https://dx.doi.org/10.1016/j.neuropsychologia.2021.107839

# Integral perception, but separate processing: The perceptual normalization of lexical tones and vowels

Kaile Zhang[a], Matthias J. Sjerps[b], Gang Peng[a,c]

[a]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR

[b]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Radboud University, Kapittelweg 29, Nijmegen 6525 EN, The Netherlands

[c]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Boulevard, Shenzhen, 518055, China

kaile.k.zhang@connect.polyu.hk, m.j.sjerps@gmail.com, gpeng@polyu.edu.hk

**Correspondence:**
Gang Peng
AG509
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong SAR
gpeng@polyu.edu.hk

## Abstract

In tonal languages, speech variability arises in both lexical tone (i.e., suprasegmentally) and vowel quality (segmentally). Listeners can use surrounding speech context to overcome variability in both speech cues, a process known as extrinsic normalization. Although vowels are the main carriers of tones, it is still unknown whether the combined percept (lexical tone and vowel quality) is normalized integrally or in partly separate processes. Here we used electroencephalography (EEG) to investigate the time course of lexical tone normalization and vowel normalization to answer this question. Cantonese adults listened to synthesized three-syllable stimuli in which the identity of a target syllable — ambiguous between high vs. mid-tone (Tone condition) or between /o/ vs. /u/ (Vowel condition) — was dependent on either the tone range (Tone condition) or the formant range (Vowel condition) of the first two syllables. It was observed that the ambiguous tone was more often interpreted as a high-level tone when the context had a relatively low pitch than when it had a high pitch (Tone condition). Similarly, the ambiguous vowel was more often interpreted as /o/ when the context had a relatively low formant range than when it had a relatively high formant range (Vowel condition). These findings show the typical pattern of extrinsic tone and vowel normalization. Importantly, the EEG results of participants showing the contrastive normalization effect demonstrated that the effects of vowel normalization could already be observed within the N2 time window (190-350 ms), while the first reliable effect of lexical tone normalization on cortical processing was observable only from the P3 time window (220-500 ms) onwards. The ERP patterns demonstrate that the contrastive perceptual normalization of lexical tones and that of vowels occur at least in partially separate time windows. This suggests that the extrinsic normalization can operate at the level of phonemes and tonemes separately instead of operating on the whole syllable at once.

Keywords: Perceptual normalization, lexical tones, vowels, ERPs

# 1. Introduction

Natural speech sounds display a considerable amount of acoustic-phonetic variability. The lack of one-to-one mapping between speech sounds and meaningful units in language poses a fundamental challenge to the accuracy and speed of speech perception. It has been shown that listeners rely on the surrounding context to interpret speech sounds, especially when the target sound is ambiguous. This process is known as “extrinsic normalization”. Although a small number of studies reported an assimilative normalization effect (e.g., Rysling et al., 2019), typically, context affects the perception of target sounds in a contrastive way. That is, in the case of vowels, listeners tend to give low vowel responses [vowels with high first formants (F1)] if the context has a relatively low F1. For example, a vowel that is ambiguous between /ɛ/ and /ɪ/ is more often perceived as the vowel /ɪ/ if the preceding sentence has a relatively high F1, and it is more frequently reported as the vowel /ɛ/ when its precursor sentence has a relatively low F1 (Ladefoged & Broadbent, 1957; see Stilp, 2019, for a review). A similar pattern has been observed in the perception of lexical tones. A Cantonese mid-level tone is more often perceived as a Cantonese low-level tone when the fundamental frequency (F0) of the preceding context is relatively high, but when the context F0 is lowered three semitones, the identical mid-level tone is typically perceived as a high-level tone (K. Zhang et al., 2017). The perception of the Mandarin level tone and rising tone is also contrastively affected by contextual cues, with high F0 contexts for more rising tone responses than low F0 contexts (J. Huang & Holt, 2009; Luo & Ashmore, 2014).

This raises the question of whether the extrinsic normalization of different speech cues involves a single process operating on the speech sound as a whole, or instead involves two different processes operating on the separate speech cues. Holistic normalization is in line with the Reverse Assessing Model which claims that the perception of Chinese spoken words takes the whole syllable as the process unit and that only information at the syllable level and above is active for immediate use (Gao et al., 2019). The better syllabic awareness than phonemic awareness of Chinese speakers supports holistic normalization (Read et al., 1986; Shu et al., 2008). However, the highly integrated tone and vowel information at syllable level largely constrains listeners’ response to the individual cue (F0 or F1) and consequently reduces their independent contribution. This is not a problem for separate normalization. Separate normalization is theoretically closer to the TRACE model (McClelland & Elman, 1986), which posits that speech perception starts from acoustic features and each activated feature combines to activate speech units in higher levels. On the one hand, normalization operating on the individual acoustic cue is more sensitive to the change on each dimension and thus yields better normalization results, but on the other hand the separable normalization requires multiple normalization processes, which is cognitive resource-consuming. This seems to be in conflict with our intuition of effortless speech perception. Holistic/separate normalization will be further investigated in the present study by comparing the time courses of lexical tone and vowel normalizations.

### 1.1. The neurological process underlying the extrinsic normalizations of lexical tones and vowels

Several EEG studies probe the online process of perceptual normalization at either segmental or suprasegmental level. For example, Sjerps et al. (2011) used EEG to investigate the time course of extrinsic vowel normalization in an active multiple-deviant oddball paradigm. They observed that the normalization process affected the electrophysiological response from the N1 time window (80-160 ms) onwards. Considering that N1 is generally associated with the acoustic processing of sound (Näätänen & Winkler, 1999), Sjerps et al. (2011) support the notion that vowel normalization is accomplished at the acoustic stage. A recent intracranial electrocorticographic (ECoG) study provided further evidence for this finding. Sjerps et al. (2019) asked Spanish neurosurgical patients to identify a similar set of ambiguous vowels in either a high or low F1 context. The authors observed specific patches of cortex that were responsive to specific vowels such as /u/ or /o/. Interestingly, the responsiveness of these patches of cortex to vowels was modulated by the contexts' F1s. Additional analyses indicated that the affected neural populations demonstrated a tuning preference for F1 information, regardless of phonemes, indicating that contexts affected acoustic-phonetic (i.e., pre-phonemic) representations of speech on the parabelt auditory cortex.

Another way to characterize the level of context normalization process is to compare the strength of context effects induced by speech and nonspeech sounds (C. Zhang et al., 2013; K. Zhang & Peng, 2018). For example, it has been demonstrated that a nonspeech context had little effect on vowel normalization, despite having the same long term average spectrum as a speech context that did induce strong normalization effects (K. Zhang et al., 2018). K. Zhang and Peng (2018) compared perceptual normalization induced by speech and nonspeech contexts and observed the first speech-nonspeech context difference in the P2 time window (130-250 ms). This event-related potential (ERP) result indicated that the cortical effect of the extrinsic vowel normalization was most dominant in the phonological processing, in which the acoustic-phonetic information was mapped onto the language-specific phonological representations. Sjerps et al. (2011, 2019) and K. Zhang and Peng (2018) gave rise to inconsistent results about the exact time course of vowel normalization, although their results overlap with each other to some degree (50-200 ms vs. 130-250 ms).

To our knowledge, only two studies have tested the time course of lexical tone normalization, reporting different findings. C. Zhang et al. (2013) also utilized the unequal context effects of speech and nonspeech sounds to test the lexical tone normalization process. They reported that extrinsic tone normalization happened in the N400 time window (250 ms-500 ms), a time window typically related to the lexical retrieval process. Shao and C. Zhang (2019) compared lexical tone perception in a blocked-talker condition with that in a mixed-talker condition. They observed differences in the N1 time window for the two conditions. However, the N1 difference in Shao and C. Zhang (2019) may reflect the detection of changes in talker but not necessarily the extrinsic normalization process. In sum, there has been a growing interest in outlining the time courses of normalization processes of tone and vowel quality, but so far this has led to partly inconsistent results. This inconsistency in turn makes it hard to conclude whether normalization of segments and suprasegments is most dominant in the same or different stages of the cortical processing hierarchy.

### 1.2. Lexical tone and vowel perception in general

While only a small number of studies explicitly investigate the online perceptual normalization process, a considerable number of studies focus on the perception of lexical tones and vowels independent of contextual influences. There is an ongoing debate as to whether lexical tones and vowels are processed as an integrative unit or as two independent units. Some studies observed partly independent processing for tone and vowel quality. Regarding the time course, lexical tones are reported to be identified later than vowels. Stimuli containing a vowel mismatch were recognized faster than those containing a lexical tone mismatch in a monitoring task (Ye & Connine, 1999). In a Chinese sentence comprehension task, target words containing rime violations (e.g., 观赚/kuan1 tʂuan4/ “look earn”) elicited N400 and P600 of larger amplitudes compared with the congruent condition (e.g., 观众/kuan1 tʂuŋ4/ “audience”), while lexical tone violations (e.g., 观肿/kuan1 tʂuŋ3/ “look swelling”) only elicited larger P600, indicating a comparatively later detection of lexical tone violations (Zou et al., 2020). Similar ERP patterns were also observed in the perception of Chinese idioms (Hu et al., 2012; X. Huang et al., 2018) and Chinese classic poems (Li et al., 2014). The brain areas involved in lexical tone perception and vowel perception also show some separation. Using a picture-word mismatch paradigm, Malins and Joanisse (2012) found that the phonological mismatch negatives triggered by tonal mismatch (e.g., picture: /xua1/ ‘flower’; sound: /xua4/

‘painting’) and cohort mismatch (e.g., picture: /xua1/ ‘flower’; sound: /xuei1/ ‘gray’) showed different cortical distributions. The meta-analysis in Liang and Du (2018) revealed that the right auditory cortex exhibited stronger activation for lexical tones than the left auditory cortex did, but the reverse was observed for phoneme process.

There is, however, considerable evidence for integrative process of vowels and lexical tones as well. Repp and Lin (1990) showed that even though listeners were told to focus on only one dimension (either lexical tones or vowels) of CV syllables and to ignore the other, their perception of the focused dimension was still significantly interfered with by the ignored dimension. Choi et al. (2017) found that mismatch negativity elicited in vowel perception and in tone perception showed no significant difference in latency and topography. Furthermore, Brown-Schmidt and Canseco-Gonzalez (2004) and Schirmer et al. (2005) also used the violation paradigm to test the access of lexical tones and vowels. They found that lexical tone violations and rime violations generated similar N400 effects, and this occurred in both a Mandarin sentence perception task and a Cantonese sentence perception task, suggesting simultaneous access to these cues. Zhao et al. (2011) observed a similar N400 pattern for lexical tone mismatch, rime mismatch, and onset mismatch as well. More importantly, syllable mismatches elicited an earlier and stronger N400 than a partial mismatch in either dimension. Therefore, Zhao et al. (2011) argued that Chinese monosyllabic word identification probably involved a syllable-based process rather than a phonemic segment-based process. If lexical tones and vowels (or even the whole syllable) are integrated to form the dominant processing unit in speech perception, it is reasonable to assume that they are normalized together as well.

Liu and Samuel (2007) and Ye and Connine (1999) suggested that the inconsistent results reviewed above were mainly caused by the stimulus presentation. If the stimuli are presented in isolation, listeners will more likely pay attention to the sub-lexical contrast. In such a condition, the processing difference between lexical tones and vowels can be detected. On the contrary, if the tasks use highly constraining contexts that can easily evoke a top-down influence from the lexical level process to the sub-lexical level process, the processing difference between lexical tones and vowels is eliminated. However, Zou et al. (2020) found that even when the target stimuli were presented in sentences that introduced strong semantic contexts, ERP patterns emerged in vowel perception and lexical tone perception still differed a lot. Therefore, the question about holistic/separate process of lexical tones and vowels is still open to be tested.

### 1.3 The present study

It is important to clarify whether segmental and suprasegmental components are normalized integrally or whether, instead, they are normalized independently in the different time windows over which these cues are often observed to be processed. This is related to a more fundamental question concerning the dominant unit of speech normalization. Previous studies (Sjerps et al., 2011, 2019; C. Zhang et al., 2013; K. Zhang & Peng, 2018) reported different time courses for vowel normalization and lexical tone normalization. However, these incongruences can be caused by differences in the stimuli, participants, and experimental paradigms used to test the context effects. To overcome this problem, the present study tested lexical tone normalization and vowel normalization in a matched design. The same group of native Cantonese speakers was asked to identify lexical tones (/wo55/ vs. /wo33/) or vowels (/wo55/ vs. /wu55/) in either a high F0/F1 context or a low F0/F1 context while EEG signals were recorded. The stimuli used in the tone and vowel normalization tasks were recorded by the same speaker and were subjected to similar manipulations. Both the lexical tone normalization and the vowel normalization employed the word identification task and similar data analysis methods. All these strategies were used in order to minimize the experimental design differences between the lexical tone normalization and the vowel normalization. The context effect in the present study was detected by observing how the context F0/F1 manipulated the perception of the targets (Sjerps et al., 2019). The EEG data analysis was time-locked to the target onset. The ERP elicited in the high context condition was compared with that in the low context condition to see in which time window(s) the perception of the acoustically identical targets started to differ from each other.

Three ERP components, N1, P2, and N400, which were reported to index the normalization process, were investigated in the present study. Based on previous studies, it was hypothesized that lexical tones and vowels were normalized partially independently, and that vowel normalization probably occurs in the N1 (Sjerps et al., 2011) and/or the P2 time window(s) (K. Zhang & Peng, 2018) but lexical tone normalization occurs in the N400 time window (C. Zhang et al., 2013). Vowel perception was expected to trigger N1 and/or P2 with different amplitudes in high and low contexts, and lexical tone perception was expected to trigger N400 with different amplitudes in two contexts. Furthermore, since the present study used the Go/NoGo paradigm (see Section 2.3.2 for details), N2 and P3, two typical ERPs in the Go/NoGo paradigm, were also expected to emerge. The amplitudes and the latencies of N2 and P3 change along with the response times and the task complexities (Gajewski & Falkenstein, 2013; Jodo & Kayama, 1992). Therefore, vowel and lexical tone normalizations were hypothesized to affect N2 and P3 differently. Specifically, vowel normalization that was expected to happen earlier would trigger N2 and/or P3 with earlier latencies as well.

## 2. Materials and methods

### 2.1. Participants

Thirty-two native Hong Kong Cantonese speakers were recruited from The Hong Kong Polytechnic University. The behavioral results reported in the present study were based on 32 subjects' data. However, the EEG results were based on the data from 20 subjects, and the reasons for exclusion are described in Section 2.4.2. All participants were right-handed and were identified using the Edinburgh handedness inventory (Oldfield, 1971). They had no self-reported hearing impairment, speech and language-related disabilities, or brain injuries. None of them had received formal training in music, linguistics, or psychology. They were given a monetary reward for their participation. The experimental procedures were approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. Informed written consent was obtained from every participant before the experiment.

### 2.2. Stimuli

The stimuli used in this experiment were recorded by a native male Cantonese speaker in a soundproof booth. He was asked to clearly and naturally read the word list, which contained /wu55/ (烏, black), /wo55/ (窩, nest), /wo33/ (喎, a modal word), /p$^h$a21/ (琶, guitar), and /tsi25/ (紫, purple), thirty times. The /wo33/-/wo55/ pair was used to generate the targets in the lexical tone normalization and the /wu55/-/wo55/ pair was used to generate the targets in the vowel normalization. The words /p$^h$a21/ (琶, guitar) and /tsi25/ (紫, purple) are both meaningful in Cantonese but their combination is meaningless. The pseudo-phrase /p$^h$a21 tsi25/ was used as the context in the lexical tone normalization and the vowel normalization since it covered a speaker's full F0 range (the lowest pitch in T21 and the highest pitch in T25) and the full F1 range (the lowest F1 in /i/ and the highest F1 in /a/). The identical context made the experimental paradigms for the lexical tone normalization and the vowel normalization more comparable. The meaningless context minimized syntactic or semantic biases.

The stimuli manipulation largely followed the procedure used by Sjerps et al. (2018). The parameters used to synthesize the lexical tone continuum, the vowel continuum, the lexical tone contexts, and the vowel contexts are visualized in Figure 1. The 17-step lexical tone continuum /wo33-wo55/ and the 17-step vowel continuum /wu55-wo55/ were synthesized by interpolation in Praat (Boersma & Weenink, 2016). To improve the naturalness, the high-frequency ranges (above 2,000 Hz), the amplitude envelopes, and the overall amplitudes of the synthesized stimuli resembled the original recordings. A representative /p$^h$a21/ was chosen based on its clarity. Its pitch contour was shifted 20 Hz up and down to form the high and low F0 contexts for the lexical tone normalization. Its F1 was shifted 100 Hz up and down to generate the high and low F1 contexts for the vowel normalization. The same processes were done to the syllable /tsi25/. These two syllables were concatenated to form the final contexts. A neutral context whose F1 and F0 were not modified was also generated for the categorical perception task. The durations of targets were normalized to 400 ms and the durations of contexts were normalized to 600 ms.

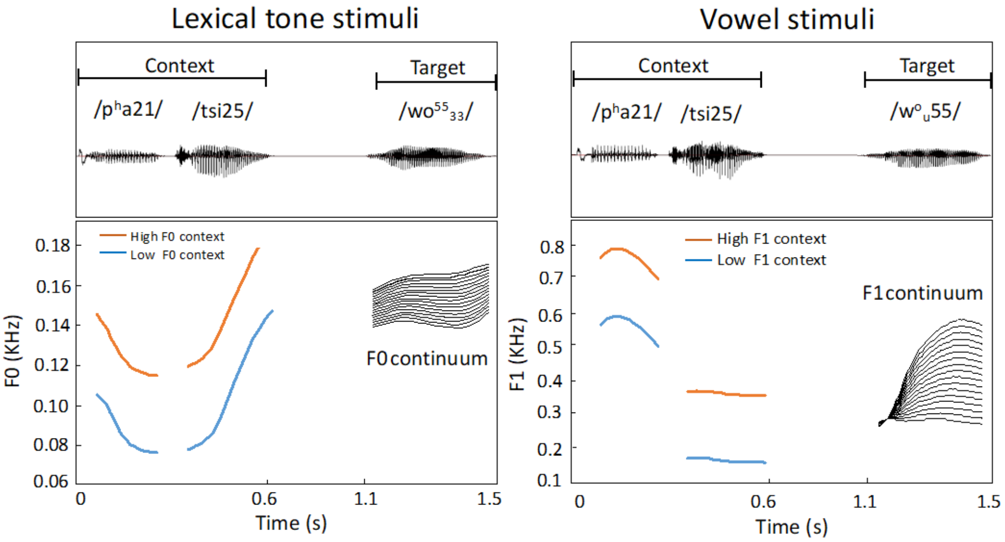


Figure 1. The parameters used to synthesize the stimuli for the lexical tone normalization (left) and the vowel normalization (right).

## 2.3. The experimental procedure

The context effect is more noticeable for the identification of ambiguous utterances near categorical boundaries, but it is weak for the perception of typical tokens (K. Zhang et al., 2018). The categorical boundaries vary from person to person. For example, some people need longer voice onset time to perceive a sound as /p$^h$/ instead of /p/ than others do. Therefore, a short categorical perception task was carried out first to identify the most ambiguous target within the 17-step continuum for each participant. Then, word identification tasks with participant-specific stimuli sets were used to evaluate the participants' extrinsic normalization. The EEG signals were recorded during the word identification tasks.

### 2.3.1. The categorical perception task

The task was carried out in Praat (Boersma & Weenink, 2016). The targets were tokens with odd numbers in the 17-step vowel or tone continuum (i.e., Step 1, Step 3…Step 17) and the context was /p$^h$a21 tsi25/ with neutral F0 and F1. In each trial, the context was played first and then a target was played after a 500 ms silence. A window with two choices (窩/wo55/ and 喎/wo33/ in the lexical tone categorical perception task; 窩/wo55/ and 烏/wu55/ in the vowel categorical perception task) was shown on the screen after the audio stimuli. Participants were asked to click a button to indicate their choice. After their response, the next trial played automatically. The sounds were played in random order with five repetitions. The target stimulus nearest to the 50% identification curve was regarded as the most ambiguous token.

### 2.3.2. The word identification task

The extrinsic normalization of lexical tones and vowels was tested using a word identification task. In each block, the context was kept constant (a blocked design). There were four blocks in total, which were presented in a pseudo-counter-balanced order across participants. Each block consisted of five types of targets: the two endpoints of the tone/vowel continuum (Step 1 and Step 17; 10 repetitions each), the most ambiguous target chosen in the categorical perception task (Step X; 60 repetitions), the stimulus before Step X in the continuum (Step X-1; 20 repetitions), and the stimulus after Step X in the continuum (Step X+1; 20 repetitions). The 120 trials were presented in random order. In each trial, a fixation (random from 400 to 600 ms) was shown on the screen first and then the context was played bilaterally to the participant via headphones. After a jittered silence (400-600 ms), the target was played. There was a 400 ms silence after the target for the purpose of EEG signal recording. A question mark was shown on the screen after the silence. Subjects were told to pay attention to any audio stimuli they heard. Once they saw the question mark, they needed to click the mouse only when they thought that the target was /wo55/, which made the word identification task resemble the Go/NoGo paradigm. The next trial was shown automatically once a response was detected or the maximum response time (1,600 ms) was reached. The present study asked listeners to make overt responses only to target /wo55/, which was different from Sjerps et al. (2019) in which listeners responded to all types of targets. The purpose of this manipulation was to make the task in the Tone and Vowel conditions identical. Participants pressed the button when the target was /wo55/ no matter whether it was the Tone condition or the Vowel condition rather than changing their response patterns when the condition changed (i.e., responding to /wo55/ and /wo33/ in the Tone condition, but responding to /wo55/ and /wu55/ in the Vowel condition).

### 2.3.3. The EEG signal recording

The EEG signals were recorded via a SynAmps 2 amplifier (NeuroScan, Charlotte, NC, U.S.A.) with a cap carrying 64 Ag/AgCI electrodes placed on the scalp at specific locations according to the extended international 10–20 system. The bipolar channels above and below the left eye monitored eyeblinks and the bipolar channels placed laterally to the outer canthus of each eye monitored horizontal eye movements. Two separate electrodes placed on each mastoid were used as the offline references. The impedance between the online reference electrode (located between Cz and CPz) and any recording electrodes was kept below 5 kΩ. Alternating current signals with a band frequency from 0.05 Hz to 400 Hz were continuously recorded and digitized with a 24-bit resolution at a sampling rate of 1,000 Hz.

## 2.4. Data analysis

### 2.4.1. Behavioral data analysis

The generalized linear mixed-effect models were fitted to participants' responses in the word identification task to statistically assess whether their perception was affected by the context F1/F0. The statistical modeling was implemented with the lme4 package (Bates et al., 2015) in R (Version 3.6.0). The logit linking function was used for the dichotomous dependent variable — response (i.e., /wo55/ vs. /wo33/ in the Tone condition and /wo55/ vs. /wu55/ in the Vowel condition). Subjects' high tone or high vowel responses (i.e., /wo55/) were coded as 1 and their low tone or low vowel responses (i.e., /wo33/ or /wu55/) were coded as 0. All predictors were centered around zero. The predictors in the models were *cues,* reflecting whether the targets to be normalized were vowels or lexical tones (vowels coded as -1 vs. lexical tones coded as 1), *shifts,* reflecting the F1/F0 manipulation of the context (high F1/F0 coded as -1 vs. low F1/F0 coded as 1), and *steps*, reflecting the step of the target on the 17-step tone/vowel continuum (Step 1 coded as -2, Step X-1 coded as -1, Step X coded as 0, Step X+1 coded as 1, vs. Step 17 [/wo55/] coded as 2). If the contrastive context effect was to emerge, subjects would give more high tone responses (i.e., /wo55/) in low F0 contexts than in high F0 contexts. This would also be the case for the vowel normalization: more /wo55/ (a vowel with high F1) would be given in low F1 contexts than in high F1 contexts. Therefore, a significant main effect of *shift* was expected.

The context effect size was also calculated to give a more intuitive representation of the extent to which speech perception was affected by the context F0/F1. The context effect size was here defined as the proportions of /wo55/ responses in low F0/F1 context minus the proportions of /wo55/ responses in high F0/F1 context. The larger the difference, the more notable the contrastive context effect.

### 2.4.2. EEG data analysis

Only three ambiguous targets (i.e., Step X-1, Step X, and Step X+1) were included in the EEG data analysis since the typical tokens were less affected by the contexts. Data for the ambiguous targets were collapsed across *steps* to improve the signal-to-noise ratio. Based on the behavioral results (see Figure 3), some participants showed no contrastive context effect (instead exhibiting the assimilative context effect): three for lexical tone normalization only, seven for vowel normalization only, and one for both normalization tasks. The neural mechanism underlying the contrastive context effect was different from that of the assimilative context effect (Rysling et al., 2019; Stilp, 2019). Considering that most previous studies about extrinsic normalization found the contrastive context effect and that the contrastive context effect rather than the assimilative context effect emerged at group level in the present study, the present study chose to investigate the neural mechanisms underlying the contrastive normalization by only including participants who showed contrastive context effect in both the Tone condition and the Vowel condition into the EEG data analysis. One participant's EEG data was missing due to a technical problem. Therefore, 20 participants' EEG data were ultimately used[1].

*EEG data preprocessing*

The preprocessing of EEG data was implemented with EEGLAB (Delorme & Makeig, 2004). Continuous data were first bandpass filtered between 0.1 Hz and 30 Hz (Tanner et al., 2015). Epochs of 900 ms time-locked to the onset of the target sounds were extracted, with the first 100 ms preceding the target sounds as the reference in the baseline correction. Epochs containing artifacts of amplitudes ±100 μV, eyeblinks, or horizontal eye movements were excluded from analysis. Eyeblinks were identified automatically by the moving window peak-to-peak amplitude function, and horizontal eye movements were detected automatically by the step-like artifact detection function. The epochs were re-referenced to the average mastoids and were downsampled to 500 Hz. The mean acceptance rate was 94.35% ($N$ = 94.35/100, $SD$ = 11.34) in the low context and 94.45% ($N$ = 94.45/100, $SD$ = 8.29) in the high context for the Tone condition. The mean acceptance rate was 95.5% ($N$ = 95.5/100, $SD$ = 6.59) in the low context and 95.4% ($N$ = 95.4/100, $SD$ = 7.39) in the high context for the Vowel condition.

*Residue iteration decomposition (RIDE)*

[1] The statistical analysis based on 31 participants' EEG data can be found in the supplementary materials. The ERPs for participants showing the assimilative context effect are also provided in the supplementary materials. However, no further statistical analysis was done on participants who showed the assimilative context effect due to the small sample size.

Even with the best experiment setting control, ERP signals vary from trial to trial in amplitude and latency. The conventional method, which averages the EEG epochs belonging to the same experimental condition, inevitably broadens the time windows of ERPs and reduces their amplitudes. In the present study, the jittered context-target intervals probably caused the variations in target processing since the context effect may be weakened by longer context-target intervals. Consequently, the target perception in trials with longer intervals should be more difficult and the ERPs might show longer latencies and lower amplitudes. On the contrary, trials with shorter intervals may trigger ERPs with shorter latencies and larger amplitudes.

To compensate for potential problems caused by the jittered latencies, we used RIDE (Ouyang et al., 2015), which was effective in recovering amplitude effects that were distorted when using the conventional averaging method (Berchicci et al., 2016; Murray et al., 2019). RIDE tries to decompose the ERP signals into different ERP component clusters: the stimulus-locked cluster (i.e., S-cluster) referring to stimulus-driven process (e.g., perception and attention), the response-locked cluster (i.e., R-cluster; not compulsory) referring to response-related process (e.g., motor execution), and the central cluster (i.e., C-cluster) referring to intermediate processes between S and R (e.g., decision-making) (Ouyang et al., 2015). RIDE estimates the latency of each component in a single trial. The latency of S-cluster is locked to the stimuli onset and the latency of R-cluster is locked to the response time. C-cluster is estimated based on the self-optimized iteration scheme. After obtaining the trial-specific latency, RIDE averaged the ERPs of the same cluster to get the latency-modified component by aligning these trials to the most probable latency (i.e., the mode latency). In this study, only S-cluster and C-cluster were decomposed from the EEG signals. A time window ranging from 0 ms to 500 ms (relative to stimulus onset) was used for the decomposition of S-cluster, and a time window ranging from 100 ms to 800 ms was used to decompose C-cluster.

*The rationale of the EEG data analysis*

In the normalization tasks, the targets in both the high and low context blocks were the same for each participant. Any difference in the perception of these acoustically identical targets can only be attributed to the context effect. In the present study, the EEG signal analysis was time-locked to the onset of the target sound. By comparing the ERPs in two contexts, we could identify at which time window(s) those contexts affect target sound perception.

## 3. Results

### 3.1. Behavioral data

The most ambiguous steps varied across participants (F0: *M* = 10.38, *SD* = 1.48; F1: *M* = 7.69, *SD* = 1.62), suggesting that it was necessary to use the participant-specific stimuli for the normalization tasks. Figure 2 shows the proportions of /wo55/ responses in different conditions (i.e., high F0, low F0, high F1, and low F1), which were averaged across participants. As can be seen, participants on average gave more /wo55/ responses in the low F0/F1 conditions than in the high F0/F1 conditions, suggesting that the target perception was affected contrastively by the context information. Participants' perception of the ambiguous targets (i.e., Step X-1, Step X, and Step X+1) varied when the context F0/F1 changed. However, their perception of the unambiguous targets (i.e., Step 1 and Step 17) remained almost constant regardless of the context F0/F1. The context effect size for each participant in the Tone condition and the Vowel condition is plotted in Figure 3 to provide a more intuitive visualization. The context effect size showed large individual differences in both the Tone condition and the Vowel condition, as indicated by the dispersion of the data points in Figure 3. A few participants showed no contrastive context effects since some data points in Figure 3 were below zero.

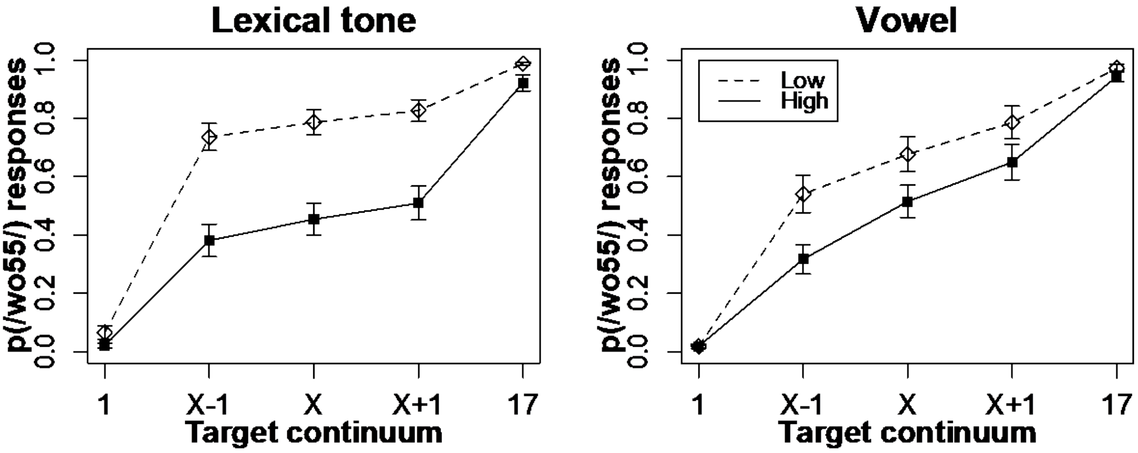


Figure 2. The overall proportion of /wo55/ response in the word identification tasks.

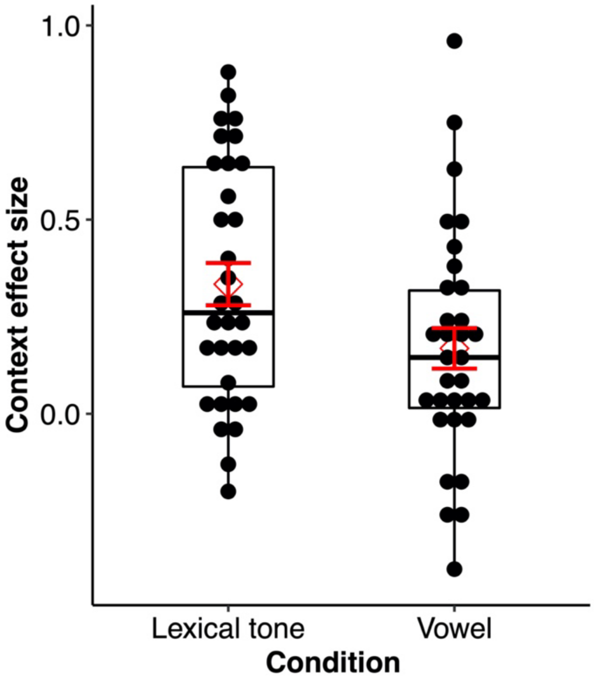


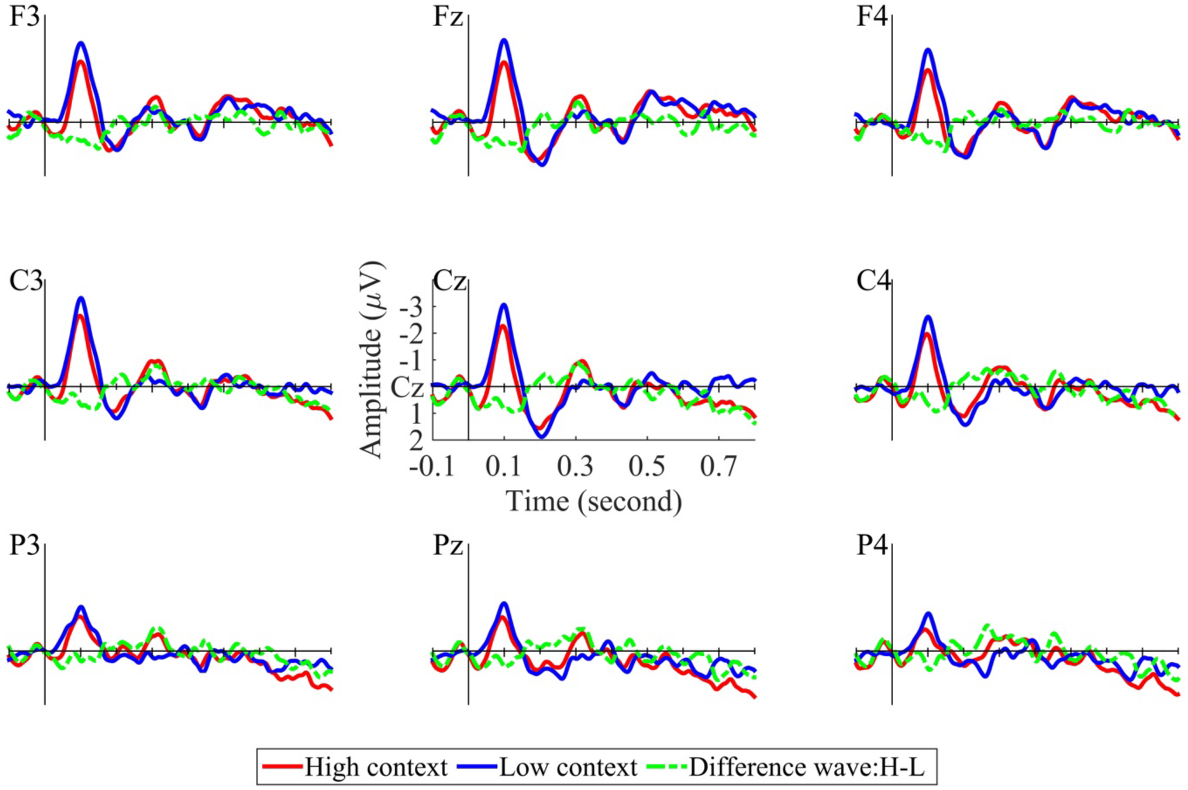
Figure 3. The context effect size in the Tone condition and the Vowel condition. Each dot represents one participant's result. The diamond is the grand average and the error bar represents the standard error of the mean.

The generalized linear mixed effect model, with fixed effects of *cues*, *steps, shifts*, and their two-way and three-way interactions, and random effects of de-correlated slopes and intercepts by *subjects* for *cues*, *steps, shifts* and their interactions, was fitted to the word identification data to statistically evaluate the patterns shown in Figures 2 and 3. The analysis revealed a main effect of *shifts* ($B$ = 1.08, $z$ = 6.27, $p$ < 0.01), suggesting that overall more /wo55/ responses were given in the low context condition than in the high context condition, and a main effect of *steps* ($B$ = 1.85, $z$ = 13.81, $p$ < 0.01), suggesting that more /wo55/ responses were observed when the targets approached the /wo55/ end of the continuum. A significant *cue* by *shift* interaction ($B$ = 0.28, $z$ = 2.34, $p$ < 0.05) was observed as the effect of context shift was larger in the perception of lexical tones than in the perception of vowels. Another significant interaction was observed between *cue* and *step* ($B$ = -0.28, $z$ = -3.83, $p$ < 0.01) as the slopes of the categorization curves were deeper in the Vowel condition.

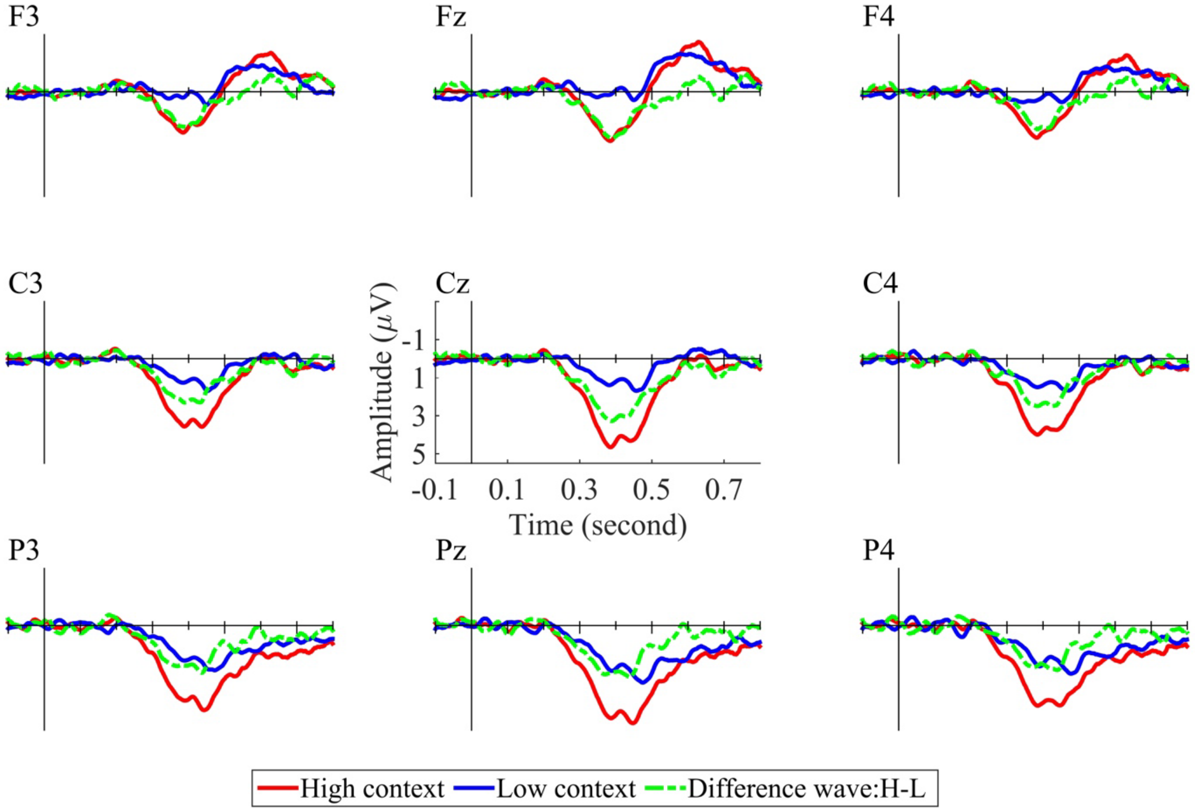
## 3.2. EEG data

### 3.2.1. ERP components

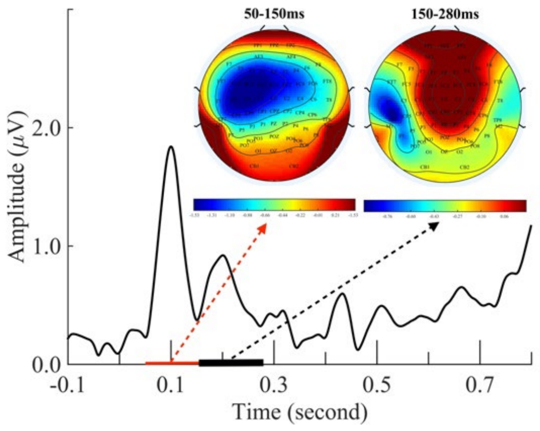
The ERP wave in each experiment condition was obtained by averaging all the accepted epochs. In the present study, participants gave overt responses in trials in which they perceived the target as /wo55/ but withheld their responses in other trials. The ERP responses to the same ambiguous stimulus may vary across trials according to the subjective perception of stimulus characteristics (Foss-feig et al., 2018). However, the present study did not further divide the accepted epochs into two types based on the presence/absence of overt responses since the reliable ERPs could hardly be obtained in conditions that had only a few trials (e.g., only 15 no-response trials in the low context block for the Tone condition). Figures 4 (a) and (b) and Figures 5 (a) and (b) show the ERP waves in different conditions that were averaged across 20 subjects. As stated in Section 1.3, N1, P2, and N400, which were reported to index the normalization process, and N2 and P3, two typical ERPs in the Go/NoGo paradigm, would be examined to see at which time windows the target perception differed due to the change of context cues. However, according to the global field powers and ERPs (Figure 4 and Figure 5), N400 did not emerge in either the Tone or Vowel conditions, and N2 was not shown in the Tone condition. Therefore, only N1, P2, and P3 were examined in the Tone Condition, and only N1, P2, N2, and P3 were examined in the Vowel Condition.



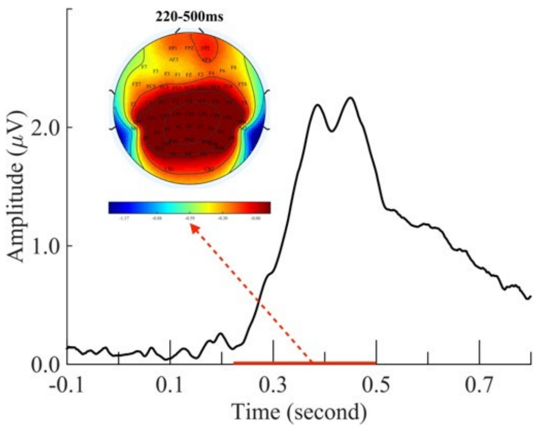
(a)



(b)

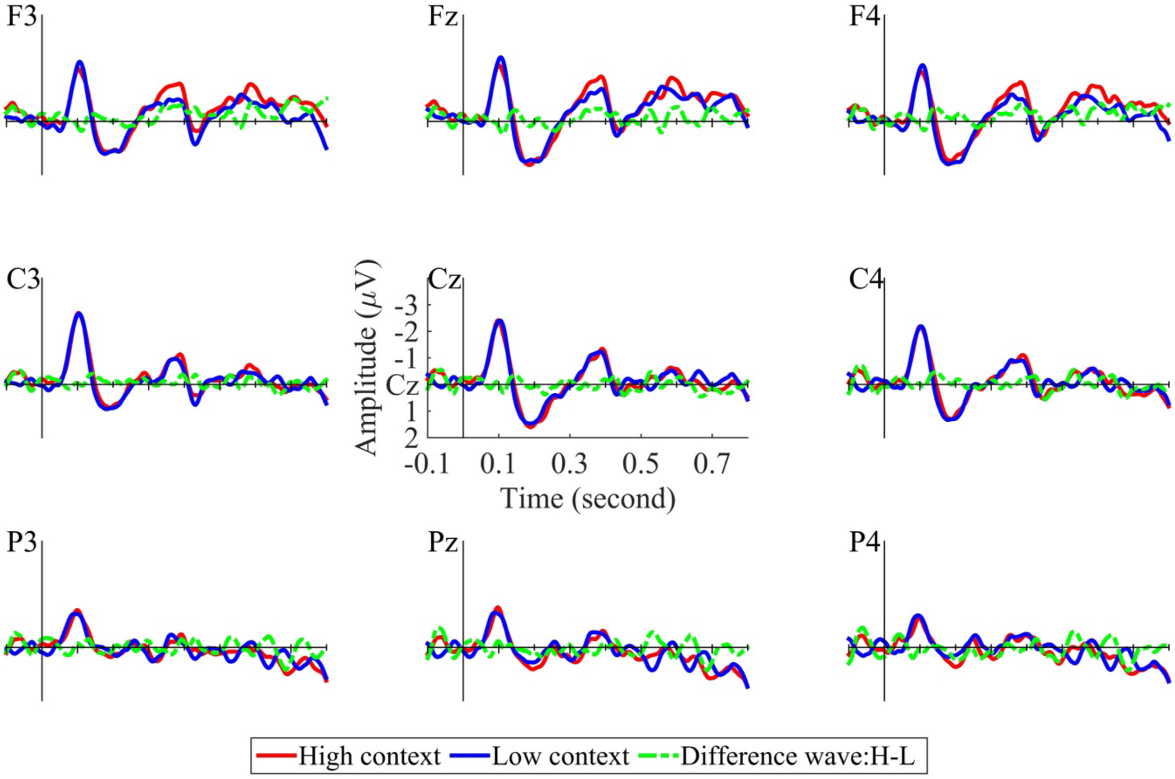


(c)

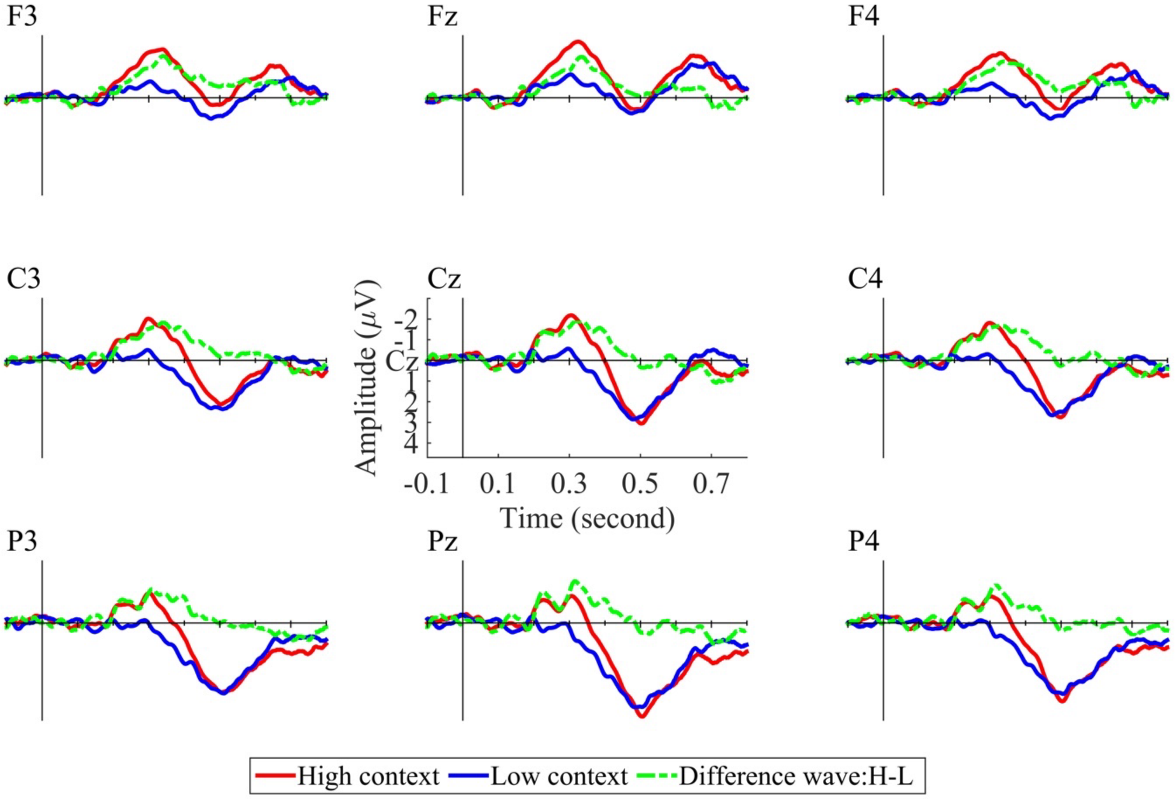


(d)

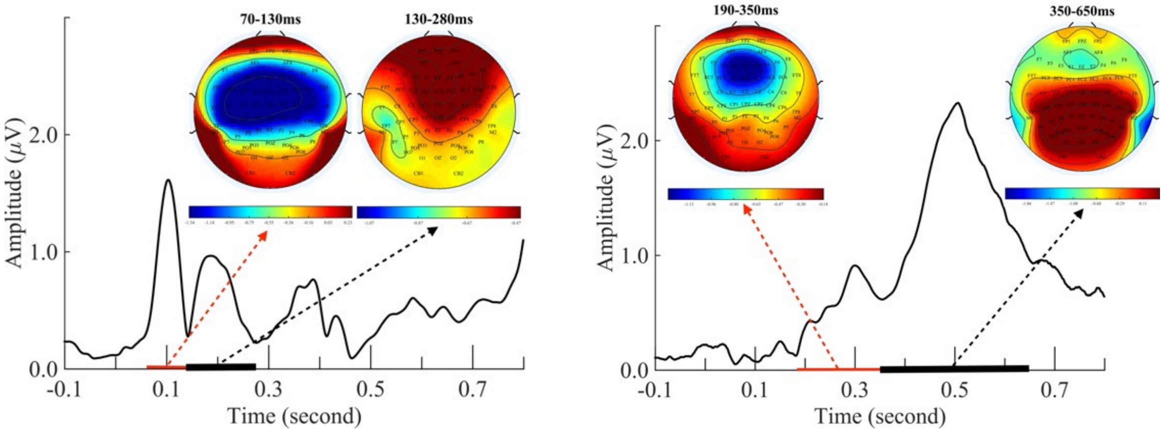
Figure 4. (a) The ERP waves of S-cluster in the Tone condition. (b) The ERP waves of C-cluster in the Tone condition. (c) The global field power of S-cluster in the Tone condition and the scalp topography maps showing the potential distribution in the N1 (left) and the P2 (right) time windows. (d) The global field power of C-cluster in the Tone condition and the scalp topography map showing the potential distribution in the P3 time window.


F3
Fz
F4
C3
Cz
C4
Amplitude (μV)
-3
-2
-1
Cz
1
2
-0.1
0.1
0.3
0.5
0.7
Time (second)
P3
Pz
P4
High context
Low context
Difference wave:H-L

(a)


F3
Fz
F4
C3
Cz
C4
Amplitude (μV)
-2
-1
Cz
1
2
3
4
-0.1
0.1
0.3
0.5
0.7
Time (second)
P3
Pz
P4
High context
Low context
Difference wave:H-L

(b)



(c) (d)

Figure 5. (a) The ERP waves of S-cluster in the Vowel condition. (b) The ERP waves of C-cluster in the Vowel condition. (c) The global field power of S-cluster in the Vowel condition and the scalp topography maps showing the potential distribution in the N1 (left) and the P2 (right) time windows. (d) The global field power of C-cluster in the Vowel condition and the scalp topography maps showing the potential distribution in the N2 (left) and the P3 (right) time windows.

Table 1. Time windows and electrodes for different ERP components.

| Condition | RIDE Cluster | ERP Component | Time window | Electrodes |
|---|---|---|---|---|
| Lexical tone | S-cluster | N1 | 50-150 ms | FC5, FC3, C5, C3, FC1, C1, FCz, Cz |
| | | P2 | 150-280 ms | F1, FC1, C1, CP1, Fz, FCz, Cz, CPz, F2, FC2, C2, CP2 |
| | C-cluster | P3 | 220-500 ms | CP3, P3, PO3, CP1, P1, CPz, Pz, POz, CP2, P2, CP4, P4, PO4 |
| Vowel | S-cluster | N1 | 70-130 ms | FC5, FC3, FC1, C5, C3, C1, FCz, Cz, FC2, C2, FC4, C4 |
| | | P2 | 130-280 ms | F1, FC1, C1, Fz, FCz, Cz, F2, FC2, C2, F4, FC4, C4 |
| | C-cluster | N2 | 190-350 ms | F1, FC1, Fz, FCz, F2, FC2 |
| | | P3 | 350-650 ms | P3, PO3, P1, Pz, POz, P2, P4, PO4 |

### 3.2.2. Time windows and electrodes for different ERP components

The time windows and electrodes for each ERP component are summarized in Table 1. The methods used to define the time windows and electrodes for each ERP component largely followed the collapsed localizer method in Luck and Gaspelin (2017). The waves in Figures 4 (c) and (d) were the global field power of S-cluster and that of C-cluster in the Tone condition. The waves in Figures 5 (c) and (d) were the global field power of S-cluster and that of C-cluster in the Vowel condition. The global field power along time of each cluster was computed as the square root of the mean square ERP values at each time point averaged across two contexts (high and low), 64 electrodes, and 20 subjects. Based on the global field power, the time range during which the ERP component showed the largest activity was selected as the time window for the corresponding ERP component. The scalp topographic maps in Figures 4 (c) and (d) showed the potential distribution at different time windows in the Tone condition. The scalp topographic maps in Figures 5 (c) and (d) showed the potential distribution at different time windows in the Vowel condition. The topographic map for each ERP component was obtained by averaging the voltage amplitude at each electrode across two contexts, 20 subjects, and the entire time window. Electrodes where the ERP component was expected to peak according to the topographic map were selected to quantify the corresponding ERP. The time windows and electrodes selected based on the collapsed data were used to measure the ERP components in each context condition separately. Considering that S-cluster mainly reflects the stimulus-driven process (Ouyang et al., 2015), the time windows and electrodes of the early ERP components, N1 and P2, were defined based on S-cluster. Since N2 and P3 are related to decision-making and response selection (Kopp et al., 1996), they were quantified based on global field powers and topographies of C-cluster.

### 3.2.2. Statistical tests

The mean amplitudes were extracted to quantify the ERP components. Since scalp distribution was not an interest of the present study, the amplitudes at different electrodes for a single ERP component were averaged to reduce the familywise error rate (Luck & Gaspelin, 2017). The behavioral data analysis used the mixed-effect model to control the within-subject trial-to-trial variability. This variability was eliminated in the epoch-average step for the EEG data. Therefore, the paired-sample t-test was conducted on the amplitudes of each ERP component to see at which time window(s) listeners' target perception differed in two contexts.

The lexical tone perception triggered more negative N1 in the low context condition ($M$ = -1.93; $SE$ = 0.33) compared with the high context condition ($M$ = -1.25; $SE$ = 0.37), $t(19)$ = 3.36, $p = 0.01$. The participants' perception of lexical tones also showed a significant difference in the P3 time window, $t(19) = 2.26$, $p = 0.04$. The P3 amplitude was significantly larger in the high context condition (*M* = 2.8; *SE* = 0.52) compared with the low context condition (*M* = 1.33; *SE* = 0.59). The vowel perception in the high context (*M* = -1.77; *SE* = 0.55) elicited more negative N2 compared with that in the low context (*M* = -0.61; *SE* = 0.61), *t*(19) = -2.17, *p* = 0.04. The amplitudes of other ERP components showed no statistically significant differences in high and low contexts.

The statistical test based on 31 participants' EEG data (see supplementary materials) revealed a significant main effect of context in the N1 time window and a marginally significant main effect of context in the P3 time window in the Tone condition, which was largely consistent with the results obtained from the 20 participants' data in the Tone condition. The absence of the main effect of context in the Vowel condition was probably caused by the mix of the contrastive and assimilative context effects. Participants who showed the assimilative context effect (the assimilative cohort) elicited different ERP patterns (see Supplementary Figure 5 b) compared with participants showing the contrastive context effect (the contrastive cohort; Figure 5 b) in the Vowel condition. A noticeably larger N2 amplitude was observed in the low context compared with the high context for the assimilative cohort in the Vowel condition, while oppositely, the contrastive cohort elicited significantly larger N2 amplitude in the high context. However, due to the small sample size (eight subjects only), no statistical test was done on the ERP data of the assimilative cohort. Further studies are needed to test if the observed N2 difference in the assimilative cohort is statistically significant.

### 3.3. ERP and behavioral data correlation

The difference waves (the dash-dotted lines in Figures 4 a and b and Figures 5 a and b) were obtained by subtracting the ERP amplitudes in the low context condition from those in the high context condition. They reflected how the cortical processing of the target sounds was affected by the context shift. The analysis detailed above revealed that ERPs in two contexts differed significantly in N1 and P3 for lexical tone perception and in N2 for vowel perception. Therefore, the amplitudes of the difference waves in these time windows were used as the neurological indexes of the context effect size. Pearson's correlation coefficients were calculated to see how the context effect size at the neurological level was correlated with that at the behavioral level. The amplitude of the difference wave in the P3 time window was highly and positively correlated with the behavioral response for the lexical tone normalization (*r* = 0.61; *p* < 0.05). Participants who showed larger context effect sizes at the neurological level also showed a larger context effect size at the behavioral level. Such a correlation was not observed in the N1 time window for the lexical tone normalization (*r* = 0.02; *p* = 0.9). The neurological context effect size, however, was not correlated with the behavioral context effect size in the N2 time window for vowel normalization (*r* = -0.1; *p* = 0.68).

## 4. Discussion

### 4.1. Lexical tone normalization in the N1 and P3 time windows

An early ERP component (N1) emerged in the lexical tone normalization. This differs from C. Zhang et al.'s (2013) observation of cortical effects of lexical tone normalization in a later time window. However, the significant context effect in N1 was likely caused by the unbalanced context-target pitch difference in two conditions, and thus N1 might not be a reliable neural marker for the lexical tone normalization in the present study. The acoustic analysis demonstrated that the mean pitch height was 138.3 Hz for the high F0 context and 100.1 Hz for the low F0 context. The target — Step 10 (averaged across subjects) — was 147 Hz. The pitch difference between the low F0 context and the target was likely more salient than the difference between the high F0 context and the target (46.9 Hz vs. 8.7 Hz). The forward energetic masking suggests that speech perception will be difficult if its precursor has acoustic energies in the same frequency region (Viswanathan et al., 2013). Therefore, the onset of the target sound in the present study was presumably more salient in the low F0 context than in the high F0 context, resulting in a larger N1 in the low F0 context. This result was consistent with Butler (1968), who reported that increased N1 amplitude was associated with increased pitch/intensity distance between the target and the preceding auditory stimuli. It was reasonable to assume that if the context-target differences in F0 were similar in high and low blocks, lexical tone normalization would elicit comparable N1 amplitudes in the two conditions. This assumption was partially verified by the vowel normalization in the present study. The context-target (Step 8 as the average target across subjects) differences in F1 were more or less the same in the two conditions (i.e., 105.15 Hz in the high F1 context and 101.6 Hz in the low F1 context). Meanwhile, statistically similar N1 amplitudes were observed across the two context conditions for the vowel normalization task.

The context F0 shifts also modulated the neural responses in the P3 time window (220-500 ms) in the lexical tone normalization, with a larger P3 amplitude in the high F0 context. The context effect size in the ERP analysis of P3 was significantly correlated with the context effect size at the behavioral level. Therefore, the lexical tone normalization in the present study was most likely implemented in the P3 time window. C. Zhang et al., (2013) observed tone normalization effects in the N400 window (250 ms-500 ms), but it was important to note that the time window of N400 largely overlapped with the P3 window (220-500 ms) in the current study.

In the present study, participants were asked to make overt responses to targets that they perceived as /wo55/ and to hold their responses if they thought that the target was /wo33/. The asymmetrical overt response made the experiment more or less resemble the Go/NoGo paradigm with a delayed response. Therefore, the ERP components elicited in the lexical tone normalization should be modulated by the properties of the Go/NoGo paradigm. In the classic Go/NoGo paradigm, participants are asked to make responses as soon as possible to one type of stimulus (usually frequent; referred to as Go trials) and withhold responses to another (usually rare; referred to as NoGo trials). A larger P3 will show in the NoGo trials compared with the Go trials in the central-anterior region around 300-500 ms after the stimuli onset (Lavric et al., 2004). The increased P3 in the NoGo trials is generally related to the inhibition of overt motor responses (Gajewski & Falkenstein, 2013) and cognitive inhibition since the increased NoGo P3 is also found when participants are asked to silently count Go trials (Smith et al., 2008). In the high F0 context block, more trials (57.5%) were perceived as /wo33/ due to the contrastive context effect. These conditions thus involved a withholding of an overt response. In the low F0 context block, however, participants only withheld their responses in a small proportion of trials (14.6%). The observed cortical differences may therefore be partially attributed to differences in cognitive inhibition. The topographic map of the difference wave in the P3 time window (see Supplementary Figure 1a) suggested that the higher P3 amplitude was shown not only in the frontal-central region but also in the parietal region. Verleger et al. (2016) asked participants to give responses only to the rare stimuli in a speech perception task. The authors decomposed the observed P3 with RIDE and found that the P3 in their study was composed of the Go P3 in the parietal region, which was related to the detection of rare stimuli and the NoGo P3 in the fronto-central region, which was related to the inhibition of the NoGo trials. In the present study, comparatively fewer trials (42% vs. 58%) were perceived as /wo55/ in the high F0 context block. Participants needed to respond to those rare trials, and thus a Go P3 due to the detection of comparatively rare stimuli was observed in the parietal region. In sum, the P3 observed in 220-500 ms in the present study could be the combination of the inhibition to the response to the /wo33/ trials and the detection of the comparatively rare /wo55/ trials. Participants must have fully or partially finished the contrastive normalization of lexical tones before they decided whether to give a response or not. It can be deduced that the contrastive normalization of lexical tones is implemented immediately before the P3 time window or partially within the P3 time window in the present study.

The stimulus /wo33/ in the Tone condition is a modal particle but /wo55/ is a content word. The difference in word class might bias listeners in favor of /wo55/ responses. This might be one of the reasons why the tone normalization effect was mild in the high context condition (42.5% /wo55/ vs. 57.5% /wo33/) but strong in the low context condition (85.4% /wo55/ vs. 14.6% /wo33/). The normalization effect in the present study was quantified by comparing the proportion of /wo55/ responses in two contexts, which to some extent eliminated the potential problems caused by word class bias since the word class difference affected participants' response in the same direction (i.e., more /wo55/ responses) regardless of the context F0 heights. Therefore, despite this bias, in the present study it was still observed that the identical lexical tone was perceived differently when the context F0 changed from low to high (i.e., the normalization effect; see Figure 2, left panel). Further studies with stimuli of the identical word class may observe a less biased response pattern.

### 4.2. Vowel normalization in the N2 time window

The manipulation of context F1 affected vowel perception in the N2 time window (190-350 ms), with a larger N2 in the high F1 context. This is a little bit later than the N1 time window (80-160 ms) in Sjerps et al. (2011), the acoustic-phonetic stage in Sjerps et al. (2019), and the P2 time window (130-250 ms) in K. Zhang and Peng (2018). Like in the lexical tone normalization, the ERP components in the vowel normalization were also modulated by the task properties in the present study. A larger N2 in the frontal-central region is elicited in NoGo trials in the Go/NoGo tasks, and thus N2 is also viewed as a neural correlate of inhibition of premature response plan or tendency (Gajewski & Falkenstein, 2013; Kopp et al., 1996). In the present study, the contrastive context effect led to fewer responses in the high F1 context (49.8%) than in the low F1 context (77.9%). Thus, participants withheld their responses more frequently in the high F1 context block, eliciting a larger N2. It can be deduced that listeners in the present study must have fully or partially finished the contrastive vowel normalization process before inhibiting a response tendency. Therefore, the contrastive normalization process of vowels was most likely accomplished immediately before or partially within the N2 time window.

The stimuli used in the Vowel condition start with an approximate /w/ which lasts around 50 ms (see the right panel in Figure 1). Therefore, the vowel information is available around 50 ms after the stimulus onset. If the vowel information had been available at the very beginning of the stimulus, like the lexical tones in the present study, we would have probably observed an N2 in 140-300 ms, which largely overlapped with the time window of P2 in K. Zhang and Peng (2018; 130-250 ms). However, unlike the lexical tone normalization, the correlation analysis revealed that the neurological context effect size in N2 was not significantly correlated with the behavioral context effect size (*r* = -0.1; *p* = 0.68). This might be caused by the relatively small context effect size (*M* = 0.281, *SE* = 0.06) in the behavioral level for the vowel normalization. Larger shifts of context F1, like 200 Hz used by Sjerps et al. (2011), would probably generate a larger context effect size. Furthermore, the small sample size (20 subjects) could be another reason for the nonsignificant correlation. Studies with more subjects and larger F1 shifts should be carried out to further investigate the time course of vowel normalization.

### 4.3. The normalization of lexical tones and vowels in the perception of tonal languages

The current study was set up to assess whether lexical tones and vowels are normalized as an integrated percept or through two at least partly separate processing streams. To that end, Cantonese speakers were asked to identify speech sounds in a lexical tone and a vowel normalization task with highly similar experimental designs. The behavioral results suggested that the majority of participants demonstrated a contrastive context effect, although a small number of them showed an assimilative context effect. The EEG data analysis focusing on the contrastive cohort revealed that contrastive normalization effects modulated the lexical tone process and the vowel process in partially different time windows. In the Tone condition, participants' inhibition to /wo33/ responses elicited a larger P3, indicating that lexical tones were normalized immediately before, or partially within, the P3 time window (220-500 ms). However, listeners tended to inhibit their responses to /wu55/ in the N2 time window, suggesting that vowel normalization occurred immediately before or partially within the N2 time window (190-350 ms). If the vowel information is available at the very beginning of the stimulus (e.g., a stimulus without the initial 50-ms /w/), the vowel normalization will possibly trigger N2 at around 140-300 ms. Although the lexical tone normalization and the vowel normalization triggered different ERPs, their time windows showed some overlap. The present study thus demonstrated that regarding the contrastive normalization, lexical tones and vowels were processed in at least *partially separate* time windows, with the normalization of vowels occurring earlier than the normalization of lexical tones. Perceptual normalization is thus a process at the sub-syllabic level that operates on phonemes and tonemes in a partially separated fashion rather than on the syllable as a whole.

Previous studies demonstrate that different brain regions and neural pathways are involved in the lexical tone process and the phoneme process (see Liang & Du, 2018, for a review). Vowel quality is generally found to be recognized earlier than lexical tones (Hu et al., 2012; Li et al., 2014; Zou et al., 2020). The modified TRACE model, which tries to illustrate the perception process of tonal languages, holds that the mental representations of tonemes and phonemes are stored separately and are processed partially independently within their own pathways (Ye & Connine, 1999). However, a batch of studies (e.g., Choi et al., 2017; Shuai & Malins, 2017; Tong et al., 2014) observe similar processing patterns for lexical tones and vowels, pointing to the simultaneous access of these cues. The present study, which finds different time courses for lexical tone normalization and vowel normalization, supports the concept of separate processing of lexical tones and vowels in tonal languages. Furthermore, the results also support the notion that lexical tones are recognized relatively later than vowel quality.

Although the present study only included three syllables differing in either F0 height or F1, considering the essence of the extrinsic normalization process (i.e., recalibrating the target acoustic cues against the context cues), the results based on these three syllables can largely be applied to the normalization of other syllables with ambiguous F0 height or F1. The perception of lexical tones and vowels, however, is affected by many factors. Contour tones rely more on pitch slope than pitch height, and they are thus not as sensitive to the context F0 height manipulation as level tones are. Therefore, the contour tone normalization might trigger a P3 with comparative smaller amplitude compared with the level tone normalization. Regarding vowel perception, not only the F1, but also the higher formants, especially F2, are important. Listeners show different sensitivities to formants at different frequency ranges. Therefore, it is hard to tell whether the normalization of vowels with an ambiguous F2 will show similar perceptual results as the normalization of vowels with an ambiguous F1. Further studies are needed to test the normalization of lexical tones and vowels with ambiguous acoustic cues in other dimensions.

## 5. Conclusion

The present study tested whether lexical tones and vowels whose acoustic signals are tightly coupled in time domain are also normalized holistically. With the highly similar experimental designs for two normalization tasks, the present study observed that the contrastive lexical tone normalization occurs in the P3 time window (220-500 ms) while the contrastive vowel normalization occurs in the N2 time window (190-350 ms), supporting a partially separate contrastive normalization process of lexical tones and vowels. The result suggested that perceptual normalization process is implemented at the sub-syllabic level rather than on the whole syllable and that lexical tones are recognized relatively later than vowels.

## Declaration of competing interest

None.

## CRediT authorship contribution statement

Kaile Zhang: Conceptualization, Methodology, Data collection, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.

Matthias J. Sjerps: Conceptualization, Methodology, Writing – review & editing, Funding acquisition.

Gang Peng: Conceptualization, Methodology, Writing – review & editing, Project administration, Funding acquisition.

## Acknowledgement:

This study was supported by the Research Grant Council of Hong Kong (GRF: 14408914). We are grateful to two anonymous reviewers for their valuable comments on earlier versions of this paper.

## References

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

Berchicci, M., Spinelli, D., & Di Russo, F. (2016). New insights into old waves. Matching stimulus- and response-locked ERPs on the same time-window. *Biological Psychology*, *117*, 202–215. https://doi.org/10.1016/j.biopsycho.2016.04.007

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]. Version 6.0.16, retrieved 10 August 2016 from http://www.praat.org/.*

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*(2), 103–135. https://doi.org/10.1023/B:JOPR.0000017223.98667.10

Butler, R. A. (1968). Effect of changes in stimulus frequency and intensity on habituation of the human vertex potential. *The Journal of the Acoustical Society of America*, *44*(4), 945–950. https://doi.org/10.1121/1.1911233

Choi, W., Tong, X., Gu, F., Tong, X., & Wong, L. (2017). On the early neural perceptual integrality of tones and vowels. *Journal of Neurolinguistics*, *41*, 11–23. https://doi.org/10.1016/j.jneuroling.2016.09.003

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. https://doi.org/10.1016/j.jneumeth.2003.10.009

Foss-feig, J. H., Stavropoulos, K. K. M., Mcpartland, J. C., & Mark, T. (2018). *Biomarker for Sensory and Communication Alterations in*. *43*(2), 109–122. https://doi.org/10.1080/87565641.2017.1365869.Electrophysiological

Gajewski, P. D., & Falkenstein, M. (2013). Effects of task complexity on ERP components in Go/Nogo tasks. *International Journal of Psychophysiology*, *87*(3), 273–278. https://doi.org/10.1016/j.ijpsycho.2012.08.007

Gao, X., Yan, T. T., Tang, D. L., Huang, T., Shu, H., Nan, Y., & Zhang, Y. X. (2019). What Makes Lexical Tone Special: A Reverse Accessing Model for Tonal Speech Perception. *Frontiers in Psychology*, *10*(December). https://doi.org/10.3389/fpsyg.2019.02830

Hu, J., Gao, S., Ma, W., & Yao, D. (2012). Dissociation of tone and vowel processing in Mandarin idioms. *Psychophysiology*, *49*(9), 1179–1190. https://doi.org/10.1111/j.1469-8986.2012.01406.x

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, *125*(6), 3983–3994. https://doi.org/10.1121/1.3125342

Huang, X., Liu, X., Yang, J. C., Zhao, Q., & Zhou, J. (2018). Tonal and vowel information processing in Chinese spoken word recognition: An event-related potential study. *NeuroReport*, *29*(5), 356–362. https://doi.org/10.1097/WNR.0000000000000973

Jodo, E., & Kayama, Y. (1992). Relation of a negative ERP component to response inhibition in a Go/No-go task. *Electroencephalography and Clinical Neurophysiology*, *82*(6), 477–482. https://doi.org/10.1016/0013-4694(92)90054-L

Kopp, B., Mattler, U., Goertz, R., & Rist, F. (1996). N2, P3 and the lateralized readiness potential in a nogo task involving selective response priming. *Electroencephalography and Clinical Neurophysiology*, *99*(1), 19–27. https://doi.org/10.1016/0921-884X(96)95617-9

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. https://doi.org/10.1121/1.397821

Lavric, A., Pizzagalli, D. A., & Forstmeier, S. (2004). When "go" and "nogo" are equally frequent: ERP components and cortical tomography. *European Journal of Neuroscience*, *20*(9), 2483–2488. https://doi.org/10.1111/j.1460-9568.2004.03683.x

Li, W., Wang, L., & Yang, Y. (2014). Chinese tone and vowel processing exhibits distinctive temporal characteristics: An electrophysiological perspective from classical chinese poem processing. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0085683

Liang, B., & Du, Y. (2018). The functional neuroanatomy of lexical tone perception: An activation likelihood estimation meta-analysis. *Frontiers in Neuroscience*, *12:495*. https://doi.org/10.3389/fnins.2018.00495

Liu, S., & Samuel, A. G. (2007). The role of Mandarin lexical tones in lexical access under different contextual conditions. *Language and Cognitive Processes*, *22*(4), 566–594. https://doi.org/10.1080/01690960600989600

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Luo, X., & Ashmore, K. B. (2014). The effect of language experience on perceptual normalization of Mandarin tones and non-speech pitch contours. *The Journal of the Acoustical Society of America*, *135*(6), 3585–3593. https://doi.org/10.1121/1.4874619

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*(8), 2032–2043. https://doi.org/10.1016/j.neuropsychologia.2012.05.002

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Murray, J. G., Ouyang, G., & Donaldson, D. I. (2019). Compensation of trial-to-trial latency jitter reveals the parietal retrieval success effect to be both variable and thresholded in older adults. *Frontiers in Aging Neuroscience*, *11:179*. https://doi.org/10.3389/fnagi.2019.00179

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, *125*(6), 826–859. https://doi.org/10.1037/0033-2909.125.6.826

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Ouyang, G., Sommer, W., & Zhou, C. (2015). Updating and validating a new framework for restoring and analyzing latency-variable ERP components from single trials with residue iteration decomposition (RIDE). *Psychophysiology*, *52*(6), 839–856. https://doi.org/10.1111/psyp.12411

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, *24*(1–2), 31–44. https://doi.org/10.1016/0010-0277(86)90003-X

Repp, B. H., & Lin, H. (1990). Integration of segmental and tonal information in speech perception: A cross-linguistic study. *The Journal of the Acoustical Society of America*, *87*, S46. https://doi.org/10.1121/1.2028239

Rysling, A., Jesse, A., & Kingston, J. (2019). Regressive spectral assimilation bias in speech perception. *Attention, Perception, and Psychophysics*, *81*(4), 1127–1146. https://doi.org/10.3758/s13414-019-01720-9

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), 1–12. https://doi.org/10.1162/0898929052880057

Shao, J., & Zhang, C. (2019). Talker normalization in typical Cantonese-speaking listeners and congenital amusics: Evidence from event-related potentials. *NeuroImage: Clinical*, *23*(April), 101814. https://doi.org/10.1016/j.nicl.2019.101814

Shu, H., Peng, H., & McBride-Chang, C. (2008). Phonological awareness in young Chinese children. *Developmental Science*, *11*(1), 171–181. https://doi.org/10.1111/j.1467-7687.2007.00654.x

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, *49*(1), 230–241. https://doi.org/10.3758/s13428-015-0690-0

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, *10:2465*. https://doi.org/10.1038/s41467-019-10365-z

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, *49*(14), 3831–3846. https://doi.org/10.1016/j.neuropsychologia.2011.09.044

Sjerps, M. J., Zhang, C., & Peng, G. (2018). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 914–924. https://doi.org/10.1037/xhp0000504

Smith, J. L., Johnstone, S. J., & Barry, R. J. (2008). Movement-related potentials in the Go/NoGo task: The P3 reflects both cognitive and motor inhibition. *Clinical Neurophysiology*, *119*(3), 704–714. https://doi.org/10.1016/j.clinph.2007.11.042

Stilp, C. (2019). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1517. https://doi.org/10.1002/wcs.1517

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Tong, X., Mcbride, C., Lee, C. Y., Zhang, J., Shuai, L., Maurer, U., & Chung, K. K. H. (2014). Segmental and suprasegmental features in speech perception in cantonese-speaking second graders: An ERP study. *Psychophysiology*, *51*, 1158–1168. https://doi.org/10.1111/psyp.12257

Verleger, R., Grauhan, N., & Śmigasiewicz, K. (2016). Is P3 a strategic or a tactical component? Relationships of P3 sub-components to response times in oddball tasks with go, no-go and choice responses. *NeuroImage*, *143*, 223–234. https://doi.org/10.1016/j.neuroimage.2016.08.049

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2013). Similar response patterns do not imply identical origins: An energetic masking account of nonspeech effects in compensation for coarticulation. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1181–1192. https://doi.org/10.1037/a0030735

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes*, *14*(5–6), 609–630. https://doi.org/10.1080/016909699386202

Zhang, C., Peng, G., & Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language*, *126*(2), 193–202. https://doi.org/10.1016/j.bandl.2013.05.010

Zhang, K., & Peng, G. (2018). The time course of the extrinsic vowel normalization: An ERP study. *Phonetic Conference of China*.

Zhang, K., Sjerps, M. J., Zhang, C., & Peng, G. (2018). Extrinsic normalization of lexical tones and vowels: Beyond a simple contrastive general auditory mechanism. *Proc. TAL2018, Sixth International Symposium on Tonal Aspects of Languages*, *June*, 227–231. https://doi.org/10.21437/tal.2018-46

Zhang, K., Wang, X., & Peng, G. (2017). Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism. *The Journal of the Acoustical Society of America*, *141*(1), 38–49. https://doi.org/10.1121/1.4973414

Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, *49*(7), 1761–1770. https://doi.org/10.1016/j.neuropsychologia.2011.02.054

Zou, Y., Lui, M., & Tsang, Y. K. (2020). The roles of lexical tone and rime during Mandarin sentence comprehension: An event-related potential study. *Neuropsychologia*, *147*(February), 107578. https://doi.org/10.1016/j.neuropsychologia.2020.107578