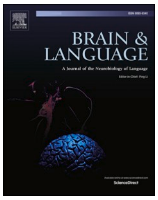
Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Research Article

The modulation of cognitive load on speech normalization: A neurophysiological perspective

Kaile Zhang, Gang Peng [*]

The Research Centre for Language, Cognition, and Neuroscience, The Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

ARTICLE INFO

Keywords:
Cognitive load
Speech variability
Context cues
Extrinsic normalization
ERP

ABSTRACT

Extrinsic normalization, wherein listeners utilize context cues to adapt to speech variability, is essential for maintaining perceptual constancy. In daily communication, distractions are ubiquitous, raising questions about the influence of cognitive load on this process, particularly at the cortical level. This study investigates how cognitive load modulates extrinsic normalization using electroencephalography (EEG). Native Cantonese speakers were asked to perceive Cantonese tones from multiple speakers with context cues in both single- and dual-task conditions. The secondary task did not hinder listeners' normalization process at the behavioral level. However, EEG data revealed significant modulations of extrinsic normalization under cognitive load. Extrinsic normalization elicited P2, N400, and LFN, suggesting that extrinsic normalization encompasses multiple perceptual adjustments at stages of phonological processing, lexical retrieval, and decision-making. Cognitive load influenced extrinsic normalization at all these stages, as evidenced by smaller P2, larger N400, and larger LFN, highlighting the active and controlled nature of this process.

1. Introduction

In daily communication, listeners encounter different speakers, and the acoustic realization of the same word from different speakers can vary significantly (Ladefoged & Broadbent, 1957). For example, the fundamental frequency (F0) is the most important acoustic correlate for lexical tone perception. However, due to generally shorter and thinner vocal folds of female speakers, the F0 of a female speaker's low tone could be higher than that of a male speaker's high tone, making intrinsic acoustic cues less reliable for lexical tone perception (Peng, 2006). In such conditions, extrinsic context serves as a crucial cue for listeners to normalize speaker variability and achieve perceptual constancy. For instance, using the Cantonese greeting "早晨" (/zou 25 san 21/, good morning), listeners figure out the pitch heights for the highest (the endpoint of T25) and the lowest Cantonese tones (the endpoint of T21) for a specific Cantonese speaker, then categorize incoming tone tokens based on this speaker-specific acoustic-phonemic (i.e., the pitch height-tonal category) framework (K. Zhang et al., 2024). Accuracy of Cantonese lexical tone perception from multiple speakers improved significantly with context cues compared to tone identification in isolation (Peng et al., 2012; Wong & Diehl, 2003). Normalizing speaker variability in speech signals with context cues is called extrinsic normalization, a specific form of talker normalization strategies (Johnson, 2005). Extrinsic normalization is a crucial speech perception strategy, enabling listeners to adjust for speech variability and recognize the same phonemes and words regardless of the speaker's age, gender, accent, or emotional state. This remarkable adaptability of the human auditory system is essential for speech communication in multiple-speaker conditions.

Daily communication scenarios frequently involve not only multiple speakers but also multitasking. For instance, drivers converse with passengers while monitoring traffic. Studies investigating the cognitive and neural mechanisms of speech normalization indicate that it is an actively controlled process and thus co-occurring distractions would theoretically deteriorate our accommodation to speech variability (Magnuson & Nusbaum, 2007). For example, Nusbaum & Magnuson (1997) found that the perception of mixed-speaker speech, which requires more frequent speaker normalization processes, is more time-consuming and less accurate than the perception of blocked-speaker speech, indicating that speech normalization relies on cognitive resources and thus is a controlled process. Wong et al. (2004) reported that while normalizing speaker variability in Cantonese tones, the middle/

[*] Corresponding author.
E-mail address: gpeng@polyu.edu.hk (G. Peng).

https://doi.org/10.1016/j.bandl.2025.105579
Received 18 December 2024; Received in revised form 7 April 2025; Accepted 7 April 2025
Available online 15 April 2025

superior temporal regions responsible for speech processing (Yi et al., 2019) and the superior parietal regions related to attentional networks (Behrmann et al., 2004) showed increased activity. The increased activity in the temporal-parietal network indicated that more cortical resources were required to process the speech variability and the involvement of the attentional network suggested that speech normalization is constrained by attentional resources. Aside from the attentional network, the speech normalization process also recruited areas associated with decision-making, such as the inferior frontal gyrus, indicating an actively controlled process of speech normalization (Myers et al., 2009; Myers & Mesite, 2014).

So far, only a few researchers have directly investigated whether cognitive load affected the extrinsic normalization process using a dual-task paradigm. However, the findings from these studies differ from the above-mentioned studies focusing on the cognitive and neural mechanisms of speech normalization. Reese and Reinisch (2022) synthesized a series of consonants ranging from /s/ to /ʃ/ to simulate production variability across multiple speakers. They embedded the target words with these ambiguous consonants in sentences produced by either a female or male speaker which provided talker-specific information. Participants were required to identify the ambiguous word while simultaneously performing a visual search task. They found that participants' performances were comparable no matter whether the visual search tasks were absent, easy, or difficult, suggesting that the co-occurring distractions did not impair listeners' ability to utilize talker-specific cues in contexts to resolve consonant variability. Bosker et al. (2017) tested the spectral and duration normalization processes in speech perception using a similar dual-task paradigm. They asked listeners to identify ambiguous vowel /ɑ/-/a:/ pairs within sentences that varied in duration or formant frequencies. Although Bosker et al. (2017) did not directly address talker variability, the underlying cognitive processes of spectral or duration normalization are similar to those involved in talker normalization with contextual cues in Reese and Reinisch (2022); both require extracting contextual information and recalibrating target sounds accordingly. They found that co-occurring distractions did not affect listeners' use of spectral and temporal cues in contexts to normalize vowel variabilities. Specifically, listeners gave more /a:/ responses in the context of short duration or low formant frequencies and more /ɑ/ responses in the context of long duration and high formant frequencies, regardless of the co-occurring visual search task. Based on Bosker et al. (2017), it is reasonable to deduce that the normalization of talker variability in vowels with context cues is less likely affected by cognitive load as well.

To date, research on the influence of cognitive load on extrinsic normalization has been limited to the perception of Dutch vowels (Bosker et al., 2017) and Austrian German consonants (Reese & Reinisch, 2022), with little attention given to suprasegmental components. In fact, suprasegmental elements, such as lexical tones, are much more vulnerable to speaker variability than consonants and vowels. The prominent speaker variabilities lie in gender and age, mainly distinguished by F0. The variation in F0 does not significantly affect the identification of vowels and consonants in non-tonal languages, but it matters in tonal languages such as Cantonese which use F0 to distinguish lexical meanings. For example, there are three level tones in Cantonese: high level (T55), mid level (T33), and low level (T22). The same base syllable /ji/ means different words when combined with different tones. It means "doctor" with /ji55/, "meaning" with /ji33/, and "two" with /ji2/ (Yip, 2002). Three level tones share similar pitch contours but differ in pitch height, and thus the variation in F0 introduced by different speakers severely affects the identification of Cantonese level tones. As a consequence, the perception of lexical tones—especially Cantonese level tones—appears to rely more heavily on extrinsic contextual cues compared with vowels and consonants. This was also confirmed by K. Zhang et al. (2018) and K. Zhang & Peng (2021), who compared the perception of vowels and tones and found that the perception of vowels relatively less relies on context cues, which may be due to the robustness of intrinsic cues. It is therefore possible that while the perception of vowels and consonants, being robust to talker variability, may be less affected by cognitive load (as suggested by Bosker et al., 2017, and Reese & Reinisch, 2022), the perception of lexical tones which relies heavily on extrinsic contexts might be more sensitive to cognitive load. To fill this research gap, the present study would test the extrinsic normalization of Cantonese level tones under different cognitive load conditions, aiming to determine whether the effects of cognitive load on the extrinsic normalization process vary across different speech components.

Another point worth mentioning is that existing studies on the effects of cognitive load on extrinsic normalization are limited to behavioral studies. Behavioral studies which yield only binary (yes/no) responses, are insufficient for revealing how cognitive load interacts with different stages of the extrinsic normalization process. The speaker adaptation with context cues is much more complex than listeners might realize. Speech perception encompasses multiple processing stages, including acoustic decoding, phonological processing, semantic retrieval, and final decision-making (McClelland & Elman, 1986; Shuai & Malins, 2017). The extrinsic normalization process—adjusting perceived ambiguous speech signals by referring to context cues—could occur at any of these stages within the speech perception hierarchy. Aravamudhan et al. (2008), Holt et al. (2001), and Laing et al. (2012) found that the spectro-temporal contrast between context and target is enough to trigger the noticeable extrinsic normalization process. Sjerps et al. (2011) also observed that extrinsic normalization elicited the N1 component, an event-related potential (ERP) component closely related to the acoustic processing of speech signals, indicating an auditory-level process of extrinsic normalization. As illustrated in the morning greeting example, listeners may use context cues to establish a speaker-specific acoustic-phonemic mapping and use it to recalibrate incoming signals, pointing to a perceptual adjustment at the phonological processing stage (Magnuson et al., 2021; Nusbaum & Magnuson, 1997). This has been supported by findings that phonological information (native vs. non-native speech) enhances the speaker adaptation process (Kang et al., 2016; K. Zhang & Peng, 2021), and that significant P2, an ERP component related to phonological processing (Cheng et al., 2014), was observed in the normalization of Cantonese tones (K. Zhang et al., 2023; K. Zhang & Peng, 2021). C. Zhang et al. (2013) reported that compared with tone perception without effective context cues, tone perception with effective contexts triggered significant N400. Considering N400 is related to lexical retrieval, C. Zhang et al. (2013) suggested that extrinsic normalization might occur when listeners retrieve lexical representations. Speaker adaptation has been observed at the final decision-making stage as well, where a comprehensive set of cues, including acoustic, phonological, and higher-level linguistic knowledge, are integrated to make perceptual adjustments before delivering an overt response (Bosker et al., 2017). Electroencephalography (EEG) studies also reported that the extrinsic normalization of lexical tones triggered a significant late positive component (LPC), an ERP component related to stimuli evaluation, contextual integration and decision making, and the normalization efficiency at the behavioral level was positively correlated with LPC amplitude (C. Zhang et al., 2013). Although most studies pinpointed the time locus of extrinsic normalization to a single stage of speech perception (for example, the acoustic processing stage in Sjerps et al., 2011), extrinsic normalization is most likely a multi-stage process encompassing both early acoustic and phonological adjustments and later semantic and cognitive adjustments, supported by a recent computational modeling study. X. Xie et al. (2023) tested several computational models simulating the extrinsic normalization process at different levels (i.e., acoustic, phonological, and cognitive adjustments) using real perceptual data. The study concluded that no single perceptual adjustment alone could adequately explain the complexities of real-world speech perception. Instead, a combined approach involving perceptual adjustments at all stages provided the best explanation for adaptive speech perception behavior (Persson & Jaeger, 2023; X. Xie et al., 2023). If extrinsic normalization indeed involves perceptual adjustments across multiple stages of speech perception, it is challenging to discern how cognitive load impacts each stage based solely on these behavioral findings. To address this research gap, the current study employs EEG, with its high temporal resolution, to capture the complete online extrinsic normalization process under varying cognitive load conditions. This approach enables us to examine how cognitive load influences extrinsic normalization at different stages of speech processing.

According to previous reports about the modulation of cognitive load on speech perception, the present study might observe that each stage of the extrinsic normalization would be affected by cognitive load. Cognitive load affects speech perception through pulse skipping mechanism — a process in which increased cognitive demands cause the auditory system to omit or "skip" certain discrete temporal events (or "pulses") within the speech signal. When these pulses, which contribute to the fine-grained spectral and temporal details of speech, are not fully processed, the precision of the acoustic information available for perception is compromised (Block et al., 2010; Chiu et al., 2019; Feng et al., 2021; Zakay & Block, 1995). For example, Heinrich et al. (2020) reported that cognitive load affected the encoding of low-level acoustic information, such as pure-tone threshold, duration, intensity, and VOT. Some EEG research also identified the effect of cognitive load on speech perception at the relatively late stage involving working memory. For example, Z. Xie et al. (2023) reported that compared to the single-task condition, the dual-task condition selectively reduced neural tracking of some linguistic features mainly at latencies $>$ 200 ms, beyond early acoustic processing. Kasper et al. (2014) reported that in a dual task paradigm, a visual task affected the auditory task at both early (as indexed by reduced N1 amplitude in unattended conditions) and late (as indexed by enhanced P3 in unattended conditions) stages of speech processing. Therefore, cognitive load probably modulates the extrinsic normalization at different adjustment stages as well. However, the degree of modulation may vary at different stages. Considering that acoustic processing is more automatic than the late stages of the speech perception process, such as phonological processing, lexical retrieval, and decision-making, cognitive load may show stronger effects on the later perceptual adjustment stages of the extrinsic normalization process but relatively weaker effects on acoustic adjustment. The present study will compare high and low cognitive load conditions to elaborate the finer interplay between cognitive load and different stages of the extrinsic normalization process.

Extrinsic normalization is an important perceptual strategy for listeners to adapt to speaker variability. Multitasking is common in daily communication. However, it remains unclear how the cognitive load introduced by multitasking modulates the extrinsic normalization process, especially for suprasegmental components. Recent computational modeling research (X. Xie et al., 2023) suggested that the extrinsic normalization likely involves perceptual adjustment across multiple speech perception stages. In contrast, existing behavioral studies cannot reveal how cognitive load interacts with extrinsic normalization at these distinct stages. To address these questions, the present study would use EEG to investigate the perception of Cantonese level tones under different cognitive load conditions. Findings of this study are expected to enrich our understanding of the neural mechanisms underlying extrinsic normalization in tonal languages in the multitasking scenario.

Considering that Cantonese level tones show stronger reliance on extrinsic contexts than the perception of segments such as vowels (K. Zhang et al., 2018; K. Zhang & Peng, 2021), it is hypothesized that the present study would observe a significant modulation of cognitive load on extrinsic normalization, differing from those reported by Bosker et al. (2017) and Reese & Reinisch (2022). Native Cantonese speakers will be asked to perceive Cantonese level tones from multiple speakers with the help of contexts. If they use context cues to guide their tone perception, the same level tone would be perceived as a high-level tone in the context of low pitch and as a low level tone in the context of high pitch. By comparing their perception of Cantonese tones in contexts of different pitch heights, the present study aims to observe if listeners use context cues to normalize lexical tone variability at the behavioral level (Francis et al., 2006; Moore & Jongman, 1997; Wong & Diehl, 2003; K. Zhang et al., 2017). The extrinsic normalization process at the cortical level cannot be easily separated by comparing tone perception in contexts of different pitch heights, since no matter in the high or low contexts, listeners performed the normalization process. However, many studies have replicated that nonspeech contexts with the same F0 information as speech contexts did not affect the perception of Cantonese tones, but speech contexts did, indicating that extrinsic normalization did not show in the nonspeech-context condition but in the speech-context condition at least for Cantonese tone perception. Therefore, the present study adopts the speech-nonspeech comparison method to observe the extrinsic normalization process at the cortical level, which was used in previous EEG studies (e.g., K. Zhang & Peng, 2021, C. Zhang et al., 2013). Since previous EEG studies reported N1 (e.g., Sjerps et al., 2011), P2 (K. Zhang et al., 2023; K. Zhang & Peng, 2021), N400 (e.g., C. Zhang et al., 2013), and LPC (e.g., C. Zhang et al., 2013) components during the extrinsic normalization process, the present study will focus on these components as well. It is hypothesized that at least one ERP component among the four would show in the extrinsic normalization of Cantonese tones. The dual-task paradigm will be used to investigate the effect of cognitive load on extrinsic normalization. To compare with previous studies, a similar visual search task to Bosker et al. (2017) and Reese & Reinisch (2022) is used as the secondary task. Listeners need to finish the Cantonese tone perception task while performing a co-occurring visual search task. Three conditions, no secondary task, easy secondary task, and difficult secondary task, are included to see how the degree of cognitive load modulates extrinsic normalization at the behavioral and cortical levels. It is hypothesized that at the behavioral level, extrinsic normalization of Cantonese tones becomes harder when the cognitive load increases, and that at the cortical level, the amplitude (s) of ERP component(s) indexing extrinsic normalization process changes and the latencies become longer as cognitive load increases.

2. Methods

2.1. Participants

Thirty-two native Hong Kong Cantonese speakers participated in the experiment (Mean age = 20.52; 15 females). They were students at the Hong Kong Polytechnic University during the experiment's timeframe and spent most of their time in Hong Kong before 18 years old. All participants had minimal exposure to professional music training, capping at three years, and self-reported normal hearing, speech, and language abilities. All participants were right-handed, as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). Comprehensive insights into the experimental procedures were provided to all participants, and informed consents were acquired prior to the commencement of the experiment. Participants received appropriate compensation for their time. The study was conducted in adherence to the ethical guidelines approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

2.2. Experimental procedure and stimuli

A dual task paradigm was used to evaluate how cognitive load affects the speech normalization process. The primary task involved a Cantonese word identification task, where participants were required to identify a target word embedded in either speech or nonspeech contexts. The secondary task was a visual search task, in which participants looked for the presence of a specific pattern. Audio stimuli were bilaterally delivered through inserted earphones, and visual stimuli were presented at the center of a monitor. The entire experiment was conducted in a soundproof booth. The experiment consisted of four blocks presented in a counter-balanced order across participants: a nonspeech-context with no secondary task block (NS-N), a speech-context with no secondary task block (SP-N), a speech-context with low-load secondary task block (SP-LL), and a speech-context with high-load secondary task block (SP-HL).

The trial procedure was illustrated in Fig. 1. For blocks without the secondary task, each trial began with a 500-ms fixation, followed by a context stimulus. A target stimulus was then played after a silence interval ranging between 300 ms and 500 ms. A prompt “Which word?” appeared on the screen 350–550 ms following the target stimulus, instructing participants to press the corresponding key to indicate the last word they heard: 醫 (/ji55/, doctor), 意 (/ji33/, meaning), or 二 (/ji22/, two). The next trial was initiated once a response was detected or if the maximum response time of 1500 ms was reached. The trial procedure for the blocks with the secondary task was largely similar except that when presenting the context stimulus, a figure appeared on the screen. Participants were asked to search if there is a white diamond in the figure. After completing the auditory task, another prompt showing on the screen asked participants to indicate whether a white diamond was present by pressing the corresponding key. The experimenter emphasized that both the word identification task and the visual search task were of equal importance; thus, they should attend carefully to both auditory context and the visual figure.

The auditory stimuli for the Cantonese word identification task replicated those used in Tao et al. (2021). Each trial incorporates a context (either speech or nonspeech) and a speech target. The speech context was the Cantonese phrase 呢個字係 (/li55 ko33 tsi22 hɐi22/, meaning “this word is …”), articulated by four native Cantonese speakers with varying pitch ranges [female high (FH), female low (FL), male high (MH), and male low (ML)]. To introduce intra-speaker variability, the F0 trajectories of the original recordings were adjusted three semitones up or down in Praat (Boersma & Weenink, 2023), forming three distinct pitch-height contexts: high-F0, mid-F0, and low-F0 speech context. Nonspeech contexts, created from triangle waves, were manipulated to mirror the pitch heights and durations of the speech counterparts, resulting in three nonspeech contexts. The speech targets consistently utilized the Cantonese syllable 意 (/ji33/, meaning) spoken by the same four speakers. Different from contexts, there was no pitch height manipulation for target speech. The manipulation resulted in 24 contexts (2 sound types × 4 speakers × 3 pitch heights) and four targets (/ji33/ × 4 speakers). When concatenating the context and target stimuli, we ensured that each trial featured stimuli from the same speaker. Cantonese tone T33 is highly ambiguous, especially when presented in isolation under mixed-talker conditions (Peng et al., 2012). However, if listeners use context cues to assist their identification (i.e., the extrinsic normalization), they tend to perceive the target word as /ji55/ in contexts of low F0 and as /ji22/ in contexts of high F0. Therefore, by testing if there is context-dependent perception of target word, we can tell whether listeners engaged in the speech normalization process. To maintain naturalness, context stimuli durations remained unchanged, whereas target stimuli durations were normalized to 450 ms to accommodate EEG signal processing. The intensity of the speech stimuli was set at 55 dB, and nonspeech stimuli at 75 dB to equate perceived loudness, as evaluated by native Cantonese listeners. Specifics on F0s and durations of the auditory stimuli are detailed in Tao et al. (2021). Additional fillers, undergoing similar pitch manipulation as the test stimuli, were also included. The context fillers were Cantonese phrases: 我而家讀 (/ŋo23 ji21 ka55 tuk2/, now I will read…) and 請留心聽 (/tshiŋ25 ləu21 sɐm55 thiŋ55/, please listen to…carefully). The target fillers were Cantonese 意 (/ji33/, meaning) or 二 (/ji22/, two).

The low- and high-load secondary tasks differed in the complexity of the visual stimuli. The visual stimuli used for the visual search task were similar to those in Bosker et al. (2017). There are two types of visual stimuli composed of object grids, containing an equal number of randomly positioned red diamonds, red/white triangles, red/white upside-down triangles, and red/white squares on a black background. A 4 by 4 grid made up the low cognitive-load condition and a grid of 8 by 8 made up the high cognitive load condition (Fig. 2). In half of the trials, one randomly selected object in the grid was replaced by a white diamond.

Each block contained 108 test trials (4 speakers × 3 pitch heights × 9 repetitions) and 36 fillers. Each repetition forms a subblock. Four speakers and three pitch heights were all presented in one subblock and presented in a random order to maximize the effect of speech variability. Before the formal tasks, practices were provided for each subject to familiarize themselves with the experimental procedures. The visual and auditory stimuli used in the practice were not used in the formal task. Speech contexts were four-syllable Cantonese phrases “呢個字係”, while the target syllables were low-, mid-, and high-level tone characters (i.e., “二”, “意”, and “醫”) in the high-, mid-, and low-F0 speech context trials respectively, to reduce ambiguity during the practice.

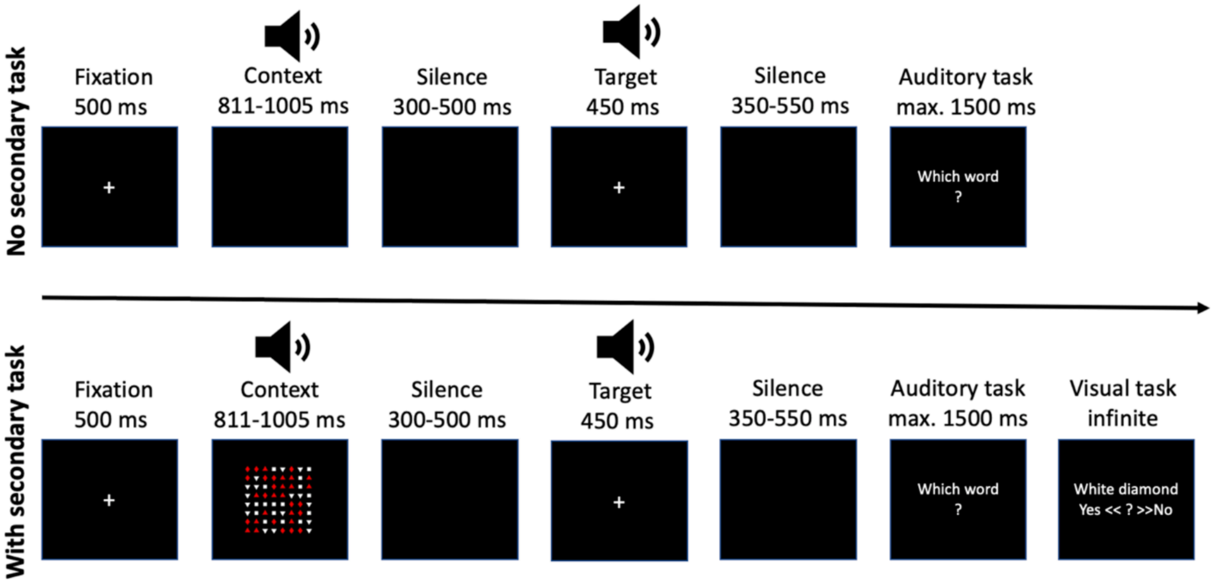


Fig. 1. The trial procedures for the cantonese word identification task.

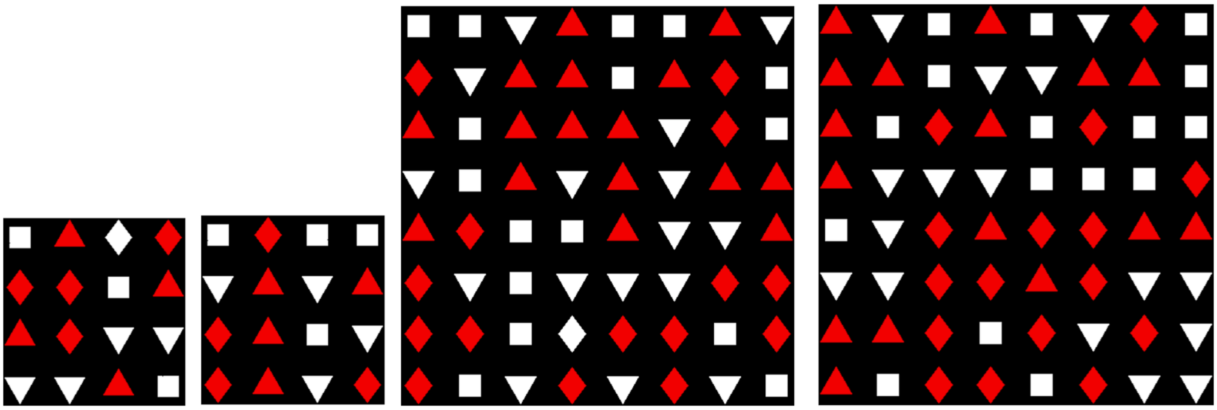


Fig. 2. The pictures used in the visual search task.

2.3. EEG signal recording

EEG signal was recorded using a SynAmps 2 amplifier (NeuroScan, Charlotte, NC, U.S) with a cap carrying 64 Ag/AgCl electrodes placed on the scalp surface at the standard locations according to the international 10–20 system. Two offline reference channels were placed at the left and right mastoids respectively. Two bipolar channels were used to record horizontal and vertical electrooculography (EOG) to monitor the horizontal and vertical eye movements, respectively. Impedance between the online reference electrode (placed between Cz and CPz) and any recording electrode was kept below 5 kΩ for all participants. EEG signals were recorded continuously at the sampling rate of 1000 Hz. The detailed preprocessing of EEG signals was presented immediately before the EEG results report (see section 3.2. for details).

3. Results

3.1. Behavioral data

3.1.1. *The primary task: The Cantonese word identification task*

Participants' response in each experimental condition is plotted in Fig. 3. As can be seen, participants' identification of the target words was almost not affected by the pitch heights of nonspeech contexts, but their responses to the target words changed along with the pitch height of speech contexts no matter with or without secondary tasks. To statistically evaluate how different conditions affect listeners' utilization of context cues to normalize target words, a multinomial logistic regression model was fitted to all participants' responses in the Cantonese word identification task, using the *nnet* package (Venables & Ripley, 2002) in R. Response category (three levels: /ji22/, /ji33/, and /ji55/; with /ji33/ as the reference level) was the dependent variable. *Pitch height* (three levels: low, mid, high; dummy coded with mid as the reference level), *condition* (four levels: NS-N, SP-N, SP-LL, and SP-HL; dummy coded with NS-N as the reference level), and *pitch height* by *condition* interaction were included as predictors. The significance of each predictor was assessed using likelihood ratio tests via the anova() function from the *car* package. The analysis revealed a significant main effect of *pitch height*, χ^2 (4) = 9310, $p < 0.001$, a significant main effect of *condition*, χ^2 (6) = 1025.6, $p < 0.001$, and a significant *pitch height* by *condition* interaction, χ^2 (12) = 2567.2, $p < 0.001$, indicating that the effects of pitch height on the Cantonese word identification were different across four experimental conditions.

To statistically evaluate how different experimental conditions modulated the normalization process, the post-hoc analysis on the *pitch height* by *condition* interaction was conducted using the *emmeans* package (Lenth, 2019) in R. If participants did the normalization process,

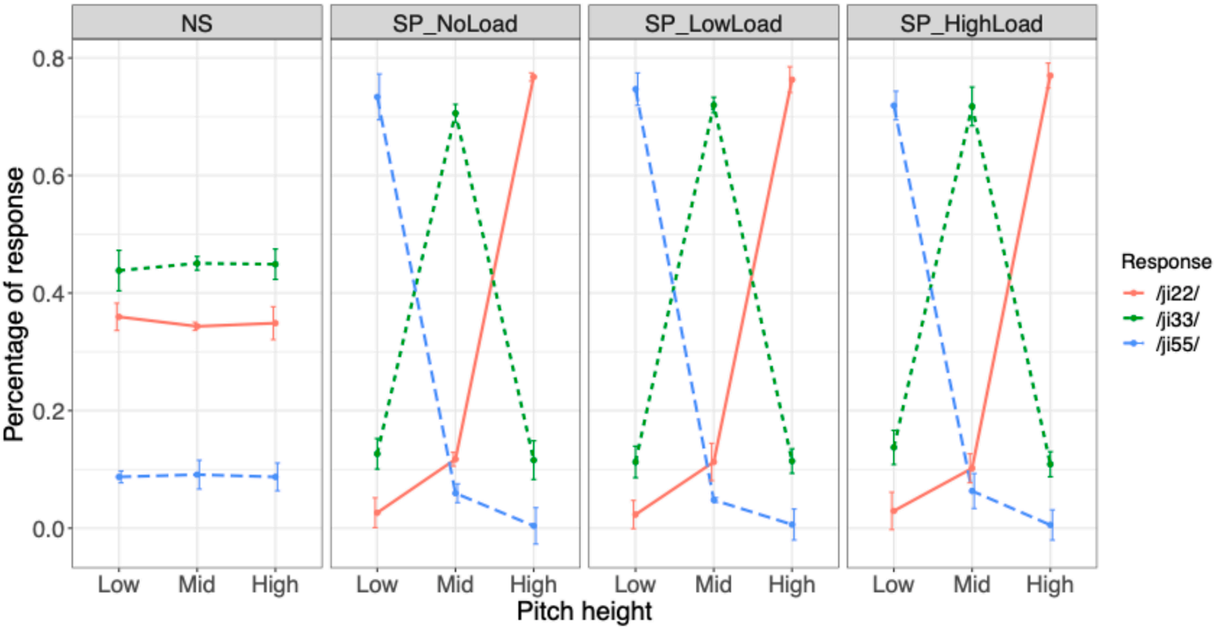


Fig. 3. The percentage of three responses in different experimental conditions.

their word identification should show a context-dependent pattern. Specifically, /ji55/ responses would be most prevalent in low-F0 contexts, /ji33/ responses most prevalent in mid-F0 contexts, and /ji22/ responses most prevalent in high-F0 contexts. Therefore, in the post-hoc analysis, we first compared the expected response with another two alternatives and see if there is a typical context-dependent perception in each condition. Bonferroni method was adopted to correct for the multiple comparison. Table 1 lists the prob and SE of each response in different conditions and the statistics of by-pair contrast. As can be seen, the analysis revealed that in SP-N, SP-LL, and SP-HL conditions, subject gave significantly more /ji55/ responses (see Table 1 for the specific prob and SE for each response and *p* value for each pairwise comparison) in low contexts than /ji33/ and /ji22/; they gave significantly more /ji33/ responses in mid contexts than /ji55/and /ji22/; they gave significantly more /ji22/ responses in high contexts than /ji33/ and /ji55/, indicating that normalization process occurs in SP-N, SP-LL, and SP-HL conditions. However, in the NS condition, /ji33/ response is the most prevalent one in all three pitch heights, indicating no reliable normalization process in the NS condition. The analysis reduplicated the previous finding that the normalization of Cantonese level tones only occurs at the speech contexts but not the nonspeech contexts (e.g., C. Zhang et al. 2013).

We further compared the percentage of expected responses in the same pitch height across SP-N, SP-LL, and SP-HL, three conditions showing significant normalization process, to evaluate if the degree of normalization process varies across three conditions with different cognitive loads. Bonferroni method was adopted to correct for the multiple comparison. The results revealed that in low-F0 contexts, the percentages of /ji55/ response in SP-N, SP-LL, and SP-HL are comparable ($ps = 1$), in mid-F0 contexts, the percentage of /ji33/ response in SP-N, SP-LL, and SP-HL are comparable ($ps = 1$), and in high-F0 contexts, the percentage of /ji22/ response in SP-N, SP-LL, and SP-HL are also comparable ($ps = 1$), indicating that the secondary tasks did not affect the normalization process at the behavioral level.

3.1.2. *The secondary task: The visual search task*

The accuracy rate of the visual search task is plotted in Fig. 4. A mixed-effects linear regression model was fitted on the accuracy of the visual search task to statistically evaluate if the accuracy of the secondary task was affected by the task difficulty. The *pitch height* (three levels: low, mid, and high), *load* (two levels: low and high) and their two-way interaction were included as the fixed effects in the model. Due to convergence issues, only the by-participant intercept and by-speaker intercept were included as the random effects. The *p*-values for each fixed factor were derived using the 'afex' package (Singmann et al., 2015). After identifying significant main effects or interactions, post-hoc pairwise comparisons were conducted using the 'lsmeans' package (Lenth, 2019), applying Tukey adjustment for multiple comparisons. The results suggested that *load* is a significant main effect, χ^2 (1) = 507.85, $p < 0.001$. The visual search accuracy was significantly higher in the low-load condition than in the high-load condition. Neither *pitch height* nor *pitch height* by *condition* interaction was significant in the analysis of the accuracy of the visual search task ($ps > 0.05$).

3.2. *ERP results*

3.2.1. *The preprocessing of EEG signals*

EEG data was processed using custom scripts in MATLAB with functions from EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). The EEG signals were filtered offline with a 0.1 Hz high-pass and a 30 Hz low-pass filters (both slopes = 12 dB/Oct) and re-referenced offline to the average mastoid recordings. Epochs from

Table 1
The results of the post-hoc analysis on the pitch height by context interactions in the Cantonese word identification task.

	Pitch height	Resp.	*prob*	*SE*	*β*	*SE*	*t*	*p*
NS	Low	/ji55/	0.099	0.009				
	Low	/ji33/	0.495	0.015	0.397	0.02	20.339	<0.001
	Low	/ji22/	0.406	0.015	0.308	0.019	16.277	<0.001
	Mid	/ji55/	0.103	0.009	−0.406	0.02	−20.581	<0.001
	Mid	/ji33/	0.509	0.015				
	Mid	/ji22/	0.388	0.014	−0.121	0.028	−4.369	<0.001
	High	/ji55/	0.099	0.009	−0.296	0.019	−15.724	<0.001
	High	/ji33/	0.507	0.015	0.113	0.028	4.072	0.0013
	High	/ji22/	0.394	0.014				
SP-LL	Low	/ji55/	0.846	0.011				
	Low	/ji33/	0.128	0.01	−0.719	0.02	−35.93	<0.001
	Low	/ji22/	0.026	0.005	−0.82	0.013	−61.998	<0.001
	Mid	/ji55/	0.054	0.007	−0.763	0.016	−47.939	<0.001
	Mid	/ji33/	0.818	0.011				
	Mid	/ji22/	0.128	0.01	−0.69	0.02	−33.988	<0.001
	High	/ji55/	0.007	0.002	−0.857	0.011	−78.411	<0.001
	High	/ji33/	0.129	0.01	−0.735	0.02	−36.904	<0.001
	High	/ji22/	0.864	0.01				
SP-N	Low	/ji55/	0.828	0.011				
	Low	/ji33/	0.143	0.01	−0.685	0.021	−32.793	<0.001
	Low	/ji22/	0.03	0.005	−0.798	0.014	−57.633	<0.001
	Mid	/ji55/	0.067	0.007	−0.733	0.017	−43.093	<0.001
	Mid	/ji33/	0.8	0.012				
	Mid	/ji22/	0.133	0.01	−0.667	0.021	−32.297	<0.001
	High	/ji55/	0.004	0.002	−0.861	0.011	−81.451	<0.001
	High	/ji33/	0.13	0.01	−0.735	0.02	−36.909	<0.001
	High	/ji22/	0.865	0.01				
SP-HL	Low	/ji55/	0.812	0.012				
	Low	/ji33/	0.155	0.011	−0.657	0.022	−30.409	<0.001
	Low	/ji22/	0.033	0.005	−0.779	0.014	−54.027	<0.001
	Mid	/ji55/	0.072	0.008	−0.741	0.017	−43.309	<0.001
	Mid	/ji33/	0.813	0.012				
	Mid	/ji22/	0.115	0.009	−0.698	0.02	−35.498	<0.001
	High	/ji55/	0.006	0.002	−0.865	0.011	−81.448	<0.001
	High	/ji33/	0.123	0.01	−0.748	0.02	−38.397	<0.001
	High	/ji22/	0.871	0.01				

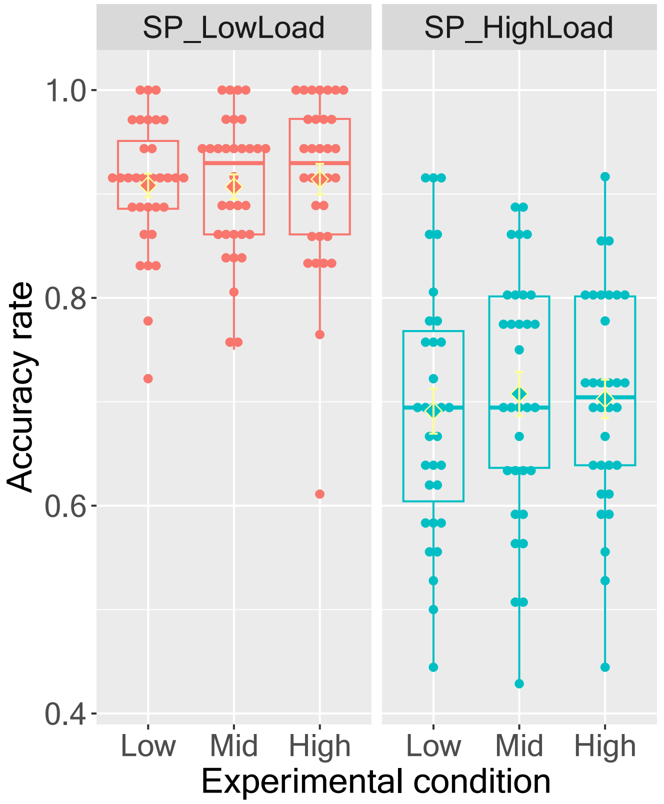


Fig. 4. The accuracy of the secondary task in different experimental conditions.

−100 to 800 ms (time locked to the onset of the target stimulus) were extracted and baseline-corrected based on the −100 – 0 ms pre-target stimulus activity. Any epochs exceeding ± 100 µV at any scalp channels were discarded. Eye blinks were detected automatically by a moving window peak-to-peak threshold criterion on the VEOG data with a threshold of 100 µV, a window size of 200 ms, and a widow step of 50 ms. Horizontal eye movements were detected automatically by a step-like threshold criterion on the HEOG data with a threshold of 40 µV, a window size of 400 ms, and a window step of 10 ms. Participants with less than 70 % accepted epochs in any bins (i.e., 4 conditions × 3 pitch heights) were excluded, resulting in 30 participants for analysis. Overall, the acceptance rate was 95.6 % (SD = 4.1 %) in NS-N, 97.3 % (SD = 3.1 %) in SP-N, 94.2 % (SD = 5.3 %) in SP-LL, and 92.2 % (SD = 5.7 %) in SP-HL.

3.2.2. *The selection of time window and electrodes for each ERP component*

The global field power (Fig. 5, left) was calculated as the root mean square of the ERP voltage and then averaged across the scalp electrodes, three pitch heights, four speakers, four conditions, and 30 subjects. ERP waves at different experimental conditions were plotted in Fig. 6. According to the global field power and the ERP waves, four ERP components were identified during the target tone perception: N1, P2, N400, and LFN. The appearance of N1, P2, and N400 were consistent with our hypothesis and previous reports (Sjerps et al., 2011; K. Zhang & Peng, 2021; C. Zhang et al., 2013), and thus they were included in the further analysis. Although the appearance of LFN was unexpected, a similar component, LPC, was reported by C. Zhang et al., (2013). It was said that LFN and LPC (especially LPC at the parietal area) could be two separate ends of the same dipole (Astle et al., 2008), and thus LFN was included in the analysis as well. The selected time window and electrodes for each ERP component are detailed in Table 2. The time window for each ERP component was identified based on the timeframe within which the ERP component appeared, as discerned through visual inspection of the global field power (refer to the boxes in Fig. 5, left). The electrodes corresponding to each ERP component were chosen based on the topographies where the ERP amplitudes were anticipated to peak (Fig. 5, right).

3.2.3. *ERP amplitude*

The mean amplitude of each ERP in different conditions is visualized in Fig. 7. A linear mixed-effects regression model was applied to the mean ERP amplitude of each component, utilizing the '*lme4*' package (Bates et al., 2015) in R. The ERP wave for each condition was calculated by averaging the EEG epochs over nine repetitions and four speakers in MATLAB. Given that nine repetitions are insufficient to acquire reliable ERPs, the EEG epochs were further averaged across four speakers, as the study did not intend to examine the participants' responses to individual speakers. The mean ERP amplitudes, which were used for model fitting, were further averaged over electrodes for two reasons. Firstly, models with electrodes as a random effect failed to converge. Secondly, given the EEG signals' low spatial resolution, analyses on each electrode might not yield meaningful insights. Consequently, the model was designed to

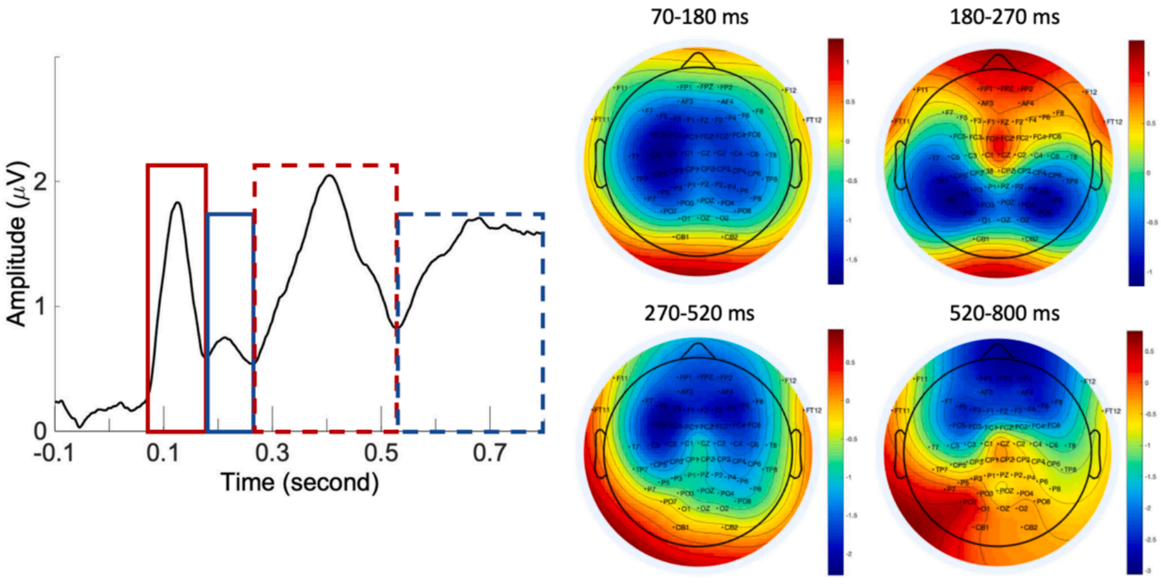


Fig. 5. The global field power (left) and topographies in different time windows (right).

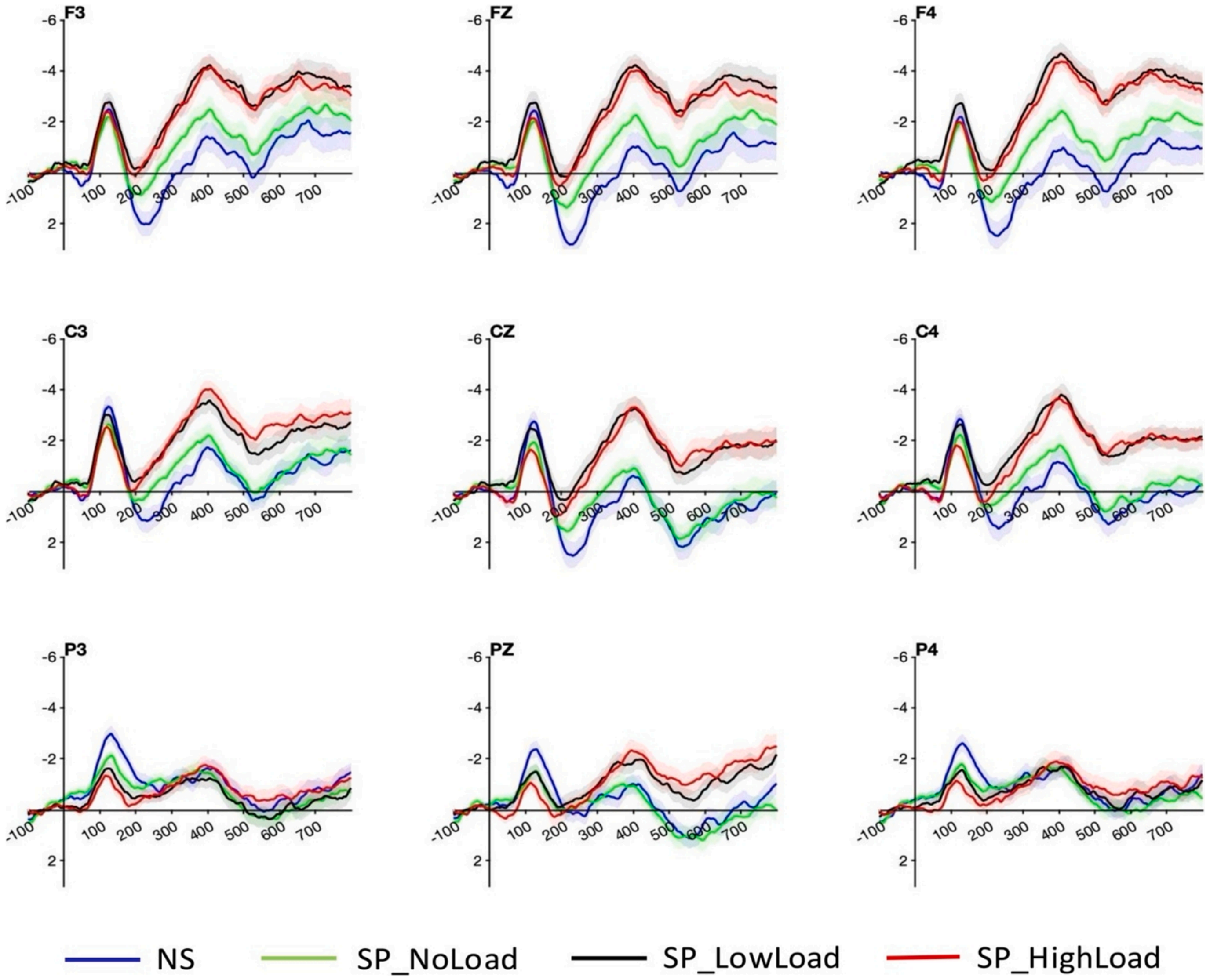


Fig. 6. The ERP waves during the target word perception in different conditions.

Table 2
The time window and electrodes for each ERP component.

ERP	Time window	Electrodes
N1	70–180 ms	F3, F1, FC5, FC3, FC1, C5, C3, C1, CP5, CP3, CP1, P5, P3, FC2, CP4, C4
P2	180–270 ms	FP1, AF3, F1, FC1, FPz, Fz, FCz, Cz, FP2, AF4, F2, FC2, F8,
N400	270–520 ms	F5, F3, F1, FC5, FC3, FC1, C5, C3, Fz, F2, F4, F6, FC4, FC6, AF4
LFN	520–800 ms	F5, FC5, FP1, AF3, F3, FC3, F1, FPz, Fz, FP2, AF4, F2, F4, F6, F8

include only *condition* (NS-N, SP-N, SP-LL, and SP-HL), *pitch height* (high, mid, low), and their two-way interactions as fixed effects. Due to convergence issues, only the by-participant intercept was included as a random effect in each model. The *p*-values for each fixed factor were derived using the 'afex' package (Singmann et al., 2016). After identifying significant main effects or interactions, post-hoc pairwise comparisons were conducted using the 'lsmeans' package (Lenth, 2019), applying Tukey adjustment for multiple comparisons. The results were summarized below.

N1: There was a significant main effect of *condition*, χ^2 (3) = 10.34, $p = 0.016$. However, the post-hoc analysis revealed that N1 amplitudes in the four conditions were not significantly different from each other ($p > 0.05$). *Pitch height* ($p = 0.908$) and *condition* by *pitch height* interaction ($p = 0.937$) were not significant.

P2: There was a significant main effect of *condition*, χ^2 (3) = 78.17, $p < 0.001$. The P2 amplitude was significantly higher in nonspeech contexts (M = 1.93, SE = 0.274) than in the other three conditions (SP-N: M = 0.768, SE = 0.23; SP-LL: M = − 0.465, SE = 0.254; SP-HL: M = − 0.164, SE = 0.298, *ps* < 0.001). The P2 amplitude was also significantly higher in SP-N than in SP-LL and SP-HL (*ps* < 0.01). The P2 amplitudes were comparable in SP-LL and SP-HL ($p = 0.719$). *Pitch height* ($p = 0.372$) and *condition* by *pitch height* interaction ($p = 0.945$) were not significant.

N400: There was a significant main effect of *condition*, χ^2 (3) = 145.44, $p < 0.001$. The post-hoc analysis revealed that NS-N (M = 0.056, SE = 0.257) triggered significantly smaller N400 compared with the other three conditions (SP-N: M = − 0.86, SE = 0.229; SP-LL: M = − 2.5, SE = 0.204; SP-HL: M = − 2.41, SE = 0.224; *ps* < 0.001). The N400 amplitude was significantly smaller in SP-N than in SP-LL and SP-HL (*ps* < 0.01). SP-LL and SP-HL showed no significant difference ($p = 0.98$). *Pitch height* ($p = 0.78$) and *condition* by *pitch height* interaction ($p = 0.925$) were not significant.

LFN: There was a significant main effect of *condition*, χ^2 (3) = 83.22,

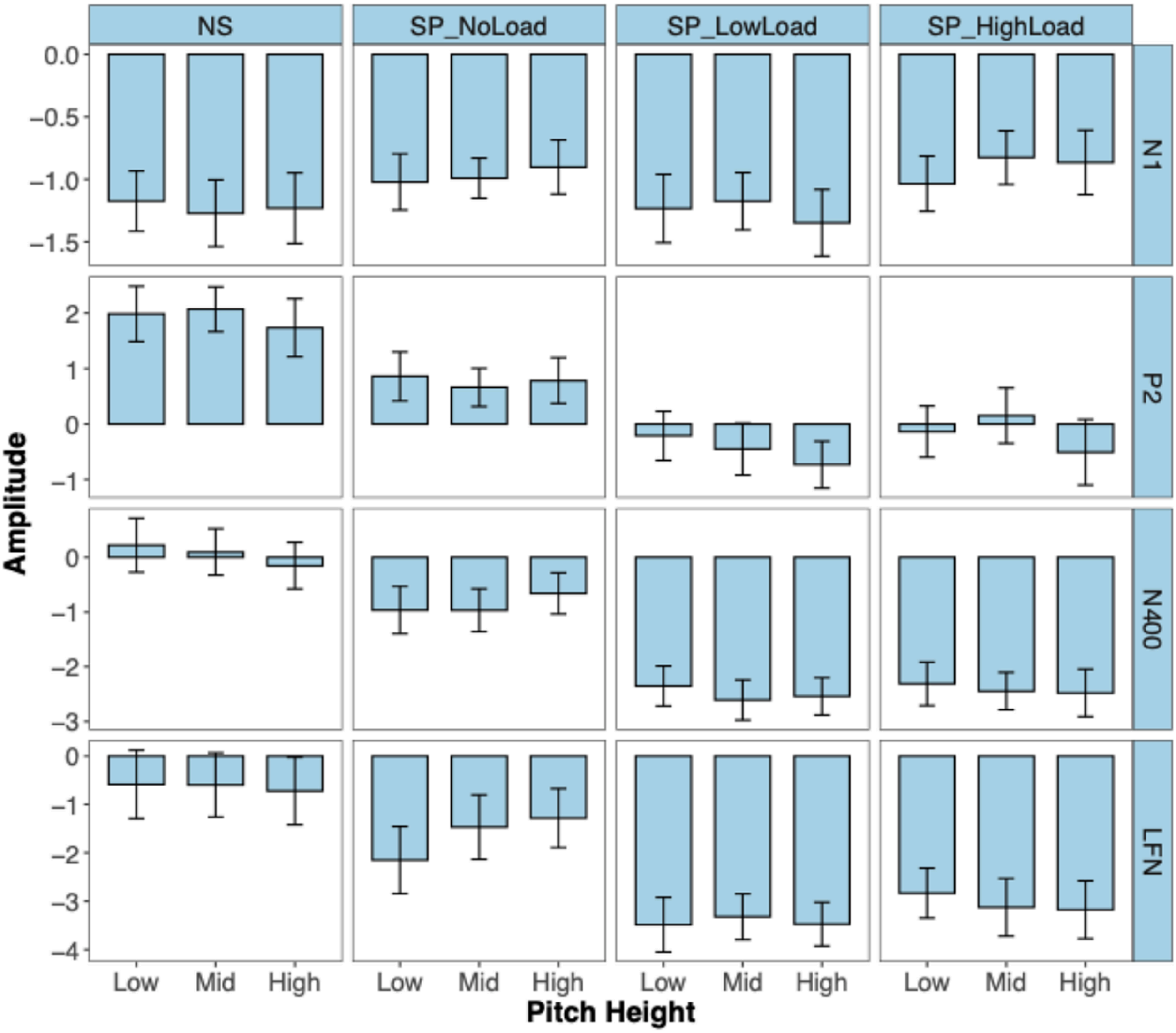


Fig. 7. The amplitude of each ERP component in different conditions.

$p < 0.001$. The NS-N (M = − 0.634, SE = 0.394) triggered significantly smaller LFN compared with the other three conditions (SP-N: M = − 1.63, SE = 0.376, $p = 0.014$; SP-LL: M = − 3.43, SE = 0.285, $p < 0.001$; SP-HL: M = −3.05, SE = 0.325, $p < 0.001$). The LFN amplitude was significantly smaller in SP-N than in SP-LL or SP-HL (*ps* < 0.001). SP-LL and SP-HL showed no significant difference ($p = 0.654$). *Pitch height* (p = 0.882) and *condition* by *pitch height* interaction (p = 0.811) were not significant.

3.2.4. ERP latency

The peak latency of each component, for every participant, was determined as the time point corresponding to either the minimal (for N1, N400, and LFN) or maximal (for P2) point of the 2nd-order polynomial curve fitted to the ERP wave. The peak latency of each ERP component in different experimental conditions is illustrated in Fig. 8. The statistical approach employed for analyzing ERP latency mirrored the one used for ERP amplitude, incorporating *pitch height*, *condition*, and their two-way interaction as fixed effects, and by-participant intercept as the random effect within each linear mixed-effects model. The results were presented subsequently.

P2: There was a significant main effect of *condition*, χ^2 (3) = 7.94, p = 0.047. However, the post-hoc analysis revealed that the latencies in four conditions were not significantly different from each other (*ps* > 0.05). *Pitch height* (p = 0. 709) and *condition* by *pitch height* interaction (p = 0.695) were not significant.

N400: There was a significant main effect of *condition*, χ^2 (3) = 10.37, $p = 0.016$. The post-hoc analysis only revealed a significant difference between NS-N (M = 0.397, SE = 0.007) and SP-HL (M = 0.418, SE = 0.005, $p = 0.023$). The latencies in other conditions were not significantly different from each other (*ps* > 0.05). *Pitch height* (p = 0.998) and *condition* by *pitch height* interaction (p = 0.15) were not significant.

No significant main effects or interactions were found in the analysis of N1 and LFN latencies.

In sum, the behavioral results suggested that listeners can successfully adapt to speaker variability when contexts are speech no matter in single- or dual- task conditions, but they cannot use cues from nonspeech contexts to assist their Cantonese word identification. More importantly, their behavioral performance of speaker adaptation was similar in speech contexts with and without secondary tasks. Subjects performed significantly poorer in the visual search task of high cognitive load. However, different from the behavioral results, the analysis of the EEG signals revealed that at the cortical level, listeners performed differently in the single- and dual- task conditions. Specifically, subjects showed reduced P2 and enhanced N400 and LFN amplitudes in the dual-task speech-context conditions compared with the single-task speech-context condition. However, the N1 components were similar across different experimental conditions.

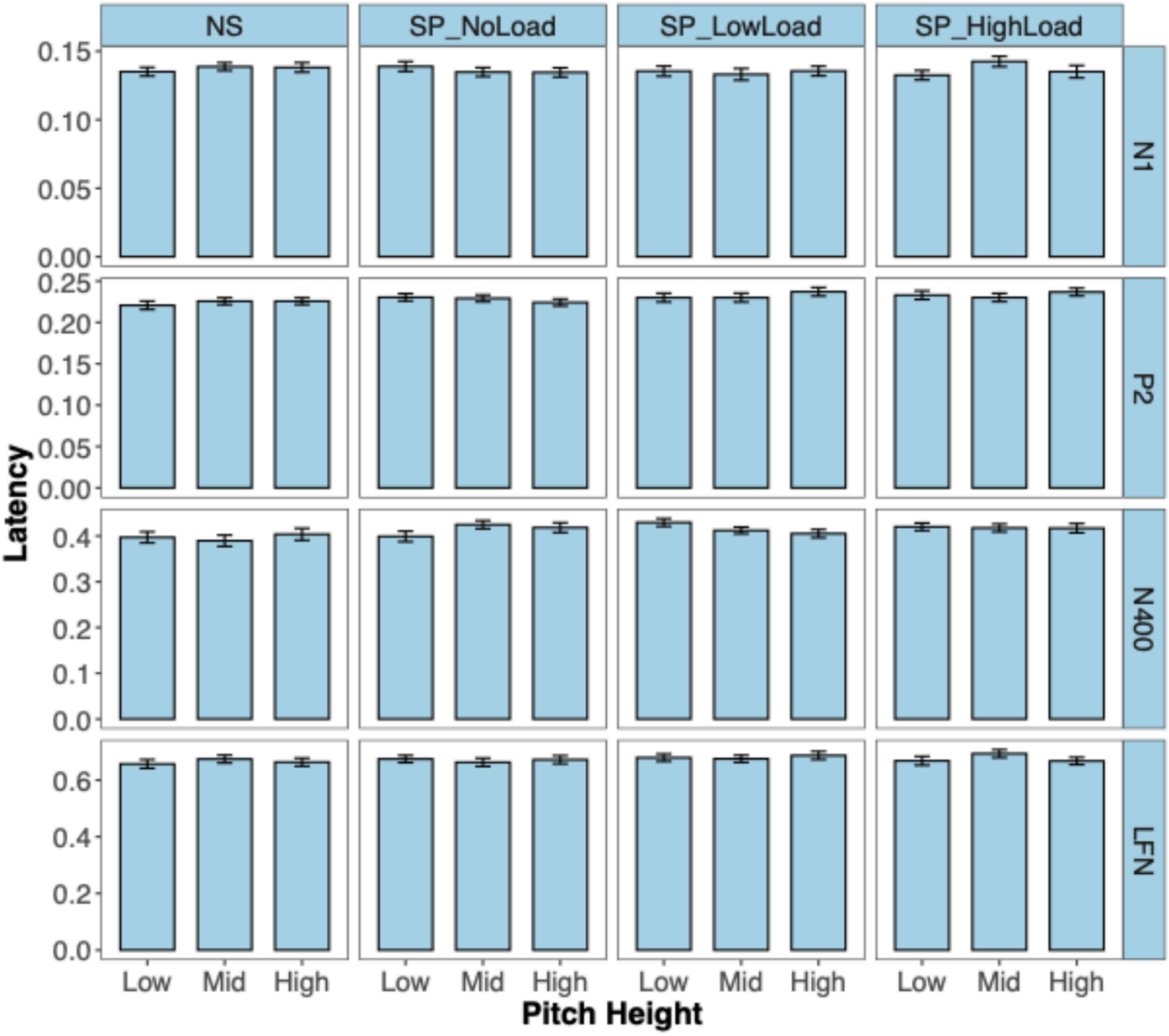


Fig. 8. The latency of each ERP component in different conditions.

4. Discussion

To investigate the neural mechanisms underlying the effect of cognitive load on extrinsic normalization process, especially for suprasegmental components, the present study recorded EEG signals while Cantonese listeners perceived Cantonese level tones from four different speakers with speech and nonspeech contexts in single- and dual- task conditions. Four ERP components emerged during the target tone perception, indicating a much more complex online process of extrinsic normalization than reported by previous EEG studies which frequently allocated the time locus of the extrinsic normalization process to a single ERP (e.g., Sjerps et al., 2011). The effects of cognitive load were also observed in the analysis of several ERP components. Although the statistical analysis incorporated all four experimental conditions into a single regression model to avoid multiple comparison issues, we adopted a two-step strategy to interpret the results. As mentioned in the Introduction, we first identified the ERP components that were related to normalization process by comparing EEG signals in speech and nonspeech contexts. Then, we investigated how cognitive load modulates speech normalization by comparing these ERPs related to normalization process across speech contexts with or without secondary tasks.

4.1. *The multiple stages of the extrinsic normalization process*

Listeners in the present study could use immediate pitch cues in speech contexts to guide their perception of target tones, regardless of the secondary task. Specifically, they gave more /ji55/ responses in low-F0 speech contexts, more /ji33/ responses in mid-F0 speech contexts, and more /ji22/ responses in high-F0 speech contexts, showing a typical context-dependent perception of Cantonese tones in speech contexts. However, their perception of target tones was almost unaffected by nonspeech contexts, as reflected by the similar perceptual heights in nonspeech contexts with different pitch heights. Therefore, the behavioral results suggested that successful speaker normalization only occurred in speech contexts but not in nonspeech contexts, replicating the speech-specific context effect on speaker normalization (Chen et al., 2023; Francis et al., 2006; C. Zhang et al., 2012; K. Zhang et al., 2017). The speech-specific context effect indicated that the extrinsic normalization process at the cortical level can be observed by contrasting the EEG signals elicited by Cantonese tone perception in speech and nonspeech contexts.

At the cortical level, the Cantonese tone perception did show significant differences in speech and nonspeech contexts. Although the present study used three speech-context blocks (i.e., SP-N, SP-LL, and SP-HL), the speaker normalization process at the cortical level was identified by contrasting NS-N with SP-N only to make the experimental conditions comparable. The ERP analysis suggested that the Cantonese tone perception in SP-N started to differ from that in NS-N from 180 ms after the stimulus onset and continued to the end of the ERP analyzing time window, covering P2, N400, and LFN. However, Cantonese tone perception triggered statistically comparable N1 in NS-N and SP-N. Considering that speech-nonspeech difference reflected the extrinsic normalization process, the ERP pattern suggested that extrinsic normalization was mainly implemented in the P2, N400, and LFN time windows, and that no significant extrinsic normalization process was observed in the N1 time window. The ERP results from the present study provided at least two important insights. First, extrinsic normalization is less likely a pure acoustic level process. N1 is mainly affected by the stimuli acoustic properties such as frequency and intensity, and thus it is regarded as reflecting the acoustic process of auditory signals (Picton et al., 1978). The absence of N1 difference in NS-N and SP-N indicated that extrinsic normalization is less likely implemented at the acoustic processing stage. Second, the significant speech-nonspeech difference in the P2, N400, and LFN time windows suggested that extrinsic normalization is more likely a joint effect of multiple adjustments at different speech perception stages.

Previous studies about the auditory word identification reported that P2 is associated with the phonological process (Cheng et al., 2014; Landi et al., 2012; C. Zhang et al., 2015). For example, P2 amplitudes differed in perceiving real words compared with phonotactically illegal nonwords (Cheng et al., 2014). BA 22, one of the cortical origins of P2 (Godey et al., 2001), showed selective activation in processing phonemic features (Mesgarani et al., 2014). Therefore, the significant speech-nonspeech difference in the P2 time window in the present study suggested that listeners adjust to the speaker differences with context cues in the phonological process stage. When listeners need to map the decoded acoustic signals to phonemes, they refer to the context cues to make the perceptual adjustment to facilitate this process.

N400 is often observed when the semantic meaning violates the expectation built from the contexts (Kutas & Federmeier, 2011). Lau et al. (2009) observed similar N400 components for the word identification in different contexts (e.g., sentence frame vs. prime word contexts; incongruent sentence endings vs. unrelated word pairs). Considering that different contexts should require different levels of integration, Lau et al. (2009) further clarified that N400 observed in the semantic violation conditions reflected the *access* stage of stored lexical information, rather than the *integration* stage of context and target lexical information. The significant speech-nonspeech difference in the N400 time window in the present study indicated that listeners made further perceptual adjustments with context cues in the lexical retrieval stage. When listeners retrieve a label from the mental lexicon to match the perceived signals, they again refer to the context cues to facilitate their lexical retrieval process.

Different from C. Zhang et al. (2013), a significant LFN instead of LPC emerged in the lexical tone normalization process in the present study. LFN and late parietal positivity (i.e., the LPC at the parietal area) could be the two separate ends of the same dipole, and thus they may index similar underlying cognitive processes (Astle et al., 2008). LFN was frequently observed in the task-switching paradigm. Switching to a new rule elicits a larger LFN, and thus LFN reflects cognitive flexibility (Rahman et al., 2022; Waxer & Morton, 2011). Rahman et al. (2022) also observed that LFN was significantly correlated with response time. Trials with later responses also showed more negative LFN, indicating that LFN was strongly related to decision-making. The significant speech-nonspeech difference in the LFN time window in the present study suggested that even at the final decision-making stage, subjects still used context cues to make perceptual adjustments to the perceived signals.

Taken together, the ERP pattern in the present study suggested that listeners first decoded the acoustic cues of the target speech in the N1 time window, and when they need to interpret the acoustic cues, for example giving them linguistic labels, they would use all the information available including context cues to assist this interpretation. Since the language-specific process started from the phonemic level, the first noticeable context effect was observed in P2 which is usually related to phonological processing. More importantly, listeners likely used context cues to make perceptual adjustments to the perceived signals at multiple stages before giving the overt response, as indexed by the speech-nonspeech context differences in multiple consecutive ERP components (i.e., P2, N400, and LFN). The multiple-stage adjustments indicate that listeners most likely maximize the usage of context cues during the Cantonese tone perception to make the most optimal response. This is the first study to provide neural evidence supporting a multi-stage adjustment model for the extrinsic normalization process, encompassing the early phonological processing stage, the intermediate lexical retrieval phase, and the final decision-making stage. This EEG result is largely consistent with the computational modeling study which found that a combined computation model involving perceptual adjustments at multiple speech processing stages matches best with the listeners' speaker adaptation data (X. Xie et al., 2023).

4.2. *The effects of cognitive load on the extrinsic normalization process*

The investigation of the effect of cognitive load on extrinsic normalization is important to understand how the brain functions in real-life listening conditions which involve both multiple speakers and multitasking. To answer this question, the present study asked listeners to identify the Cantonese level tones from four different speakers with context cues in either single- or dual-task conditions. However, different results were observed at the behavioral and cortical levels. At the behavioral level, the Cantonese tone perception was significantly affected by speech contexts no matter in the single- or dual- task condition, indicating a successful speaker normalization process. More importantly, listeners' Cantonese tone perception in the single-task condition and that in the dual-task conditions were comparable, and even the more difficult visual search task did not significantly affect the Cantonese tone perception compared with the single-task condition. Therefore, the behavioral results suggested that extrinsic normalization of Cantonese tones was not affected by the co-occurring visual search task, which was consistent with previous behavioral studies using similar experiment paradigms (e.g., Bosker et al., 2017; Reese & Reinisch, 2022). Although compared with vowels and consonants used in Bosker et al. (2017) and Reese & Reinisch (2022), the perception of Cantonese level tones relies on more extrinsic contexts (K. Zhang et al., 2018), the present study failed to find a significant effect of cognitive load on listeners' utilization of context cues at the behavioral level, indicating that the effect of cognitive load on extrinsic normalization might not vary across segmental and suprasegmental components.

However, we cannot simply conclude that extrinsic normalization is an automatic process, since at the cortical level, a significant modulation of the visual search task on the extrinsic normalization process was observed. P2, N400, and LFN, ERP components indexing extrinsic normalization in the present study, showed significant differences in the single- and dual- task blocks. The P2 amplitude was significantly smaller in the dual-task block than in the single-task block. Meanwhile, the amplitudes of N400 and LFN were significantly larger in the dual-task block than in the single-task block. N1, which was not notably affected by extrinsic normalization, was not affected by the visual search task either. The presence of a visual search task inherently increases cognitive demand compared to the no-task condition. In the present study, participants responded first to the auditory target word and then to the visual task. Consequently, during the auditory target word listening period—the window used for EEG analysis—participants were still internally managing the visual task (e.g., memorizing, recalling, and preparing a response for the upcoming visual task). These internal processes consume cognitive resources that would otherwise be available for processing the auditory target, thus increasing the overall cognitive load. Meanwhile, the pictures, which were only presented during the context phase, likely disrupts the extraction of critical context cues necessary for lexical tone normalization. Cognitive load theory posits that incomplete information increases mental effort, as listeners must compensate by inferring or reconstructing missing cues (Sweller, 2011). This additional processing demand further elevates cognitive load. When cognitive load increased (relative to no secondary task condition), three ERP components, reflecting distinct stages of normalization process, changed accordingly. Therefore, we interpret these changes as evidence that cognitive load modulates the normalization process at the cortical level. Besides, we cannot entirely rule out the possibility that attention contributed to the observed effects as well, although several measures were taken to minimize its influence. Specifically, pictures were presented exclusively during the context stimulus phase; no visual stimuli appeared during the target word presentation; and the ERP analysis was time-locked to the onset of the target. These strategies reduce the likelihood that participants' attention was directed toward the pictures during target processing. Nevertheless, it is possible that the allocation of attention to the visual stimuli during the context phase leads to less available contextual information for target normalization, which may have contributed to the observed ERP differences as well.

The ERP difference was only observed in conditions with or without secondary tasks, but two secondary tasks, differing in difficulty, showed similar effects on the normalization process at the cortical level. This might suggest that the cognitive load imposed by our visual search tasks may not have varied sufficiently to differentially impact normalization. A careful inspection of the visual search task provided evidence for this interference. The figures in the visual search task were only presented during the context phase, and when processing the target word, both easy and difficult secondary tasks imposed similar cognitive load by requiring internal management of visual task processing (e.g., memory of the decision and response preparation). This resulted in statistically comparable effects on the cortical normalization process. If we had used secondary tasks, such as digital recall tasks, with varying cognitive loads, we likely would have observed corresponding effects of task difficulty on the normalization process. Although the critical load effect is not captured by the distinction between easy and hard secondary tasks due to the above-mentioned reason, the presence versus absence of a secondary task indeed revealed a significant cognitive load effect.

Overall, the results suggested that extrinsic normalization was affected by the visual search task at the cortical level, but such effects were not observed at the behavioral level. It is important to note that behavioral measures and ERP signals capture different aspects of speech processing. Behavioral data reflect the final decision outcome, whereas ERPs index the dynamic, stage-by-stage processing leading up to that outcome. The normalization process, as a multistage processes, may exhibit compensatory mechanisms across stages, stabilizing behavioral outcomes despite variations in intermediate processing. Thus, the absence of behavioral differences does not preclude the modulation of cognitive load on the normalization process at the cortical level.

The ERP results suggested that all time windows related to the extrinsic normalization process were modulated by increased cognitive load introduced by the visual search tasks. P2, as an auditory evoked potential, was related to the phonological processing (Cheng et al., 2014; Landi et al., 2012; Mesgarani et al., 2014; C. Zhang et al., 2015). Directing attention toward auditory stimuli enhanced the P2 amplitude (Picton et al., 1971). Therefore, the smaller P2 in the dual-task condition in the present study might be due to the divided attention caused by the co-occurring visual search task. Amplitude reductions might reflect smaller post-synaptic potentials in the same neurons and/or activation of fewer neurons in a population (Kutas & Federmeier, 2011). Fewer neuron activations might result in a shallower phonological processing at the P2 time window. It can be deduced that listeners in the dual-task blocks in the present study cannot be fully engaged in the perceptual adjustment with context cues at the phonological processing stage due to the divided attention.

N400 reflects lexical retrieval (Lau et al., 2009), and its amplitude becomes smaller when lexical retrieval is easier (Kutas & Federmeier, 2011). The larger N400 amplitude in the dual-task blocks in the present study indicated that even if behavioral accuracy is maintained, the underlying neural processes of lexical retrieval may be taxed more heavily under the dual-task condition than the single-task condition. It is possible that due to the secondary task, listeners do not have enough cognitive resources to use context cues effectively to facilitate lexical retrieval. Another potential reason for the lexical retrieval difficulty might be the shallower phonological processing in the P2 time window. That is, listeners in the dual-task blocks cannot use context cues to adjust for speaker variability in the phonological processing stage as successfully as they did in the single-task block, and thus they might not be able to give a clear phonological label to the perceived signal. The ambiguous phonological labels in turn made lexical retrieval more difficult in the N400 time window. This possibility was somewhat supported by the correlation analysis between P2 amplitude reduction and N400 amplitude increasement. We first calculated the amplitude difference between SP-N and SP-LL, and between SP-N and SP-HL for both P2 and N400 to index the cognitive-load manipulations on P2 and N400. The Pearson correlation analysis suggested that the P2 amplitude difference was significantly correlated with the N400 amplitude difference (SP-LL: $r = 0.85$, $p < 0.001$; SP-HL: $r = 0.8$, $p < 0.001$). The greater the reduction in P2 amplitude in the dual-task blocks compared to the speech-context single-task block (i.e., SP-N), the larger the increase in N400 amplitude, suggesting that the phonological process constrained the lexical retrieval process in the present study.

LFN, most frequently observed in the task-switching paradigm, reflects cognitive flexibility in response adjustment and is closely related to the decision-making process (Rahman et al., 2022; Waxer & Morton, 2011). In the task-switching paradigm, switching to a new rule, which is more challenging than following the old rule, elicits a larger LFN. The enhanced LFN in the dual-task condition in the present study suggested that additional neural resources are required to make the final perceptual adjustment with context cues due to the distraction of the secondary task. Although this increased neural processing need did not translate to a deficit in behavioral performance, it signaled a more challenging normalization process during the decision-making stage at the cortical level. Aside from the distraction from the secondary task, the difficulty in making perceptual adjustments in the LFN time window could also be partially caused by the shallower phonological processing in the P2 time window. Similarly, we also calculated the LFN amplitude differences between SP-N and SP-LL, and between SP-N and SP-HL to index the load modulation on LFN. Significant correlations between the P2 amplitude reduction and LFN amplitude increase were observed (SP-LL: $r = 0.59$, $p < 0.001$; SP-HL: $r = 0.64$, $p < 0.001$). The larger the P2 amplitude reduction due to the secondary task, the larger the LFN amplitude increase, indicating that the final decision making was constrained by the early phonological processing as well.

In summary, the present study is the first to investigate how cognitive load modulates the extrinsic normalization process at the cortical level. The ERP differences observed between the single- and dual- task conditions suggest that the increased cognitive load introduced by visual search tasks continuously influences the extrinsic normalization process at the cortical level, starting from the early phonological processing stage and continuing to the final decision-making stage. Cognitive load made listeners less engaged in the perceptual adjustment with context cues at the phonetic/phonological stage. Furthermore, due to the increased cognitive load, perceptual adjustment at the lexical retrieval stage became more difficulty and required greater cognitive effort to adapt at the decision-making stage. Additionally, a shallower phonological process might further increase the difficulty in lexical retrieval and decision making. These EEG results complement previous behavioral studies, offering deeper insight into the neural dynamics involved in the management of talker variability under multitasking conditions. Although the interpretations of our EEG findings are grounded in established literature, EEG data alone cannot definitively quantify the "difficulty" of cortical processing. Future studies employing methods capable of directly measuring metabolic or energy consumption during task performance could further clarify the relationship between these ERP modulations and the underlying neural processing demands.

4.3. The insignificant effects of cognitive load on extrinsic normalization at the behavioral level

Although cognitive load affected extrinsic normalization at the cortical level, behaviorally listeners still successfully adjusted speaker differences in Cantonese level tones with context cues, which might be partially due to the invulnerable process of context speech. The ERP analysis revealed comparable N1s in the single- and dual- task conditions, suggesting that the acoustic process of the target speech was not affected by the secondary task. The unaffected N1 also means that the acoustic process of speech signals is an automatic process (Näätänen et al., 1978; Näätänen & Picton, 1987). Consequently, the acoustic process of the context speech was probably not or only mildly affected by the secondary task. In such a condition, the acoustic information, such as the speakers' F0 range conveyed by speech contexts, could be used for perceptual adjustments at the later speech processing stages (i.e., the P2, N400, and LFN time windows). This might be one of the reasons why the present study and previous studies with similar experiment designs failed to find a significant effect of cognitive load on extrinsic normalization at the behavioral level (e.g., Bosker et al. 2017; Reese & Reinisch, 2022).

The insignificant effect of cognitive load on extrinsic normalization at the behavioral level may also lie in the target processing stage. Aside from the context cue extraction, the perceptual adjustment at the target processing stage is another important part of extrinsic normalization. Most studies presented the distractors only in the context processing stage, and subjects needed to decide whether the distractors have the designated properties (e.g., with/without a white diamond in the present study). In most cases, the response to the secondary task is after the word identification task. Therefore, subjects mainly need to keep an answer to the secondary task (yes or no) during the target speech perception. Probably, some subjects could recall the visual stimuli or make response preparation, which may affect the target speech perception as well. The storage of an answer in the working memory and the potential processing lagging of the secondary task did significantly affect the online target speech processing, as reflected by the smaller P2, larger N400, and larger LFN in the present study. However, they might not be strong enough to totally block the perceptual adjustments during the target speech perception. Feng et al. (2021) presented six graphic symbols sequentially in the secondary task and asked subjects to recall their orders. Subjects in Feng et al. (2021) had to keep six meaningless symbols in their working memory, which required much more cognitive resources than the visual search tasks in the present study, and indeed Feng et al. (2021) observed a significant effect of cognitive load on Mandarin tone perception. Therefore, it is possible that if we change to a secondary task that consumes more cognitive resources, we may observe an impaired extrinsic normalization of Cantonese tones at the behavioral level as well.

SP-LL and SP-HL conditions showed no significant difference at the behavioral and cortical levels, which might be due to the above-mentioned two reasons as well. First, considering the automaticity of acoustic processing, the acoustic cues of the context speech can be extracted no matter in the high-load or low-load blocks. Besides, as discussed in Section 4.2, easy and difficult visual search tasks essentially imposed similar cognitive load, such as memorizing a yes/no decision or internally preparing responses for the secondary task. Future studies may employ secondary tasks, such as digital recall, with varying cognitive loads to assess how different loads influence the normalization process.

5. Conclusion

The present study offers novel insights into how cognitive load influences the extrinsic normalization of Cantonese tones from a neurophysiological perspective. Our results demonstrate that extrinsic normalization is not a singular event but involves a series of processing stages—from initial phonological encoding to higher-level decision making—as evidenced by the sequential activation of P2, N400, and LFN components. This finding provides the first neural evidence for the multiple-stage model of the extrinsic normalization process. Importantly, we show that cognitive load exerts its influence across all of these stages, a nuance that has not been captured in previous behavioral studies. Under dual-task conditions, the observed ERP pattern, a diminished P2 response alongside augmented N400 and LFN amplitudes, suggests that the secondary task impairs the efficiency of online perceptual adjustments, reinforcing the idea that extrinsic normalization is an actively controlled process. Although cognitive load did not alter the acoustic processing of speech signals (as evidenced by consistent N1 responses across conditions), it had a pronounced impact on higher-level speech processes that rely on working memory, including phonological processing, lexical retrieval, and decision-making. These distinctions reinforce the notion that extrinsic normalization is sensitive to the availability of cognitive resources. Future studies need to be carried out to test whether more demanding distractor tasks might influence speaker normalization at the behavioral level.

CRediT authorship contribution statement

Kaile Zhang: Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Gang Peng:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was supported by the National Natural Science Foundation of China (NSFC: 12304526), the Departmental General Research Fund from the Hong Kong Polytechnic University (P0052546), and the Research Grants Council of Hong Kong (GRF: 15607518)

Data availability

Data will be made available on request.

References:

Aravamudhan, R., Lotto, A. J., & Hawks, J. W. (2008). Perceptual context effects of speech and nonspeech sounds: The role of auditory categories. *The Journal of the Acoustical Society of America, 124*(3), 1695–1703. https://doi.org/10.1121/1.2956482

Astle, D. E., Jackson, G. M., & Swainson, R. (2008). Fractionating the cognitive control required to bring about a change in task: A dense-sensor event-related potential study. *Journal of Cognitive Neuroscience, 20*(2), 255–267. https://doi.org/10.1162/jocn.2008.20015

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

Behrmann, M., Geng, J. J., & Shomstein, S. (2004). Parietal cortex and attention. *Current Opinion in Neurobiology, 14*(2), 212–217. https://doi.org/10.1016/j.conb.2004.03.012

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*(3), 330–343. https://doi.org/10.1016/j.actpsy.2010.03.006

Boersma, P., & Weenink, D. (2023). *Praat: doing phonetics by computer [Computer program]. Version 6.3.09*, retrieved 2 March 2023 from http://www.praat.org/.

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language, 94*, 166–176. https://doi.org/10.1016/j.jml.2016.12.002

Chen, F., Zhang, K., Guo, Q., & Lv, J. (2023). Development of achieving onstancy in lexical tone identification with contextual cues. *Journal of Speech, Language, and Hearing Research, 66*, 1148–1164. https://doi.org/10.1044/2022_JSLHR-22-00257

Cheng, X., Schafer, G., & Riddel, P. M. (2014). Immediate auditory repetition of Words and Nonwords: An ERP study of lexical and sublexical processing. *PLoS ONE, 9*(3), Article E91988. https://doi.org/10.1371/journal.pone.0091988

Chiu, F., Rakusen, L. L., & Mattys, S. L. (2019). Cognitive load elevates discrimination thresholds of duration, intensity, and f for a synthesized vowel. *The Journal of the Acoustical Society of America, 146*(2), 1077–1084. https://doi.org/10.1121/1.5120404

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

Feng, Y., Meng, Y., Li, H., & Peng, G. (2021). Effects of cognitive load on the categorical perception of mandarin tones. *Journal of Speech, Language, and Hearing Research, 64* (10), 3794–3802. https://doi.org/10.1044/2021_JSLHR-20-00695

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America, 119*(3), 1712–1726. https://doi.org/10.1121/1.2149768

Godey, B., Schwartz, D., De Graaf, J. B., Chauvel, P., & Liégeois-Chauvel, C. (2001). Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: A comparison of data in the same patients. *Clinical Neurophysiology, 112*(10), 1850–1859. https://doi.org/10.1016/S1388-2457(01)00636-8

Heinrich, A., Ferguson, M. A., & Mattys, S. L. (2020). Effects of Cognitive Load on Pure-Tone Audiometry Thresholds in Younger and Older Adults. *Ear & Hearing, 41*(4), 907–917. https://doi.org/10.1097/AUD.0000000000000812

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *The Journal of the Acoustical Society of America, 109*(2), 764–774. https://doi.org/10.1121/1.1339825

Johnson, K. (2005). Speaker Normalization in Speech Perception. In D. B. Pisoni, & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Blackwell Publishing. https://doi.org/10.1002/9780470757024.ch15.

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication, 77*, 84–100. https://doi.org/10.1016/j.specom.2015.12.005

Kasper, R. W., Cecotti, H., Touryan, J., Eckstein, M. P., & Giesbrecht, B. (2014). Isolating the Neural Mechanisms of Interference during Continuous Multisensory Dual-task Performance. *Journal of Cognitive Neuroscience, 26*(3), 476–489. https://doi.org/10.1162/jocn_a_00480

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

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

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology, 3*, 1–9. https://doi.org/10.3389/fpsyg.2012.00203

Landi, N., Crowley, M. J., Wu, J., Bailey, C. A., & Mayes, L. C. (2012). Deviant ERP response to spoken non-words among adolescents exposed to cocaine in utero. *Brain and Language, 120*(3), 209–216. https://doi.org/10.1016/j.bandl.2011.09.002

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language, 111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

Lenth, R. (2019). Emmeans: estimated marginal means. In *R package version 1.4.2*. https://cran.r-project.org/package=emmeans.

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. https://doi.org/10.3389/fnhum.2014.00213

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, and Psychophysics.*. https://doi.org/10.3758/s13414-020-02203-y

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

Mesgarani, N., Cheung, C., Johnson, K., & Edward, C. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*, 1006–1010. https://doi.org/10.1126/science.1245994

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America, 102*(3), 1864–1877. https://doi.org/10.1121/1.420092

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance. *Psychological Science, 20* (7), 895–903. https://doi.org/10.1111/j.1467-9280.2009.02380.x

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language, 76*(2), 80–93. https://doi.org/10.1016/j.jml.2014.06.007

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica, 42*(4), 313–329. https://doi.org/10.1016/0001-6918(78)90006-9

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology, 24*(4), 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x

Nusbaum, H., & Magnuson, J. S. (1997). Talker Normalization : Phonetic Constancy as a Cognitive Process. In K. A. Johnson, & J. W. Mullennix (Eds.), *Talker variability and speech processing* (pp. 109–132). Academic Press. https://doi.org/10.1121/1.2028337.

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

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of mandarin and cantonese. *Journal of Chinese Linguistics, 34*(1), 134–154.

Peng, G., Zhang, C., Zheng, H., Minett, J. W., & Wang, W.-S.-Y. (2012). The effect of intertalker variations on acoustic – perceptual mapping in Cantonese. *Journal of Speech, Language, and Hearing Research, 55*, 579–596. https://doi.org/10.1044/1092-4388(2011/11-0025)language

Persson, A., & Jaeger, T. F. (2023). Evaluating normalization accounts against the dense vowel space of Central Swedish. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1165742

Picton, T. W., Woods, D. L., & Proulx, G. B. (1978). Human auditory sustained potentials. II. Stimulus relationships. *Electroencephalography and Clinical Neurophysiology, 45*(2), 198–210. https://doi.org/10.1016/0013-4694(78)90004-4

Picton, T. W., Hillyard, S. A., Galambos, R., & Schiff, M. (1971). Human auditory attention: A central or peripheral process? *Science, 173*(3994), 351–353. https://doi.org/10.1126/science.173.3994.351

Rahman, A. A., Tan, H. K., Loo, S. T., Malik, A. B. A., Tan, K. H., Gluckman, P. D., Chong, Y. S., Meaney, M. J., Qiu, A., & Rifkin-Graboi, A. (2022). Cognitive flexibility in preschoolers: A role for the late frontal negativity (LFN). *Cognitive Development, 63*, Article 101200. https://doi.org/10.1016/j.cogdev.2022.101200

Reese, H., & Reinisch, E. (2022). Cognitive load does not increase reliance on speaker information in phonetic categorization. *JASA Express Letters, 2*(5). https://doi.org/10.1121/10.0009895

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., 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

Sweller, J. (2011). Cognitive Load Theory. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 55, pp. 37–76). Elsevier Inc. https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

Tao, R., Zhang, K., & Peng, G. (2021). Music Does Not Facilitate Lexical Tone Normalization: A Speech-Specific Perceptual Process. *Frontiers in Psychology, 12*, 1–14. https://doi.org/10.3389/fpsyg.2021.717110

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth Edi). https://www.stats.ox.ac.uk/pub/MASS4/.

Waxer, M., & Morton, J. B. (2011). Multiple processes underlying dimensional change card sort performance: A developmental electrophysiological investigation. *Journal of Cognitive Neuroscience, 23*(11), 3267–3279. https://doi.org/10.1162/jocn_a_00038

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46*(2), 413–421. https://doi.org/10.1044/1092-4388(2003/034)

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*(7), 1173–1184. https://doi.org/10.1162/0898929041920522

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex, 166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003

Xie, Z., Brodbeck, C., & Chandrasekaran, B. (2023). Cortical Tracking of Continuous Speech Under Bimodal Divided Attention. *Neurobiology of Language, 4*(2), 318–343. https://doi.org/10.1162/nol_a_00100

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron, 102*(6), 1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023

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

Zakay, D., & Block, R. (1995). An attentional gate model of prospective time estimation. *Time and the Dynamic Control of Behavior, 5*, 167–178.

Zhang, C., Peng, G., & Wang, W.-S.-Y. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *The Journal of the Acoustical Society of America, 132*(2), 1088–1099. https://doi.org/10.1121/1.4731470

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, C., Xia, Q., & Peng, G. (2015). Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics, 33*, 149–162. https://doi.org/10.1016/j.jneuroling.2014.07.002

Zhang, K., Li, D., & Peng, G. (2024). Achieving perceptual constancy with context cues in second language speech perception. *Journal of Phonetics, 103*, Article 101299. https://doi.org/10.1016/j.wocn.2024.101299

Zhang, K., & Peng, G. (2021). The time course of normalizing speech variability in vowels. *Brain and Language, 222*, Article 105028. https://doi.org/10.1016/j.bandl.2021.105028

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

Zhang, K., Tao, R., & Peng, G. (2023). The advantage of the music-enabled brain in accommodating lexical tone variabilities. *Brain and Language, 247*, Article 105348. https://doi.org/10.1016/j.bandl.2023.105348

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