http://dx.doi.org/10.1016/j.bandl.2021.105028

The time course of normalizing speech variability in vowels

Kaile Zhang[a] and Gang Peng[a,b]

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

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

kaile.k.zhang@connect.polyu.hk, 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

To achieve perceptual constancy, listeners utilize contextual cues to normalize speech variabilities in speakers. The present study tested the time course of this cognitive process with an event-related potential (ERP) experiment. The first neurophysiological evidence of speech normalization is observed in P2 (130–250 ms), which is functionally related to phonetic and phonological processes. Furthermore, the normalization process was found to ease lexical retrieval, as indexed by smaller N400 (350–470 ms) after larger P2. A cross-language vowel perception task was carried out to further specify whether normalization was processed in the phonetic and/or phonological stage(s). It was found that both phonetic and phonological cues in the speech context contributed to vowel normalization. The results suggest that vowel normalization in the speech context can be observed in the P2 time window and largely overlaps with phonetic and phonological processes.

Keywords:

Speech variability, perceptual normalization, time course, P2, vowels

1. Introduction

1.1. Speech variability and extrinsic perceptual normalization

Vowel perception is largely determined by formant frequencies which are generated by vocal tract resonance. Speakers differ significantly in vocal cavity configuration. While producing a phonologically identical vowel, women tend to produce tokens of higher formant frequency than men due to their comparatively shorter vocal tract. Therefore, there is no one-to-one map between formant frequency and vowel category (Peterson & Barney, 1952). Speech variability sometimes blurs the boundaries among different vowel categories, forming an obstacle for accurate speech perception. For example, the vowel /ɪ/ produced by one speaker may be acoustically closer to another speaker's production of /ɛ/.

In addition to intrinsic formants, contextual information also affects how listeners interpret target words, a process that has been called "extrinsic normalization" (Ainsworth, 1975; Nearey, 1989). Ladefoged and Broadbent (1957) synthesized six versions of the sentence '*Please say what this word is*' as contexts which had either high or low first formant (F1) and/or second formant (F2), and listeners were asked to identify the /b-V-d/ syllable embedded in these contexts. Their results showed that the identical /b-V-d/ syllable was more frequently identified as 'bit' in the high-F1 context and as 'bet' in the low-F1 context. Since this pioneering work, extrinsic normalization has been reduplicated by many studies on vowel perception (e.g., Sjerps et al., 2011; Watkins, 1991) and also extended to the perception of other speech components, like consonants (e.g., Lotto & Kluender, 1998; Mann, 1980) and lexical tones (e.g., Lin & Wang, 1984; Wong & Diehl, 2003). In most cases, contextual information affects target perception in a contrastive way. That is, a target speech cue tends to be perceived as being low if the preceding contextual cue is high. Although both preceding and following contexts affect the perception of target sounds, vowel perception is more notably affected by preceding contexts (Sjerps et al., 2018).

Extrinsic context facilitates the identification of high-variability speech. Johnson and Sjerps (2018) showed that vowel identification with extrinsic context was more accurate than vowel identification in isolation (86% vs. 77.3% for speech materials in Peterson & Barney, 1952; 78.1% vs. 62.9% for speech materials in Hillenbrand et al.,

1995). Similar results have also been reported for lexical tone perception. For example, Wong and Diehl (2003) found that Cantonese speakers could only identify 48.6% of trials correctly in a Cantonese level tone perception task if the stimuli from multiple speakers were presented in isolation. However, with the help of context, listeners achieved an accuracy rate of over 90%.

1.2. The hierarchical process of speech perception

People with normal hearing and language ability hardly notice that speech perception is composed of multiple stages of processing. There is an ongoing debate about speech processing units (Samuel, 2020). The speech processing hierarchy in some speech perception models, like TRACE, is composed of linguistic units, such as acoustic features, phonemes, and words (McClelland & Elman, 1986). However, other models, for example, adaptive resonance theory (Goldinger & Azuma, 2003; Grossberg et al., 1997), indicate that speech processing units are particular "chunks" which reflect the probabilistic exposure pattern of a listener. Some chunks can be non-linguistic units, but others coincide with linguistic units. It is beyond the scope of the present study to test the reasonability of each theory. Considering that the well-defined linguistic units are potential speech processing units in both the TRACE model and adaptive resonance theory, the present study uses linguistic units, ranging from acoustic features to phonetic features, phonemes, and words, as the processing units in different stages of speech perception.

According to the TRACE model (McClelland & Elman, 1986), incoming acoustic signals are first interpreted as linguistically meaningful units (i.e., phonemes) in the pre-lexical stage. Phonemes activated at the pre-lexical level are combined to trigger the word identification and language comprehension processes at higher levels. However, acoustic signals are not directly linked to abstract phonemes. Instead, the pre-lexical stage contains at least three speech processes: acoustic, phonetic, and phonological (Phillips, 2001). The hierarchical speech perception process is supported by various neuroimage studies. Obleser et al. (2007) compared listeners' perception of different stop consonants and spectrally rotated incomprehensible analogues. They found that the superior temporal region immediately posterior to Heschl's gyrus (HG; including the planum temporale) activated across all types of stimuli, suggesting that these areas were involved in a non-categorical

prelinguistic step of sound analysis. By contrast, the anterolateral superior temporal gyrus/sulcus (STG/S) operates on a categorical basis, preferring overlearned speech sounds. Studies on lexical tone perception have also reported a hierarchical acoustic-phonological processing stream that originates in the core auditory region and projects to the mid-left middle temporal gyrus (left mMTG). Zhang et al. (2011) tested the hierarchical processing of lexical tones with an oddball paradigm. Three stimuli from an 11-step Mandarin Tone 2-Tone 4 continuum — Step 3, Step 7, and Step 11— severed as the across-category deviant (Step 3), the standard (Step 7), and the within-category deviant (Step 11), respectively. The across-category lexical tone variations elicited stronger activation than within-category lexical tone variations in the left mMTG, reflecting abstract phonological processing. The dorsal and posterolateral superior temporal areas, especially the right middle STG, showed more activation for the acoustic variation (i.e., changing from the standard to the within-category deviant), indicating that right middle STG or nearby regions play an important role in the acoustic processing of lexical tones. The above-mentioned studies provide neuroimaging evidence for the differentiation between acoustic and speech-specific processes. Andics (2006) tried to separate acoustic-phonetic process from phonological process. By varying contextual cues, he caused the acoustically identical consonant (the same acoustic-phonetic process) in a CV syllable to be perceived as different consonantal categories (different phonological processes). Andics' neural imaging results suggested that the cortical regions corresponding to acoustic-phonetic process and those corresponding to phonological process were anatomically different, supporting the functional distinction between processing levels. Specifically, the acoustic-phonetic process activates the brain areas in the left hemisphere, including the STG/S, MFG, and SFG. Some brain areas in the left hemisphere are also involved in the phonological process, like the STG and IFG. However, the phonological process also activates areas in the right hemisphere, like the anterior STG, the IFG, and the right fusiform gyrus.

1.3. Perceptual normalization mechanisms and time loci of the normalization process

It is as yet unclear in which stage(s) of speech perception hierarchy the normalization process is implemented. Different theories are proposed to explain the

mechanism underlying extrinsic normalization; their perspectives on the signal properties used for, and the time locus of, extrinsic normalization vary greatly.

1.3.1. General auditory contrast enhancement mechanism and acoustic normalization

Extrinsic normalization is not limited to speech materials. Nonspeech also affects the interpretation of target words (Aravamudhan et al., 2008; Holt, 2005, 2006; Lotto et al., 2003; Wade & Holt, 2005). A preceding context composed of sine wave of high pitch leads to more responses of /ga/ (the alternative with greater energy in the low-frequency region) in the perception of the /ga/–/da/ continuum, while more /da/ responses (the alternative with greater energy in the high-frequency region) have been observed in low-pitch contexts (Holt, 2005; Lotto & Kluender, 1998). The contrastive perceptual behavior has been observed even in birds (Japanese quails; Holt et al., 2001). Based on these findings, a general auditory contrast enhancement mechanism was proposed to account for the extrinsic normalization process (Holt & Lotto, 2002; Lotto & Kluender, 1998). The general auditory account indicates that extrinsic normalization is a contrastive encoding of targets relative to the acoustic properties of contexts. While perceiving target speech, listeners are more sensitive to the acoustic properties that are suppressed in contexts and thus show a contrastive perceptual pattern. The contrastive encoding operates on the general auditory level where all types of sounds (e.g., speech and nonspeech) are processed in a similar way. The prerequisite of this process is *spectral contrast*. Watkins and Makin (1994, 1996) suggested that the normalization process first computes the long-term average spectrum (LTAS) of the preceding speech and then reversely filters the following speech with the properties of the preceding speech's LTAS.

Based on this mechanism, the normalization process occurs in the acoustic processing stage operating on the general auditory level, as supported by some neuroimaging studies (e.g., Sjerps et al., 2011b). Sjerps et al. (2011b) embedded extrinsic vowel normalization into an active multiple-deviant oddball paradigm. Due to context effect, the ambiguous vowel /ɪ˔ɛ/ (the standard stimulus) was more frequently perceived as /ɪ/ in the high-context block and as /ɛ/ in the low-context block. Therefore, the typical vowel /ɪ/ (one of the deviant stimuli) could be hard to detect in the high-context block but easier in the low-context block. The perception of the same target vowel may elicit different

event-related potential (ERP) patterns in different contexts due to detectability. Sjerps et al. (2011b) tested context effect by comparing the perception of the deviants (i.e., the typical vowels /ɪ/ and /ɛ/) in high-context and low-context blocks. They observed that the normalization process elicited the first reliable electrophysiological response in the N1 time window (80–160 ms). Considering that N1 is generally associated with the acoustic process (Näätänen & Winkler, 1999), Sjerps et al. (2011b) showed that speech normalization can be observed as early as the acoustic processing stage.

1.3.2. Context tuning mechanism and phonetic/phonological normalization

Unlike the general auditory contrast enhancement mechanism, the context tuning mechanism describes speech normalization as an active recalibration that requires speech-specific information and higher-level cognitive processes. The context tuning mechanism suggests that listeners use extrinsic contextual information to compute a speaker-specific mapping of acoustic patterns onto abstract linguistic units and then recalibrate ambiguous target speech with that mapping (e.g., Joos, 1948; Magnuson & Nusbaum, 2007). The speaker-specific mapping connecting acoustic patterns and linguistic units requires speech-specific information, indicating a normalization process at the speech-specific processing stage. Studies that compare speech and nonspeech context effects also support the necessity of speech-specific information in the normalization process. Zhang et al. (2012, 2017) investigated the perception of ambiguous Cantonese mid-level tones in speech and nonspeech contexts. The mid-level tone was perceived as a high-level tone in a speech context of low-fundamental frequency (F0) and as a low-level tone in a speech context of high F0, but the contrastive perception did not show in a nonspeech context which was composed of a triangle wave and had the same pitch information as the speech contexts. Even though some studies have found a statistically significant context effect of nonspeech, the effect size is much smaller than in speech contexts (Sjerps et al., 2011a). The unequal effect of speech and nonspeech contexts suggests that acoustic information may contribute to the normalization process but a reliable normalization effect requires speech-specific information, and that normalization also occurs after the acoustic processing stage. However, it is still unclear whether the speech-specific information required to compute

the mapping is cross-linguistic general phonetic information, language-specific phonological (phonemic) information, or both.

Phonetic processing stage

Sjerps et al. (2019) tested the neurological processing of extrinsic vowel normalization with a different paradigm using the electrocorticographic (ECoG) signal. They asked Spanish neurosurgery patients to identify an ambiguous target varying from /sofo/ to /sufu/ in the sentence "a veces se halla" ("at times she feels rather …") of either high F1 or low F1. The ECoG signal was collected simultaneously with a 256-electrode cortical surface array placed on the peri-Sylvian region. Sjerps et al. identified specific patches of cortex that were responsive to either /u/ or /o/. Interestingly, their responsiveness to the target vowels was modulated contrastively by the preceding context's F1, suggesting that these patches of cortex were involved in the normalization process. Additional analyses of the affected neural population showed that neurons showing a stronger response to low-F1 targets also responded more strongly to low-F1 contexts. This result indicates that the neural populations involved in normalization demonstrate a tuning preference for higher or lower F1 ranges in general, not exclusively for a discrete vowel category. Since F1 information is cross-linguistical general, Sjerps et al. suggested that normalization affects the pre-phonemic representation of speech (F1 characteristics) on the auditory cortex (AC), supporting the general auditory contrast enhancement mechanism.

However, high/low F1 ranges also correspond to phonetic features [± high]. Mesgarani et al. (2014) demonstrated that the neuron population in the posterior and middle BA22 (part of AC where the normalization process was observed in Sjerps et al., 2019) encodes phonetic features (i.e., dorsal, coronal, labial, etc.), including [± high] and [± front] for vowel categorization. Furthermore, the time window of the normalization process in Sjerps et al. (2019) was around 60–190 ms at one sample electrode (see Figure 2. c in Sjerps et al. 2019) and around 130–300 ms at another (see Figure 3. b in Sjerps et al. 2019). The time window of the first sample electrode was consistent with the typical time window of N1 (Woods, 1995) but the second somewhat overlapped with the time window of P2 (Crowley & Colrain, 2004). Magneto-encephalography and stereo-electro-encephalography have shown that one of the cortical origins of P2 lies in the Brodmann area (BA 22; Godey et al., 2001), in which the neuron populations show selective activation

in response to phonetic features (Mesgarani et al., 2014). Phonetic features are products of the phonetic process, during which acoustic signals are mapped onto phonetic representations. P2 originating in the BA22 thus should reflect the phonetic process. Therefore, another interpretation of the results observed by Sjerps et (2019) is that vowel normalization occurs in the phonetic processing stage, in which acoustic signals are mapped onto phonetic features. An acoustic-phonetic mapping is necessary for phonetic categorization during the normalization process. Context, which is used to establish the acoustic-phonetic mapping, should contain at least acoustic and phonetic information.

Phonological processing stage

Another possibility is that normalization occurs in the phonological processing stage, in which listeners classify stimuli into language-specific phonological categories. Joos (1948) suggested that when first hearing a few words from a new speaker, listeners quickly establish the vowel pattern of that speaker. For example, from the first greeting, 'How do you do?' the listeners would know that the /a/ was pronounced with the low central articulation gesture, the /u/ with the speaker's highest and strongest back articulation, and the /j/ with a higher and more forward articulation. The vowel pattern then serves as a reference to place the incoming signals in a speaker-specific vowel space.

Kang et al. (2016) asked native English and native French speakers to identify sounds on the /s–ʃ/ continuum when followed by /a/, /u/, or /y/. The ambiguous /^{s}ʃ/ sound is more likely to be perceived as /s/ before a rounded vowel like /u/ and /y/ and as /ʃ/ before an unrounded vowel like /a/. The /y/ sound exists in French but not in English. Kang et al. found that both English and French speakers showed context effect for the vowel /u/ but only French speakers showed it for /y/. Their results suggest that perceptual normalization is a language-specific rather than a cross-linguistic general effect, and that it requires an acoustic-phonological but not an acoustic-phonetic mapping.

To our knowledge, no ERP study has tested the normalization process in the phonological processing stage. Perceptual normalization in this stage probably also triggers the P2 component. P2 has more than one functional meaning. Previous studies on auditory word identification (e.g., Cheng et al., 2014; Landi et al., 2012; Zhang, Xia, et al., 2015) reported that P2 is associated with the phonological process. In a Chinese word identification task, participants showed significantly larger P2 for homophonic and rhyme

pairs than phonologically unrelated pairs. Real words triggered a significantly smaller ERP amplitude than phonotactically illegal nonwords in the P2 time window (Cheng et al., 2014).

Phonetic and phonological processing stages

Zhang et al. (2015) reported that both phonetic and phonological information is helpful for perceptual normalization, although their contributions are unequal. Zhang et al. (2015) compared the effects of different contexts on the perception of ambiguous Cantonese mid-level tones, including nonspeech (triangle waves; spectral contrast), reversed speech (normal speech reversed in the time domain; acoustic-phonetic information), meaningless speech (meaningless word sequences; acoustic-phonological information), and meaningful speech (normal speech; acoustic-phonological, syntactic, and semantic information). The contrastive context effect was observed in all contexts, but the effect size differed significantly across different contexts. Meaningful speech (55.9%), which had the richest speech cues, was most effective for lexical tone normalization. It was followed by meaningless speech (44.1%), which provided the acoustic-phonological mapping. Reversed speech (14.2%) was also helpful, but the effect of nonspeech (0.6%–0.8%) was negligible. Zhang et al. (2015) suggested that perceptual normalization was a product of multiple levels of processing and that the primary contribution came from phonological information. This suggests that speech normalization is a continual process which occurs in both phonetic and phonological stages.

1.4. The present study

Vowel normalization has been observed in the N1 time window, during which acoustic signals are processed, supporting the idea of a general auditory contrast enhancement mechanism (Sjerps et al., 2011b). Meanwhile, phonetic/phonological information also plays an important role in the perceptual normalization of lexical tones (Zhang et al., 2015) and consonants (Kang et al., 2016), indicating that in addition to the acoustic stage, normalization probably also occurs in the speech-specific processing stages, which is in line with the context tuning mechanism. However, the effectiveness of phonetic/phonological knowledge has not been proven in vowel normalization. Relatively few studies have investigated the effect of phonetic cues on vowel normalization. In

general, previous studies tend to separate the pre-lexical process into phonological process and acoustic-phonetic (or just acoustic) process (e.g., Andics, 2006; Sjerps & Smiljanić, 2013), overlooking the unique contribution of phonetic cues. In regard to phonological knowledge, Sjerps and Smiljanić (2013) found that while perceiving vowels, Dutch, English, Spanish, and English-Spanish bilinguals showed comparable normalization effects regardless of the languages of the precursor contexts, indicating that phonological knowledge is not significant in vowel normalization. However, it is worth noting that even though the contexts used in Sjerps and Smiljanić (2013) were from different languages, they contained vowels – /a/, /i/, and /e/ – that were shared by the languages under study. It was possible that with the help of these familiar vowels, participants were able to successfully estimate the speakers' vowel spaces and normalize the target vowels. Therefore, it remains unclear whether phonological information is useful or not for vowel normalization.

As reviewed above, behavioral studies are inconclusive regarding the effectiveness of phonetic/phonological information on vowel normalization. More importantly, no ERP studies have probed the vowel normalization process in the phonetic/phonological stage. Therefore, it is still unknown whether vowel normalization in speech contexts also occurs in the phonetic/phonological processing stage that are more speech-specific. The present study aims to clarify the time course of vowel normalization. To answer the question of whether vowel normalization in speech contexts also occurs in the phonetic/phonological stage, two experiments were conducted with behavioral and electroencephalographic (EEG) measurements.

Perceptual normalization in general results in contrastive context effects. Listeners give more high-vowel (vowel of low F1) responses if contexts have high F1s, and give more low-vowel responses if contexts have low F1s. Therefore, at the behavioral level, the normalization effect can be detected by comparing listeners' perceptual results in two contexts (i.e., high vs. low). However, the situation is different at the neurological level. The behavioral alternation in two contexts is essentially triggered by the same cognitive process (i.e., normalization) at the neurological level. Therefore, the contrast between the neurological response in high context and that in low context is probably not due to the

normalization process but the neurological encoding of different words or the motor response tendency for different choices.

A more feasible way to test the neurological process of perceptual normalization is to utilize the unequal effect of speech and nonspeech contexts (Zhang et al., 2013). Zhang et al. (2012) found that the reliable lexical tone normalization was observed in speech context but not in nonspeech context. Therefore, Zhang et al. (2013) suggested that the comparison of ERP patterns elicited by target perception in speech and nonspeech contexts could reflect lexical tone normalization process at the neurological level. Using a similar experimental design to that used in Zhang et al. (2012), Zhang et al. (2013) asked subjects to perceive ambiguous Cantonese mid-level tones in speech and nonspeech contexts of different pitch heights. The lexical tone perception elicited three ERP components: N1 (100–220 ms), N400 (250–500 ms), and LPC (500–800 ms). The first systematic speech-nonspeech context difference was found in the N400 time window. Therefore, Zhang et al. suggested that lexical tone normalization occurred in the N400 time window, it overlapped with the lexical retrieval process. Although Zhang et al. (2013) tested lexical tone normalization, their method is informative for testing vowel normalization. Normalization in a nonspeech context takes place in the acoustic stage since nonspeech contexts contain only acoustic information. However, speech contexts have rich information, including spectral contrast, phonetic features, and phonological categories. It is as yet unclear which information is used by listeners to construct speaker-specific mapping and recalibrate ambiguous targets. As suggested by Zhang et al. (2015), the normalization process in speech contexts probably occurs in different stages, including the acoustic and phonetic/phonological stages. The post-acoustic normalization process in speech contexts possibly elicits ERP component(s) that is/are not shown in nonspeech contexts or affect(s) the amplitude(s) of certain ERP component(s).

Experiment 1 tested the neurological normalization process using the unequal effects of speech and nonspeech contexts. Participants were asked to perceive ambiguous vowels in both speech and nonspeech contexts and EEG signals were recorded simultaneously. The differences in ERP patterns in speech and nonspeech contexts should capture the moment when the speech-specific normalization process occurs. It was hypothesized that if vowel normalization in speech context also occurs in the

phonetic/phonological stage, a P2 component would be observed in speech contexts but not in nonspeech contexts, or the P2 amplitude in speech contexts would differ from that in nonspeech contexts. However, P2 has manifold functional meanings. It indexes the pre-phonemic stage of the speech process like the phonetic process, and it also reflects the phonological process. A P2 difference can hardly show whether perceptual normalization occurs in the phonetic stage and/or the phonological stage. A cross-language vowel identification task was conducted in Experiment 2 to tease out phonetic process from phonological process. Experiment 2 was expected to would give a more precise interpretation of the ERP results in Experiment 1.

Experiment 1: The EEG Experiment

2. Methods

2.1. Subjects

The EEG experiment largely followed the experimental design of Zhang et al. (2013), who compared tone perception in speech and nonspeech contexts with 16 subjects and observed a significant context effect. Therefore, in the present study it was intended to recruit at least 16 subjects. Finally, 18 native Cantonese speakers (nine males) participated in Experiment 1 with a small monetary reward. One subject's data were excluded from analysis since more than 20% of trials (83/400) received no response. All of the participants were right-handed undergraduates at The Hong Kong Polytechnic University and reported no brain injuries, hearing impairments, or speech and language disorders. Informed written consent was obtained from every participant in compliance with the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

2.2. Stimuli

The speech stimuli used in Experiment 1 were adapted from the vowel stimuli in Sjerps et al. (2018). For details on the methods used to generate the speech stimuli, please refer to Sjerps et al. (2018). The target stimuli in Experiment 1 were a 17-step /bo55/–

/bu55/ continuum (F0: 170Hz, F1: 641–292 Hz). The differentiation of the vowels /o/ and /u/ mainly relies on F1 values. Therefore, only F1 values differed across the target continuum. Each target was normalized to 200 ms in duration. The /bo55/–/bu55/ continuum was generated by trimming the fricative consonant /f/ in the /fo55/–/fu55/ continuum used in Sjerps et al. (2018). The fricative consonant was cut to improve the timing accuracy in the stimulus onset detection in the EEG experiment. Theoretically, this manipulation should result in the vowels /o55/–/u55/. However, native Cantonese speakers perceived these syllables as being closer to /bo55/–/bu55/. Both /b/ and /f/ are labial consonants and their transitions (i.e., the onset of the F2 in the following vowel) point to similar loci, which is important for perceiving the place of consonants' articulation. Even though the visible transitions of /fo55/ and /fu55/ were cut, the following F2 may still contain some information about the transitions, which causes listeners to perceive the syllables as if they started with a labial consonant. When the fricative part has been removed, the 'phony consonant' in such a condition tends to be perceived as the plosive consonant /b/. A native male Cantonese speaker evaluated nativeness and clarity of the trimmed target stimuli (i.e., /bo55/ and /bu55/) on a Likert scale from one to seven, with seven sounding the most native/clear. Based on his evaluation (/bo55/: 6/7 for nativeness and 6/7 for clarity; /bu55/: 5/7 for nativeness and 5/7 for clarity), the target stimuli were suitable for the perception experiment.

The speech context in Experiment 1 is the meaningless phrase /p^{h}a21 tsi25/ (琶紫, guitar purple), which covers a speaker's full F1 range – /a/ for the highest F1 and /i/ for the lowest F1. This facilitates the estimation of a speaker's vowel space. Contexts of either high or low F1 were synthesized with the source-filter method in Praat (Boersma & Weenink, 2016). The formant trajectories of the original recording of /p^{h}a21 tsi25/ were extracted with the To Formant (Burg) function in Praat. Five formants were estimated between 0 and 5,500 Hz. The F1 trajectory was shifted 100 Hz up or down to resemble speakers with different vocal tract configurations. The source signal was obtained from a representative vowel /o55/ produced by the same speaker using the linear predictive coding (LPC) procedure in Praat. Formant patterns with modified F1s and the source signal were combined, resulting in a high-F1 context and a low-F1 context. The nonspeech contexts were composed of iterated rippled noise (IRN). IRN was generated by a cascade of delay

and add operations which gave it a spectral ripple and temporal regularity (Yost, 1996). Therefore, the IRN had comparable temporal and spectral complexity to speech sound. The nonspeech context was manipulated to have the same LTAS and pitch height as its speech counterpart in Praat (Boersma & Weenink, 2016). The duration of each context was normalized to 565 ms.

2.3. Procedure

Context is useful for the identification of ambiguous utterances near categorical boundaries, but the effect is weak for the perception of typical tokens (Sjerps et al., 2018). The phonological boundaries vary from person to person. Therefore, a short categorical perception task was carried out first to identify the most ambiguous target within the 17-step vowel continuum for each participant. Then the participant-specific stimuli sets were used to evaluate their normalization process with a word identification task. Because the stimuli used in Experiment 1 were not naturally produced /bo55/ and /bu55/, participants familiarized themselves with these syllables before the task. In the familiarization phase, the end points of the target continuum (i.e., Step 1 and Step 17) were embedded in the original speech context (i.e., context without F1 manipulation). All participants reported that they were familiar with these stimuli after several exposures.

2.3.1. The categorical perception task

The categorical perception task was carried out in a soundproof booth using a Praat ExperimentMFC identification test design (Boersma & Weenink, 2016). The targets used in this task were the odd-numbered stimuli on the 17-step vowel continuum (i.e., $Step_{2n-1}$, $n \in [1, 9]$). They were embedded in the original speech context. In each trial, a target stimulus was played before the context with a 500 ms interval of silence in between. A window with two rectangles labeled either /bo55/ or /bu55/was shown on the screen after the context stimulus. Subjects were asked to click on the corresponding rectangle in the window to indicate their choice. Each target was repeated five times, played in random order. The target stimulus near to the 50% point of the identification curve was regarded as the most ambiguous target sound.

2.3.2. The word identification task

The word identification task was implemented in E-prime 2.0. In each block, the context was kept constant (i.e., a blocked design). Therefore, there were four blocks in total: speech high, speech low, nonspeech high, and nonspeech low. Four blocks were presented in a pseudo-counterbalanced order across participants. Each block consisted of five types of targets: two end points of the 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 the most ambiguous target in the vowel continuum (Step X-1; 20 repetitions), and the one after it (Step X+1; 20 repetitions). The 120 trials were presented in random order. The trial procedure is illustrated in Figure 1. In each trial, the target was first played to participants bilaterally via headphones after a 500 ms fixation. Then the context was played after a jittered silence (from 400 ms to 600 ms). A question mark was shown on the screen after each audio stimulus. Participants were told to pay attention to any audio stimuli they heard and to choose which syllable the first sound was once they saw the question mark. They were asked to press the corresponding buttons on the mouse to show their identification. The maximum response time (RT) allowed was 1,250 ms. In Experiment 1, the target was played before the context in each trial, which was adapted from the experiment design in Sjerps et al. (2011b). Due to the blocked design, the context in each trial served as the context of the following trial. We tried in this way to enlarge the context-target interval which in turn reduced the possible carry-on effect from context EEG to target EEG.

[Figure 1 here]

2.3.3. EEG signal recording

EEG signals were recorded during the word identification task. Participants sat in a silent and comfortable room. The EEG signals were collected 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 which were placed above and below the left eye recorded the vertical electrooculography (EOG) to monitor eyeblinks. The horizontal EOG was recorded to monitor horizontal eye movements by the bipolar channels placed lateral to the outer canthus of each eye. Two separate electrodes placed on the mastoids were used as 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 between 0.05 Hz and 400 Hz were continuously recorded and digitized at 24-bit resolution at a sampling rate of 1,000 Hz.

3. Results

3.1. Behavioral results

The categorical boundary of /bo55/–/bu55/ varied from participant to participant (M = 9.06, SE = 0.46), indicating the necessity of using participant-specific stimuli for the normalization tasks. The proportion of /bo55/ responses in each context condition is plotted in Figure 2. Comparison of the left- and right-hand panels of Figure 2 reveals an unequal effect of speech context and nonspeech context. In speech contexts, participants gave more /bo55/ responses in the low-F1 condition than they did in the high-F1 condition. However, no such perceptual difference between high and low contexts can be observed in the nonspeech condition, as participants gave similar responses to the targets, regardless of the spectral manipulation of nonspeech contexts. The left-hand panel of Figure 2 also suggests that the normalization effect size (as indicated by the separation between the dashed line and the solid line) is more noticeable for the ambiguous targets than the typical targets. In speech contexts, participants' perception of ambiguous targets (i.e., Step X-1, Step X, and Step X+1) varied with the change of context's F1. However, the participants' perception of the unambiguous targets (i.e., Step 1 and Step 17) remained almost constant, regardless of the F1 of speech contexts.

[Figure 2 here]

A mixed-effects logistic regression model was built with the lmer4 package (Aguiar & Sala, 1998) in R (Version 3.6.0) to statistically evaluate the patterns shown in Figure 2. The high-F1 response (i.e. /bo55/) was coded as 1 and the low-F1 response (i.e. /bu55/) was coded as 0. All the predictors were centered around zero, with nonspeech coded as -1 and speech coded as 1 for the predictor *context*, high context coded as -1 and low context

coded as 1 for the predictor *shift*, and Step 1 (/bu55/ end in the vowel continuum) coded as -2, Step X-1 coded as -1, Step X coded as 0, Step X+1 coded as 1, and Step 17 (/bo55/ end in the vowel continuum) coded as 2 for the predictor *step.* The interstimulus interval (ISI; i.e., the context-target interval) was composed of RT in the last trial (M = 331, SD = 162) and the fixation duration (500 ms). RT varies a lot from trial to trial and from participant to participant. It is possible that longer ISI may weaken context effect. To control the effect of ISI, log ISI was also included in the data analysis (Davidson & Martin, 2013). Trials in which log ISI was more than three standard deviations from the mean were removed (0.9% trials; Magnuson et al., 2021). The mixed-effect logistic regression model included *context*, *shift*, *step*, *ISI*, and the possible two-way, three-way, and four-way interactions as the fixed effects, and the de-correlated slopes and intercepts by subjects for all the fixed effects that were not involved *ISI* as the random effects. According to previous studies (e.g., Ladefoged & Broadbent, 1957), participants' responses to targets that are acoustically identical change in different contexts if the normalization effect emerges. Specifically, due to the contrastive context effect, more /bo55/ (the vowel of high F1) responses should therefore be given in low-F1 contexts than in high-F1 contexts in the present study. Therefore, the predictor *shift* was expected to be a significant main effect. Considering the unequal effect of speech and nonspeech contexts observed in Zhang et al. (2012), it was predicted that participants' target perception would be contrastively affected by F1 shift of speech context but not by spectral shift of nonspeech context. Therefore, the *context* by *shift* interaction should also be significant.

Table 1. The fixed effects for the mixed-effects logistic regression model on the word identification task in Experiment 1

Fixed effects	Estimate	SE	*z*	*p*	
(Intercept)	0.58	1.7	0.34	0.73	
Context	1.77	1.61	1.1	0.27	
Shift	1.48	1.63	0.91	0.36	
Step	6.14	1.92	3.2	0.001	**
ISI	-0.22	0.24	-0.91	0.36	
Context:Shift	3.62	1.62	2.24	0.03	*
Context:Step	0.22	1.89	0.11	0.91	
Shift:Step	1.03	1.92	0.54	0.59	
Context:ISI	-0.34	0.24	-1.45	0.15	

Shift:ISI	-0.13	0.24	-0.55	0.58	
Step:ISI	-0.62	0.28	-2.17	0.03	*
Context:Shift:Step	3.05	1.88	1.62	0.11	
Context:Shift:ISI	-0.45	0.24	-1.89	0.06	.
Context:Step:ISI	-0.01	0.28	-0.02	0.98	
Shift:Step:ISI	-0.18	0.29	-0.63	0.53	
Context:Shift:Step:ISI	-0.48	0.28	-1.73	0.08	.

The statistics in Table 1 show that subjects overall gave more /bo55/ responses when the target stimuli approached the /bo55/ end of the target continuum (positive *step*). As expected, interaction was observed between *context* and *shift*. The positive *context* by *shift* interaction suggested that the shifts in context F1 affected subjects' vowel perception more significantly in speech contexts than in nonspeech contexts, reduplicating the unequal effects of speech and nonspeech contexts. To evaluate whether *shift* was significant in both speech and nonspeech contexts, the mixed-effects logistic regression model was fitted to the perception data in the speech context condition and nonspeech context condition, respectively. The fixed effects were *step*, *shift*, and *shift* by *step* interaction, and the random effects were intercepts by subjects and slopes by subjects for *step*, *shift*, and *shift* by *step* interaction. The analysis of perception data in the speech context condition revealed a main effect of *shift* (B = 0.956, $z = 3.009$, $p < 0.01$), indicating that participants gave more /bo55/ responses when the context F1 was low. But *shift* was not significant in the nonspeech context condition (B = 0.02, $z = 0.056$, $p = 0.96$), suggesting that participants' responses were not significantly affected by the spectral shifts of nonspeech. A negative significant *step* by *ISI* interaction was observed, indicating that the identification curve of /bo55/ was less steep when ISI was longer. The three-way *context* by *shift* by *ISI* interaction was marginally significant. To better understand how ISI affects normalization effect in different contexts, the mixed-effects logistic regression model was fitted to the perception data in the speech context condition and nonspeech context condition, respectively. The fixed effects were *step*, *shift*, *ISI* and their two-way and three-way interactions, and the random effects were intercepts by subjects and slopes by subjects for *step*, *shift*, and *shift* by *step* interaction. The analysis in the speech context condition revealed a marginally significant *shift* by *ISI* interaction (B = -0.66, $z = -1.9$, $p = 0.06$), indicating that context effect size showed a decreasing tendency as ISI increased, but such a tendency was not

statistically significant at the 0.05 level. The two-way *shift* by *ISI* interaction was not significant in the nonspeech context condition (B = 0.35, z = 1.05, p = 0.29), suggesting that the normalization effect was not significantly affected by ISI in the nonspeech condition.

3.2. EEG results

Only trials containing ambiguous targets (i.e., Step X-1, Step X, and Step X+1) were included in the EEG data analysis because the perception of typical tokens did not show the contrastive context effect (see Step 1 and Step 17 in Figure 2). Although the normalization effect emerged during the perception of Step X-1 and Step X+1, no reliable ERP could be obtained for Step X-1 or Step X+1 due to the small number of epochs involved (only 20 trials each). To improve the signal to noise ratio of ERPs, the EEG epochs of three ambiguous targets were pooled (100 epochs in total) and averaged. Since two end points were excluded and three ambiguous steps were averaged, *step* was not a predictor in the EEG data analysis. Since the EEG data analysis averaged the epochs in the same experimental condition, the ISI information in each trial was lost. Furthermore, the behavioral results suggested that the normalization effect was not significantly affected by ISI. Therefore, the EEG data analysis did not include ISI either. Vowel normalization has been found in the acoustic processing stage (Sjerps et al., 2011). Both speech and nonspeech contexts provide acoustic information. Therefore, N1 components are expected to show in both speech and nonspeech contexts with comparable amplitudes. It was hypothesized above that vowel normalization also occurs in the phonetic/phonological stage. Since speech context also contains phonetic/phonological information, it is likely that P2 is observed in speech contexts but not in nonspeech contexts, or that P2 amplitudes differ in two contexts.

3.2.1 EEG data preprocessing

The preprocessing of EEG data was implemented using the EEGlab toolbox (version: 14.1.1; Delorme & Makeig, 2004). The signal was a bandpass filtered between 0.5 Hz and 30 Hz. Epochs of 900 ms time-locked to the onset of the target words were extracted from each trial, with the first 100 ms preceding to the target word as the reference in the baseline correction. Epochs that contained eyeblinks or horizontal eye movements

were excluded from further analysis. The eyeblinks were detected automatically by a moving window peak-to-peak amplitude method, with a voltage threshold of 100 μV, a window size of 200 ms, and a window moving step of 50 ms. The horizontal eye movements were detected automatically by the step-like artifact detection method, with a voltage threshold of 40 μV, a window size of 400 ms, and a window moving step of 10 ms. The mean acceptance rate of the epochs was 96.3% (N = 96.3/100, SE = 0.8) in the speech high context condition, 95.4% (N = 95.4/100, SE = 1.3) in the speech low context condition, 94.9% (N = 94.9/100, SE = 1.6) in the nonspeech high context condition, and 93.6% (N = 93.6/100, SE = 2.3) in the nonspeech low context condition. ERPs were calculated by averaging the accepted epochs in each condition per participant. These ERPs were finally re-referenced against the average mastoid.

3.2.2. Electrodes and time windows for different ERP components

[Figure 3 here]

The method for selecting analysis parameters largely followed the collapsed localizer method used in Luck and Gaspelin (2017). The global field power over time shown in Figure 3. a was computed as the root mean square of the ERP value at each time point averaged across 64 electrodes, 4 contexts, and 17 participants. The global field power and the ERP waves indicated that, in addition to the expected ERP component – N1 and P2 – N400 also emerged during target vowel perception. The present study also included N400 in the EEG data analysis. The analysis of N400 is largely exploratory. According to the visual inspection of the global field power amplitude in Figure 3. a, the time range during which the ERP component emerged was selected as its time window: 60 – 120 ms for N1, 130 – 250 ms for P2, and 350 – 470 ms for N400. The time window selected for each ERP component is consistent with the typical time window reported in previous studies (N1 in Näätänen & Winkler, 1999; P2 in Crowley & Colrain, 2004; and N400 in Kutas & Federmeier, 2011). The topographic maps in Figure 3. a showed the potential distributions in different time windows. The topographic map for each ERP component was obtained by averaging the voltage amplitude at each electrode across 4 contexts, 17 subjects, and the entire time window. Electrodes where the ERP component was expected

to peak according to the topographic map were selected to quantify it: F3, FC3, C3, CP3, Fz, FCz, Cz, CPz, F4, FC4, C4, and CP4 for N1; FC1, C1, CP1, FCz, Cz, CPz, FC2, C2, and CP2 for P2; and F1, FC1, C1, Fz, FCz, Cz, F2, FC2, and C2 for N400.

3.2.3 Statistical tests

The mean amplitude was calculated to quantify each ERP component. The ERP amplitudes of electrodes in the same scalp region (i.e., left, mid, and right) were averaged to simplify the statistical analysis. A three-way repeated-measures ANOVA with *context* (speech vs. nonspeech), *shift* (high vs. low), and *position* (left, mid, vs. right) as the within-subject factors was conducted on the amplitudes of each ERP component, to see in which time window(s) the normalization effect emerged and whether the normalization process showed any cortical lateralization. The Greenhouse–Geisser method was employed to correct violations of sphericity, where appropriate. Considering that the neurological normalization process in the present study was detected by comparing vowel perception in speech and nonspeech contexts, it was hypothesized that *context* would be observed as a main effect in one or multiple ERP component(s). Since the analysis of ECoG signals in Sjerps et al. (2019) revealed neurological normalization processes by comparing vowel perception in high and low speech contexts, a *context* by *shift* interaction might also be observed in the present study.

N1

None of the three predictors (i.e., *context*, *shift*, and *position*) achieved significance at the 0.05 level. Their two-way and three-way interactions were not significant either.

P2

The analysis revealed a significant main effect of *context*, $F(1, 16) = 8.6$, $p = 0.01$, $\eta_p^2 = 0.35$ and a significant main effect of *position*, $F(2, 32) = 11.88$, $p < 0.001$, $\eta_p^2 = 0.43$. Speech context ($M = 1.74$; $SE = 0.44$) triggered significantly larger P2 than nonspeech context ($M = 1$; $SE = 0.48$). The P2 amplitude in the mid region ($M = 1.61$; $SE = 0.45$) was significantly larger than in the left region ($M = 1.24$; $SE = 0.44$; $p = 0.001$) and in the right region ($M = 1.26$; $SE = 0.45$; $p = 0.001$). The P2 amplitude in the left region and that in the right region were statistically comparable ($p = 1$). There was also a significant *context* by *position* interaction, $F(2, 32) = 3.7$, $p = 0.036$, $\eta_p^2 = 0.19$. The simple main effect analysis of the *context* by *position* interaction suggested that P2 amplitude in speech context was

significantly higher than that in nonspeech context in the left region (speech: M = 1.6; SE = 0.44; nonspeech: M = 0.89; SE = 0.46, p = 0.014), in the mid region (speech: M = 2.02; SE = 0.45; nonspeech: M = 1.21; SE = 0.48, p = 0.007), and in the right region (speech: M = 1.6; SE = 0.44; nonspeech: M = 0.93; SE = 0.49, p = 0.011). Even though all three regions showed a significant speech-nonspeech difference, the difference was more noticeable in the mid region.

N400

The analysis revealed a significant main effect of *context*, $F(1, 16) = 11.53$, p = 0.004, $\eta_p^2 = 0.42$. The N400 amplitude was significantly more negative in nonspeech contexts (M = -2.5; SE = 0.48) than speech contexts (M = -1.59; SE = 0.44). Other main effects and interactions were not statistically significant at the 0.05 level.

To test whether the N400 difference was driven by P2, the mean amplitude of N400 was reanalyzed with a post-P2 baseline correction (250–350 ms). The mean amplitude of ERP waves in the 250–350 ms time window was first subtracted from the mean amplitude of the original N400. The three-way ANOVA revealed a significant main effect of *position*, $F(2, 15) = 5.949$, $p = 0.013$, $\eta_p^2 = 0.19$. The amplitude of post-P2-corrected N400 was more negative in the mid region (M = -1.291; SE = 0.422) than in the left region (M = -1.171; SE = 0.405, $p < 0.05$). However, *context* was neither a significant main effect nor involved in any significant interactions.

4. Discussion

Although a number of studies (e.g., Lotto & Holt, 2006) have observed the normalization effect in nonspeech contexts, Experiment 1 has shown that nonspeech contexts had little effect on ambiguous vowel perception at the group level. The exploration of the individual data revealed that around half the participants showed the expected normalization effect in nonspeech contexts but the other half did not, indicating a large individual difference in the nonspeech condition (see Supplementary Figure 1). Consistent with previous studies on the normalization process, a reliable normalization effect was observed in speech contexts. The behavioral results suggest that spectral contrast may help some listeners to normalize vowels, but reliable vowel normalization requires speech-specific information.

Three ERP components were tested in the present study to identify the time locus of vowel normalization process: N1, P2, and N400. The primary AC is the predominant neural generator of N1, and thus N1 is closely related to acoustic information encoding (Näätänen & Winkler, 1999). Although the present study detected the neurological normalization process by comparing ERPs in speech and nonspeech contexts, the absence of speech-nonspeech difference in the N1 time window does not mean that there is no normalization process occurring in the acoustic processing stage (albeit rather weak and highly inconsistent across subjects). Speech and nonspeech contexts provide similar information at the acoustic level (i.e., spectral contrast). The similar N1s suggest that normalization in the acoustic stage is comparable in speech and nonspeech contexts.

A significant speech-nonspeech context difference was observed in the P2 and N400 time windows, indicating that the same target stimulus was processed differently in speech and nonspeech contexts in these two time windows. It is worth noting that speech and nonspeech contexts differ in many ways, and thus the observed ERP differences in speech and nonspeech contexts reflect a number of things including -but not limited to- normalization. Experimental conditions that only differ in normalization process could be ideal for the present study but difficult (or even impossible) to find, since speech perception is affected by many factors, and they may also interact with each other. For example, the normalization process may be modulated by different perception strategies with two context types (speech vs. non-speech). Nonetheless, the experimental design in the present study still enables us to locate the time window of the vowel normalization process, due to the different functional meanings underlying P2 and N400. The results suggested that the potential processing differences in speech and nonspeech contexts (including normalization process) in the present study were implemented in either P2 (130 – 250 ms) or N400 (350 – 470 ms) time window. N400 is functionally related to word retrieval, semantic memory, and other post-lexical level processes (Kutas & Federmeier, 2011). The normalization process which maps acoustic signals to the abstract linguistic units should be finished before these post-lexical level processes. Therefore, the robust normalization effect in speech context that was observed in the behavioral response largely occurred in the P2 time window.

Considering that P2 is associated with phonetic/phonological processing, vowel normalization also occurs in the phonetic/phonological processing stage(s). The P2 observed in the present study was maximal in the central and central-frontal areas, with a latency of 130–250 ms. This latency and scalp distribution is consistent with that reported in previous studies (for a review of P2, see Crowley & Colrain, 2004), indicating that the P2 in the present study is reliable and typical. However, as stated in Section 1.4, due to the multiple functional meanings of P2, it is hard to tell whether the normalization process observed in the P2 time window was implemented in the phonetic stage and/or the phonological stage. This question will be further investigated in Experiment 2.

As found by Zhang et al. (2013), the target vowel perception in nonspeech contexts elicited greater negative N400 amplitude than speech contexts. N400 amplitudes reflect the ease of conceptual retrieval from long-term memory (the lexical retrieval in speech perception; Brouwer & Hoeks, 2013). The results of Experiment 1 suggest that lexical retrieval was easier in speech contexts than nonspeech contexts. Analysis of the post-P2 corrected N400 suggests that N400 amplitudes were noticeably affected by P2. It is possible that, in the P2 time window, speech context provides phonetic and/or phonological information for listeners to normalize speech variability and that acoustic signals are thus successfully mapped onto clear phonetic and/or phonological representations. The clear phonetic and/or phonological representations facilitate lexical retrieval in the N400 time window. Nonspeech contexts, however, cannot help listeners to transform acoustic signals into correct phonetic and/or phonological representations. Without sufficient phonetic and/or phonological information, lexical representations cannot be easily activated (McClelland & Elman, 1986).

Experiment 2: Cross-Language Vowel Identification

The comparison between speech and nonspeech contexts in Experiment 1 separated acoustic and speech-specific processes, but phonetic and phonological processes were still mixed. To give a more precise interpretation of the P2 observed in the EEG experiment, a cross-language vowel identification task was conducted in Experiment 2. The rationale of Experiment 2 is that if vowel normalization occurs in the phonetic stage, it is reasonable to assume that phonetic cues in context will affect target vowel normalization, and similarly,

if normalization occurs in the phonological stage, phonological cues in context are expected to modulate target vowel perception.

In Experiment 2, native English speakers were asked to perceive the ambiguous vowel /ᵘ ɔ/ in the native context /ɛ i/ and the nonnative context /œ y/. The front unrounded vowels /ɛ/ and /i/ exist in English and thus contain both phonetic and phonological information for English speakers. English does not have the front rounded vowels /œ/ and /y/, but the phonetic features [+rounded], [-back], and [+high] do exist in English. Therefore, they provide phonetic information but not phonological information for English speakers. Extrinsic normalization with these two contexts could result in three possible results for English speakers: (1) only the native context works; (2) both native and nonnative contexts work, but the native context works significantly better; or (3) both native and nonnative contexts work, and they work equally well. If the first possibility holds, the normalization effect should be attributed to the phonological information and the normalization process at speech-specific level occurs only in the phonological processing stage. If the second possibility is true, the normalization effect triggered by the nonnative vowels can largely be attributed to phonetic information. Meanwhile, since the native context works better than the nonnative context, one can say that phonological information also helps vowel normalization. Therefore, if the result is consistent with the second prediction, the vowel normalization process occurs at both phonetic and phonological levels. But if the result reflects the third possibility, it means that the additional phonological information provided by the native context is almost useless and vowel normalization is implemented in the phonetic stage. However, it is possible that the unequal effects of the two contexts are caused by vowels' signal properties rather than whether they are part of listeners' native phonological systems. That is, some vowels are inherently more effective than others in facilitating the normalization process. To monitor such a possibility, native Cantonese speakers were recruited as the control group. All four vowels exist in Cantonese, and thus the two contexts provide both phonetic and phonological information for Cantonese speakers. If the normalization effects of the two contexts are comparable for native Cantonese speakers, the observed difference between them for English speakers (if any) can only be attributed to nativeness.

5. Methods

5.1. Subjects

Experiment 2 largely followed the design of Sjerps et al. (2018), which observed a significant normalization effect in perceiving /u/-/o/ vowel pair by comparing high and low contexts with 16 subjects. Therefore, Experiment 2 also aimed to recruit at least 16 participants for each group. Sixteen native English speakers were recruited from the University of California, Berkeley community. Five of them were English monolinguals and the others were also fluent in Spanish, Tagalog, Japanese, or Hindi. Since the vowel systems of these other languages do not contain /œ/ and /y/ (the vowels used in the nonnative context), these bilinguals/multilinguals were included. Sixteen native Cantonese speakers were recruited from The Hong Kong Polytechnic University as the control group. All Cantonese participants spoke Cantonese, English, and Mandarin. None of the 32 participants had any self-reported hearing impairment or speech and language disorder. They were paid for their participation. The experimental procedure was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley and the Human Subjects Ethics Sub-committee at The Hong Kong Polytechnic University. Written informed consent was obtained from each participant before the experiment began.

5.2. Stimuli

Two male native Cantonese speakers were asked to read 36 Cantonese words (CV syllables; see Supplementary Table 1) three times in a soundproof recording booth. These words covered the vowels /ɛ/, /i/, /œ/, /y/, /u/, and /ɔ/. The lexical tones of Cantonese were likely to make the stimuli sound foreign to English speakers. To find the most natural lexical tone-vowel combination for English speakers, each vowel was first recorded with six Cantonese long tones and then a native speaker was asked to make a nativeness judgment on these versions. To improve the reliability of the nativeness judgment task, some vowels read by native English speakers were used as fillers. Two male native English speakers were asked to make recordings of /ɛ/, /i/, /u/, and /ɔ/ embedded in /b-V-d/ syllables three times in a quiet room. One repetition of each word (for both English and Cantonese)

was chosen based on its clarity and naturalness. The vowels were cut from the embedded syllables and then normalized to 0.4 ms.

A native English speaker was asked to rate the nativeness of the recordings (2 Cantonese speakers × 4 vowels × 6 lexical tones + 2 English speakers × 4 vowels) on a Likert scale from one to seven, with seven sounding the most native. Only vowels that exit in English and Cantonese were used in the rating task. The tokens with the highest scores [/ɛ22/ (5.7/7), /i22/ (6.3/7), /ɔ55/ (5.7/7), and /u55/(5/7) from the first Cantonese speaker] were chosen for Experiment 2. It is worth noting that all the selected tokens were level tones. To be consistent, for the vowels /y/ and /œ/, which do not exist in English, one token with level tones and with the best clarity and naturalness was chosen from the same speaker: /y33/ and /œ33/.

The selected tokens /ɛ22/ and /i22/ were concatenated as unrounded contexts, and /œ33/ and /y33/ were concatenated as rounded contexts. The selected tokens /u55/ and /ɔ55/ were used to generate a 17-step /u–ɔ/ continuum, forming targets in the vowel identification task. The pitches of the contexts (both /ɛ22 i22/ and /œ33 y33/) were shifted up 20 Hz to match the pitch height of the targets (/u55 ɔ55/). Due to tone merger, the speaker's production of /ɛ22 i22/ and /œ33 y33/ had roughly similar pitch heights. To make the stimuli comparable for English and Cantonese participants, the formants of /ɛ/ were shifted to the average values of the English /ɛ/ (cf. Hillenbrand, Getty, Clark, & Acoust, 1995) and the Cantonese /ɛ/ (cf. Zee, 2003). Similar processes were also carried out for /i/, /u/, and /ɔ/. The formants of /y/ and /œ/ were shifted to the values reported in Zee (2003), to make the stimuli more representative. The F1s of the context /ɛ i/ and /œ y/ were shifted 100 Hz up/down to form the high/low contexts. The procedure to manipulate context F1 was the same as the method used in Experiment 1. The 17-step /u–ɔ/ continuum was generated by interpolation in Praat (Boersma & Weenink, 2016). The parameters used to generate the stimuli are visualized in Figure 4.

[Figure 4 here]

5.3. Procedure

Before the experiment, a typical /u/ and a typical /ɔ/ were played ten times to participants to familiarize them with the two vowels. After the familiarization, a short test was conducted to see whether the participants could identify the typical /u/ and /ɔ/ sounds. A categorical perception task following the same procedure as the one used in Experiment 1 was carried out to identify the most ambiguous step along the /u–ɔ/ continuum for each participant.

The normalization process was measured by vowel identification task with a participant-specific stimuli set. OpenSesame (Mathôt et al., 2012) was used to control the stimuli presentation and data collection in the vowel identification task. Participants were asked to identify vowels in contexts with different F1 manipulations. A blocked design was employed. There were four blocks in total: /ɛ i/ of high F1, /ɛ i/ of low F1, /œ y/ of high F1, and /œ y/ of low F1. For the English native participants, /œ y/ blocks were always played first so that their vowel perception for these blocks would be less affected by the acoustically similar vowels /ɛ i/. The four blocks were pseudo-counterbalanced for the Cantonese subjects. Each block contained 100 trials (10 for Step 1 and Step 17, 20 for Step X-1 and Step X+1, and 40 for Step X). They were played in a random order. In each trial, a context was played first. The phonetic transcripts of contexts were also shown on a computer screen for subjects' reference. For the English speakers, no phonetic transcripts were used for the /œ y/ blocks, instead the symbols 'V*V**' were used. After a random jittered silence of 350 ms to 450 ms, the target was played to subjects with two choices (/u/ or /ɔ/) presented on the screen. Subjects were asked to make their judgment after the target sound was played. The next trial was played automatically once a response was detected or the maximum RT (2,000 ms) was reached.

5.4. Data analysis

A mixed-effects logistic regression model was fitted to the participants' responses in the vowel identification task. A response of /ɔ/ was coded as 1 and a response of /u/ was coded as 0. All predictors were centered around zero. The predictor *step* was coded

increasingly from /u/ to /ɔ/, with Step 1 as -2, Step X -1 as -1, Step X as 0, Step X + 1 as 1, and Step 17 as 2. The predictor *context* was coded negatively (-1) for the rounded context (i.e., /œ y/) and positively (1) for the unrounded context (i.e., /ɛ i/). The high F1 context was coded as -1 and the low context was coded as 1 for the predictor *shift.* The predictor *language* (i.e., the participants' native language) was coded as -1 for English and as 1 for Cantonese. Since there were too many possible interactions in this four-way design, the predictor *step* was included only as a main effect. The possible three-way and two-way interactions and the main effects of *context*, *shift*, and *language* were included in the model as fixed effects. The random effects in the model were de-correlate intercepts by subjects and slopes by subjects for *step*, *shift*, *context*, and *shift* by *context* interaction.

Based on Kang et al. (2016), the present study hypothesized that English speakers' vowel perception would not be affected by F1 shifts in the rounded contexts, but could be affected by F1 shifts in the unrounded contexts. It was also hypothesized that Cantonese speakers' vowel perception could be equally affected by rounded and unrounded contexts. Therefore, a significant *language* by *shift* by *context* interaction was expected to show in the results.

6. Results

The ambiguous steps for English speakers (M = 8; SE = 0.32) and Cantonese speakers (M = 7.75; SE = 0.36) were close to each other, $t(30) = 0.542$, $p = 0.592$. Figure 5 shows that among English speakers the normalization effect varied across vowel roundedness, with a comparatively larger normalization effect in unrounded contexts (as indicated by the greater separation between the solid and dashed lines in the bottom left panel of Figure 5) and a comparatively smaller normalization effect in rounded contexts (as indicated by the smaller separation between the solid and dashed lines in the bottom right panel of Figure 5). However, the normalization effects were comparable for Cantonese speakers in both rounded- and unrounded-vowel contexts.

[Figure 5 here]

Table 2. The fixed effects for the mixed-effects logistic regression model on the vowel identification task in Experiment 2

Fixed effects	Estimate	SE	*z*	*p*	
(Intercept)	1.1	0.34	3.28	<0.01	**
Step	1.79	0.12	14.53	<0.001	***
Context	0.01	0.09	0.08	0.94	
Shift	0.88	0.14	6.11	<0.001	***
Language	0.45	0.34	1.36	0.18	
Context:Shift	0.2	0.07	2.72	0.01	**
Context:Language	-0.01	0.09	-0.13	0.9	
Shift:Language	-0.2	0.14	-1.36	0.17	
Context:Shift:Language	-0.13	0.07	-1.75	0.08	.

The mixed-effect logistic regression model was fitted to statistically evaluate the patterns shown in Figure 5. The statistics of the fixed effects are summarized in Table 2. The positive significance of *Intercept* suggested that participants tended to give more /ɔ/ responses in general. Participants also gave more /ɔ/ responses overall when the stimuli approached the /ɔ/ end of the vowel continuum (positive *step*) and when the F1 of the context was shifted low (positive *shift*). The positive *context* by *shift* interaction indicated that, in general, the /ɛ i/ context was more likely to trigger the contrastive context effect than the /œ y/ context. The *language* by *context* by *shift* interaction was marginally significant.

To understand the marginally significant *language* by *context* by *shift* interaction, the mixed-effects logistic regression model was fitted to the perception data of Cantonese participants and of English participants, respectively. The fixed effects were *step*, *context*, *shift*, and *context* by *shift* interaction, and the random effects were intercepts by subjects and slopes by subjects for *step*, *context*, *shift*, and *shift* by *context* interaction. Analysis of the Cantonese participants' perception data revealed a main effect of *shift* (B = 0.689, z = 4.892, $p < 0.001$), indicating that Cantonese participants gave more /ɔ/ response when contexts were low. The *context* by *shift* interaction was not significant (B = 0.072, z = 1.001, $p = 0.317$), indicating that the normalization effect was statistically comparable in

the rounded and unrounded contexts for Cantonese speakers. Analysis of the English participants' data revealed a main effect of *shift* (B = 1.065, $z = 4.269$, $p < 0.001$), showing a contrastive context effect. Furthermore, an interaction was observed between *context* and *shift* (B = 0.32, $z = 2.557$, $p = 0.011$), suggesting that the normalization effect is significantly larger for unrounded vowels (native contexts) than rounded vowels (nonnative contexts).

Another observation that can be drawn from Figure 5 is that the English native subjects were affected by context shifts more noticeably than the Cantonese subjects, especially in the /ɛ i/ blocks. To test whether this difference was statistically significant, the mixed-effects logistic regression model was fitted to the perception data in the /ɛ i/ blocks and that in the /œ y/ blocks. The fixed effects were *step*, *language*, *shift*, and *language* by *shift* interaction, and the random effects were intercepts by subjects and slopes by subjects for *step* and *shift*. The analysis revealed a marginally significant *language* by *shift* interaction in the /ɛ i/ blocks (B = -0.43, $z = -1.8$, $p = 0.07$). There is a tendency for a larger normalization effect among English speakers than Cantonese speakers in the /ɛ i/ blocks, but it is not statistically significant at the 0.05 level. The *language* by *context* interaction was not significant in /œ y/ blocks (B = -0.06, $z = -0.43$, $p = 0.67$), indicating a comparable normalization effect for Cantonese and English speakers in the /œ y/ blocks.

7. Discussion

To evaluate the contribution of phonetic and phonological information to extrinsic vowel normalization, Experiment 2 asked native English speakers and native Cantonese speakers to perceive the ambiguous vowels /ᵘ ₎/ in the context /ɛ i/ and /œ y/. The results revealed two important findings. First of all, the contexts composed of nonnative vowels /œ/ and /y/ successfully triggered a contrastive context effect among the English speakers, indicating that phonetic information played a crucial role in extrinsic vowel normalization. Second, both contexts worked equally successfully in the Cantonese speakers' perceptual normalization, but among the English speakers, the context composed of native vowels /ɛ/ and /i/ showed significantly larger normalization effects than the context composed of nonnative vowels /œ/ and /y/. The unequal effect between native and nonnative contexts suggests that phonological information makes additional contributions to normalization

results. Therefore, Experiment 2 shows that the perceptual normalization of vowels relies on both phonetic and phonological information.

8. General discussion

To investigate the time course of the vowel normalization process, the present study compared listeners' electrophysiological responses when they perceived vowels in speech and nonspeech contexts in Experiment 1. The EEG signals suggested that the robust vowel normalization in speech context triggered a larger P2 amplitude around 130 to 250 ms after vowel onset. Previous studies (e.g., Godey et al., 2001; Mesgarani et al., 2014; Zhang et al., 2015) suggest that P2 has manifold functional meanings and is observed in both phonetic and phonological processes. To further specify in which speech processing stage(s) vowel normalization happened, a cross-language vowel perception task was conducted in Experiment 2. The results suggest that both phonetic and phonological information contribute to vowel normalization. Therefore, the EEG patterns in Experiment 1 and the vowel perception results in Experiment 2 together reveal that the vowel normalization process in the speech context can be detected in the P2 time window (130–250 ms after vowel onset) and largely overlaps with phonetic and phonological processes.

Previous studies have observed vowel normalization in the acoustic stage at both behavioral (e.g., Holt, 2006) and neurological levels (e.g., Sjerps et al., 2011b). However, the behavioral results of Experiment 1 suggest that the normalization process in the acoustic stage is weak and highly inconsistent across listeners. The EEG experiment, which tested the normalization process by comparing speech context with nonspeech context, was not ideal to detect neurological normalization processes in the N1 time window, because speech and nonspeech contexts provide similar acoustic information. Therefore, based on the EEG data, it is hard to tell whether the neurological normalization process in the acoustic stage is more robust than observed from the behavioral response.

Sjerps et al. (2019) did not give an explicit time window for the vowel normalization process. However, based on ERP waves in different contexts at one sample electrode (see Figure 3. b in Sjerps et al. 2019), it could be inferred that the normalization process was implemented around 130–300 ms after vowel onset, at least at some electrodes in their study, which largely overlaps with the time course of vowel normalization found

in the present study (i.e., 130–250 ms). Sjerps et al. (2019) located vowel normalization in the acoustic-phonetic stage mainly based on the finding that the neural populations involved in the vowel normalization process are attuned to either high or low F1 ranges, not specific vowel categories. However, Experiment 2 in the present study suggests that although cross-linguistic general phonetic cues facilitate vowel normalization, the normalization process prefers native vowels that contain language-specific phonological information, indicating that the process also occurs after the pre-phonemic stage. Experiment 1 in the present study adopted a similar paradigm to that used in Zhang et al. (2013) to investigate the extrinsic normalization of vowels. The results suggest that vowel normalization may occur earlier than lexical tones (P2 in the present study vs. N400 in Zhang et al., 2013), which is consistent with the findings in Zhang et al. (2021).

Few studies have investigated how phonetic information affects our speech perception. With the aid of phonetic features in the nonnative context of /œ-y/, English listeners in Experiment 2 successfully recalibrated ambiguous target stimuli. The phonetic features [+high], [-back], and [+round] to some extent describe how to produce /œ/ and /y/. The effectiveness of phonetic information in vowel normalization partially supports the close relationship between speech production and perception. Bruderer et al. (2015) found that infants' discrimination of Hindi /d̪/–/ɖ/ contrast was impeded if their tongue tip movement and tongue tip position were constrained by a teether. Their study suggests that the connection between perception and production of speech cues is established at a very early stage of our language development, and that production cues affect our speech perception. The phonetic features in the nonnative contexts which reflect the production cues probably facilitate listeners' reconstruction of the speaker's vocal tract properties, and thus eases the perceptual normalization process. The additional contribution of phonological information suggests that language experience aids perceptual normalization. Phonetic features are universal, but phonological categories are language-specific. The identified phonetic features are submitted to the phonological stage for further processing according to hierarchical speech processing theories (e.g., the TRACE model in McClelland & Elman, 1986). When mapping these phonetic features onto phonological categories, with the help of language experience, listeners can construct reference points of greater precision for their speech perception. Phonological information seems to be more

important for bilingual/multilingual speakers. According to the speech learning model (Flege et al., 1997), L1 and L2 phonological systems are represented in the same mental space. Therefore, normalization results can be improved by referring to the appropriate phonological system.

9. Conclusion

Normalizing speech variabilities with contextual cues is a useful strategy for accurate speech perception. Vowel normalization has been detected in the acoustic stage by previous studies. The present study has investigated whether this cognitive process also occurs in the phonetic/phonological stage by an ERP experiment and a cross-language vowel identification experiment. The ERP results show that robust vowel normalization in the speech context triggers a larger P2, an ERP component which is functionally related to both phonetic and phonological processes. The cross-language vowel perception experiment revealed that vowel normalization was affected by both phonetic and phonological cues. Together, these results suggest that vowel normalization might occur in the acoustic stage, but robust vowel normalization in the speech context is observed in the P2 time window (130–250 ms) and significantly overlaps with the phonetic and phonological processes of speech perception.

Acknowledgment:

This study was supported by a grant from the Research Grants Council of Hong Kong (GRF: 14408914). The data of the English subjects in Experiment 2 were collected by the first author during her study visit at the University of California, Berkeley. We are grateful to Dr. Keith Johnson for his application for ethical approval at the University of California, Berkeley and his suggestions on the experimental designs.

Open Practices Statement:

The stimuli, data, and analysis scripts are available through the Open Science Framework (https://osf.io/f8pj6/).

References

Aguiar, M. R., & Sala, O. E. (1998). Interactions among grasses, shrubs, and herbivores in Patagonian grass-shrub steppes. *Ecologia Austral*, *8*(2), 201–210. https://doi.org/10.18637/jss.v067.i01

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 103–113). Academic Press. https://ci.nii.ac.jp/naid/10012630848/en/

Andics, A. (2006). Distinguishing between prelexical levels in speech perception: an adaptation-fMRI study. *Nijmegen CNS*, *1*, 47–66.

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

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/*.

Brouwer, H., & Hoeks, J. C. J. (2013). A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7:758*. https://doi.org/10.3389/fnhum.2013.00758

Bruderer, A. G., Kyle Danielson, D., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

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),

E91988. https://doi.org/10.1371/journal.pone.0091988

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology*, *115*(4), 732–744. https://doi.org/10.1016/j.clinph.2003.11.021

Davidson, D. J., & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychologica*, *144*(1), 83–96. https://doi.org/10.1016/j.actpsy.2013.04.016

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

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

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

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3–4), 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural Dynamics of Variable-Rate Speech Categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 481–503. https://doi.org/10.1037/0096-1523.23.2.481

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler kimberlee. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of*

America, *97*(5), 3099–3111.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312. https://doi.org/10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5), 2801–2817. https://doi.org/10.1121/1.2354071

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1), 156–169. https://doi.org/10.1016/S0378-5955(02)00383-0

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., & Sjerps, M. (2018). Speaker normalization in speech perception. In *UC Berkeley Phonetics and Phonology Lab Annual Report (2018)*.

Joos, M. (1948). Acoustic Phonetics. *Language*, *24*(2), 1–136.

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

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

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

Lin, T., & Wang, S. Y. W. (1984). Shengdiao Ganzhi Wenti (Tone perception). *Zhongguo Yuyan Xuebao (Chinese Linguistics)*, *2*(February), 59–69.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183. https://doi.org/10.3758/BF03193667

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*(4), 602–619. https://doi.org/10.3758/BF03206049

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *The Journal of the Acoustical Society of America*, *113*(1), 53–56. https://doi.org/10.1121/1.1527959

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

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

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412. https://doi.org/10.3758/BF03204884

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

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. *Science*, *343*(February), 1006–1010. https://doi.org/10.1126/science.1245994

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

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex*, *17*(10), 2251–2257. https://doi.org/10.1093/cercor/bhl133

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Joual of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Phillips, C. (2001). Levels of representation in the electrophisiology of speech. *Cognitive Science*, *25*, 711–731. https://doi.org/10.1207/s15516709cog2505_5

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*(June 2019), 104070. https://doi.org/10.1016/j.jml.2019.104070

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. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*(4), 1195–1215. https://doi.org/10.3758/s13414-011-0096-8

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). 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., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics*, *41*, 145–155. https://doi.org/10.1016/j.wocn.2013.01.005

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

Wade, T., & Holt, L. L. (2005). Effects of later-occurring nonlinguistic sounds on speech categorization. *The Journal of the Acoustical Society of America*, *118*(3), 1701–1710. https://doi.org/10.1121/1.1984839

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *90*(6), 2942–2955. https://doi.org/10.1121/1.401769

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *The Journal of the Acoustical Society of America*, *96*(3), 1263–1282. https://doi.org/10.1121/1.410275

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *The Journal of the Acoustical Society*

of America, *99*(6), 3749–3757. https://doi.org/10.1121/1.414981

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)

Woods, D. L. (1995). The component structure of the N1 wave of the human auditory evoked potential. *Electroencephalography and Clinical Neurophysiology. Supplement*, *44*(May), 102–109.

Yost, W. A. (1996). Pitch of iterated rippled noise. *J Acoust Soc Am*, *100*(1), 511–518. https://doi.org/10.1121/1.415873

Zee, E. (2003). Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1117–1120.

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., Peng, G., Wang, X., & Wang, W. S. (2015). Cumulative effects of phonetic context on speech perception. *Proceedings of the 18th International Congress of Phonetic Sciences*.

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

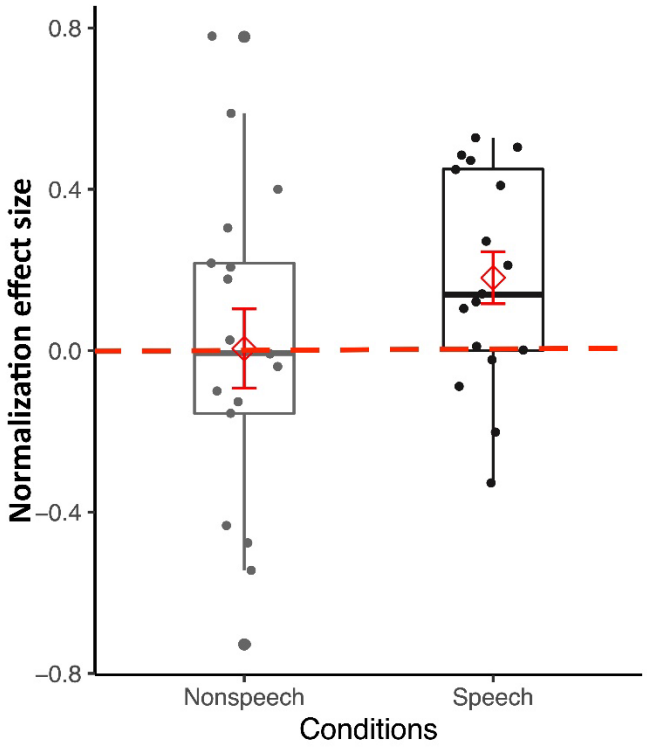
Zhang, C., Xia, Q., & Peng, G. (2015). Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics*, *33*, 149–162. https://doi.org/10.1016/j.jneuroling.2014.07.002

Zhang, K., Sjerps, M. J., & Peng, G. (2021). Integral perception, but separate processing: The perceptual normalization of lexical tones and vowels. *Neuropsychologia*, *156*, 107839. https://doi.org/10.1016/j.neuropsychologia.2021.107839

Supplemental materials

Supplementary Table 1. The word list used for recordings in Experiment 2

ji55	ji25	ji33	ji21	ji23	ji22
醫	椅	意	移	議	二
jy55	jy25	jy33	jy21	jy23	jy22
於	瘀	酗	漁	羽	寓
ʃɛ55	ʃɛ25	ʃɛ33	ʃɛ21	ʃɛ23	ʃɛ22
賒	寫	卸	蛇	社	射
hœ55	tœ25	kœ33	khœ21	tœ23	tʃœŋ22
靴	朵	鋸	瘸	唾	杖
wu55	wu25	wu33	wu21	fu23	wu22
烏	捂	惡	湖	婦	互
wɔ55	wɔ25	wɔ33	wɔ21	tʃhɔ23	wɔ22
窩	捰	喎	和	坐	禍



Supplementary Figure 1. The normalization effect size in nonspeech condition and the speech condition. The normalization effect size is the proportions of /bo55/ response in low context minus the proportions of /bo55/ responses in high context. Each dot represents one participant's result. The diamond is the grand average and the error bar represents the standard error of the mean.

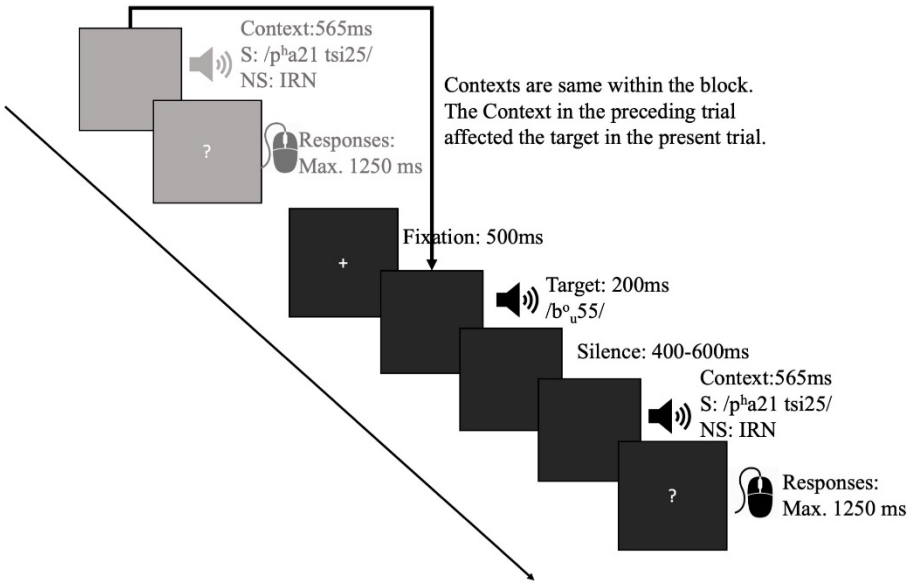


Figure 1. The trial procedure of the word identification task in Experiment 1. The gray frames relate to the previous trial and the black frames are in the present trial.

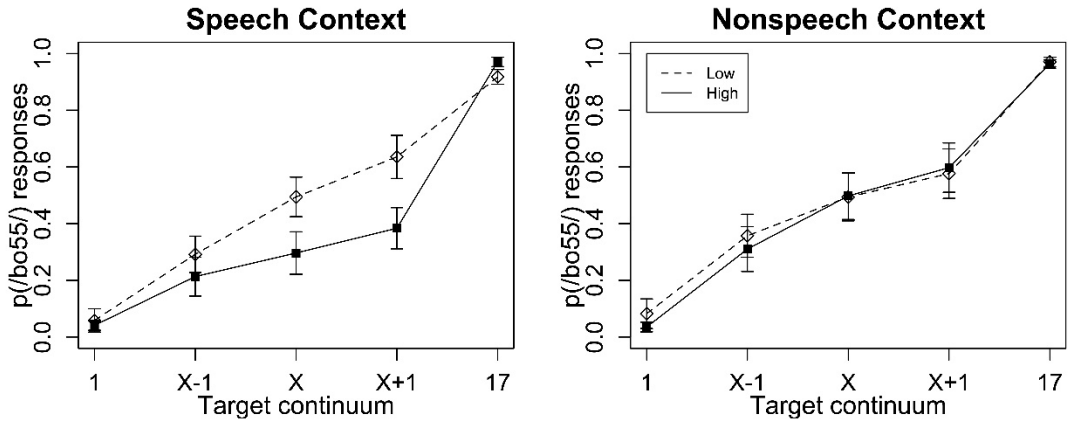
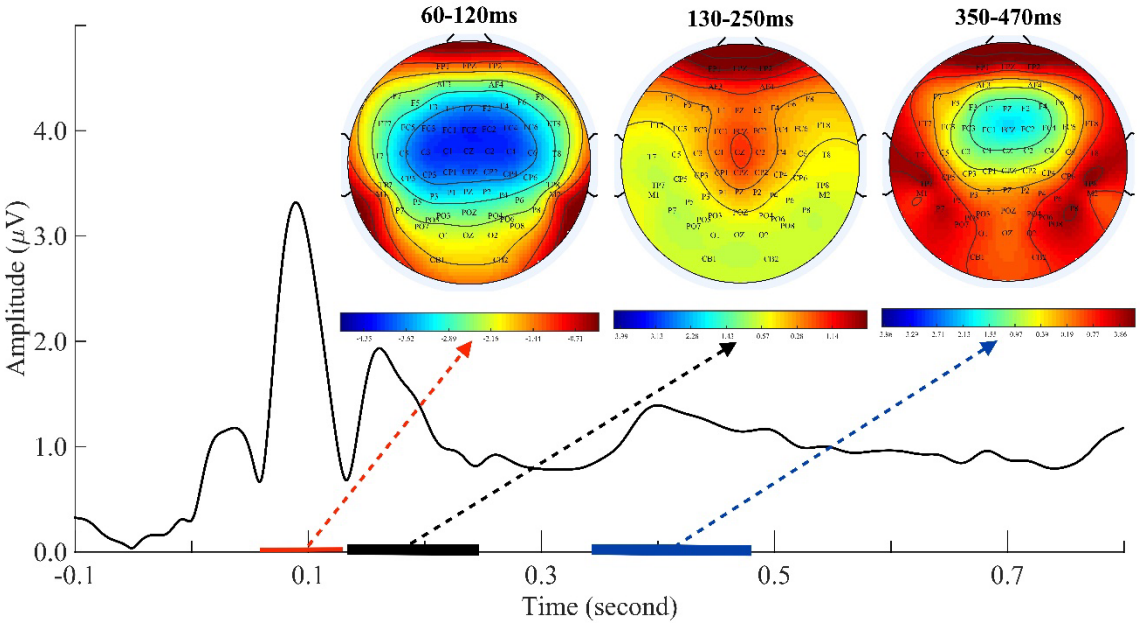
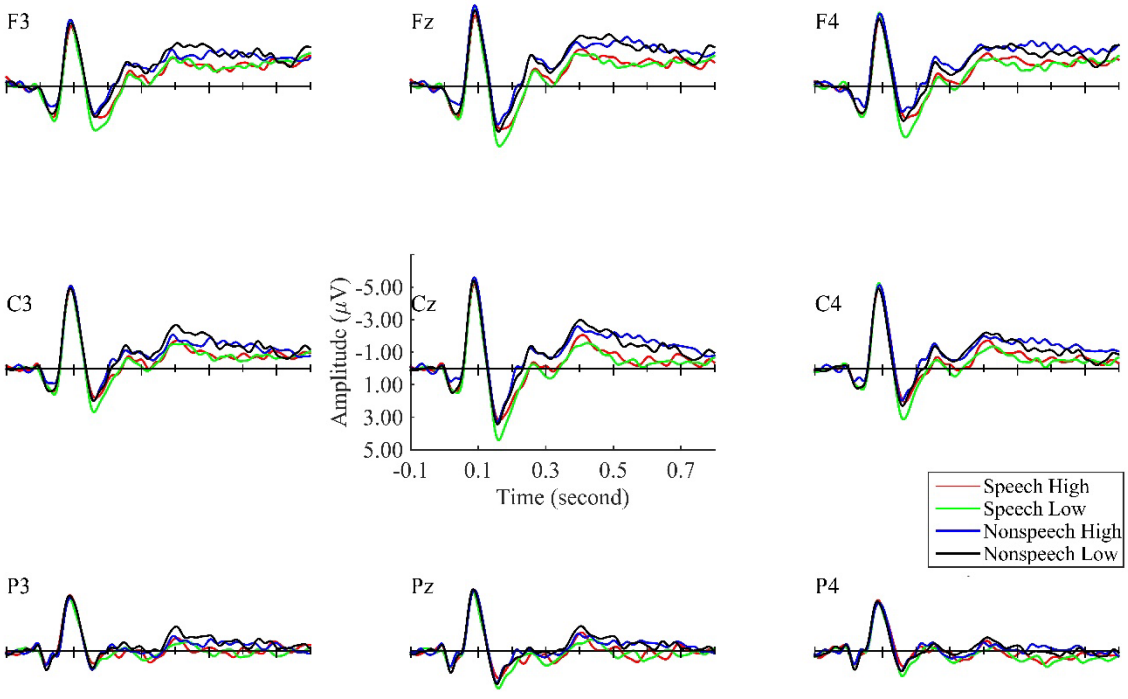


Figure 2. The proportion of /bo55/ responses in different conditions.


60-120ms
130-250ms
350-470ms
Amplitude (μV)
4.0
3.0
2.0
1.0
0.0
-0.1
0.1
0.3
0.5
0.7
Time (second)

(a)


F3
Fz
F4
C3
Cz
C4
P3
Pz
P4
Amplitude (μV)
-5.00
-3.00
-1.00
1.00
3.00
5.00
-0.1
0.1
0.3
0.5
0.7
Time (second)
Speech High
Speech Low
Nonspeech High
Nonspeech Low

(b)

Figure 3. (a) the global field power and the topographic maps in the N1 (left), P2 (middle), and N400 (right) time windows; (b) ERPs at nine electrodes in the four contextual conditions.

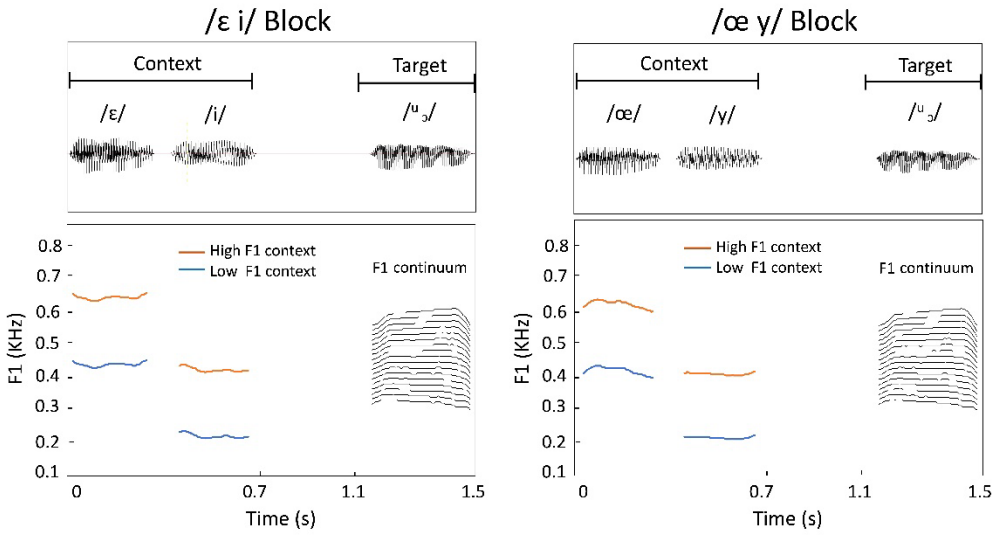


Figure 4. The parameters used to synthesize the stimuli in Experiment 2.

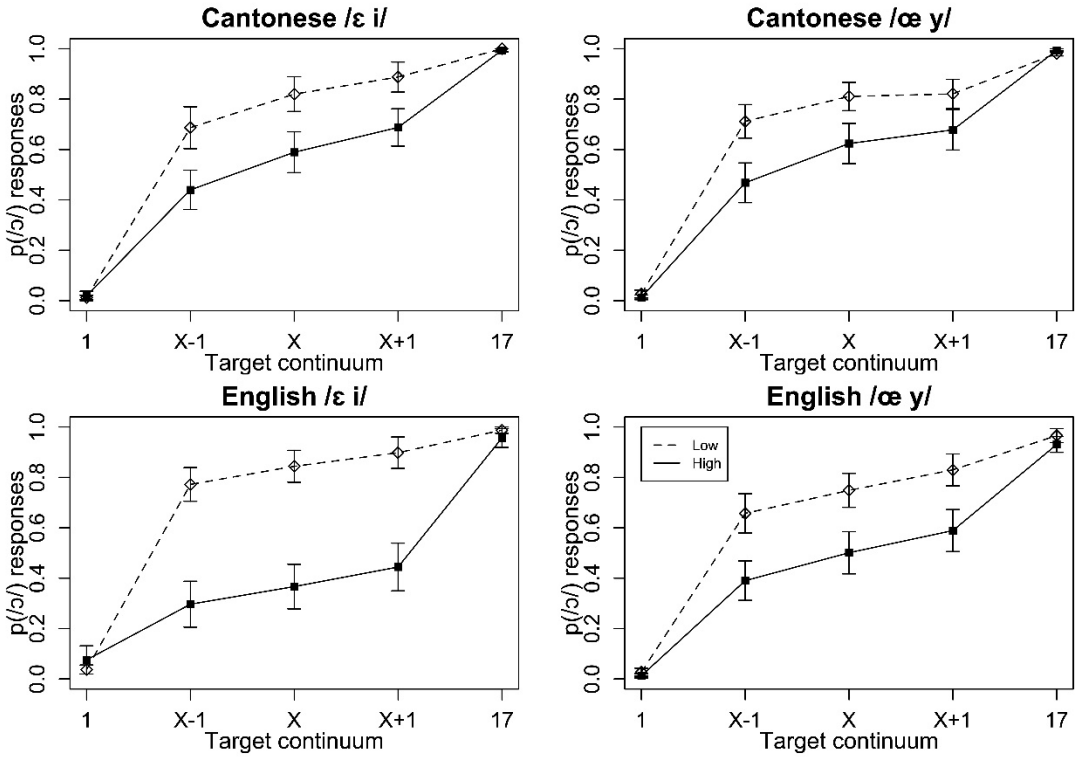


Figure 5. The proportion of /ɔ/ responses in different contexts by Cantonese (top) and English (bottom) subjects.