https://dx.doi.org/10.1016/j.wocn.2024.101299

Achieving perceptual constancy with context cues in second language speech perception

Kaile Zhang[a], Defeng Li[b], Gang Peng[c*]

[*] Corresponding author: gpeng@polyu.edu.hk

[a] Centre for Cognitive and Brain Sciences, University of Macau, Taipa, Macau Special Administrative Region, China

[b] Centre for Studies of Translation, Interpreting and Cognition, Faculty of Arts and Humanities, University of Macau, Taipa, Macau Special Administrative Region, China

[c] Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong Special Administrative Region, China

Abstract

Context cues are useful for listeners to normalize speech variability and achieve perceptual constancy. It remains unknown whether this normalization strategy is language-independent and can be generalized directly from the perception of first language (L1) to second language (L2). To answer this question, Experiment 1 in the present study asked Mandarin learners of Cantonese to perceive ambiguous Cantonese tones with context cues. The results revealed a significant Cantonese-tone normalization process in Mandarin learners, but the effect size was smaller than native speakers, suggesting that speech normalization required language-specific knowledge and thus it was refined gradually during L2 acquisition. The results also showed that even with effective context cues, Mandarin learners tended to give more high level tone responses, a tone also in Mandarin, implying that L1 phonological system interacts with immediate L2 context during L2 speech normalization. Experiment 2 revealed that L2 immersion but not overall L2 proficiency or L2 phonological proficiency facilitated L2 normalization process, indicating that L2 speech normalization improved with perceptual practice and needed more high-level L2 knowledge than L2 phonology.

Keywords

Speech variability, perceptual normalization, context cues, L2 acquisition, lexical tones, language immersion

Achieving perceptual constancy with context cues in second language speech perception

1.0 Introduction.

Speech signals produced by different talkers vary a lot due to individual anatomic configurations of vocal tracts (Peterson & Barney, 1952). Even the same speaker's speech changes noticeably in different psychological conditions. The lack of one-to-one mapping between highly variable speech signals and the abstract linguistic representations is a fundamental difficulty for human speech perception (K. Johnson, 2005; Liberman et al., 1967; Sjerps et al., 2019). Listeners took more time to identify words presented in mixed-talker conditions than in blocked-talker conditions since the mixed-talker condition required frequent adaptation to new talkers (Nusbaum and Morin, 1992). Sometimes, talker variability even blurs the boundary of two similar phonemes. For example, fundamental frequency (f0) is the main acoustic cue to identify lexical tones, but the f0 of a female speaker's low tone could be even higher than the f0 of a male speaker's high tone, resulting in more mistakes for tone perception in multiple-talker condition. Talker variability impacts not only native speakers but also second language (L2) learners. While there has been considerable research into how native speakers manage talker-related speech variability, few studies have examined the perceptual strategies L2 learners employ to cope with this variability—this gap will be addressed by the current study. The investigation on this question would enhance the understanding of the challenges inherent in L2 speech perception. The following part of the introduction would first review relevant studies in L1 and L2 speech normalization, and then present the research plan for the present study.

1.1 Normalizing talker variability with context cues

Although talker variability presents challenges for fast and accurate speech perception, in most cases, native listeners are able to communicate successfully with different talkers. One way to achieve perceptual constancy is to remove talker-related differences from each speech token and obtain a talker-independent linguistic pattern for phonemic identification (i.e., speech perception with normalization process) (K. Johnson, 2005). This perceptual strategy was supposed by recent ECoG studies which found that the talker-independent speech cues were represented at the auditory cortex (Oganian et al., 2023; Sjerps et al., 2019). Research has

shown that talker variability can be reduced through the use of contextual cues—a mechanism referred to as extrinsic normalization (Ainsworth, 1975; Nearey, 1989). For instance, a level tone from a multiple-speaker corpus can hardly be identified as either high tone or low tone, but the identification rate improved dramatically once it was presented in a speech context (Peng et al., 2012; Wong & Diehl, 2003). Contextual cues generally influence the perception of the target speech in a contrastive manner. Specifically, an ambiguous tone is more likely to be perceived as low when preceded by a high-pitched context and as high when preceded by a low-pitched context, an effect known as the contrastive context effect (Moore & Jongman, 1997; K. Zhang et al., 2021).

Although extrinsic normalization has been widely demonstrated to be effective in overcoming talker variability, it is inconclusive about the underlying cognitive mechanism. Understanding the cognitive mechanism of extrinsic normalization is crucial for the present study, as various theoretical models yield contrasting predictions regarding L2 learners' normalization performance (refer to Section 1.2 for details). Prior to delving into L2 speech normalization, the present study would first explore the different theoretical frameworks about extrinsic normalization. The existing theories regarding extrinsic normalization fall into two broad categories, distinguished by the level at which the process occurs: either at a general-auditory level or at a speech-specific level.

Some researchers have posited that the extrinsic normalization relies on the acoustic properties of speech signals and operates at the general-auditory level (e.g., Holt & Lotto, 2002). In the extrinsic normalization process, the auditory system first assesses the acoustic properties of the preceding context (e.g., long-term average spectrum, f0, and durations) and then automatically encodes the target speech cue in contrast to the acoustic properties of the preceding context (Watkins & Makin, 1994, 1996). The neural adaptation within the central auditory system might be the biological basis for the acoustically contrastive encoding account of extrinsic normalization. Neurons in the primary auditory cortex, attuned to different frequency regions, become less responsive to frequencies similar to those in the preceding context due to adaptation. Conversely, neurons that were less active during the context phase become more sensitive to the frequencies of subsequent sounds, leading to the spectral contrast perception (Stilp, 2019). Supporting evidence for this general-auditory level process comes from studies like Holt et al. (2001) which demonstrated that even birds trained to recognize speech sounds exhibit contrastive perceptual behavior.

Other scholars are of the view that extrinsic normalization necessitates language-specific knowledge and functions at the level of speech-specific language processing. Nusbaum and Morin, (1992); Nusbaum & Magnuson, (1997) proposed that listeners could use context information to compute a talker-specific mapping between acoustic patterns and abstract phonemic representations, and ambiguous target cues were identified by referring to the talker-specific acoustic-phonemic mapping. For example, in the Cantonese greeting "早晨" (/zou 25 san 21/, good morning), the highest F0 aligns with the highest pitch (the end point of T25) and lowest F0 matches the lowest pitch in Cantonese (the end point of T21), which outline a speaker's acoustic space for the phonemic system (the tonal system in this example) (K. Zhang et al., 2023). The subsequent tones were perceived by referring to the talker-specific acoustic-phonemic space. Based on the acoustic-phonemic mapping account of the extrinsic normalization, language-specific knowledge is required since listeners at least need to know phonological categories to construct the talker-specific acoustic-phonemic mapping. Beyond phonological information, other higher-level linguistic knowledge, such as semantic or syntactic information, is also helpful for the normalization process, as C. Zhang et al. (2015) demonstrated that contexts composed of meaningful speech elicited better normalization effects than meaningless word sequencies.

1.2. Speech normalization in L2 speech perception

Talker variability also posed difficulties for L2 speech perception. Perceiving high-variability L2 speech could be a dual challenge for L2 learners, as they have to cope with talker variability and imperfect L2 knowledge (Lecumberri et al., 2010). Based on the cognitive mechanisms of extrinsic normalization outlined previously, opposite predictions would be drawn concerning L2 learners' handling of L2 speech variability. Should extrinsic normalization be a general-auditory level process, it would likely be language-independent, allowing the normalization skills developed in L1 to be directly applied to L2. This would result in comparable performances in L2 leaners and native speakers in terms of managing L2 speech variability. Conversely, if extrinsic normalization operates at a speech-specific level that relies on language-specific information, it would be language-dependent. In this case, L2 learners' proficiency in extrinsic normalization is expected to incrementally reach a native-like standard as their L2 proficiency increases.

Albeit the well-known challenge of speech variability in L2 speech perception, only a limited amount of research investigated how L2 learners deal with speech variability, yielding different results. The central debate surrounding these studies concerns whether the normalization process is language-independent and can be applied directly to L2 based on prior experience with L1 speech, echoing the dispute between two theoretical frameworks reviewed above. Bradlow & Pisoni, (1999) found that talker variability had a comparable effect on English word identification in native and non-native English speakers (7.2% vs. 7.9%) which was quantified by calculating the difference in listeners' word identification rates between the multiple-talker and single-talker conditions. Similar results were observed in studies examining the perception of Mandarin tones (C. Y. Lee et al., 2013) and Mandarin fricatives (C.-Y. Lee et al., 2012), where native and non-native speakers showed equivalent performance decrements in mixed-talker condition (i.e., high talker variability) relative to the blocked-talker condition (i.e., low talker variability). These findings led Bradlow and Pisoni (1999) and C. Y. Lee et al. (2013) to conclude that the ability to deal with talker variability was a language-independent skill that transferred easily from L1 to L2. However, Antoniou et al. (2015) reported a more severe impact of speech variability on L2 learners than on native speakers in an English word-monitoring task, in terms of signal detection accuracy and recognition time, suggesting a poorer speech normalization ability of L2 learners. Additionally, Bent et al., (2010) found that Korean listeners identifying English vowels from multiple speakers achieved a 22% lower accuracy rate than native English speakers, with the error pattern analysis indicating a strong influence of production variability within the normal range for native English talkers. This suggests that non-native listeners may be more susceptible to a cross-talker variability due to difficulties in forming native-like vowel categories. The limited cognitive resources available for L2 learners, particularly those with low L2 proficiency, also matter since speech normalization is constrained by cognitive faculties like attention and working memory (Kapadia & Perrachione, 2020; Nusbaum & Magnuson, 1997; Wong et al., 2004).

Bradlow and Pisoni (1999), C. Y. Lee et al. (2013), and Antoniou et al. (2015) investigated L2 speech normalization using a blocked- vs. mixed- talker (or the single- vs. multiple- talker) paradigm. Talker identity shifts more frequently in the mixed-talker condition, requiring increased talker normalization. However, the stimuli in these studies were words presented in isolation but not in contexts, which meant that no explicit external information was available for speech normalization. As a result, the contrasts between blocked and mixed-talker conditions in these studies demonstrate the extent to which talker variability impacts L2

speech perception, rather than how effectively L2 learners utilize contextual cues to mitigate this variability. To better quantify the extrinsic normalization strategy employed by L2 learners, language tasks such as the identification of multiple-talker speech within contexts could be more appropriate, which would be adopted by the present study. So far, no study has directly tested how well L2 learners could use context cues to overcome talker variability. The most related study was done by C. Y. Lee et al. (2009) who tested if contexts could help native Mandarin speakers and English learners of Mandarin to identify incomplete Mandarin tones, such as missing the central part. The authors found that contexts were helpful for both groups, but native speakers demonstrated a greater improvement in the presence of context cues compared to L2 learners. Although C. Y. Lee et al. (2009) did not explicitly address the question about speech normalization, it is reasonable to infer that L2 learners also cannot use context cues as effectively as native speakers to normalize talker variability.

1.3 The present study

Extrinsic normalization has been extensively investigated in the context of native speech perception. However, it remains unclear whether L2 learners can utilize contextual cues to accommodate talker variability with the same effectiveness as native speakers. This inquiry is particularly significant as it sheds light on whether extrinsic normalization, as a crucial strategy for achieving perceptual constancy, is a general auditory-level process or a language-dependent process that is gradually acquired during L2 learning, similar to L2 learners' cue weight strategy in speech perception. Further investigation into this matter could also inform L2 education by highlighting the importance of training this ability in curriculum design.

The present study aimed to probe this question by investigating Mandarin learners' extrinsic normalization of Cantonese tones. This approach differed from C. Y. Lee et al. (2009) and C. Y. Lee et al. (2013) who tested the English learners' perception of Mandarin tones in several aspects. Firstly, the present study utilized Cantonese tones as testing material rather than Mandarin tones. Mandarin has four lexical tones: high level (T55), high rising (T35), falling rising (T214), and high falling (T51), each characterized by unique pitch heights and contours, thus relying primarily on intrinsic cues for differentiation. When these intrinsic cues are available, the influence of context is relatively minimal. In contrast, the discrimination of Cantonese tones, particularly the three level tones, depends heavily on extrinsic contexts. Cantonese has three level tones: high level (T55), mid level (T33), and low level (T22). A base syllable with different level tones has different meanings. For example, the base syllable /ji/

means ‘doctor’ with T55, ‘meaning’ with T33, and ‘two’ with T22. Three level tones share a similar pitch contour (i.e., level) but different pitch heights, and thus, they can only be distinguished by pitch heights. However, individual talkers' vocal fold configurations can cause them to produce identical tones at different pitch heights, compromising the reliability of this intrinsic cue. Therefore, context cues are indispensable for interpreting Cantonese level tones, making them more suitable than Mandarin tones for testing the extrinsic normalization process (e.g., Wong & Diehl, 2003; C. Zhang et al., 2012; K. Zhang et al., 2017).

Second, Mandarin speakers instead of English speakers (tonal vs. nontonal language speakers) were invited. C. Y. Lee et al. (2013) found that both English speakers and native Mandarin speakers were equally affected by talker variability while perceiving Mandarin tones. Consequently, they concluded that the ability to deal with talker variability could be transferred easily from L1 to L2. However, since English speakers have no experience with lexical tone normalization, it is not rigorous to conclude that the normalization strategies can be directly transferred from L1 to L2 based on English speakers’ normalization of Mandarin tones. Mandarin learners, as speakers of a tonal language, are proficient in the tone normalization process, at least in Mandarin. Recruiting Mandarin learners of Cantonese instead of non-tonal language speakers made it more feasible to test whether the normalization process for lexical tones acquired in L1 can be directly generalized to L2 tone perception. Similar to Cantonese speakers, Mandarin learners exhibit considerable confusion with Cantonese tones that have similar contours but different pitch heights, such as the three level tones (T55, T33, and T22) and two rising tones (T25 and T23) (K. Zhang et al., 2018). Owing to the influence of their native language (i.e., Mandarin T55), Mandarin learners often misperceive Cantonese T33 and T22 as T55 when the stimuli are presented in isolation. Therefore, extrinsic contexts may be even more crucial for Mandarin learners to disambiguate the similar Cantonese tones.

Two experiments were carried out to test the L2 speech normalization process. Experiment 1 aimed to test whether extrinsic normalization process is language-independent and can be generalized easily to L2 based on L1 language use experience. Experiment 2 followed up on the results of Experiment 1 and aimed to identify the factors that contribute to L2 normalization. L2 learners’ speech normalization performance was assessed in Experiment 1 by testing their ability to use context cues to accommodate multiple-talker speech, as opposed to the blocked- vs. mixed- talker contrast utilized by Bradlow and Pisoni (1999), C. Y. Lee et al. (2013), and Antoniou et al. (2015). Mandarin learners of Cantonese were asked to perceive

Cantonese level tones produced by multiple talkers in speech contexts with different pitch heights. The preceding context typically influences speech perception in a contrastive manner. For instance, a target tone is perceived as low if it follows a context spoken by a high-pitched speaker, whereas the same target tone is perceived as high following a context from a low-pitched speaker. Therefore, if Mandarin learners interpreted target tones by referring to context cues, they should give more T55 responses (i.e., high tone) in low-f0 contexts, more T33 responses in mid-f0 contexts, and more T22 (i.e., low tone) responses in high-f0 contexts. An experimental condition with minimal context influence was included as a baseline. To keep experimental conditions more comparable, the present study used a nonspeech-context condition rather than the isolated condition to evaluate Cantonese tone perception without effective context information, since Cantonese speakers' tone perception in nonspeech contexts was similar as in isolation (Francis et al., 2006; C. Zhang et al., 2013; K. Zhang & Peng, 2021). Native Cantonese speakers were recruited as the comparison to see if L2 learners could use the extrinsic normalization strategy as effectively as native speakers did. Based on previous findings (C. Y. Lee et al., 2009), the present study hypothesized that Mandarin learners of Cantonese would exhibit a smaller normalization effect compared to native Cantonese speakers. Most Mandarin participants from Experiment 1 also participated in Experiment 2 which measured several potential factors related to L2 normalization. The combination of two experiments aimed to shed light on the generalizability and factors influencing the extrinsic normalization process in L2 speech perception.

EXPERIMENT 1: MANDARIN LEARNERS' EXTRINSIC NORMALIZATION OF CANTONESE TONES

2.0 Materials and Methods

2.1 Participants

To ensure that the Mandarin participants in the study had adequate knowledge of Cantonese and could accurately complete the tasks, a short screening test was conducted. The screening test comprised two parts. In the first part, a native Cantonese speaker who specialized in linguistics read six sentences, three of which were used as contexts in the present study. The Mandarin participants were asked to translate these sentences into Mandarin. In the second part of the screening test, the Mandarin participants were asked to produce six Cantonese words containing three target syllables (/ji22/, /ji33/, /ji55/), and their pronunciation was evaluated by

the same native Cantonese speaker. Only Mandarin participants who correctly translated the Cantonese sentences and correctly produced the Cantonese words were invited to participate in the Cantonese-tone normalization task.

Ultimately, 30 young adult native Mandarin speakers (female = 16, male = 14) who were learning Cantonese participated in the Cantonese-tone normalization task. All Mandarin participants were born in Mainland China. While some spoke Chinese dialects in addition to Mandarin, they all identified Mandarin as their native language. Of the 30 participants, five were not born in Northern China (Guizhou, Jiangxi, Hubei, Chengdu, and Shenzhen, respectively). However, none of their local Chinese dialects contains more than one level tone, and the subject born in Shenzhen speaks Mandarin instead of Cantonese. Thus, they were included as well. Additionally, 30 young adult native Cantonese speakers (female = 15, male = 15) born in Hong Kong were invited as the comparison group. None of the participants had received more than three years of professional music training. All participants reported no hearing, speech, or language difficulties. All participants were right-handed which was assessed using the Edinburgh Handedness Inventory (Oldfield, 1971). The study was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. Comprehensive insights into the experimental procedures were provided to all participants, and informed consent was acquired prior to the commencement of the experiment.

2.2 Stimuli

The stimuli used in the Cantonese tone normalization task were the same as those in Tao et al. (2021) and K. Zhang et al. (2017). Each trial consists of either a speech context or a nonspeech context and a speech target. The speech context is a Cantonese phrase 呢個字係 (/li55 ko33 tsi22 hɐi22/, this word is …), which was produced by four native Cantonese speakers of different pitch ranges (a female of high pitch, a female of low pitch, a male of high pitch, and a male of low pitch). The f0 trajectories of the original recordings were further raised or lowered three semitones in Praat (Boersma & Weenink, 2023) to introduce the intra-talker variability and also to trigger the contrastive context effect, resulting in three contexts of different pitch heights: high-f0 speech context, mid-f0 speech context, and low-f0 speech context. The nonspeech contexts were composed of triangle waves. The pitch heights of nonspeech contexts were manipulated to match the pitch heights of their speech counterparts, resulting in three types of nonspeech contexts: high-f0 nonspeech context, mid-f0 nonspeech

context, and low-f0 nonspeech context. The speech targets were the Cantonese syllable 意 (/ji33/, meaning) which were produced by the same four speakers. The task also included some fillers produced by the female speaker of low pitch and the male speaker of high pitch, which underwent the same pitch manipulation as the test stimuli. The context fillers were Cantonese phrases: 我而家讀 (/ŋo23 ji21 ka55 tuk2/, now I will read...) and 請留心聽 (/ts^{h}iŋ25 ləu21 sɐm55 t^{h}iŋ55/, please listen to…carefully). The target fillers were Cantonese 意 (/ji33/, meaning) or 二 (/ji22/, two). The speaker of the context and target stimuli was matched in each trial. Detailed f0 information of the auditory stimuli used in the present study can be found in Table 1 in Tao et al. (2021). The durations of context stimuli were kept unchanged, while the duration of the target stimuli was normalized to 450 ms. The intensity of speech stimuli was normalized to 55dB, but the intensity of nonspeech stimuli was normalized to 75 dB to match the perceived loudness of speech stimuli, which was evaluated by the native Cantonese speakers.

2.3 Experimental procedure

The Cantonese tone normalization task which contained two blocks was conducted in a sound-proofed booth: a speech-context block and a nonspeech-context block. Each block was composed of 108 test trials (4 speakers × 3 pitch heights × 9 repetitions) and 36 fillers. The trial procedure was illustrated in Figure 1. The audio stimuli were delivered bilaterally to subjects via inserted earphones. In each trial, a 500-ms fixation was played first, which was followed by a context stimulus. A period of silence jittered between 300 ms and 500 ms appeared between context and target. After the target sound, a question mark appears on the screen for 2000 ms. Subjects were asked to press the corresponding keys on the keyboard to indicate whether the last word they heard was 醫 (/ji55/, doctor), 意 (/ji33/ meaning), or 二 (/ji22/, two). The next trial appears after the elimination of the question mark.

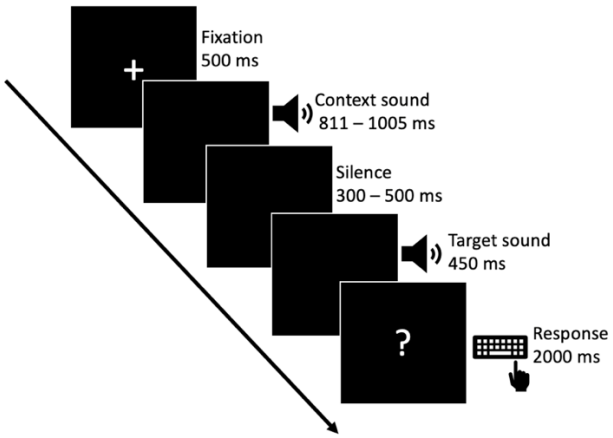

Figure 1. ***The trial procedure of the Cantonese-tone normalization task***

3.0 Results

Figure 2 illustrates the percentages of three response choices (i.e., 医 /ji55/ "doctor", 意 /ji33/ "meaning", and 二 /ji22/ "two") in each experimental condition. To statistically evaluate how listeners used context cues to interpret target tones, a multinomial logistic regression model was fitted to all participants' responses in the Cantonese-tone normalization task, using the *nnet* package (Venables & Ripley, 2002) in *R*. The dependent variable was *Response category* with three levels (/ji22/, /ji33/, and /ji55/). The predictors included *Pitch height* (low, mid, and high), *context* (nonspeech and speech), *group* (Cantonese and Mandarin), and their possible two-way and three-way interactions. The by-subject intercept was also included. All the predictors were centered around 0 (*group*: -1 for Cantonese and 1 for Mandarin; *pitch height*: -1 for low, 0 for mid, and 1 for high; *context*: -1 for nonspeech and 1 for speech). The reference level for *response category* was set to /ji33/. The fitted model could explain significantly more variance than the baseline model that only had by-subject intercept (AIC difference = 4533.441, $p < 0.001$).

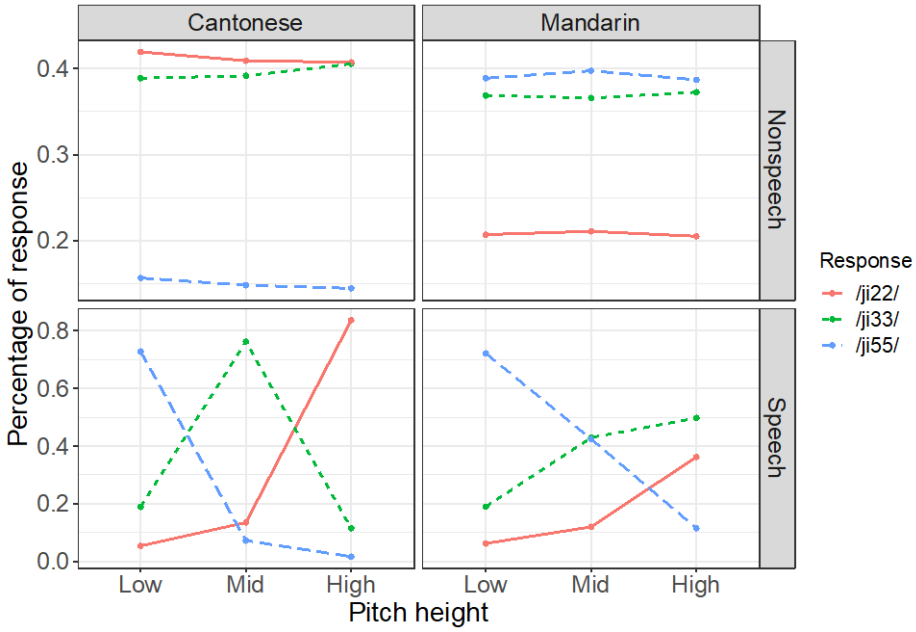

Figure 2. ***The percentage of three responses in each condition in the Cantonese-tone normalization task***

The statistics of the multinomial logistic regression model were summarized in Table 1. The significant main effect of *context* indicated that participants gave significantly fewer /ji55/ responses (relative to /ji33/) ($\beta = -0.233$, $p < 0.001$) and fewer /ji22/ responses (relative to /ji33/) ($\beta = -0.347$, $p < 0.001$) in speech contexts than in nonspeech contexts. The significant main effect of *group* suggested that Mandarin learners, on average, provided more /ji55/ responses (relative to /ji33/) ($\beta = 0.692$, $p < 0.001$) and fewer /ji22/ responses (relative to /ji33/) ($\beta = -0.127$, $p < 0.001$) than Cantonese speakers, indicating an influence of Mandarin T55. The significant main effect of *pitch height* indicated that participants on average gave significantly fewer /ji55/ responses (relative to /ji33/) ($\beta = -1.054$, $p < 0.001$) and more /ji22/ responses (relative to /ji33/) ($\beta = 0.739$, $p < 0.001$) when pitch height changed from low to high, illustrating a contrastive context effect. However, this effect varied between groups, as indicated by the significant *pitch height* by *group* interaction. With an increase in context pitch height from low to high, Mandarin learners gave more /ji55/ responses (relative to /ji33/) ($\beta = 0.415$, $p < 0.001$) and fewer /ji22/ responses (relative to /ji33/) ($\beta = -0.507$, $p < 0.001$) than Cantonese speakers, suggesting a diminished contrastive context effect among Mandarin learners. Additionally, the contrastive context effect varied between speech and nonspeech contexts, as revealed by the significant *pitch height* by *context* interaction. When pitch height increased from low to high, participants gave fewer /ji55/ responses (relative to /ji33/) ($\beta = -$

1.021, $p < 0.001$) and more /ji22/ responses (relative to /ji33/) ($\beta = 0.759$, $p < 0.001$) in speech contexts compared to nonspeech contexts, indicating a more pronounced contrastive context effect in speech contexts. There was also a significant *context* by *group* interaction. When contexts changed from nonspeech to speech, Mandarin learners, compared to Cantonese speakers, gave more /ji55/ responses (relative to /ji33/) ($\beta = 0.197$, $p < 0.001$) and more /ji22/ responses (relative to /ji33/) ($\beta = 0.188$, $p < 0.001$).

Table 1 ***The multinomial logistic regression model for the Cantonese-tone normalization task.***

	/ji55/ (relative to /ji33/)				
	Estimate	SE	95% CI	*z*	*p*
Intercept	-0.355	0.016	[-0.386 -0.323]	-22.255	<0.001
pitch height	-1.054	0.039	[-1.13 -0.977]	-27.014	<0.001
context	-0.233	0.032	[-0.296 -0.171]	-7.316	<0.001
group	0.692	0.032	[0.63 0.755]	21.717	<0.001
pitch height : context	-1.021	0.039	[-1.098 -0.945]	-26.184	<0.001
pitch height : group	0.415	0.039	[0.339 0.491]	10.639	<0.001
context : group	0.197	0.032	[0.134 0.259]	6.166	<0.001
pitch height : context : group	0.386	0.039	[0.31 0.463]	9.904	<0.001
	/ji22/ (relative to /ji33/)				
	Estimate	SE	95% CI	*z*	*p*
Intercept	-0.31	0.014	[-0.337 -0.283]	-22.525	<0.001
pitch height	0.739	0.036	[0.668 0.81]	20.378	<0.001
context	-0.347	0.028	[-0.401 -0.293]	-12.605	<0.001
group	-0.127	0.028	[-0.181 -0.073]	-4.606	<0.001
pitch height : context	0.759	0.036	[0.688 0.83]	20.933	<0.001
pitch height : group	-0.507	0.036	[-0.578 -0.436]	-13.995	<0.001
context : group	0.188	0.028	[0.134 0.242]	6.816	<0.001
pitch height : context : group	-0.523	0.036	[-0.594 -0.452]	-14.421	<0.001

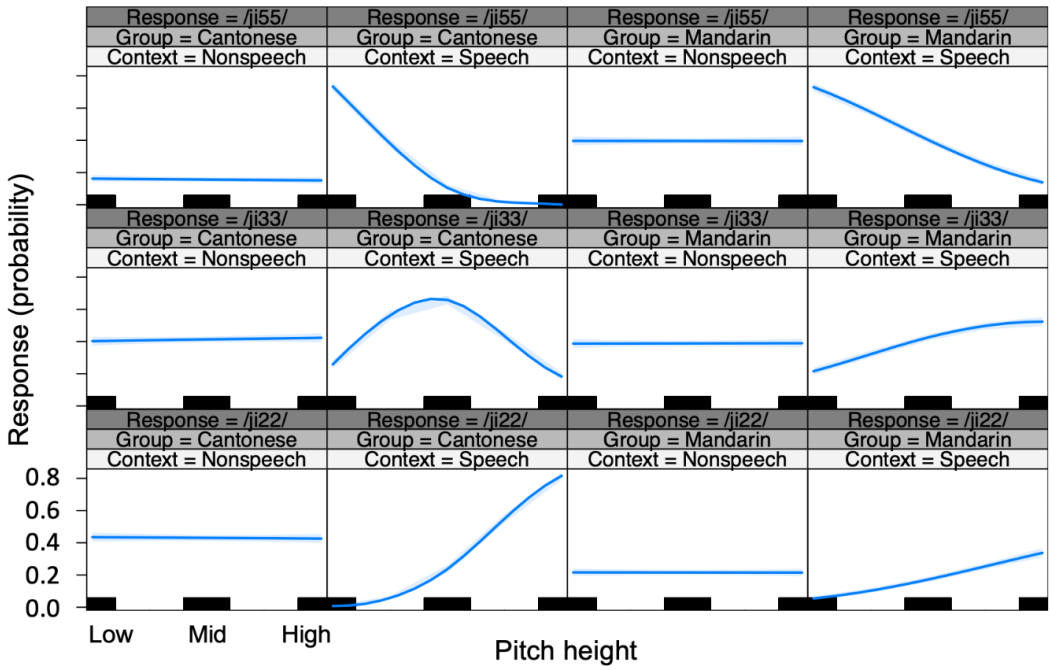

Figure 3. ***The visualization of pitch height by context by group interaction in the Cantonese-tone normalization task.***

The significant *pitch height* by *context* by *group* interaction was depicted in Figure 3, using the *effects* package (Fox & Hong, 2009) in R. It shows the predicted probability of each response in different experimental conditions. First of all, to ascertain how context pitch heights affect the tone perception in each group, a post hoc analysis with Bonferroni adjustment was conducted by comparing each response across three pitch heights for each group, using *emmeans* package (Lenth, 2019) in R. The analysis revealed significant context-dependent tone perception in both groups within speech contexts. Specifically, native Cantonese speakers gave more /ji55/ responses in the low-f0 speech contexts (*prob* = 0.75; *SE* = 0.01) compared to high-f0 (0.017; 0.004) and mid-f0 speech contexts (0.0745; 0.008) (*ps* < 0.001). Native Cantonese speakers gave more /ji33/ responses in mid-f0 speech contexts (0.786; 0.0126) than high-f0 (0.118; 0.01) and low-f0 speech contexts (0.194; 0.0122) (*ps* < 0.001). Native Cantonese speakers gave more /ji22/ responses in high-f0 speech contexts (0.8641; 0.01) than in mid-f0 (0.139; 0.01) and low-f0 speech contexts (0.055; 0.007) (*ps* < 0.001). In speech contexts, Mandarin learners gave more /ji55/ responses in the low-f0 condition (0.718; 0.013) than in high-f0 (0.127; 0.01) and mid-f0 condition (0.426; 0.015) (*ps* < 0.001). Mandarin learners gave more /ji22/ responses in high-f0 speech contexts (0.359; 0.014) than in speech contexts with mid f0 (0.128; 0.01) and low f0 (0.076; 0.008) (*ps* < 0.001). Mandarin learners gave more /ji33/ responses in mid-f0 speech contexts (0.445; 0.015) contexts than in speech contexts with low

f0 (0.205; 0.012) (p < 0.001). However, Mandarin learners' /ji33/ responses did not differ significantly between mid-f0 (0.445; 0.015) and high-f0 (0.5138; 0.015) speech contexts (p > 0.05), the only pair deviating from the contrastive context effect. In non-speech contexts, no significant response differences were observed when context pitch heights varied from low to high for either group (ps > 0.05).

Second, to elucidate group differences, post hoc analysis was conducted on the significant *pitch height* by *context* by *group* interaction by comparing responses between the two groups across each pitch height condition. Significant group differences were observed in both speech- and nonspeech-context conditions. In speech contexts, Mandarin learners gave fewer expected responses than Cantonese speakers. Specifically, in high-f0 speech contexts, Mandarin speakers gave fewer /ji22/ responses (the expected response in high-f0 contexts) than Cantonese speakers (β = - 0.51, SE = 0.02, p < 0.001). This was also the case in mid-f0 speech contexts: Mandarin speakers gave fewer /ji33/ responses than Cantonese speakers (β = - 0.34, SE = 0.02, p < 0.001). Only In low-f0 speech contexts, /ji55/ responses from Mandarin learners and native Cantonese speakers were comparable (β = - 0.032, SE = 0.02, p = 1). In nonspeech contexts, Mandarin learners more frequently perceived the target words (i.e, /ji33/ from multiple speakers) as /ji55/ (*prob* = 0.396; SE = 0.015) and /ji33/ (0.388; 0.015) rather than /ji22/ (0.216; 0.013) (*ps* < 0.001), but native Cantonese speakers were more likely perceive the target words (i.e, /ji33/ from multiple speakers) as /ji33/ (0.403; 0.015) and /ji22/ (0.431; 0.015) rather than /ji55/ (0.156; 0.011) (ps < 0.001).

In sum, the results of the Cantonese-tone normalization task suggested that both groups were not affected by the pitch manipulation of nonspeech contexts, but in the speech-context condition, both groups' tone perception changed significantly according to the pitch heights of the contexts, indicating an extrinsic normalization process in Mandarin learners and Cantonese speakers. However, Mandarin learners gave significantly fewer expected responses than native Cantonese speakers especially in mid-f0 speech contexts (0.445 vs. 0.786), and in high-f0 speech contexts (0.359 vs. 0.8641), indicating a smaller normalization effect in Mandarin learners.

EXPERIMENT 2: THE FACTORS THAT AFFECT EXTRINSIC NORMALIZATION OF L2 SPEECH

The significant but smaller extrinsic normalization of Cantonese tones in Mandarin learners revealed in Experiment 1 suggested that extrinsic normalization process is most likely acquired gradually in L2 acquisition and cannot be directly transferred from L1 to L2 (see section 6.1 for more discussion). This finding raises the question of which factors affect the development of L2 extrinsic normalization. According to the acoustic-phonemic mapping mechanism, learners' familiarity with the Cantonese tonal system is the most probable factor influencing the development of L2 extrinsic normalization. Mandarin learners who were sophisticated in Cantonese tones were more likely to establish a precise acoustic-phonemic mapping to recalibrate ambiguous target speech cues. This is partially supported by studies showing that the development of speech normalization in children is constrained by the maturation of the phonological system. For examples, although 6-month-old children can recognize phonemes from different talkers (Kuhl, 1976), this ability still do not achieve the adult-like level at the age of six (Bent & Atagi, 2017). Mandarin-speaking children cannot use context cues to identify ambiguous tones until the age of six (F. Chen et al., 2023), which is consistent with their maturation in the categorical perception of Mandarin tones (F. Chen et al., 2017; Feng & Peng, 2023).

Aside from acoustic and phonemic information, higher level linguistic knowledge also contributes to the normalization process. Meaningful contexts triggered 11% higher normalization effect than contexts composed of meaningless word sequence (C. Zhang et al., 2015). Therefore, overall L2 proficiency may also affect the development of L2 speech normalization process. L2 learners' insufficient knowledge of the sociolinguistic variation of L2 was correlated with their poor performance in accommodating high-variability L2 speech (Tamati & Pisoni, 2014). L2 immersion, in such a condition, is essential to boost the knowledge of L2 sociolinguistic variation. Only 21-day Spanish exposure led to an improvement in acquiring the L2 fine-phonetic details associated with stop voicing (Casillas, 2020). Long-term Finnish exposure even enabled L2 learners to distinguish the Finnish vowel contrast at the pre-attentive level (Winkler et al., 1999). The social contacts with native speakers significantly affected how the regional variant of Spanish aspirated /s/ was identified (Schmidt, 2015). L2 learners of English who lived in America for several years showed much better talker accommodation in identifying English words than L2 learners who were never exposed to a native English environment (Antoniou et al., 2015). Therefore, L2 immersion probably also boosts the accommodation of L2 speech variability at suprasegmental level.

Experiment 2 moved one step forward and investigated how L2 phonological proficiency, overall L2 proficiency, and L2 immersion would affect the development of L2 extrinsic normalization process. The proficiency in the Cantonese tonal system was assessed through a Cantonese tone identification task and a Cantonese tone production task. The overall Cantonese proficiency and Cantonese immersion were evaluated using the language history questionnaire (LHQ3) (Li et al., 2020). All these factors were hypothesized to significantly affect Mandarin learners' normalization of Cantonese tones, with proficiency in the Cantonese tonal system probably having a larger effect.

4.0 Material and methods

4.1 Participants

The Mandarin learners of Cantonese who took part in Experiment 1 were invited back to participate in a Cantonese tone identification task and a Cantonese tone production task to assess their proficiency in Cantonese tones. In addition, they were asked to complete the LHQ 3 to report their Cantonese acquisition and usage (Li et al., 2020). Finally, out of the 30 Mandarin participants in Experiment 1, 27 took part in Experiment 2.

4.2 The Cantonese tone identification task

The stimuli used in the Cantonese tone identification task were identical to those used in a previous study by K. Zhang et al., (2018). All stimuli (see Table 2) were produced by a native Cantonese female speaker. The f0 trajectories of six Cantonese tones are depicted in Figure 4, generated from stimuli using /jan/ as the base syllable. The trial procedure is illustrated in Figure 5. During each trial, after a 500-ms fixation, a Cantonese syllable was presented bilaterally to the subjects through inserted earphones. After the auditory stimuli, a picture containing six Cantonese words, their corresponding tone letters, and tone numbers was displayed on the screen. The six words on the screen had the same base syllable but different tones, and the subjects were required to press the corresponding key from 1 to 6 on the keyboard to indicate which word they heard. The next trial appeared after the subjects' response or when the maximum response time of 2500 ms was reached. Thirty-six Cantonese syllables (see Table 2) covering six Cantonese long tones were repeated five times and presented randomly to subjects.

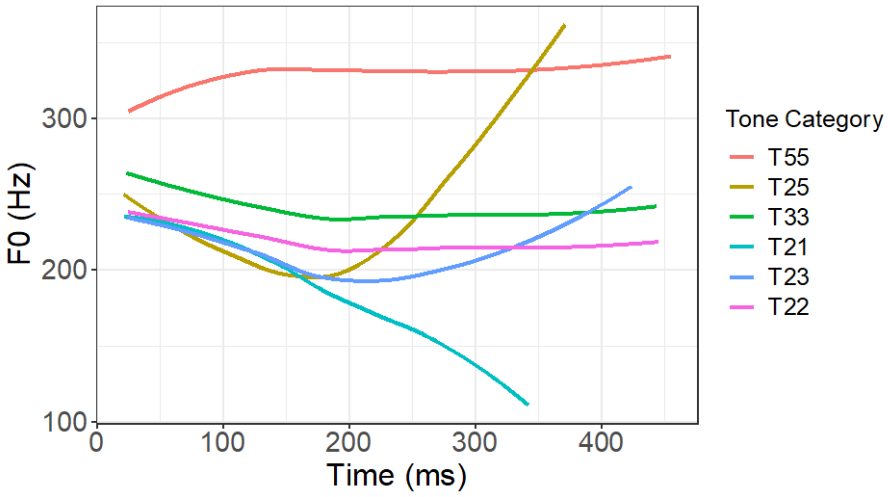

Figure 4. ***The f0 trajectories of six Cantonese tones.***

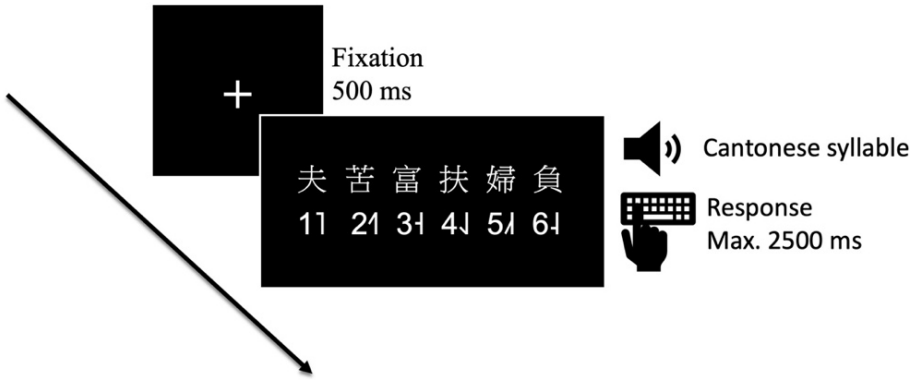

Figure 5. ***The trial procedure of the Cantonese tone identification task***

Table 2 ***Thirty-six Cantonese tonal syllables***

	/fan/	/fu/	/jan/	/ji/	/se/	/si/
T55	婚	夫	因	醫	些	詩
T25	粉	苦	隱	倚	寫	史
T33	訓	富	印	意	卸	嗜
T21	焚	扶	人	兒	蛇	時
T23	奮	婦	引	耳	社	市
T22	份	負	孕	二	射	事

4.3 The Cantonese tone production task

The Cantonese tone production task utilized in the present study was identical to the Cantonese tone production task used in K. Zhang et al. (2018). The 36 Cantonese words listed in Table 2 were repeated twice and presented in a random order. During each trial, a traditional Chinese character, its Jyutping, and its tone letter were displayed on the screen. Mandarin learners were instructed to produce the Cantonese words naturally and clearly, and their production was recorded using Praat software installed on a Macbook Pro laptop in a soundproof booth.

4.4 LHQ 3

The Mandarin learners were instructed to complete the LHQ3 in the laboratory (Li et al., 2020). The LHQ3 is a widely used online tool developed by Li et al. (2020) for assessing the linguistic background and language proficiency of L2 learners. It is based on frequently asked questions from published studies, covering topics such as the age of acquisition, duration of language usage, self-reported language proficiency, and scores from standardized language proficiency tests. The full version of LHQ contains 27 questions, but it also offers users the flexibility to select questions of interest and create customized versions. The present study utilized this customized version, including questions related to the age of acquisition, years of usage, self-rated language proficiency, self-rated strength of foreign accent, hours spent using each language per day across various activities and with different people, and the frequency of language mixing. The Mandarin learners were asked to report their acquisition and usage of both their L1 (Mandarin) and Cantonese. Other L2 languages, such as English, were not included.

5.0 Results

The objective of Experiment 2 was to investigate how various factors influenced the Mandarin learners' normalization process. Consequently, the results section only reported scores that assessed the Cantonese tone proficiency, overall L2 proficiency, and L2 immersion, without conducting any further statistical analyses within each task.

5.1 Cantonese tone identification task

The accuracy rate of the Cantonese tone identification task was averaged across six base syllables and plotted in Figure 5 (a). Each dot in Figure 5 (a) represents one Mandarin

learner's result. The confusion matrix for Mandarin learners' identification of Cantonese tones is presented in Table 3. The data indicate a strong L1 influence; Mandarin learners tend to identify T33 as T55, which is also present in Mandarin, rather than T22, which is acoustically more similar to T33 (refer to Figure 4). Even 17% of T22 stimuli were perceived as T55.

Table 3 ***The confusion matrix of Mandarin learners' Cantonese tone identification***

Stimuli / Response	T55	T25	T33	T21	T23	T22
T55	83%	3%	30%	2%	2%	17%
T25	4%	62%	5%	3%	47%	6%
T33	9%	4%	49%	7%	5%	46%
T21	1%	7%	3%	61%	9%	6%
T23	1%	22%	6%	7%	32%	7%
T22	1%	1%	6%	18%	2%	16%

5.2 Cantonese tone production task

Mandarin learners' Cantonese tone production was evaluated by six native Cantonese speakers who specialized in linguistic-related subjects and had completed at least one course in phonetics. Tokens from the same Mandarin learner were presented in one block during the evaluation. Two tokens (/fan55/ and /fan25/) from native Cantonese speakers (one male and one female) were used as fillers (matched in gender in each block) to monitor the rater's performance. Finally, there were 27 blocks and each block contained 74 tokens (36 syllables × 2 repetitions + 2 fillers). Twenty-seven blocks were presented randomly, and 74 tokens within each block were also presented in random order. Six native Cantonese speakers rated each token on a 1-5 Likert scale, with 1 for the worst and most non-native-like pronunciation and 5 for the best and most native-like pronunciation. They were asked to insist on their standards throughout the whole evaluation.

All raters completed the task with great care, as demonstrated by their high ratings of the two fillers produced by native Cantonese speakers (mean = 4.67 out of 5). The Interrater reliability was high among six raters [ICC (2, 6) = 0.96, $p < 0.01$], and thus all six raters' rating scores were included. The rating scores for each Cantonese tone were averaged across six base syllables and six raters. The result of each tone for each Mandarin learner were represented by a single dot in Figure 5 (b).

5.3 The aggregate scores from LHQ 3

LHQ 3 utilizes aggregated scores to represent the proficiency and immersion levels of participants in the languages they acquired. The scores are derived from four components: reading, writing, listening, and speaking (Li et al., 2020). Proficiency level is the weighted sum of participants' self-evaluation of their proficiency levels on each component, while immersion level is based on their ages, age of acquisition, and duration of use of the language in each component. For the detailed calculation formula, please see https://lhq-blclab.org/static/docs/aggregate-scores.html. Although Mandarin and Hong Kong Cantonese adopt distinct Chinese writing systems, people educated in one writing system can somewhat understand another due to the similarities. Therefore, reading and writing were excluded while calculating aggregated scores. Figure 5 (c) illustrates the overall Cantonese proficiency and Cantonese immersion aggregated scores for each Mandarin learner.

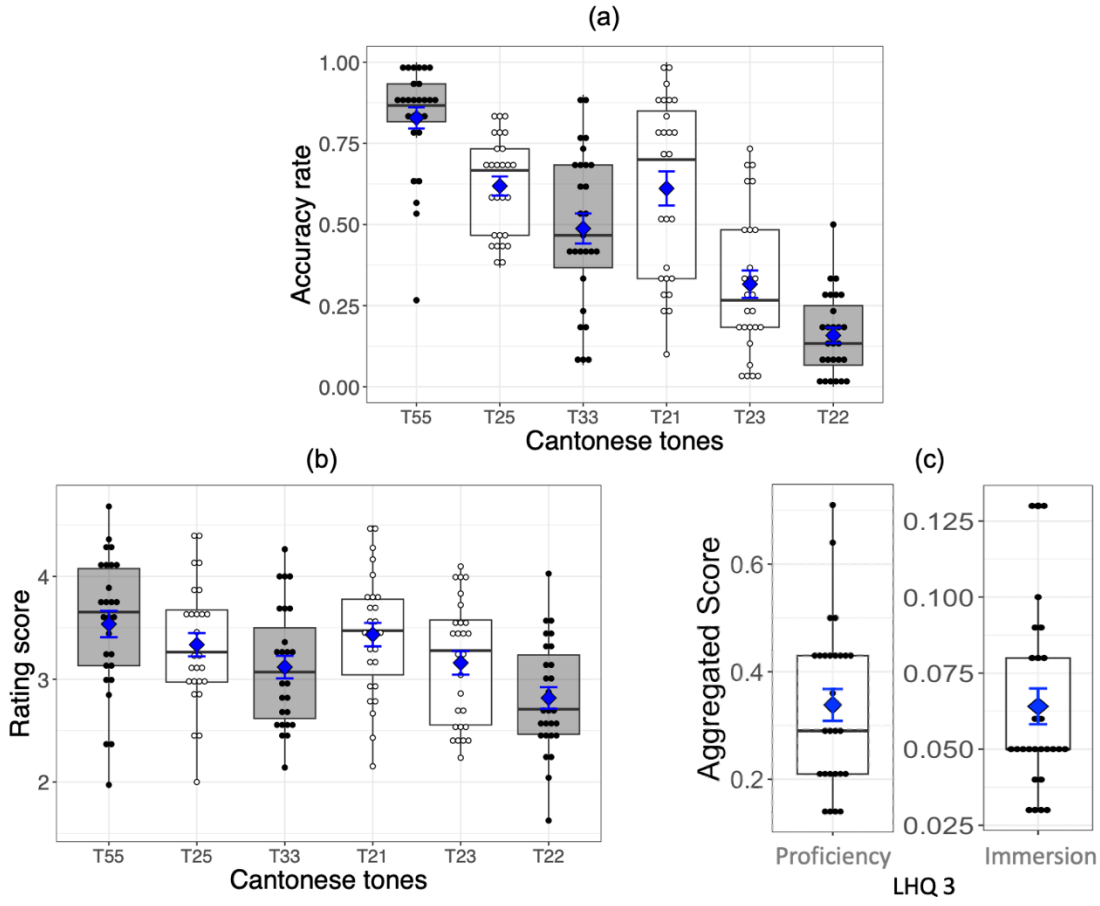

Figure 6. ***(a) Mandarin learners' accuracy rate in the Cantonese tone identification task, (b) the rating score of Mandarin learners' Cantonese tone production, and (c) the aggregated***

scores from LHQ 3. The boxes in black in panel (a) and (b) represent the tonal categories used in E1.

5.4. The effects of different factors on Mandarin learners' Cantonese tone normalization process

A generalized linear mixed-effect model was fitted using the *lme4* package (Bates et al., 2015) in R to evaluate how Mandarin learners' Cantonese tone normalization was affected by their Cantonese tone proficiency, overall L2 proficiency and L2 immersion. According to the contrastive context effect, listeners were anticipated to give /ji55/ responses for low-f0 contexts, /ji33/ responses for mid-f0 contexts, and /ji22/ responses for high-f0 contexts. Thus, a normalization accuracy of 1 was assigned to each trial when participants gave the expected response and 0 otherwise. The dependent variable in the model was normalization accuracy, whereas the fixed factors in the model were Cantonese tone identification accuracy, Cantonese tone production score, L2 proficiency score and L2 immersion score. All continuous independent variables were centered around 0. Only the by-subject intercept was included as a random effect due to convergence issues.

Considering the correlation among the predictors (Cantonese tone perception score and Cantonese tone production score: *Pearson's* $r = 0.4$; $p < 0.05$; L2 immersion score and Cantonese tone production score: $r = 0.41$; $p < 0.05$; L2 proficiency score and L2 immersion score: $r = 0.65$; $p < 0.01$), the variance inflation factor (VIF) was calculated for each predictor in the logistic regression model to detect the multilinearity problem. All VIF values were less than 5 (1.6 for Cantonese tone identification accuracy, 1.86 for Cantonese tone production score, 1.9 for L2 proficiency score, and 2.75 for L2 immersion score), indicating a low risk of multicollinearity, and thus all factors were retained in the model (James et al., 2021).

Results from the logistic regression analysis revealed that L2 immersion score significantly predicted L2 normalization accuracy ($\beta = 9.16$, 95% CI = [1.75, 16.63], $SE = 3.61$, $z = 2.53$, $p = 0.01$). Specifically, Mandarin learners with more extensive L2 immersion demonstrated more effective normalization of talker variability. However, Cantonese tone identification ($\beta = 0.78$, 95% CI = [- 0.48, 2.06], $SE = 0.62$, $z = 1.26$, $p = 0.21$), Cantonese tone production ($\beta = - 0.03$, 95% CI = [- 0.37, 0.3], $SE = 0.17$, $z = - 0.21$, $p = 0.84$), and overall L2 proficiency score ($\beta = -0.08$, 95% CI = [- 1.31, 1.15], $SE = 0.6$, $z = - 0.13$, $p = 0.89$) did not significantly impact Mandarin learners' Cantonese-tone normalization.

6.0 Discussion.

6.1. Extrinsic normalization process is a language-dependent process.

Extrinsic contexts are useful for listeners to normalize speech variability and achieve perceptual constancy. To investigate whether L2 learners could utilize context cues to normalize L2 speech variability as they did in their native languages, Experiment 1 in this study examined Mandarin learners' ability to identify Cantonese level tones produced by multiple speakers in nonspeech and speech contexts. Neither Mandarin learners nor native Cantonese speakers were affected by the pitch manipulation in nonspeech contexts, confirming the ineffectiveness of nonspeech contexts in accommodating speech variability. In the absence of effective context cues (i.e., in the nonspeech-context condition), both groups were severely confused T33 produced by multiple speakers with another two Cantonese level tones (see the upper panels in Figure 2), highlighting the importance of context cues in Cantonese tone perception. When effective context cues were available (i.e., in the speech-context condition), the perception of the target tone changed significantly in response to the pitch heights of the contexts in both groups. Specifically, native Cantonese speakers gave more /ji55/ responses in low-f0 speech contexts than in high-f0 speech contexts (0.75 vs. 0.017) and more /ji22/ responses in high-f0 speech contexts than in low-f0 speech contexts (0.86 vs. 0.06). More importantly, Mandarin learners demonstrated similar perceptual patterns, giving more /ji55/ responses when contexts changed from high f0 to low f0 (0.72 vs. 0.13) and more /ji22/ responses when contexts changed from low f0 to high f0 (0.4 vs. 0.08). These results suggest that Mandarin learners exhibited a significant extrinsic normalization process during Cantonese tone perception, similar to native Cantonese speakers, when effective context cues are available.

However, the normalization effect was significantly smaller in Mandarin learners than in native Cantonese speakers. Except in low-f0 contexts, Mandarin learners gave significantly fewer expected responses based on the contrastive contexts effect. Specifically, they gave fewer /ji33/ responses in mid-f0 speech contexts than native Cantonese speakers did (0.45 vs. 0.79), and fewer /ji22/ responses in high-f0 speech contexts than native Cantonese speakers did (0.36 vs. 0.86). The significantly smaller normalization effect suggested that Mandarin learners could not use context cues as effectively as native Cantonese speakers could. This finding supported that extrinsic normalization requires language-specific knowledge and thus should be acquired gradually during L2 learning, rather than directly generalized from L1 to

L2. The finding of Experiment 1 was also consistent with previous research reporting that contexts composed of native phonemes were more effective than those composed of nonnative phonemes in consonant normalization (Kang et al., 2016) and vowel normalization (K. Zhang & Peng, 2021). Together with the findings from the developmental study of child language which showed that children only at the age of six started to show extrinsic normalization process (F. Chen et al., 2023), the present study suggested that extrinsic normalization was a language-dependent process and was developed gradually in language acquisition of both L1 and L2. The importance of language-specific knowledge in extrinsic normalization further supports the notion that extrinsic normalization is implemented at the speech-specific level. To our knowledge, this is the first study to address this dispute (i.e., the general-auditory level process vs. the speech-specific level process) from the perspective of L2 acquisition, thereby providing new evidence to shed light on the cognitive mechanisms underlying extrinsic normalization.

6.2. Language immersion facilitates the extrinsic normalization of L2 speech.

Following up the results of Experiment 1, Experiment 2 aimed to examine how L2 phonological proficiency, overall L2 proficiency, and L2 immersion affected the extrinsic normalization process in L2. The logistic regression model with all these factors showed that only L2 immersion had a significant and positive impact on the extrinsic normalization accuracy. This means that L2 learners with more extensive L2 immersion can use context cues to cope with talker variability more effectively. Previous studies have found that language immersion enhances L2 learners' ability to identify L2 words from multiple talkers (Antoniou et al., 2015) and native speakers' ability to perceive nonnative-accented speech (Bradlow & Bent, 2008; Xie et al., 2017; Zheng & Samuel, 2020). These findings have been explained by the perceptual re-organization hypothesis, which proposes that language exposure leads to the updating of mental representations of linguistic units based on speech inputs, by adjusting category boundaries or internal category structure (Bradlow & Bent, 2008; Norris et al., 2003; Xie et al., 2017). Apart from retuning the mental representations of intrinsic cues, the present study suggests that the capacity to utilize extrinsic contextual cues in speech perception serves as an additional factor modulating the interplay between language immersion and speech normalization. The extrinsic normalization may benefit from L2 immersion due to perceptual practice. Mandarin learners have to cope with talker variability in Cantonese tones when they communicate with native speakers, so longer Cantonese immersion also implies more practice

in the extrinsic normalization process. This frequent practice enables L2 learners to improve their ability to use context cues to accommodate speech variability and thus achieve better normalization outcomes.

The regression analysis of Experiment 2 in this study showed that Cantonese tone proficiency did not significantly predict the extrinsic normalization results. This contradicted previous studies that found that familiarity with context phonemes facilitates the extrinsic normalization process. For instance, Kang et al. (2016) demonstrated that the rounded vowel /y/ in French did not influence English speakers' perception of the /s/-/ʃ/ continuum, but French speakers exhibited the contrastive context effect in such a condition. K. Zhang & Peng (2021) reported a significant reduction of the context effect on English speakers' vowel normalization in nonnative speech contexts compared to native speech contexts. A possible explanation for this discrepancy is that phonological knowledge is indeed necessary for speech normalization, but it may have a limited impact. Speech normalization might involve multiple processing stages (Xie et al., 2023). Besides perceptually adjusting the perceived signals against the acoustic-phonemic mapping at the first stage, speech normalization also includes a second stage — cognitive adjustment — where listeners make final decisions on the perceived signals with high-level information (Bosker et al., 2017). Semantic and syntactic information might play a role at this stage, as evidenced by C. Zhang et al. (2015), which showed that meaningful contexts elicited an 11% higher normalization effect than contexts composed of meaningless word sequences. The high-level information (including but not limited to semantic and syntactic information) varied greatly among Mandarin learners in this study as indicated by their overall L2 proficiencies. In addition, the cognitive adjustment also relied on listeners' cognitive ability, which also varied individually (Carroll & Maxwell, 1979; Ou & Law, 2017). Thus, the potential between-subject variability in the cognitive-adjustment stage might obscure the effect of Cantonese tone proficiency on the final normalization results.

Overall L2 proficiency was also not a significant predictor of the L2 learners' normalization process. However, an exploratory logistic regression analysis that only included L2 proficiency score, Cantonese tone perception score, and Cantonese tone production score as fixed factors and normalization accuracy as the dependent variable showed a significant main effect of L2 proficiency score ($\beta = 0.96$, 95% CI = [- 0.02, 1.95], $SE = 0.49$, $z = 1.98$, $p = 0.048$). This effect disappeared when L2 immersion score was added to the model. Despite the significant correlation between L2 proficiency score and L2 immersion score ($r = 0.65$, p

< 0.01), the insignificance of L2 proficiency score in the larger model might not result from multicollinearity as the VIF value for each factor was less than 5. It is more likely that L2 proficiency score and L2 immersion score measure a similar underlying construct, leading to a situation where they share variance in explaining the dependent variable (i.e., normalization accuracy). In this case, when L2 immersion score is added to the model, it may account for some of the variance that was previously attributed to the L2 proficiency score, thus making the latter non-significant. This study assessed the overall L2 proficiency using the LHQ 3, which calculated the aggregated scores of language proficiency based on self-rated proficiency levels on listening and speaking. The self-rated proficiency may not be as reliable as standard language tests since different people has different standards. Future studies can use standard language tests to better measure overall L2 proficiency, which may explain more underlying variance of the L2 normalization process that was not captured by L2 immersion.

6.3 L1 phonological representations interact with immediate context cues in the L2 speech normalization process.

Mandarin learners' performance in Experiment 1 was strongly influenced by their L1 tonal system. In the absence of effective context cues (i.e., in the nonspeech-context condition), native Cantonese speakers mostly confused T33 with T22 (around 43%) and occasionally with T55 (around 16%), which was mainly due to the intrinsic f0 similarity between T22 and T33 (Peng, 2006). However, unlike native Cantonese speakers, Mandarin learners more frequently misperceived T33 as T55 (around 40%), which is also a Mandarin tone, but relatively less as T22 (around 21%), whose intrinsic f0 is closer to T33, showing a strong L1 influence. Similar results can also be observed in Table 3 which showed the confusion matrix of Mandarin learners' perception of Cantonese tones in isolation. According to the Perceptual Assimilation Model, L2 learners tend to assimilate L2 phonemes to similar phonemes in their L1 and use the mental representations of L1 phonemes to perceive the similar L2 phonemes (Best & Tyler, 2007). Although the screening results suggested that all Mandarin participants could produce three Cantonese level tones, the mental representations of three Cantonese level tones were not robust. In such a condition, they probably relied on the mental representation of Mandarin T55 to perceive the similar Cantonese level tones, resulting in a high proportion of confusion between T33 and T55 (around 40%).

When effective context cues were available (i.e., in the speech-context condition), Mandarin learners' Cantonese tone perception was influenced by both immediate context cues

and L1 tonal representations. A clear contrastive context effect was observed in Mandarin learners, as they gave more T22 in high-f0 (35.9%) than in low-f0 (12.8%) speech contexts, and more T55 in low-f0 (71.8%) than in high-f0 (12.7%) speech contexts. Meanwhile, noticeable L1 influences were also observed. First, Mandarin learners overall gave more T55 responses in speech contexts than native Cantonese speakers did (see Figure 2). The interaction between L1 influence and immediate context cues was particularly evident in high-f0 speech contexts. L1 influence typically resulted in a higher number of T55 responses, but the high-f0 speech contexts theoretically led to more T22 responses. The interplay of L1 influence and immediate context cues produced the greatest proportion of T33 responses within the high-f0 speech contexts among the Mandarin group. To our knowledge, this was the first study that examined the influence of L1 tone information on L2 tone processing within the framework of speech normalization. The joint effect of L1 phonological influence and immediate context in the present study is consistent with S. Chen et al. (2022) who modeled mental representations in speech normalization of prosodic cues. S. Chen et al. (2022) first estimated the statistical distributions of acoustic cues associated with Cantonese level tones from a speech corpus and then tested how the distributional information that listeners established based on experience and the immediate context cues affected native Cantonese speakers' tone normalization. Their results suggest that speech normalization was a complex process that integrated both the distributional information of target tones and the immediate context information.

7.0 Conclusions

The present study tested if L2 learners can generalize the extrinsic normalization process that they used in L1 to L2 to accommodate L2 speech variability by investigating Mandarin learners' perception of Cantonese tones. A significant but weaker normalization process in Mandarin learners than native Cantonese speakers indicates that the extrinsic normalization process cannot be generalized directly from L1 to L2 and requires gradual learning. Such a finding suggests that L2 learners and teachers should pay attention to the training of this skill in L2 acquisition. The significant positive impact of language immersion on L2 learners' extrinsic normalization process highlights the importance of language immersion for developing L2 speech perceptual strategies. Future studies may include speakers of nontonal languages to explore whether having a tonal-language background is an advantage for L2 tone normalization.

Acknowledgements

This work was supported by the Research Grants Council of Hong Kong [GRF: 15607518].

Reference List

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/

Antoniou, M., Wong, P. C. M., & Wang, S. (2015). The effect of intensified language exposure on accommodating talker variability. *Journal of Speech, Language, and Hearing Research*, *58*(3), 722–727. https://doi.org/10.1044/2015_JSLHR-S-14-0259

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

Bent, T., & Atagi, E. (2017). Perception of nonnative-accented sentences by 5-to 8-year-olds and adults: The role of phonological processing skills. *Language and Speech*, *60*(1), 110–122. https://doi.org/10.1177/0023830916645374

Bent, T., Kewley-Port, D., & Ferguson, S. H. (2010). Across-talker effects on non-native listeners' vowel perception in noise. *The Journal of the Acoustical Society of America*, *128*(5), 3142–3151. https://doi.org/10.1121/1.3493428

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins Publishing Company.

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

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

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. https://doi.org/10.1121/1.427952

Campbell, K. L., & Tyler, L. K. (2018). Language-related domain-specific and domain-

general systems in the human brain. *Current Opinion in Behavioral Sciences*, *21*, 132–137. https://doi.org/10.1016/j.cobeha.2018.04.008

Carroll, J. B., & Maxwell, S. E. (1979). Individual Differences in Cognitive Abilities. *Annual Review of Psychology*, *30*(1), 603–640. https://doi.org/10.1146/annurev.ps.30.020179.003131

Casillas, J. V. (2020). The Longitudinal Development of Fine-Phonetic Detail: Stop Production in a Domestic Immersion Program. *Language Learning*, *70*(3), 768–806. https://doi.org/10.1111/lang.12392

Chen, F., Peng, G., Yan, N., & Wang, L. (2017). The development of categorical perception of Mandarin tones in four-to seven-year-old children. *Journal of Child Language*, *44*(6), 1413–1434. https://doi.org/10.1017/S0305000916000581

Chen, F., Zhang, K., Guo, Q., & Lv, J. (2023). Development of achieving onstancy in lexical tone identification with contextual cues. *Journal of Speech, Language, and Hearing Research*, 1–17. https://doi.org/10.1044/2022_JSLHR-22-00257

Chen, S., Zhang, C., Lau, P., Yang, Y., & Li, B. (2022). Modelling representations in speech normalization of prosodic cues. *Scientific Reports*, *12*(1), 1–21. https://doi.org/10.1038/s41598-022-18838-w

Feng, Y., & Peng, G. (2023). Development of categorical speech perception in Mandarin-speaking children and adolescents. *Child Development*, *94*(1), 28–43. https://doi.org/10.1111/cdev.13837

Fox, J., & Hong, J. (2009). Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package. *Journal of Statistical Software*, *32*(1). https://doi.org/10.18637/jss.v032.i01

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

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

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

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

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, *204*(February), 104393. https://doi.org/10.1016/j.cognition.2020.104393

Kuhl, P. K. (1976). Speech perception in early infancy: Perceptual constancy for vowel categories. *The Journal of the Acoustical Society of America*, *60*(S1), S90–S91. https://doi.org/10.1121/1.2003600

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lee, C.-Y., Zhang, Y., Li, X., Tao, L., & Bond, Z. S. (2012). Effects of speaker variability and noise on Mandarin fricative identification by native and non-native listeners. *The Journal of the Acoustical Society of America*, *132*(2), 1130–1140. https://doi.org/10.1121/1.4730883

Lee, C. Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*(1), 1–15. https://doi.org/10.1016/j.wocn.2008.08.001

Lee, C. Y., Tao, L., & Bond, Z. S. (2013). Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners. *Speech, Language and Hearing*, *16*(1), 46–54. https://doi.org/10.1179/2050571X12Z.0000000003

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

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism*, *23*(5), 938–944. https://doi.org/10.1017/S1366728918001153

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

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

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

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nusbaum and Morin. (1992). Paying attention to difference among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Speech Production, and Linguistic Structure* (pp. 113–134). IOS Press, Amsterdam.

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

Oganian, Y., Bhaya-Grossman, I., Johnson, K., & Chang, E. F. (2023). Vowel and formant representation in the human auditory speech cortex. *Neuron*, *111*(13), 2105-2118.e4. https://doi.org/10.1016/j.neuron.2023.04.004

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

Ou, J., & Law, S. P. (2017). Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Attention, Perception, and Psychophysics*, *79*(3), 945–963. https://doi.org/10.3758/s13414-017-1283-z

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

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

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

Schmidt, L. B. (2015). Not all forms of dialect contact are the same: Effects of regional media, travel, and social contacts on the perception of Spanish aspirated-/s/. *Borealis – An International Journal of Hispanic Linguistics*, *4*(1), 99. https://doi.org/10.7557/1.4.1.3284

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

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

Tamati, T. N., & Pisoni, D. B. (2014). Non-native Listeners' Recognition of High-Variability Speech Using PRESTO. *Journal of the American Academy of Audiology*, *25*(09), 869–892. https://doi.org/10.3766/jaaa.25.9.9

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

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

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

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe,

V., Ilmoniemi, R. J., & Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology*, *36*(5), 638–642. https://doi.org/10.1017/S0048577299981908

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

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

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

Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 206–217. https://doi.org/10.1037/xhp0000285

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

Zhang, K., Peng, G., Li, Y., Minett, J. W., & Wang, W. S. Y. (2018). The effect of speech variability on tonal language speakers' second language lexical tone learning. *Frontiers in Psychology*, *9*(OCT), 1–13. https://doi.org/10.3389/fpsyg.2018.01982

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

Zhang, K., Tao, R., & Peng, G. (2023). The Advantage of the Music-Enabled Brain in

Accommodating Lexical Tone Variabilities. *Brain and Language*, *247*(October), 105348. https://doi.org/Available at SSNR: http://dx.doi.org/10.2139/ssrn.4354281

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

Zheng, Y., & Samuel, A. G. (2020). The Relationship Between Phonemic Category Boundary Changes and Perceptual Adjustments to Natural Accents. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(7), 1270–1292. https://doi.org/10.1037/xlm0000788