refers to the phonological contrast signaled by f0 contours in tone languages, like Mandarin.

category differences and better than Mandarin listeners at discriminating within-category differences. This pattern of results fits with the hypothesis that listeners' responses are determined, at least in part, by their phonological categories: since English speakers do not have tone categories, they do not show categorical perception of tones.

In line with the results of Xu et al. (2006), Peng et al. (2010) examined the influence of a tone versus non-tone native language (Mandarin and German), as well as the influence of different tone inventories (Mandarin and Cantonese) on categorization of Mandarin tones, using two-tone continua (rising to level and level to falling). Identification results showed that listeners from all three language groups had similar category boundaries along the continua. However, discrimination results showed that tone-language listeners had higher accuracy than German listeners at the category boundary. Moreover, German listeners showed a discrimination peak that was different from tone language listeners. Specifically, tone language listeners' discrimination peaks were approximately centered along the f0 continua, corresponding to the phonological boundary in Mandarin; in contrast, German listeners' discrimination peak was located towards the ends of the rising and falling continua.

A possible explanation for the different discrimination peak for German listeners is that even though they do not have tone categories, they have structured an *ad hoc* category boundary based on task demands or low-level psychoacoustic factors. In particular, they may place a boundary at the point at which the f0 contour switches direction from rising to falling (which is not necessarily the same as the phonological boundary). In this case, the peak would not reflect linguistic processing of the tones. Similar psychoacoustic effects have been observed for non-speech acoustic continua (Miller et al., 1976; Pisoni, 1977). Peng et al.'s results also reveal the influence of different tone systems. In the level-falling continuum, Cantonese listeners showed a different discrimination peak than Mandarin listeners, suggesting that Cantonese listeners were also influenced by a low falling tone that only exists in Cantonese.

Together, the results of these studies suggest that listeners' categorization of f0 contours is likely determined by a combination of factors. When stimuli span an acoustic dimension that corresponds to a phonological difference in the listener's language, they show a category boundary at that location. When listeners do not have categories along that dimension, they may either show no peak in their discrimination (indicating no category boundary; Xu et al., 2006) or an *ad hoc* category boundary based on task demands or low-level auditory differences between stimuli (Peng et al., 2012). However, it is unclear how to separate the contribution of these two sources of information in traditional behavioral tasks designed to measure categorical perception.

## Electrophysiological measures of categorical perception

Given the limitations of the behavioral paradigm described above, several studies using electrophysiological measures of neural activity have also investigated how listeners categorize speech sounds varying along acoustic continua, with some finding evidence for categorical perception, but others finding no evidence for it. Many of these studies have investigated the mismatch negativity (MMN) ERP response. The MMN is elicited around 100–200 ms, and is argued to reflect change detection or discrimination in auditory processing at a pre-attentional stage (Näätänen, 2001; Woldorff, Hackley, & Hillyard, 1991). Typically, these studies present listeners with a standard (frequently-occurring) speech stimulus and measure responses to deviant stimuli that are exemplars of the same phoneme (within-category deviants) or a different phoneme (across-category deviants).

Some studies examining segmental features have shown evidence of categorical perception using the MMN. Dehaene-Lambertz (1997) examined the categorization of vowel features by looking at one phonetic boundary that is present and another that is absent in the listeners' native language. A large mismatch negativity was induced by native between-category deviants, but not by non-native or within-category deviants. This result was argued to suggest that phonological categorization serves as the basis for auditory mismatch detection. Phillips et al. (2000) found similar effects consistent with categorical perception in an MEG experiment measuring the magnetic counterpart to the MMN.

However, other studies have failed to find evidence of categorical perception in the MMN. Sharma et al. (1993) reported that, for stimuli varying in voice onset time (VOT), an acoustic cue for word-initial voicing, the MMN was observed both for within- and between-category oddballs. Furthermore, the MMN did not differ in latency, amplitude, or area for within- and between-category deviants, indicating equal discrimination of both sounds. Sharma et al. argued that the MMN appears to be sensitive to subtle acoustic differences and reflects the processing of acoustic aspects of the speech stimulus, but not phonetic processing of categories. Other work has also shown that the MMN is sensitive to within-category phonetic differences (Joanisse, Robertson, & Newman, 2007).

The MMN has also been used to study categorical perception for lexical tone contrasts. Xi et al. (2010) presented listeners with stimuli varying in lexical tone in an oddball paradigm and measured MMN responses to across- and within-category deviant tones. They found that, in a passive-listening task, across-category contrasts elicited a larger MMN than within-category contrasts at left frontal-central electrode sites, whereas the MMN did not differ between the two deviants for electrodes located on the right side of the scalp. However, this

result needs further examination since the MMN is usually measured over frontal and central electrode sites.

In a subsequent study, Zhang, Xi, Wu, Shu, and Li (2012) examined categorical perception in an active task, focusing on the N2b and P3 ERP components instead. These responses are typically elicited in an oddball paradigm, where one task-relevant stimulus occurs infrequently relative to other stimuli. The N2 component generated in response to rare auditory targets can be separated into three subcomponents: the MMN (also referred to as the N2a), the central N2b, and the posterior N2c. The N2b, in particular, is argued to reflect degree of perceptual deviation, where larger N2b responses reflect the mismatch between a stimulus and mental template (Folstein & Van Petten, 2008). Auditory oddball tasks also elicit a P3 response, which indexes categorization, with larger amplitude P3s reflecting better exemplars of a category, including phonological categories (Toscano et al., 2010).

Zhang et al. (2012) found differences for both across- and within-category deviants in the N2b and P3 components, and like the results of Xi et al. (2010), these effects varied with scalp topography. Across-category deviants elicited larger N2 and P3 responses at left-parietal sites than at right parietal sites, whereas the within-category deviants elicited similar N2 and P3 responses over both sites. In addition, Zhang et al. found a longer latency P3 for the across- versus within-category oddball, which they suggested was in accordance with category-driven perception. However, it stands to reason that across-category oddballs should be easier to distinguish from the standard stimulus, which might lead us to expect a shorter P3 latency.

## Categorical versus non-categorical effects

Overall, the studies described above show mixed evidence for categorical perception using ERP paradigms. Moreover, a number of behavioral experiments have argued against categorical perception, primarily by demonstrating that other tasks reveal listeners' sensitivity to within-category phonetic differences. For example, Massaro and Cohen (1983) used a rating task to show that listeners' have graded sensitivity to stimuli varying along acoustic continua between phonological categories differing in place (/b/ vs. /d/), voicing (/b/ vs. /p/), and vowel quality (/i/ vs. /I/). Similarly, McMurray, Tanenhaus, and Aslin (2002) used the visual-world eye-tracking paradigm to demonstrate effects of gradient acoustic differences on lexical access.

More recently, ERP evidence has suggested a rethinking of the categorical perception paradigm altogether. Toscano et al. (2010) presented listeners with an auditory oddball task with stimuli varying along VOT continua between the words *beach-peach* and *dart-tart*. Participants listened for one of the four words on different blocks of the experiment.

Toscano et al. measured two ERP components thought to serve as potential indices of perceptual processing and categorization. Specifically, they measured the fronto-central auditory N1, a response that has been shown to vary with the acoustic properties of the stimulus (Picton, Woods, & Proulx, 1978), and the P3, a measure of categorization. Results revealed that listeners are sensitive to fine-grained VOT differences at both stages of processing. In particular, the N1 showed a continuous mapping to VOT, even when participants indicated stimuli belonged to different phonological categories. In contrast, the P3 was influenced by phonological categories, but in a gradient fashion: P3 amplitude was largest for exemplars most consistent with the target word, and gradually decreased as the VOT approached the category boundary. Thus, the N1 data suggest that early perception is non-categorical, while later categorization stages show sensitivity to both within-category phonetic differences and listeners' phonological categories. This study provides strong evidence that listeners are sensitive to fine-grained acoustic differences and that this sensitivity is maintained through late perceptual and categorization stages.

Overall, these studies provide evidence that fine-grained acoustic differences within phonetic categories have significant effects on listeners' perception and categorization. They suggest that speech perception may be fundamentally continuous (Massaro & Cohen, 1983) and that the effect of phonological information is a product of categorization and task demands, not perceptual encoding (Pisoni & Lazarus, 1974; Toscano et al., 2010). Moreover, previous reports by Xi et al. (2010) and Zhang et al. (2012) arguing for categorical perception of tone need further scrutiny. For example, the sample size used in the Zhang et al. study was small (N=13). In addition, the behavioral data showed that participants were above chance at detecting within-category oddballs (85%), suggesting within-category contrasts were not entirely lost. In addition, the ERP analysis used a high-pass filter setting (0.3 Hz) that has since been shown to distort effects, timing, polarity, and scalp topography for slow effects such as the N4 and P6 (Tanner, Morgan-Short, & Luck, 2015) and could potentially affect the N2 and P3 as well. As such, it is worthwhile to reexamine N2b and P3 responses to lexical tone for native Mandarin listeners.

Examining corresponding responses in native English listeners also provides an opportunity to study the broader theoretical issue surrounding the nature of the categorization processes driving these effects. Given that listeners' knowledge of phonological categories has an effect on their responses in these tasks, we may be able to separate the contributions of phonological knowledge and low-level auditory perception by examining ERP components associated with perceptual processing and categorization with tone continua presented to both native tone-language listeners and non-native listeners. The different brain responses observed by Zhang et al., for

example, might reflect differences in perceptual salience of the stimuli (e.g., a change in the slope of the  $f_0$  contour from positive to negative), and not specifically linguistically-defined phonological categories. The present study will examine this question.

## Summary and goals

The current study investigates the discrepant findings in work examining lexical tone perception by replicating and extending the approach used by Zhang et al. (2012), but includes a larger sample size of native Mandarin speakers ( $N=20$  vs.  $N=13$ ). Additionally, naïve native English speakers are included to evaluate the influence of language experience on the perception of tones. We investigate perceptual deviation detection using the N2b, and phonological categorization using the P3 component. If these responses reflect listeners' phonological category structure, we expect to see effects related to tone categories for the Mandarin group only. English speakers should show no difference in processing the two oddballs (which are equally acoustically different from the standard), since they have no phonological categories along the tone dimension.

Alternatively, if we find processing similarities between the Mandarin and English speakers, this would suggest that the N2/P3 responses do not reflect linguistically-defined tone categories. There are two possible outcomes in this case. One possibility is that neither Mandarin nor English speakers will show differences in responses to the two oddballs. This would suggest that both groups of listeners are categorizing the stimuli based on their acoustic difference from the standard. It is also possible that both the English and Mandarin speakers show processing differences between the two oddballs. This would suggest that listeners are sensitive to acoustic differences along the tone dimensions besides those related to the acoustic distance between the oddballs and the standard, but again, these responses would not be related to linguistically-defined categories.

## Method

### Design

The experiment was a 3 (tone stimulus; standard, across-category deviant, within-category deviant)  $\times$  2 (native language; Mandarin vs. English) mixed design. Standard stimuli were presented 800 times, and each deviant stimulus was presented 100 times each in a pseudo-random order (see Procedure), for a total of 1,000 trials across the experiment. Each participant was tested in a single session, lasting around 2 h.

## Participants

Twenty-one native speakers of Mandarin and 20 native speakers of English participated in the experiment. The native Mandarin speakers all spoke Mandarin as their primary language and had no significant exposure to another language before the age of 12 years. The participants in the English group were all native speakers of American English without experience with Mandarin or any other tone language. All participants were right-handed, reported no current or history of hearing impairment, and reported no history of neurological impairment or use of psychoactive medication. The data from one Mandarin-speaking participant was excluded due to excessive artifacts in the raw EEGs, leaving 40 participants in the final analysis (Mandarin: 11 females, nine males; age 21–28 years,  $M=23$  years; English: 14 females, six males; age 18–24 years,  $M=20$  years). Participants also completed a music background questionnaire and a musical ability assessment (Zentner & Strauss, 2017). The two groups did not differ significantly in terms of years of musical experience ( $t(38)=-0.56$ ,  $p=0.709$ ) or average musical ability (mean Mini-PROMS score;  $t(38)=0.99$ ,  $p=0.165$ ), though the Mandarin speakers scored better on the Beats sub-test of the Mini-PROMS (Mandarin:  $M=13.4$ ,  $SD=3.5$ ; English:  $M=10.9$ ,  $SD=3.7$ ;  $t(38)=1.84$ ,  $p=0.037$ ). Descriptive statistics for each group's music experience and average ability are presented in Table 1. All participants provided informed consent in accordance with University of Illinois IRB protocols and received cash or course credit as compensation.

## Stimuli

Stimuli were identical to those used by Zhang et al. (2012) and Xi et al. (2010; used with permission).<sup>3</sup> A 10-interval lexical tonal continuum was synthesized with the monosyllable /ba/ from the high rising tone /ba2/ to the falling tone /ba4/. These speech stimuli were normed by Xi et al. (2010), and both discrimination and identification functions showed typical categorization responses (see Xi et al., 2010, for further details on stimulus construction and norming). One across-category pair and one within-category pair were chosen from the continuum. Specifically, step 7 served as the standard (which was within the /ba4/ category), with step 3 serving as the across-category deviant (/ba2/) and step 11 as the within-category deviant (/ba4/). Importantly, steps 3 and 11 are acoustically equidistant from step 7.

## Procedure

Upon arriving in the lab, participants were asked to complete a language background questionnaire and an abridged version

<sup>3</sup> We thank Ping Li for providing us with the original stimulus files.

**Table 1** Music experience and ability by group

L1	Number of musicians	Years of musical training	Average Mini-PROMS score
Mandarin	13/20 (65%)	$M=4.1, SD=5.2$	$M=21.9, SD=5.9$
English	13/20 (65%)	$M=4.7, SD=5.0$	$M=19.6, SD=4.5$

of the Edinburgh Handedness Inventory. Each participant was seated in a comfortable chair in front of a computer monitor and was randomly assigned to an experimental list. Participants were asked to try to relax and to minimize blinks and movements during stimulus presentation. The experiment began with a familiarization phase, during which the standard tone (step 7) was played 10 times for the participant. All auditory stimuli were presented via Etymotic ER-3a insert earphones, with volume individually adjusted to each participant's most comfortable level. Seven oddball blocks were presented to each participant with four pseudo-randomized orders presented across participants (five participants per list order per language group). The standard was presented 80% of the time, and the within-category deviant and across-category deviant occurred pseudo-randomly with a probability of 10% each. Any two adjacent deviants were separated by at least three standard trials.

Each trial consisted of the following series of events. A fixation cross appeared at the center of the screen, followed by the auditory stimulus. The interval between onset of the fixation cross and onset of the speech stimulus was jittered randomly between 300 and 450 ms. Participants were instructed to press one button for the standard tone that they had heard during the familiarization phase and a different button with their other hand for any other sound. The hand assigned to the standard tone was counterbalanced across participants; the prestimulus fixation cross remained on the screen through the auditory stimulus until a response was registered. After participants responded, the next trial began automatically with an inter-trial interval (ITI) that jittered randomly between 900 and 980 ms. The ITI showed a blank screen, and participants were encouraged to use this time to blink. Between each block, participants were given a short break. At the end of each block, feedback about cumulative accuracy in discriminating the standard tone from the two oddballs was provided on screen.

**Table 2** Mean proportion correct and reaction time (RT) for Mandarin speakers (standard errors of the mean reported in parentheses)

Tone type	Proportion correct	RT
Standard	.999 (.00)	470.3 ms (87.50)
Across	.957 (.04)	591.0 ms (94.21)
Within	.923 (.07)	605.1 ms (109.69)

## Data acquisition and analysis

Continuous EEG was recorded from 28 tin scalp electrodes mounted in an elastic cap (Electro-cap International), in accordance with the extended 10–20 system (Jasper, 1958; FP1, FP2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, and O2). Eye movements were monitored with electrodes placed below the left eye (referenced offline to FP1) and at the outer canthus of each eye (referenced offline in a right-to-left bipolar montage). Scalp electrodes were referenced online to the left mastoid and re-referenced offline to the algebraic mean of activity over the left and right mastoids. Impedances at scalp and mastoid sites were held below 10 k $\Omega$ , and impedances were kept below 15 k $\Omega$  for the ocular electrodes.

Continuous EEG was amplified with a BrainAmpDC bioamplifier system (Brain Products, Gilching, Germany) and digitized with a 1,000-Hz sampling rate and an online analog 250 Hz low-pass filter and 10-s time constant ( $\approx 0.016$  high-pass). All offline processing was carried out with EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes. A 0.1- to 30-Hz bandpass filter ( $-6$ -dB cutoff, 12-dB/octave roll-off) was applied to the continuous EEG offline.

ERPs were averaged time-locked to the onset of the speech stimulus for each participant over each electrode for each tone, relative to a 200-ms pre-stimulus baseline. Raw EEG data were visually inspected for artifacts, and trials characterized by excessive eye movements, drift, alpha, or other artifacts were rejected. All data from any participant with greater than 25% rejected trials in any tone condition were excluded from analysis, resulting in the rejection of one participant (see above). For the remaining 40 participants, independent component analysis (ICA) artifact correction was used to remove components that were related to blinks and saccades. These artifacts were identified by manual inspection of component scalp topographies and time courses. After removing artifactual components, the EEG data were re-screened for artifacts, and an average of 10% of trials were excluded across the three tones. Only trials with correct behavioral responses were included in the ERP analysis, and only standard tones following other standard tones were included. This resulted in an average of 567 standard trials, 89 across-category oddballs and 91 within-category oddballs per participant.

**Table 3** Pairwise contrasts for Mandarin speakers predicted by the mixed effects models using *lsmeans* (standard errors of the mean are reported in parentheses; p-values are corrected for multiple comparisons using Tukey’s method)

Tone pair	Accuracy		RT	
	<i>lsmeans</i> Estimate (log odds difference)	p-value	<i>lsmeans</i> Estimate (ms difference)	p-value
Across-standard	-3.449 (.30)	<.001	120.7 (7.47)	<.001
Within-standard	-4.044 (.31)	<.001	135.1 (12.24)	<.001
Across-within	0.595 (.19)	<.01	-14.44 (8.69)	.246

ERPs were quantified using mean amplitudes within time windows chosen *a priori* (150–300 and 300–600 ms), corresponding to the N2b and P3 time windows, respectively. These time windows were chosen to be in approximately the same time range as those used by Zhang et al. (2012), and they were well aligned with the components in our dataset. Repeated measures of ANOVA were conducted for each time window separately for midline (Fz, Cz, Pz, Oz) and lateral sites. Native language (L1) was included as a between-subjects factor for all analysis. For midline sites, tone category and electrode were included as repeated-measures factors. Data for lateral sites were grouped into six electrode pairs with six on each hemisphere: left hemisphere (F3, FC1, C3, CP1, P3, O1) and right hemisphere (F4, FC2, C4, CP2, P4, O2). For lateral sites, tone, electrode pair, and hemisphere (right, left) were included as repeated-measures factors. The Greenhouse–Geisser correction for inhomogeneity of variance was applied to all repeated measures with more than one degree of freedom in the numerator. In such cases, the uncorrected degrees of freedom and the corrected p-value are reported.

**Results**

**Behavioral results**

Analysis of behavioral data focused on accuracy of tone discrimination (standard vs. oddball) and reaction time (RT). All trials were included for accuracy analysis. For RT analysis, only correct trials were included in the analysis. In addition, trials with RTs under 100 ms were considered as absolute outliers, leading to 0.455% of total trials being rejected for Mandarin speakers, and 0.341% for English speakers.

**Table 4** Mean proportion correct and reaction time (RT) for English speakers (standard errors of the mean are reported in parentheses)

Tone type	Proportion correct	RT
Standard	.992 (.23)	477.4 ms (60.56)
Across	.894 (.14)	575.4 ms (59.27)
Within	.956 (.04)	558.0 ms (62.20)

Relative outlier cutoffs were then calculated for each participant. Outlier cutoffs are defined as +/- 2.5 SDs from each participant’s mean RT, leading to an additional 1.46% rejection rate for Mandarin speakers, and 1.21% for English speakers. Pair-wise contrasts for behavioral measures were computed using the least squares means method using the *lsmeans* package (Lenth, 2016) for R. Estimates were computed from linear mixed effects models for RTs and logistic mixed effects models for accuracy. Tone condition was a fixed effect (sum-coded); random effects were intercepts for participants and by-participant slopes for tone condition. We fit mixed-effects models for the entire dataset to examine differences across the two groups of talkers; likelihood ratio tests were used to compute p-values for these models. We also fit separate regression models for each L1 group; p-values for these follow-up analyses were corrected for multiple comparisons using Tukey’s method.

Overall, listeners were accurate in identifying the tone categories. Mixed-effects models examining tone condition and L1 revealed a main effect of tone condition (accuracy:  $\chi^2(2)=74.88, p <.001$ ; RT:  $\chi^2(2)=89.98, p <.001$ ) and a tone  $\times$  L1 interaction (accuracy:  $\chi^2(2)=26.61, p <.001$ ; RT:  $\chi^2(2)=10.65, p <.01$ ), suggesting that the pattern of accuracy and RT for specific tones differed for the two groups.

Behavioral results for Mandarin speakers are reported in Tables 2 and 3. Mandarin speakers’ behavioral responses were very good, with near-perfect accuracy for all tone conditions. However, there were significant differences in accuracy between tone conditions, and significant differences in RT between the standard tone and the two oddballs (oddball RTs were slower than standard RTs). For accuracy, there was a small but reliable advantage for the across- versus the within-category oddball, as would be predicted based on the listeners’ phonological category structure. However, target accuracy for the within-category deviant was still very high (92%). Notably, no reliable RT difference was found between the oddball tones (see Table 2).

<sup>4</sup> The model including the tone  $\times$  L1 interaction for accuracy did not converge, so we conducted an analysis excluding the by-participant random slope for tone condition. This model did converge, and the analysis revealed the same pattern of effects.

Behavioral results for English speakers are reported in Tables 4 and 5. English speakers' behavioral responses demonstrated a different pattern from the Mandarin speakers, with better discrimination accuracy for the within-category deviant than for the across-category deviant and faster RTs for within-versus across-category deviants. This suggests that, for English speakers – who do not have phonological categories along this acoustic dimension – the within-category oddball was more perceptually discriminable from the standard than the across-category oddball. All pair-wise contrasts for accuracy and RT were statistically significant.

## ERP results

Overall, the speech sounds elicited typical ERP responses to auditory stimuli. Grand mean waveforms comparing brain responses to the standard, across-category oddball, and within-category oddball stimuli are presented in Figs. 1 and 2 for Mandarin and English speakers, respectively. Visual inspection of the data reveals that oddballs elicited larger N2b and P3 amplitudes for both groups, and that these effects had stereotypical scalp topographies (Fig. 3). In addition, N2b amplitudes were larger (more negative) for the across- versus within-category oddballs for both participant groups, and P3 amplitudes were larger (more positive) for within- versus across-category oddballs for both groups.

## N2b amplitude

Statistical analysis of the N2b amplitudes revealed a main effect of tone stimulus: the two oddballs elicited a larger N2b relative to the standard (Table 6). There was a significant Tone  $\times$  Electrode interaction over midline and lateral sites, and a significant Tone  $\times$  Hemisphere  $\times$  Electrode interaction over lateral sites, suggesting that N2b amplitude was greater over central sites. Across the two groups of participants, there were significant interactions between Tone and L1, and between Tone, Electrode, and L1 over midline and lateral sites, and a Tone  $\times$  Hemisphere  $\times$  L1 interaction over lateral sites. However, these effects appeared to be caused by differences between the two groups in ERP responses to the standard tone, not to the two oddball tones. Therefore, we further investigated N2 amplitudes between the two oddballs across both groups (Table 7), including L1, electrode, and hemisphere as within-group factors.<sup>5</sup>

Follow-up analyses examining the two oddballs revealed that N2b amplitude did not differ between two groups of listeners, and that the N2b was symmetrically distributed across the scalp (Table 7). There was a main effect of tone over

midline and lateral sites: N2b was greater for the across-category deviant (midline: mean= $-2.393\mu\text{V}$ , SE=.352; lateral: mean= $-2.982\mu\text{V}$ , SE=.355) than the within-category deviant (midline: mean= $-1.549\mu\text{V}$ , SE=.375; lateral: mean= $-2.239\mu\text{V}$ , SE=.360). This contrasted with Zhang et al. (2012), where no main effect of deviant type was found. In addition, Zhang et al. (2012) found an interaction between deviant type and hemisphere, which we did not observe in the current study.

## P3 amplitude

The overall omnibus ANOVA for the 300- to 600-ms time window (measuring mean P3 amplitude) showed a main effect of Tone and a Tone  $\times$  Electrode interaction for both the midline and lateral analyses: the two oddballs elicited greater P3 amplitude relative to the standard (Table 6), and this effect was maximal over parietal sites. This is a typical P3 scalp topography. There was also a Tone  $\times$  Electrode  $\times$  L1 interaction for both the midline and lateral sites, reflecting the fact that the oddballs elicited larger P3 effects in the English group than the Mandarin group. Moreover, over the lateral sites, there was a Tone  $\times$  Hemisphere  $\times$  Electrode interaction.

Because our primary question involves whether P3 differed between the two deviant tones (cf. Zhang et al., 2012), we also conducted a planned pair-wise ANOVA comparing the only the two deviant tones (Table 7). We found a main effect of Tone: P3 amplitude was greater for the within-category deviant (midline: mean= $2.884\mu\text{V}$ , SE=.616; lateral: mean= $2.105\mu\text{V}$ , SE=.582) than the across-category deviant (midline: mean= $2.274\mu\text{V}$ , SE=.573; lateral: mean= $1.518\mu\text{V}$ , SE=.539). In contrast, Zhang et al. (2012) did not find a significant main effect of deviant type.

Three-way interactions also showed two effects of interest: over both midline and lateral sites, the Tone  $\times$  Electrode  $\times$  L1 interaction was statistically significant; over lateral sites, there was a reliable Tone  $\times$  Hemisphere  $\times$  Electrode interaction. Based on the results of Zhang et al. (2012), we expected to see qualitative hemispheric differences between two groups of listeners. Particularly, Mandarin speakers should show a left-dominant P3 for the across-category oddball, but not for the within-category oddball, whereas English speakers' P3 responses between the two oddballs would be symmetrical.

These two interactions are further examined in Figs. 4 and 5 over midline and lateral electrodes, respectively. Inspection of the Tone  $\times$  Electrode  $\times$  L1 interaction is shown in Fig. 4. P3 amplitude was larger for the within-category deviant over central and parietal electrodes in both groups. In addition, English speakers showed larger P3 amplitude to both deviants relative to Mandarin speakers. In general, the effect of tone was qualitatively similar across the two groups, with the exception of electrode Fz: Mandarin speakers showed no difference between the two deviants at Fz, but English speakers' responses were more positive to the within-category deviant at this site.

<sup>5</sup> Note that this is equivalent to conducting an analysis on standard-minus-deviant difference waves, as the same standard waveform would be subtracted from each deviant waveform.



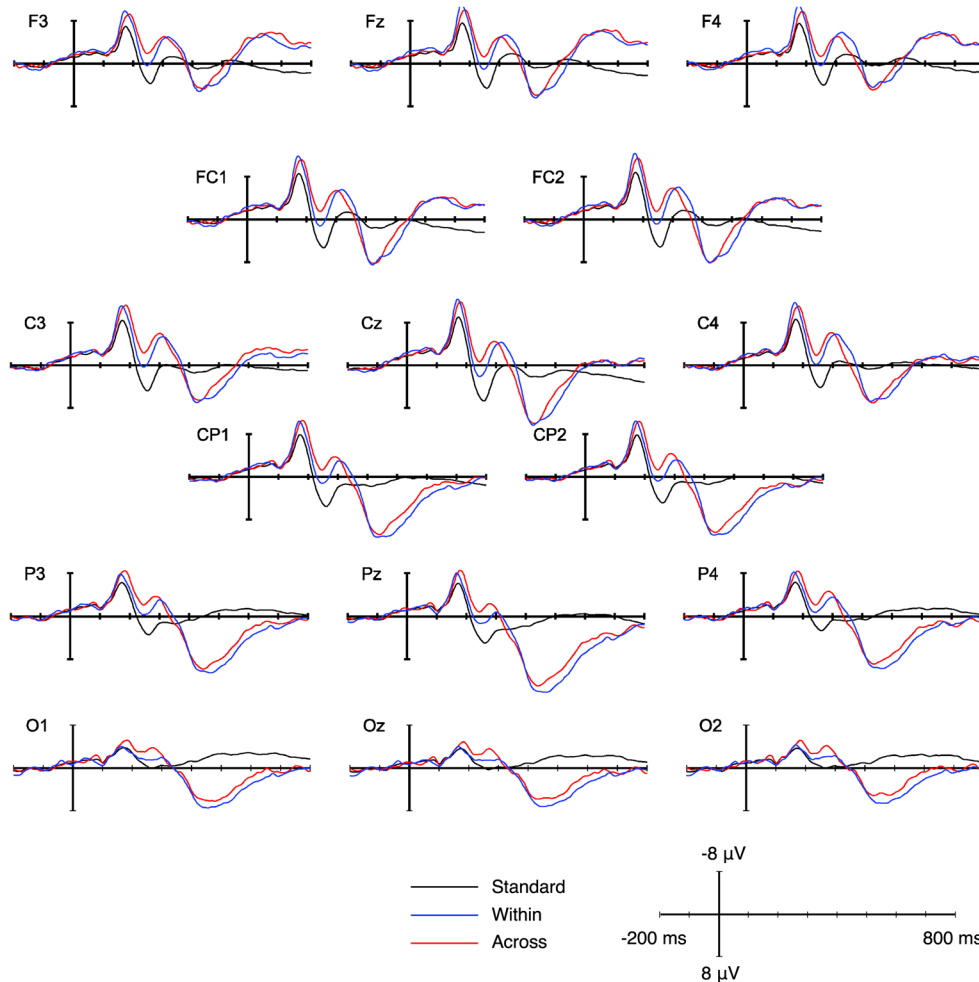
**Table 5** Pairwise contrasts for English speakers predicted by the mixed effects models using *lsmeans* (standard errors of the mean are reported in parentheses; p-values are corrected for multiple comparisons using Tukey’s method)

Tone pair	Accuracy		RT	
	<i>lsmeans</i> estimate (log odds difference)	p-value	<i>lsmeans</i> estimate (ms difference)	p-value
Across-standard	-3.606 (.45)	<.001	97.92 (9.43)	<.001
Within-standard	-2.825 (.34)	<.001	80.57 (11.57)	<.001
Across-within	-0.781 (.24)	<.01	17.35 (5.82)	.020

English speakers also showed slightly larger differences in amplitude as a function of tone type at Cz, and smaller differences in amplitude at Oz. However, since the topography of the P3 component is typically measured over parietal scalp sites (cf. Zhang et al., 2012), there is no clear evidence for any qualitative difference in P3 responses to the two deviant tone types across the two groups in the present data. This suggests that English listeners still show P3 responses even without phonological categories along this acoustic

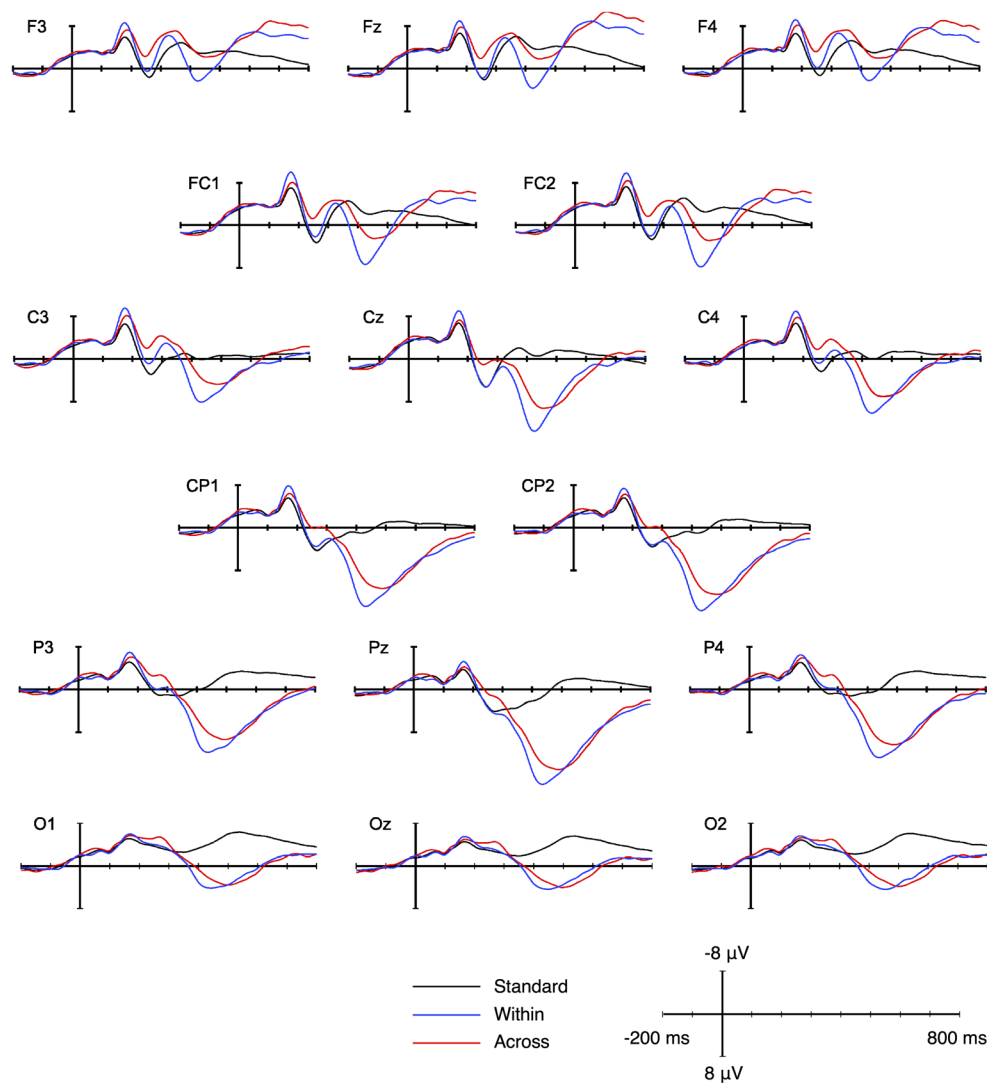
dimension, possibly reflecting ad-hoc categories (e.g., standard-like vs. non-standard-like) due either to task demands or low-level perceptual differences. Therefore, the P3, as measured in this experiment, is not simply reflecting phonological categories, since the same effect was found in both L1 groups.

We also found a Tone × Hemisphere × Electrode interaction at lateral sites as shown in Fig. 5. Between the two deviants, the P3 amplitude was larger for the within-category



**Fig. 1** Grand mean waveforms for Mandarin speakers elicited by the across-category deviants (black line), within-category deviants (red line), and the standards (blue line) at frontal, central, and posterior scalp sites.

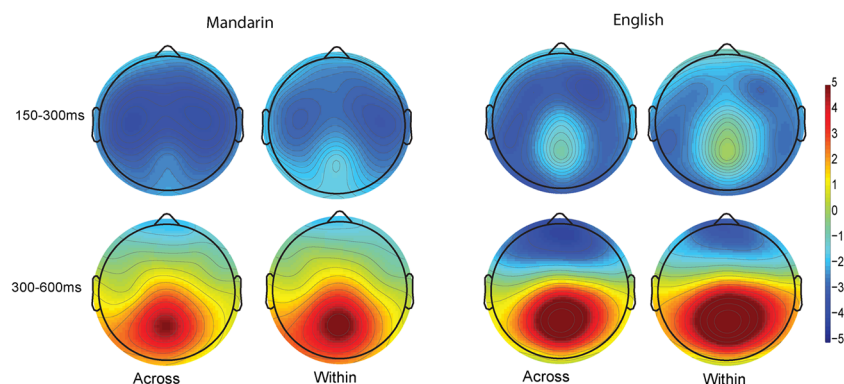
Negative voltage is plotted up; 200 ms of pre-stimulus and 800 ms of post-stimulus activity are depicted



**Fig. 2** Grand mean waveforms for English speakers elicited by the across-category deviants (black line), within-category deviants (red line), and the standards (blue line) at frontal, central, and posterior scalp sites

deviant in both hemispheres, in contrast with the results from Zhang et al. (2012), which found that Mandarin listeners showed a left-dominant P3 for the across-category oddball. In the current study, both the within- and across- category

deviants showed slightly more positive-going amplitudes over the right central site (C4) than the left central site (C3); however, because this occurred for both deviant tones, there is no evidence in our data that the P3 showed a left hemisphere bias



**Fig. 3** Topographic map for Mandarin and English speakers in 150–300 ms and 300–600 ms time windows

**Table 6** Omnibus ANOVA results in the 150–300 ms and 300–600 ms time windows

	150–300 ms				300–600 ms			
	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
<i>Midline</i>								
Tone	<b>2,76</b>	<b>46.352</b>	<b>2.969</b>	<b>&lt;.001</b>	<b>2,76</b>	<b>46.662</b>	<b>12.277</b>	<b>&lt;.001</b>
L1	1,38	.159	49.924	.692	1,38	.331	100.80	.568
Tone × L1	<b>2,76</b>	<b>5.397</b>	<b>2.969</b>	<b>.007</b>	2,76	2.041	12.277	.151
Tone × elec	<b>6,228</b>	<b>11.104</b>	<b>.392</b>	<b>&lt;.001</b>	<b>6,228</b>	<b>54.108</b>	<b>1.241</b>	<b>&lt;.001</b>
Tone × elec × L1	<b>6,228</b>	<b>4.198</b>	<b>.392</b>	<b>.007</b>	<b>6,228</b>	<b>5.922</b>	<b>1.241</b>	<b>&lt;.001</b>
<i>Lateral</i>								
Tone	<b>2,76</b>	<b>64.824</b>	<b>7.944</b>	<b>&lt;.001</b>	<b>2,76</b>	<b>32.594</b>	<b>32.826</b>	<b>&lt;.001</b>
L1	1,38	.000	146.42	.983	1,38	.270	269.98	.606
Tone × L1	<b>2,76</b>	<b>3.921</b>	<b>7.944</b>	<b>.024</b>	2,76	2.033	32.826	.150
Tone × hemi	2,76	.464	.501	.622	2,76	.724	2.635	.418
Tone × hemi × L1	2,76	.560	.501	.566	2,76	.863	2.635	.373
Tone × elec	<b>10,380</b>	<b>15.915</b>	<b>.455</b>	<b>&lt;.001</b>	<b>10,380</b>	<b>52.180</b>	<b>1.490</b>	<b>&lt;.001</b>
Tone × elec × L1	<b>10,380</b>	<b>2.797</b>	<b>.455</b>	<b>.038</b>	<b>10,380</b>	<b>5.520</b>	<b>1.490</b>	<b>.001</b>
Tone × hemi × elec	<b>10,380</b>	<b>2.634</b>	<b>.072</b>	<b>.027</b>	<b>10,380</b>	<b>3.230</b>	<b>.434</b>	<b>.048</b>
Tone × hemi × elec × L1	10,380	.865	.072	.503	10,380	.449	.434	.630

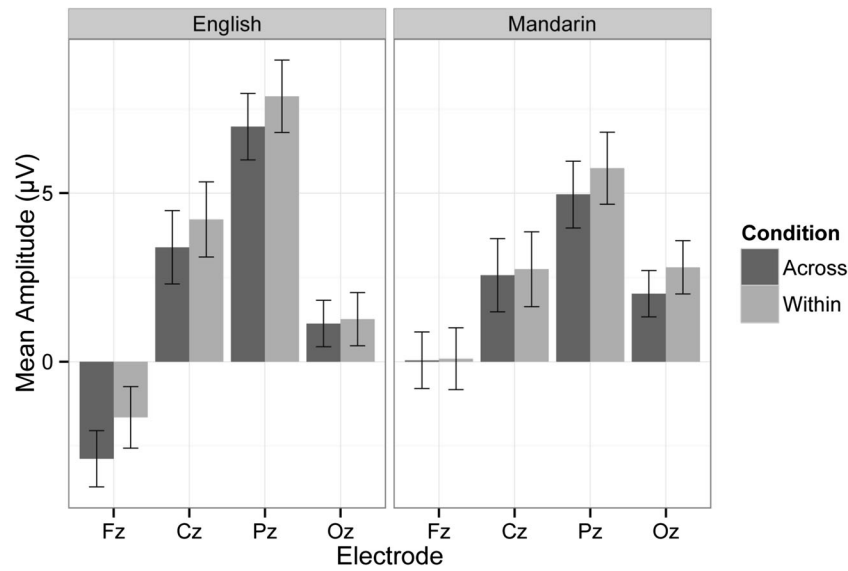
specifically for the across-category deviant (cf. Xi et al., 2010; Zhang et al., 2012). Critically, hemispheric differences were not present over parietal electrodes. Note also that there was no interaction between tone, hemisphere and L1, suggesting similar P3 brain responses in the two groups of listeners.

The lack of an interaction between deviant type and hemisphere for either component is not likely to be due to a lack of statistical power. To determine this, we computed the expected

effect size for the P3 component, which had a smaller effect size than the N2b in Zhang et al. (2012). Effect sizes were estimated from the F-statistics reported by Zhang et al. using the method described by Lakens (2013). This yields a  $\eta_p^2$  of 0.37. Based on this, we performed a post-hoc power analysis using G\*Power (Faul, Erdfelder, Lang, & Buchner., 2007), which showed that the current study has a power of >0.99 for detecting this effect with the Mandarin speakers (N=20).

**Table 7** ANOVA in the 150–300 ms and 300–600 ms time window (deviants only)

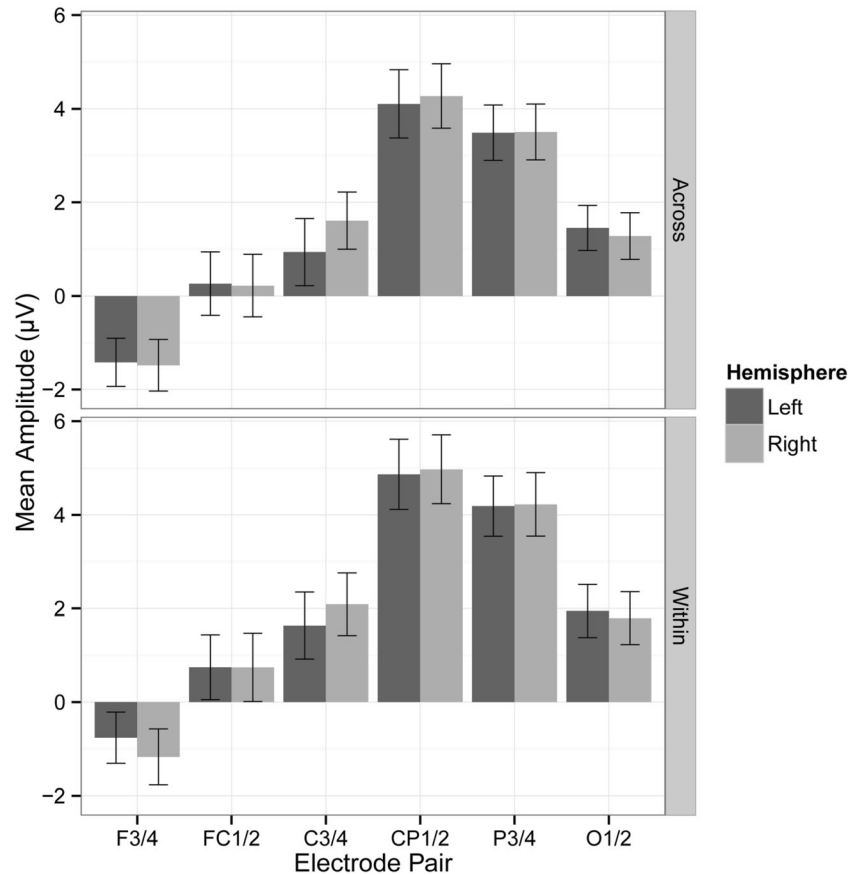
	150–300 ms				300–600 ms			
	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
<i>Midline</i>								
Tone	<b>1,38</b>	<b>16.690</b>	<b>3.418</b>	<b>&lt;.001</b>	<b>1,38</b>	<b>5.953</b>	<b>4.998</b>	<b>.019</b>
L1	1,38	.797	38.808	.378	1,38	.005	108.3	.945
Tone × L1	1,38	.046	3.418	.832	1,38	.424	4.998	.519
Tone × elec	3,114	1.817	.388	.175	3,114	.917	.637	.376
Tone × elec × L1	3,114	2.261	.388	.119	<b>3,114</b>	<b>4.827</b>	<b>.637</b>	<b>.021</b>
<i>Lateral</i>								
Tone	<b>1,38</b>	<b>15.344</b>	<b>8.644</b>	<b>&lt;.001</b>	<b>1,38</b>	<b>5.453</b>	<b>15.182</b>	<b>.025</b>
L1	1,38	.199	113.8	.658	1,38	.003	287.0	.959
Tone × L1	1,38	.044	8.644	.836	1,38	1.430	15.182	.239
Tone × hemi	1,38	.153	.394	.698	1,38	1.300	.377	.261
Tone × hemi × L1	1,38	1.134	.394	.294	1,38	.155	.377	.696
Tone × elec	5,190	1.758	.408	.180	5,190	.615	.787	.502
Tone × elec × L1	5,190	1.591	.408	.211	<b>5,190</b>	<b>4.522</b>	<b>.787</b>	<b>.023</b>
Tone × hemi × elec	5,190	1.142	.061	.334	<b>5,190</b>	<b>2.887</b>	<b>.086</b>	<b>.035</b>
Tone × hemi × elec × L1	5,190	1.109	.061	.347	5,190	1.248	.086	.295



**Fig. 4** Tone × electrode × L1 interaction in P3 over midline electrodes (deviants only; error bars show standard error of the mean)

Summarizing, the P3 effects were qualitatively similar over parietal electrodes (where the P3 is measured) across the two L1 groups, suggesting that native Mandarin and English speakers process the oddball stimuli similarly. Within-category deviants showed larger P3 amplitudes than

across-category deviants in both L1 groups, counter to the predictions of a categorical perception account. Moreover, no left hemisphere bias was found for either the N2b or P3 in the across-category oddball relative to the within-category oddball. Again, this result does not fit with the



**Fig. 5** Tone × hemisphere × electrode interaction for P3 amplitude (averaged over L1 group) over lateral electrodes (deviants only; error bars show standard error of the mean)

categorical perception account presented by Zhang et al. (2012).

## Discussion

The current study aimed to investigate whether perception of Mandarin tones is categorical and dependent on language experience and phonological knowledge. Based on previous ERP evidence that has supported categorical perception of tones, we expected native Mandarin listeners in this experiment to show the same pattern of results as Zhang et al. (2012), which found left-dominant P3 responses to the across-category deviant and symmetrically-distributed N2b responses to the within-category deviant. Such a result could be consistent with categorical perception of tones. However, we failed to replicate this result. We found that Mandarin speakers did not show any topographic differences in the N2b/P3 components between the two deviants. In addition, the within-category deviant elicited larger P3 amplitudes relative to the across-category deviant. This contrasts with the predictions of the categorical perception account, which holds that across-category oddball tones should be easier to categorize, resulting in larger amplitude P3 effects. Moreover, ERP results from the naïve native English-speaking listeners were qualitatively similar to the Mandarin speakers', suggesting this effect was not modulated by language experience.

The current ERP results indicate that both groups of listeners are sensitive to within- and between-category differences in tones, arguing against the idea that tone perception is categorical. In a similar task, P3 amplitude has been shown to be affected by distance between a category boundary and target stimulus, with larger a P3 elicited by stimuli that are further away from phonological boundary (Toscano et al., 2010). Therefore, if listeners' responses in this task were derived from underlying phonological categories, we would expect Mandarin speakers to demonstrate a larger P3 to the across-category deviant and a smaller P3 to the within-category deviant. However, this is not what was found. P3 amplitude was larger to the within-category deviant, again contrasting with predictions made by the categorical perception account. Also, assuming categorical perception, Mandarin speakers' P3 response should have been qualitatively different from naïve non-native English speakers, reflecting learned phonological processing. In contrast, we found larger P3 in response to the within-category deviant in both Mandarin and English speakers, with no clear qualitative differences between the groups. This suggests that the relative ease of categorization and greater saliency of the within-category deviant was not a product of specific learned phonological categories. Alternatively, it may be due to differences in the direction of the slope of the  $f_0$  contour, which is positive for the across-category tone, but negative for the standard tone and the

within-category deviant tone. It is also possible that English listeners map the tones onto existing English intonation categories, though it is unclear how these categories differ from Mandarin lexical tones.

Additionally, prior work using electrophysiological evidence to argue for categorical perception of Mandarin tone used topographic differences in brain responses elicited by within- versus across-category deviants. Studies using both the pre-attentive MMN and attention-related N2/P3 components<sup>6</sup> showed a left-hemisphere bias for across-category oddballs, which Xi et al. (2012) took to reflect left-hemisphere-based phonological processing; no hemispheric bias was found for within-category oddballs, which the authors took to reflect general acoustic processing (Xi et al., 2010; Zhang et al., 2012). We found no such hemispheric bias for the N2b/P3 complex for either the within-category or across-category oddballs, using the exact same stimuli as those used by both Xi et al. and Zhang et al.

Note also that the inference from topography to source generator made by prior authors is problematic, as ERP scalp topographies provide no information about the underlying location or configuration of source generators (see Luck, 2014). Indeed, depending on the orientation of the dipole generator of an effect, "paradoxical" scalp topographies may be present, where left hemisphere generators can give rise to right hemisphere-dominant effects (see e.g., Kutas, Van Petten, & Besson, 1988, for discussion of paradoxical topographies for other language-related effects). Thus, even to the extent that the prior work may replicate under other circumstances, the conclusion that a left hemisphere-dominant effect reflects a processing generator in the left hemisphere (and by extension, a linguistic process), may not be warranted.

There are several possible explanations for the discrepancy between our ERP findings and those from Zhang et al. (2012), including different sample sizes and the use of different methodologies in EEG analysis. First, the current study included a larger sample than Zhang et al., allowing us to better estimate population parameters and reduce the chances of spurious effects based on sampling. Methodologically, the Zhang et al. study used a common average reference for ERP analysis, in which data were re-referenced offline to the average of all the electrodes. Common average reference tends to minimize noise when recording from a large set of electrodes. However, it might change the waveforms and the scalp

<sup>6</sup> The MMN (i.e., N2a) and N2b have similar latencies and are generated by similar experimental conditions, though the N2b requires that the subject attends to the stimulus, whereas the MMN does not. Given this, responses to auditory oddballs in the current study may reflect a mixture of the MMN and N2b components (as well as the N2c; Folstein & Van Petten, 2008). However, the critical result (i.e., similar responses for both Mandarin and English speakers) is not dependent on which sub-component of the N2 is driving the effect, so our interpretation of the results remains the same whether the N2 differences are characterized as being generated by the MMN, the N2b, or some combination of the two components.

distribution depending on the recording electrodes, and also, it can distort the polarity of peaks, leading to misinterpretations (Luck, 2014). In contrast, we used an averaged mastoid reference, which is commonly used in other studies using the oddball paradigm, making our results more generalizable to the broader N2b/P3 literature.

In addition, Zhang et al. applied a 0.3-Hz high-pass filter to their data, which could potentially impact ERP components such as the N2 and P3 (Tanner et al., 2015; see also Acunzo, MacKenzie, & van Rossum, 2012). Particularly, it has been shown that using high-pass filter cutoffs around 0.3 Hz can produce “inverted spreading,” reducing true effects and inducing new artifactual peaks into the waveform. Moreover, inappropriate filtering like this alters the scalp topographies of effects of interest. Therefore, it is possible that the left-dominant effect found in the previous study is an artifact induced by filtering. The current study overcame these issues by using a lower cutoff value for the high-pass filter, which has been shown to maximize statistical power and minimize the potential for spurious artifacts for P3-like effects (Tanner et al., 2015).

In contrast to the ERP data, our behavioral results showed clear cross-linguistic differences. Specifically, Mandarin listeners’ behavioral data reflected effects of phonological category structure (e.g., higher accuracy for across- than for within-category oddballs). This suggests a discrepancy between ERP and behavioral responses for Mandarin speakers: they showed a greater P3 amplitude for the within-category oddball, yet their behavioral results indicated that they were more accurate for the across-category than the within-category tone. However, it is worth noting that the proportion of correct responses for the two oddballs were both very high, indicating that native Mandarin speakers could easily discriminate both oddball stimuli from the standard tone. We suggest that for native Mandarin speakers, the earlier latency of the P3 peak relative to the mean RT suggests that additional decision-making and motor control processes may come into play between initial perception of a speech sound and the eventual behavioral response. Since all three stimuli are meaningful lexical items in Mandarin, the effect of categorization might occur downstream and influence behavioral responses, but not occur at a purely perceptual level. On the other hand, naïve English speakers would not recognize these stimuli as meaningful lexical items, and therefore, there was no effect of lexicalization on their behavioral results. This speculation is borne out in the data, as the English participants’ behavioral results (both accuracy and RT) mimicked their ERP results, showing both a larger P3 and greater accuracy/faster RT for the within-category deviant.

Thus, Mandarin listeners’ behavioral results are affected by phonological categories and/or lexical factors (i.e., the tone categories in Mandarin), but these category-based effects are not due to categorical perception of the tones. It remains

unclear whether knowledge of phonological categories does not influence low-level auditory perception at all or whether it simply does not result in categorical effects. Note also that this does not mean that Mandarin does not contain information about tone categories in the statistical distributions of  $f_0$  contours (if it did not,  $f_0$  contour would not provide a good cue to lexical tone). Rather, the results suggest that these statistical distributions do not cause low-level auditory perception to operate in a categorical way. Thus, Mandarin listeners’ behavioral responses are influenced by tone categories, even though the underlying perceptual response are continuous.

Overall, our electrophysiological results suggest that lexical tone perception among native Mandarin speakers is not categorical. There were no hemispheric differences between within- and across-category oddball tones in an attentional categorization task. Moreover, our ERP results suggest that within-category oddballs were more distinct from the standard tone during categorization, as reflected by larger P3 amplitudes compared to across-category oddballs, in contrast to predictions made by the categorical perception account. Additionally, there were no meaningful differences in ERP responses between experienced native Mandarin speakers in the categorization task and naïve native English speakers, suggesting that language experience played a minimal role in shaping these early categorization-related neural responses. Category-based effects emerged in behavioral measures for native Mandarin speakers (e.g., faster RT and higher accuracy for the across-category oddball), which, we argue, reflect later-occurring lexical or phonological processing factors.

**Acknowledgements** We would like to thank Jessica Philipp, Justin Brook, and Amanda Kim for assistance with data collection. We would also like to thank Jerome Packard for helpful discussion in the early stages of this project. This project was part of Yang Agnes Gao’s undergraduate honors thesis in the Department of Linguistics at the University of Illinois at Urbana-Champaign. This work was partially supported by NSF BCS-1431324 to DT.

## References

- Acunzo, D. J., MacKenzie, G., & van Rossum, M. C. W. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. *Journal of Neuroscience Methods*, *209*, 212–218.
- Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, *8*(4), 919–924.
- 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*(1), 9–21.
- Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.
- Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, *45*(1), 152–170.

- Gandour, J. T. (1978). The perception of tone. *Tone: A Linguistic Survey*, pp. 41-76.
- Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218-1227.
- Jasper, H. (1958). Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*, *10*, 370-375.
- Joanisse, M. F., Robertson, E. K., & Newman, R. L. (2007). Mismatch negativity reflects sensory and phonetic speech processing. *NeuroReport*, *18*(9), 901-905.
- Laken, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.
- Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *J Stat Softw*, *69*(1), 1-33.
- Lieberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, 497-516.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.
- Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213.
- Luck, S. J. (2014). *An Introduction to the Event-related Potential Technique*. MIT press.
- Massaro, D.W., & Cohen, M.M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*, 15–35
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33-B42.
- Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, *50*(1–3), 271–285.
- Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*(1), 1-21.
- Peng, G., Zheng, H. Y., Gong, T., Yang, R. X., Kong, J. P., & Wang, W. S. Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, *38*(4), 616-624.
- Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. *Journal of Cognitive Neuroscience*, *12*, 1038–1055.
- Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(1), 129.
- Sharma, A., Kraus, N., McGee, T., Carrell, T., & Nicol, T. (1993). Acoustic versus phonetic representation of speech as reflected by the mismatch negativity event-related potential. *Electroencephalography and Clinical Neurophysiology*, *88*, 64–71.
- Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*, 1532-40.
- Woldorff, M. G., Hackley, S. A., & Hillyard, S. A. (1991). The effects of channel-selective attention on the mismatch negativity wave elicited by deviant tones. *Psychophysiology*, *28*(1), 30-42.
- Xi, J., Zhang, L., Shu, H., Zhang, Y., & Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience*, *170*(1), 223-231.
- Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063-1074.
- Yip, M. (2002). *Tone*. Cambridge University Press.
- Zentner, M. & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1440*, 33-45.
- Zhang, L., Xi, J., Wu, H., Shu, H., & Li, P. (2012). Electrophysiological evidence of categorical perception of Chinese lexical tones in attentive condition. *NeuroReport*, *23*(1), 35-9.