Accommodating talker variability in noise with context cues: The case of Cantonese tones

Kaile Zhang, Gang Peng*

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

*** Correspondence concerning this article should be addressed to:**

Name: Gang Peng

Address: Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong Special Administrative Region, China

E-mail: gpeng@polyu.edu.hk

Conflicts of Interest Statement:

The authors certify that there are no conflicts of interest.

Abstract

Purpose: Listeners often rely on context cues to manage talker variability in speech and achieve perceptual constancy, a process known as extrinsic normalization. However, everyday communication typically involves both talker variability and noise, and the interaction between these factors is not well understood. This study examined the effects of different noise types and levels on listeners' ability to use contextual cues for adapting to talker variability, and additionally explored the role of attentional control in this process.

Method: Thirty-seven young native Cantonese speakers participated in a speech perception task to identify Cantonese tones from four different talkers, using speech contexts provided either in quiet or noisy environments. The study tested various signal-to-noise ratios (SNRs; 10, 5, 0, -5, and -10 dB) and noise types [babble noise (BN) and babble-modulated noise (BMN)]. Attentional control was measured using the Stroop Color-Word test.

Results: Listeners were able to use context cues to adapt to talker variability in Cantonese tones at SNRs of 0 dB and above. The effectiveness of using context cues decreased as the SNR lowered. BN created more difficulty for extrinsic normalization than BMN at -5 and -10 dB SNRs. Notably, listeners with lower Stroop interference scores demonstrated better extrinsic normalization in BMN and at 10 and 0 dB SNRs.

Conclusion: Listeners can effectively use context cues to adapt to talker variability in Cantonese tones under low to moderate noise conditions. However, high noise levels significantly hinder this ability. BN presents greater challenges than BMN at lower SNRs, likely due to increased informational masking. Attentional control plays a crucial role in facilitating extrinsic normalization in specific noise conditions.

Keywords: Talker variability; Context cues; Noise; Attentional control; Cantonese tones

Accommodating talker variability in noise with context cues: The case of Cantonese tones

1. Introduction

Consider a social gathering where individuals engage in discussions with a group of friends. Owing to various physical and psychological factors, the acoustic properties of the same word from distinct talkers exhibit considerable variability. In addition to talker-related differences, listeners at such events are exposed to environmental noise and nearby conversations. Consequently, listeners confront two challenges in day-to-day communication: talker variability and background noise. Although numerous studies have examined how listeners deal with talker variability or background noise in speech perception, limited research has been conducted on listeners' abilities to simultaneously overcome these dual challenges. The present study aims to fill in this research gap by investigating listeners' capacity to accommodate talker variability in noise. The focus would be given to Cantonese tones. The subsequent section will review existing literature on talker variability in speech perception and speech perception in noise before delineating the research plan for the current study.

1.1. Talker variability and context effects on speech perception

Talkers exhibit variability in factors such as gender, age, and accent, which contribute to substantial differences in speech signals produced by distinct individuals (Peterson & Barney, 1952). Talker variability inhibits a fast and accurate speech perception. Nusbaum & Magnuson (1997) reported that listeners’ word identification was less accurate and took longer time when the words were presented in the changing-talker condition than in the fixed-talker condition.

Talker variability in speech signals can sometimes obscure the boundaries between two acoustically similar phonemes. For example, fundamental frequency (F0) is the primary acoustic cue for lexical tone perception. However, sometimes, the F0 of a male talker's high tone could be lower than the F0 of a female talker's low tone (Peng, 2006), making the intrinsic F0 less reliable for tone categorization. Research has shown that listeners use extrinsic contexts, which surround the target word, to estimate talker-specific spectro-temporal features, and subsequently, they use this information as a reference to rescale incoming acoustic signals, effectively reducing talker-related speech variability, a process known as extrinsic normalization (Johnson, 2005; Nearey, 1989). For instance, from a daily greeting "早晨" (/zou 25 san 21/, good morning) of a native Cantonese speaker, listeners will soon identify the F0 of the highest tone point and the F0 of the lowest tone point in Cantonese, and they can use this talker-specific acoustic-phonemic mapping to calibrate the incoming tones (K. Zhang et al., 2024).

The efficacy of context cues in mitigating talker variability has been observed in the perception of vowels (K. Zhang & Peng, 2021), consonants (Holt & Lotto, 2002), and lexical tones (P. C. M. Wong & Diehl, 2003; K. Zhang et al., 2021). Context cues are especially indispensable for the perception of Cantonese level tones. There are three level tones in Cantonese: high level (T55), mid level (T33), and low level (T22), which share similar pitch contours and mainly differ in pitch heights (Yip, 2002). Due to the inter- and intra-talker variability, the intrinsic F0 was less effective to differentiate three level tones, but the identification of Cantonese level tones significantly improves when presented in speech contexts (Peng et al., 2012; P. C. M. Wong & Diehl, 2003). In most cases, context cues affect target perception in a contrastive manner. An ambiguous lexical tone is more likely perceived as high tone if its preceding context has a low F0 and as a low tone if its preceding context has a high F0, which is known as the contrastive context effect (Moore & Jongman, 1997; C. Zhang et al., 2013; K. Zhang et al., 2018). Consequently, the contrastive context effect has been widely used in previous studies (e.g., Chen et al., 2023; Ladefoged & Broadbent, 1957; Moore & Jongman, 1997; Sjerps et al., 2011) as an indicator to examine whether listeners can interpret target speech signals by referring to context cues (i.e., the extrinsic normalization process). Similarly, the present study will employ the contrastive context effect to investigate the extrinsic normalization process in noise.

Context provides useful information at multiple levels, such as acoustic, phonemic, and semantic, to facilitate the extrinsic normalization process. The spectro-temporal acoustic information in the context is a prerequisite for extrinsic normalization, as only contexts modulated to exhibit contrastive spectral (e.g., F0 or formants) or temporal (e.g., speech rate) features can trigger normalization of the target speech (P. C. M. Wong & Diehl, 2003; C. Zhang et al., 2013). Sometimes, even nonspeech sounds with contrastive formants or F0 can induce the normalization process (Aravamudhan et al., 2008; Huang & Holt, 2011). As a result, some researchers suggested that extrinsic normalization partially relies on the auditory system's contrastive encoding of target speech relative to the spectro-temporal structure of the context speech (Holt et al., 2001; Huang & Holt, 2009). In addition to acoustic cues, language-specific information such as phonemic content significantly enhances the normalization effect. Contexts composed of native phonemes elicit stronger normalization effects than those made up of nonnative phonemes (K. Zhang & Peng, 2021). English speakers even are unable to use contexts composed of French /y/ sound to normalize talker differences when identifying English consonants (Kang et al., 2016). Children's ability to perform extrinsic normalization is also constrained by their phonemic knowledge (Chen et al., 2023). Semantic information also plays a critical role. Contexts consisting of meaningful speech are more helpful for the extrinsic normalization of Cantonese tones than contexts comprising meaningless word sequencies (C. Zhang et al., 2015). While most behavioral studies emphasize the reliance of extrinsic normalization on either low-level acoustic cues (Holt & Lotto, 2002; Huang & Holt, 2011) or high-level language-specific information (e.g., phonemic and semantic) (Nusbaum & Magnuson, 1997; Viswanathan et al., 2009), a recent study with computational modeling supported that context cues at all these levels matter for extrinsic normalization. Xie et al. (2023) compared several computational models simulating extrinsic normalization and found that the model incorporating perceptual adaptive processing at all levels of speech processing best predicted human perception data. The reliance on context cues at different levels suggests that extrinsic normalization may vary in sensitivity to different types of noise, which will be discussed further in Section 1.2.

1.2. Perceiving high-variability speech in noise and attentional control

Although talker variability and noise frequently cooccur in daily communication, only a few researchers have explored their intertwine. Buchan et al. (2008) discovered that sentence perception in both single-talker and multi-talker conditions was equally affected by babble noise (BN). Similarly, the impact of BN on the perception of CV syllables was comparable when they were presented in the single- or multiple- talker conditions (Hazan et al., 2013). It seems that BN is not an additional burden for perceiving sentences and CV syllables in the high-variability conditions. However, when it comes lexical tones, the results were different. Lee et al. (2013) reported that Mandarin tone perception was more severely affected by the speech-shaped noise when the Mandarin tones were presented in changing-talker conditions compared to blocked-talker conditions. Speech perception in changing-talker conditions requires more frequent talker adaptation than in blocked-talker conditions. The results from Lee et al. (2013) suggested that speech-shaped noise posed a noticeable challenge to the adaptation to talker variability in lexical tones. It is worth noting that lexical tones in Lee et al.

(2013) were presented in isolation. As mentioned before, contexts cues are effective for listeners to overcome talker variability in tone perception (Moore & Jongman, 1997; Peng et al., 2012; P. C. M. Wong & Diehl, 2003). It remains unknown if listeners can successfully accommodate lexical tone variability in noise when context cues are provided, which is the focus of the present study.

Aside from the possibility that the perception of high-variability lexical tone is more vulnerable to noise relative to other speech segments, the different results among Buchan et al. (2008), Hazan et al. (2013), and Lee et al. (2013) may be attributed to differences in noise levels and noise types employed. Lee et al. (2013) which reported a negative effect of noise on perceiving high-variability Mandarin tones, utilized lower signal-to-noise ratios (SNR) than Hazan et al. (2013) (-5, -10, and -15 dB vs. 0 dB). Generally, speech perception worsens at lower SNRs. In addition to noise levels, these studies also used different noise types (BN vs. speech-shaped noise). Corbin et al. (2016) and Wang et al. (2023) found that two-talker BN has a more detrimental effect on word identification tasks than speech-shaped noise. However, when BN includes many speakers (e.g., 12 speakers), its negative impact on speech perception becomes less severe compared to speech-shaped noise (Jin & Liu, 2012), which may result from diminished informational masking (IM) in BN composed of multiple overlapping speech streams (Calandruccio et al., 2017). Therefore, it is possible that the BN in Buchan et al. (2008), which was generated from the speech of 20 speakers, likely exerted a relatively weaker masking effect than speech-shaped noise. The afore-mentioned studies suggested that another potential explanation for the absence of noise effects on the perception of high-variability speech in Hazan et al. (2013) and Buchan et al. (2008) may be due to the high SNRs (relative to -15 to -5 dB in Lee et al., 2013) and the less disturbing BN (due to involving to many speakers) adopted in their studies.

Noise level and noise type are two crucial factors to consider when investigating speech perception in noise. The intelligibility of speech decreases as SNR decreases, but not in a linear fashion. The identification accuracy of words (P. C. M. Wong et al., 2009) and phonemes (Phatak & Allen, 2007; Qi et al., 2017) in noise is comparable to those in quiet conditions when the SNR is at or near 0 dB. However, the identification accuracy begins to decline when the energy of noise surpasses that of the speech signals (Phatak & Allen, 2007; P. C. M. Wong et al., 2009). Therefore, to better understand the effect of noise on talker accommodation, it is necessary to include multiple SNRs both above and below 0 dB, which would be addressed in the present study. Noise type also matters, as different noise types affect speech perception in distinct ways. Some noises, such as white noise, pink noise, or speech-shaped noise, are nonspeech and thus not intelligible. They primarily affect speech perception through energetic masking (EM), where the energy of the target signals and noise overlap across time and frequency regions, rendering portions of the target signal inaudible. BN is composed of speech from different speakers, and sometimes its components are intelligible, especially when SNR is low and when BN is from a few speakers. When both target signals and noise are intelligible, listeners can hardly separate the signals from similar background noise (Brungart et al., 2001; Wang & Xu, 2021). In such conditions, IM occurs, which acts beyond the peripheral auditory process (Kidd et al., 2008). IM has multiple facets and encompasses factors that reduce the intelligibility of target signals once EM has been accounted for (Cooke et al., 2008). EM and IM are two main ways that noise affects speech perception, and typically investigated using speech-shaped noise and multi-talker babbles, respectively (Wang et al., 2023). It has been shown that BN and speech-shaped noise at the same SNR have different effects on Mandarin and Cantonese tone perception mainly due to the different effects of EM and IM on tone perception (Wang et al., 2023; Wang & Xu, 2020; P. Wong et al., 2018). As discussed earlier in Section 1.1, extrinsic normalization depends on multiple context cues. Acoustic cues may be more susceptible to EM, while language-specific cues may be more vulnerable to IM. Therefore, it is necessary to include both BN and noise composed of nonspeech to gain a comprehensive understanding of talker accommodation of lexical tone variability in noise, which would be considered in the experimental design of the present study.

Apart from understanding how noise affects the extrinsic normalization process, it is also meaningful to explore the potential cognitive factors that contribute to the extrinsic normalization in noise. Many studies have indicated that attentional control is one of the key factors affecting speech perception in noise (see Porto et al., 2023 for a review). Attentional control refers to the ability to overtly or covertly select task-relevant information to process while ignoring other distractions (Anderson, 2021). Competing noises often act as attentional lures, requiring listeners to control their attention and prevent it from drifting away from the target speech signals (Tierney et al., 2019). It has been demonstrated that individuals with better attentional control can, to some extent, ignore distracting noise and perceive speech signals more effectively in noise (Dryden et al., 2017; Price & Bidelman, 2021; Stenbäck et al., 2021). Therefore, it is possible that attentional control also contributes to listeners' utilization of context cues to normalize talker variability in noise, which would be tested in the present study.

1.3 The present study

The perception of high-variability lexical tones is more difficult in noise than in the quiet conditions (Lee et al., 2010, 2013). Although speech cues such as the speakers' pitch heights (Holt, 2006) and pitch ranges (C. Zhang et al., 2012), the phonemic categories (Kang et al., 2016; K. Zhang & Peng, 2021), and the semantic information (C. Zhang et al., 2015) provided by surrounding contexts facilitate listeners' perception of high-variability lexical tones in quiet conditions, no researchers have investigated if listeners can use these speech cues in extrinsic contexts to accommodate lexical tone variability in noise. The present study would investigate if listeners can cope with talker variability using context cues in noise through a Cantonese tone perception task. Previous studies by Lee et al. (2010, 2013) examined the perception of high-variability speech in noise with Mandarin tones. However, Mandarin tones have distinct pitch contours, making them less susceptible to talker variability. In contrast, Cantonese tones, particularly the three level tones (T22, T33, and T55), rely primarily on pitch height for differentiation and are therefore more vulnerable to speaker variability (Peng et al., 2012). This makes Cantonese tones more suitable for testing the talker normalization process. Additionally, compared to Mandarin, the distribution of Cantonese tones is more condensed along the pitch height dimension (i.e., three level tones), requiring finer-grained processing of pitch height for accurate identification. Investigating the perception of Cantonese tones in noise will provide deeper insights into how noise affects the processing of fine pitch height cues in lexical tone perception.

Noise level and noise type are two important factors affecting speech perception in noise. To have a comprehensive understanding of how noise affects listener's utilization of context cues to accommodate tone variability, the present study would use multiple SNRs (i.e., 10 dB, 5 dB, 0 dB, -5 dB, and -10 dB) and two noise types [i.e., babble-modulated speech-shaped noise (BMN) and BN]. These SNRs encompass three scenarios: noise being greater than context speech, noise being equal to context speech, and noise being less than context speech. This approach allows for a comprehensive examination of how different SNR levels influence the extent to which context cues are utilized for tone perception. Meanwhile, the present study also adopted two different noise types, BMN and BN, which can test both the effect of EM and IM of noise on the utilization of context cues. The present study used BMN instead of speech-shaped noise (Lee et al., 2013; Wang et al., 2023). BMN matches BN in both long-term average spectrum (LTAS) and temporal envelope, whereas speech-shaped noise only matches in LTAS (Tang et al., 2018). Consequently, the EM exerted by BMN is more comparable with the EM exerted by BN than speech-shaped noise. The strict control of the EM in the two noise types offers the present study an opportunity to examinate how EM and IM affect the extrinsic normalization process (Liu et al., 2021).

Listeners' attentional control would be assessed by the Word-Color Stroop task, and then their performance would be entered into the regression model to evaluate if attentional control contributes to the extrinsic normalization of lexical tones in noise. There are many assessments to test attentional control, such as the Word-Color Stroop task, Flanker task, Simon task, Continuous Performance Test, and Attention Network Test (Burgoyne et al., 2023). The Word-Color Stroop task was chosen because it shares similarities with the speech perception in noise task. Subjects in both tasks will be simultaneously presented with language-related information and information from other modalities, and they need to choose one aspect to focus on while ignoring another. Therefore, it could better predict listeners' performance in the Cantonese tone perception in noise task. Meanwhile, the main task (i.e., Cantonese tone perception in noise) is relatively long (i.e., 2 noise types x 5 SNRs + 1 quiet condition). The Word-Color Stroop task is relatively short and easy to conduct, making it the best choice for the present study.

Our hypothesis is that the impact of noise on listeners' ability to use context cues to accommodate talker variability varies depending on noise type, noise level, and attentional control. Three specific predictions are outlined here. Since previous studies suggested that word identification accuracy in noise is comparable to quiet conditions at 0 dB SNR but declines as noise energy exceeds speech signal levels (Phatak & Allen, 2007; P. C. M. Wong et al., 2009), we predicted that listeners could still use context cues to accommodate Cantonese tone variability when SNR is above or equal to 0 dB, but fail to do so when SNR is below 0 dB. Compared to speech-shaped noise, BN introduces additionally IM on speech perception. The extent of IM, however, depends on the number of talkers in the babble. Previous studies indicate that IM from two-talker BN significantly impairs Mandarin tone perception (Wang et al. (2023), whereas IM effects diminish substantially in twelve-talker (Jin & Liu, 2012) and twenty-talker BN (Buchan et al., 2008). To ensure detectable IM effects on the talker normalization process in the present study, six-talker BN was selected because it generates remarkable IM compared with most multi-talker babbles (Liu et al., 2021; Simpson & Cooke, 2005). As Wang et al. (2023) found that two-talker BN with effective IM disrupted Mandarin word perception more than speech-shaped noise, we predicted that the extrinsic normalization of Cantonese tones varies across noise types and is more difficult in six-talker BN used in the present study which produces additional IM than speech-shaped noise. Since attentional control is reported to facilitate speech perception in noise (Dryden et al., 2017; Price & Bidelman, 2021; Stenbäck et al., 2021), we predicted that people with better attentional control would show better utilization of context cues to normalize talker variability in Cantonese tones as well.

2. Materials and Methods

2.1 Subjects

Forty young native Cantonese speakers were initially recruited for the present study. However, three of them (1 male and 2 females) had mild to moderate hearing impairment, as assessed by the pure-tone air-conduction hearing screening test (average threshold > 20 dB HL for 125 to 8,000 Hz at either left or right ear) and thus were excluded. As a result, 37 participants (12 males and 25 females; M_{age} = 20.9, SD_{age} = 2.5) with normal hearing at both ears (average pure-tone threshold ≤ 20 dB HL for 125 to 8,000 Hz) were included in the final data analysis. All the 37 participants reported that Cantonese was their first language, and they spend most of their time in Macau, China, where Cantonese is the dominant language. None of them reported a history of language or speech disorder. All of them were right-handed according to the Edinburgh handedness scale (Oldfield, 1971). All participants were well-informed about the experimental procedures and signed consent forms before the experiment started. A small remuneration was given to each participant as compensation for their time.

2.2 Cantonese word identification tasks

2.2.1 Stimuli

The stimuli were largely adapted from C. Zhang et al. (2013). The auditory stimuli in each trial were composed of two parts: context and target. Three types of contexts were used in the present study: speech context (SP) in quiet, SP in BN, and SP in BMN. To introduce inter-talker variability, four native Cantonese speakers with different pitch heights [female high (FH), female low (FL), male high (MH), male low (ML)] were invited to produce the original speech materials. The context was a four-syllable Cantonese phrase "呢個字係" (/li55 ko33 tsi22 hɐi22/, "This word is"). The target was the Cantonese character "意" (e.g., /ji33/, "meaning"). The context and target in each trial were from the same talker. The original F0 contour of each speech context was raised or lowed by three semitones using Pitch-Synchronous Overlap and Add (PSOLA) method in Praat (Boersma & Weenink, 2023), resulting in three speech contexts of different pitch heights for each speaker: high-F0, mid-F0, and low-F0 contexts. The manipulation of the F0 contours was to trigger the context-dependent interpretation of target tones, and to introduce intra-talker variability. Four talkers with varying pitch heights make tone identification challenging. Under such conditions, if listeners correctly identify the target tones, it suggests they might be relying on contextual cues. However, it is difficult to rule out the possibility that their responses are based solely on intrinsic cues of the targets. To evaluate whether listeners rely on contextual cues, we manipulated the F0 contours of the contexts and included the most ambiguous target tone, /ji33/, in each trial. Since the target tone for a single talker remains constant, if listeners identify it as different words, they must use the contextual cues. Moreover, speech variability occurs not only between speakers but also within a single speaker. A speaker's speech signals can vary due to different emotional or biological states and at different times of the day (Audibert & Fougeron, 2022; Stevens, 1971). It is essential to consider both inter- and intra-talker variability when discussing the talker normalization process. Manipulating the pitch cues in the context also allows us to introduce intra-talker variability into the study. The duration of each speech context was kept unchanged (811-1005 ms), but the intensity was normalized to 55 dB. The duration of the target was normalized to 450 ms and the intensity to 55 dB. The experiment also included fillers. The contexts in the filler trials were four-syllable Cantonese phrases "我而家讀" (/ŋo23 ji21 ka55 tuk2/, "Now I will read") from FL and MH speakers and "請留心聽" (/tshiŋ25 ləu21 sɐm55 thiŋ55/, "Please listen carefully to") from FH and ML speakers. The target in the fillers trials were "意" (/ji33/) from FL and MH speakers or "二" (/ji22/) from FH and ML speakers. The duration and intensity manipulations were also applied to the filler trials.

The six-talker BN was a ten-second recording of news from six talkers (three males and three females) in Mandarin. The present study used Mandarin instead of Cantonese to generate the BN primarily to avoid potential confounding effects from both the speech context and Cantonese BN. Listeners can occasionally discern words in the six-talker BN, especially at high SNRs. Previous research suggests that speech from different speakers can serve as contexts for the normalization process (Laing et al., 2012). Thus, listeners might utilize the tonal information from the Cantonese BN instead of the intended context speech for normalization. Impaired normalization in such cases could stem from either the masking effect of BN or the choice of inappropriate contexts. By contrast, listeners are less likely to choose Mandarin BN for the normalization process, and thus the impaired normalization can be caused by the masking effect alone. Meanwhile, Mandarin, as a tonal language, can generate interference at both suprasegmental and segmental levels. Although BN composed of the same language as the target signal produces a stronger masking effect than BN composed of other languages (Chen et al., 2013), Lew et al. (2024) found that for bilingual speakers, familiarity with the masking language is more important than the similarity between masking and target languages. The participants in the present study were university students from Macau China, where both Cantonese and Mandarin are official languages, and most of them are balanced bilinguals. Therefore, the masking effects elicited by Mandarin BN should closely approximate (if not equal) those of Cantonese BN. In addition, Cantonese is predominantly spoken in Hong Kong, Macau, and Guangdong Province of China, where most of the population is Cantonese-Mandarin bilingual. It is common that some people talk in Cantonese while others surrounding them are in Mandarin. Therefore, the use of Mandarin BN is not only practical but also ecologically valid. The BMN was generated by applying the LTAS and the temporal envelope of the six-talker BN on a ten-second white Gaussian noise in Praat (Boersma & Weenink, 2023). Each speech context (4 talkers × 3 shifts + 2 fillers × 2 talkers) was mixed with BMN and BN which were randomly chosen from the original ten-second noises in MATLAB. Five SNRs were used for each noise type: 10, 5, 0, -5, and -10 dB SNR. The target words were presented in quiet.

2.2.2 Experimental procedure

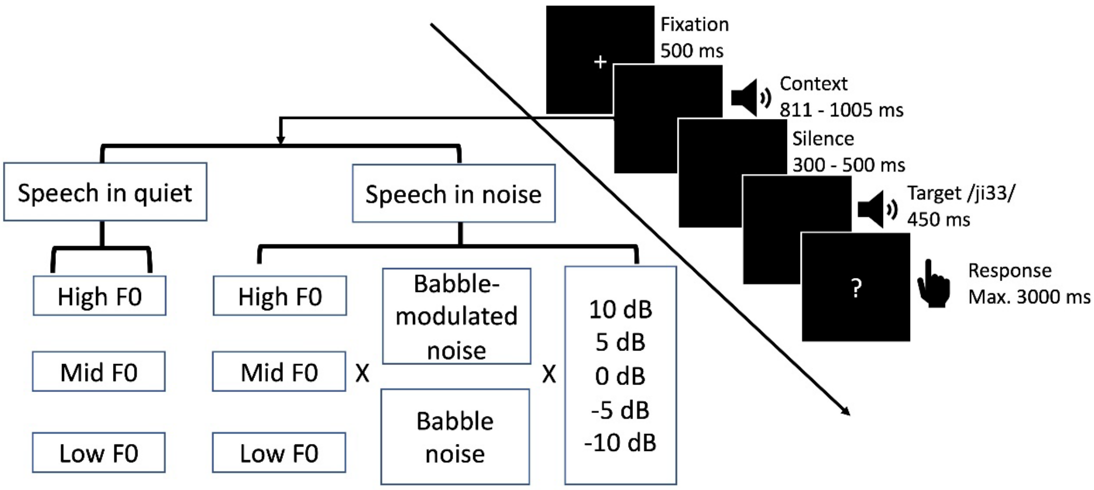

Figure 1. The experiment procedure for the Cantonese word identification task

All participants completed the Cantonese word identification task in a quiet room. The Cantonese word identification task was implemented in 11 sessions, and each with a different context type: SP in quiet, BMN at 10 dB SNR, BMN at 5 dB SNR, BMN at 0 dB SNR, BMN at -5 dB SNR, BMN at -10 dB SNR, BN at 10 dB SNR, BN at 5 dB SNR, BN at 0 dB SNR, BN at -5 dB SNR, and BN at -10 dB SNR. The eleven sessions were presented in a pseudo-randomized order across subjects. In each session, there were five blocks, and each block contained 16 non-repeated trials (4 talkers × 3 shifts + 2 fillers × 2 talkers) which were presented in a random order. The auditory stimuli were presented bilaterally to subjects via a Sennheiser HD headphone. The trial procedure was illustrated in Figure 1. In each trial, the context sound was played first. After a silence jittered between 350-450 ms, the target sound was played which was followed by a question mark. Subjects were asked to press the corresponding buttons on the keyboard to indicate which word the last syllable was after seeing the question mark. There are three choices available for listeners: 醫 (/ji55/, doctor), 意 (/ji33/, meaning), and 二 (/ji22/, two). The maximum response time (RT) for each trial was three seconds. The next trial was played automatically after detecting a response or reaching the maximum RT.

2.3 The Stroop Color-Word Test

The traditional Chinese characters – 紅 (red), 綠 (green), 藍 (blue) – printed in either red, green, or blue color were used as the stimuli in the Stroop Color-Word Test. In each trial, one of the color words was shown on the screen. Subjects were asked to identify the ink color of the word instead of the meaning of the word by pressing the corresponding button. No time limitation was set. The next trial was shown automatically after detecting a response. The inter-trial-interval was 500 ms. The key response and RT of each trial were recorded. A short practice session, consisting of six congruent trials and six incongruent trials, was carried out before the formal test. The formal test contained 60 trials (30 congruent trials and 30 incongruent trials).

2.4 Statistical Analyses

The data were analyzed in two steps. Step 1 aimed to reveal if listeners could use context cues to perceive Cantonese tones (i.e., the emergence of the extrinsic normalization) in each context condition. Step 2 would zoom in to speculate how noise type, noise level, and attentional control affected the use of context cues to accommodate Cantonese tone variability.

To statistically evaluate if there was a context-dependent perception of target tones (i.e., the extrinsic normalization process) in each condition, a multinomial logistic regression model was fitted to all participants' responses in the Cantonese word identification task, using the *nnet* package (Venables & Ripley, 2002) in R (i.e., Step 1 analysis). Response category (three levels: /ji22/, /ji33/, and /ji55/; with /ji33/ as the reference level) was the dependent variable. *Pitch height* (three levels: high F0, mid F0, and low F0; dummy coded with mid F0 as the reference level), *noise condition* (11 levels: SP in quiet, BMN at 10 dB SNR, BMN at 5 dB SNR, BMN at 0 dB SNR, BMN at -5 dB SNR, BMN at -10 dB SNR, BN at 10 dB SNR, BN at 5 dB SNR, BN at 0 dB SNR, BN at -5 dB SNR, and BN at -10 dB SNR; dummy coded with SP in quiet as the reference level), and *pitch height* by *noise condition* interaction were included as the fixed effects. Due to convergence problems, only the by-subject intercept was included as the random effect. The significance of each predictor was assessed using likelihood ratio tests via the anova() function from the *car* package. The main purpose of Step 1 analysis was to reveal if there was an extrinsic normalization process in each listening condition. Therefore, including each noise condition instead of noise type and SNR into the model can more intuitively answer this question.

To further evaluate the impact of noise level, noise type, and attentional control on the use of context cues to normalize talker variability in lexical tones, a generalized linear mixed-effects model was fitted to participants' tone perception in noise conditions using the *lmer4* package (Bates et al., 2015) in R (i.e. Step 2 analysis). The dependent variable was *accuracy rate*. If participants made the expected response, the accuracy rate for that trial was coded as 1; otherwise, it was coded as 0. Compared with response category, accuracy rate can more intuitively reveal the effect size of context cues on lexical tone normalization process. Attentional control was indexed by the Stroop interference. The Stroop interference was calculated by subtracting the mean RT in congruent trial from the mean RT in incongruent trial (Scarpina & Tagini, 2017). The larger the Stroop interference, the worse the attentional control. The model fitting started with the maximum model which included all the relevant factors and their possible two-way, three-way, and four-way interactions as the fixed effects, and by-subject and by-speaker (i.e., FH, FL, MH, and ML) slopes and intercepts as the random effects:

pitch height (three levels: high F0, mid F0, and low F0; dummy coded with mid F0 as the reference level), *SNR* (five levels: 10, 5, 0, -5, and -10 dB SNR; dummy coded with -10 dB as the reference level), *noise type* (two levels: BMN and BN; dummy coded with BMN as the reference level), and *Stroop interference* (mean centered). The model selection started from simplifying the random effects in a stepwise way until the model converged. The model was then iteratively simplified by removing fixed effects that did not significantly contribute to the model's explanatory power. The final regression model included *pitch height*, *SNR*, *noise type*, *Stroop interference,* and their possible two-way interactions as the fixed effects and by-subject intercept and by-speaker intercept as the random effects. The significance of main effects and interactions was assessed using Type II Wald chi-square tests from the Anova() function in the *car* package.

3. Results

In this section, we first present the statistical results from the multinomial logistic regression model, which examined whether an extrinsic normalization process occurred in each experimental condition. We then provide a comprehensive analysis of the statistical results from the generalized linear mixed-effects model, which tested the effects of noise type, noise level, and attentional control on the Cantonese lexical tone normalization process.

3.1 The emergence of extrinsic normalization process in different listening conditions.

The percentage of three responses (i.e., /ji22/, /ji33/, and /ji55/) in each experimental condition is illustrated in Figure 2. As can be seen, the perceptual patterns changed a lot across different listening conditions. The final output of the multinomial logistic regression model which was conducted to reveal if there was an extrinsic normalization process in each noise condition was summarized in Supplementary Material (Table 1S). The analysis revealed significant main effects of *pitch height* [χ^2 (4) = 2613.57, $p < .001$] and *noise condition* [χ^2 (20) = 82.23, $p < .001$]. There is also a significant *pitch height* by *noise condition* interaction [χ^2 (40) = 892.24, $p < .001$], indicating that participants' perceptual patterns varied across different contexts. The post-hoc analysis of the significant *pitch height* by *noise condition* interaction was conducted using the *emmeans* package (Lenth, 2019) in R to determine the emergence of extrinsic normalization processes of target tones in each noise condition (e.g., SP in quiet, BMN at 10 dB SNR etc.). If participants perceived the target tone based on context pitch cues, /ji55/ responses would be most prevalent in low-F0 contexts, /ji33/ responses most prevalent in mid-F0 contexts, and /ji22/ responses most prevalent in high-F0 contexts. Consequently, in the post-hoc analysis, this study compared the percentage of expected responses (rows in bold in Table 2) with the percentages of the other two responses in each pitch height for each noise condition, applying Bonferroni adjustment to address the issue of multiple comparisons.

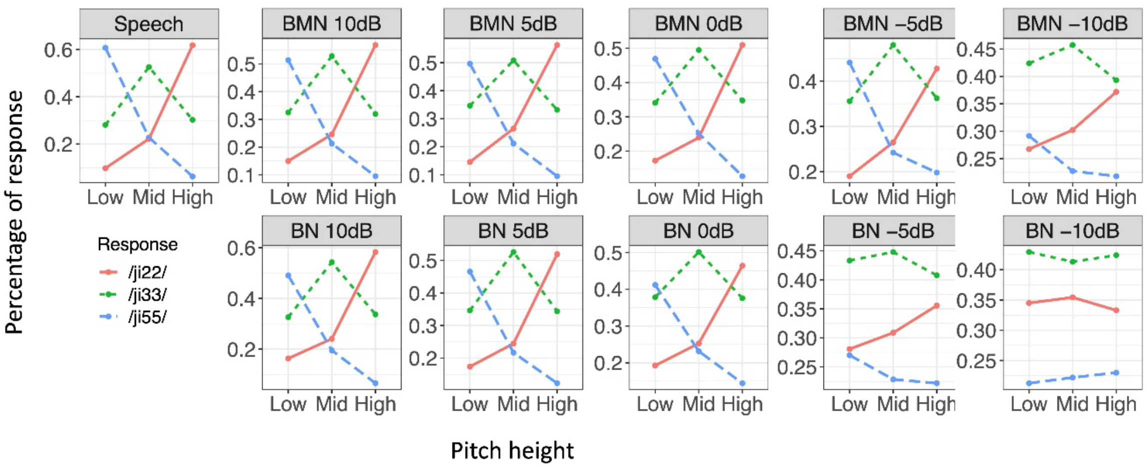

Figure 2. The percentage of three responses in each experimental condition in the Cantonese word identification task.

Table 1. The results of the post hoc analysis on the pitch height by noise condition interaction in the Cantonese word identification task

context	PH	Resp.	*prob*	*SE*	*p*	context	PH	Resp.	*prob*	*SE*	*p*
BMN	high	/ji55/	0.1	0.01	< .001	BN	high	/ji55/	0.07	0.01	< .001
10 dB	high	/ji33/	0.33	0.02	< .001	10 dB	high	/ji33/	0.34	0.02	< .001
	high	/ji22/	0.58	0.02			high	/ji22/	0.59	0.02	
	mid	/ji55/	0.22	0.02	< .001		mid	/ji55/	0.2	0.01	< .001
	mid	/ji33/	0.54	0.02			mid	/ji33/	0.55	0.02	
	mid	/ji22/	0.25	0.02	< .001		mid	/ji22/	0.25	0.02	< .001
	low	/ji55/	0.52	0.02			low	/ji55/	0.5	0.02	
	low	/ji33/	0.33	0.02	< .001		low	/ji33/	0.33	0.02	< .001
	low	/ji22/	0.15	0.01	< .001		low	/ji22/	0.17	0.01	< .001
BMN	high	/ji55/	0.1	0.01	< .001	BN	high	/ji55/	0.12	0.01	< .001
5 dB	high	/ji33/	0.34	0.02	< .001	5 dB	high	/ji33/	0.35	0.02	< .001
	high	/ji22/	0.57	0.02			high	/ji22/	0.53	0.02	
	mid	/ji55/	0.21	0.02	< .001		mid	/ji55/	0.22	0.02	< .001
	mid	/ji33/	0.52	0.02			mid	/ji33/	0.53	0.02	
	mid	/ji22/	0.27	0.02	< .001		mid	/ji22/	0.25	0.02	< .001
	low	/ji55/	0.5	0.02			low	/ji55/	0.47	0.02	
	low	/ji33/	0.35	0.02	< .001		low	/ji33/	0.35	0.02	.02
	low	/ji22/	0.15	0.01	< .001		low	/ji22/	0.18	0.01	< .001
BMN	high	/ji55/	0.13	0.01	< .001	BN	high	/ji55/	0.15	0.01	< .001
0 dB	high	/ji33/	0.35	0.02	< .001	0 dB	high	/ji33/	0.38	0.02	0.029
	high	/ji22/	0.52	0.02			high	/ji22/	0.47	0.02	
	mid	/ji55/	0.26	0.02	< .001		mid	/ji55/	0.23	0.02	< .001
	mid	/ji33/	0.5	0.02			mid	/ji33/	0.51	0.02	
	mid	/ji22/	0.24	0.02	< .001		mid	/ji22/	0.26	0.02	< .001
	low	/ji55/	0.48	0.02			**low**	**/ji55/**	**0.42**	**0.02**	
	low	/ji33/	0.35	0.02	.007		**low**	**/ji33/**	**0.38**	**0.02**	**1**
	low	/ji22/	0.18	0.01	< .001		**low**	**/ji22/**	**0.2**	**0.01**	**< .001**
BMN	high	/ji55/	0.2	0.01	< .001	BN	**high**	**/ji55/**	**0.23**	**0.02**	**< .001**
-5 dB	high	/ji33/	0.37	0.02	< .001	-5 dB	**high**	**/ji33/**	**0.41**	**0.02**	**1**
	high	/ji22/	0.43	0.02			**high**	**/ji22/**	**0.36**	**0.02**	
	mid	/ji55/	0.25	0.02	< .001		mid	/ji55/	0.23	0.02	< .001
	mid	/ji33/	0.49	0.02			mid	/ji33/	0.45	0.02	
	mid	/ji22/	0.27	0.02	< .001		mid	/ji22/	0.31	0.02	< .01
	low	**/ji55/**	**0.45**	**0.02**			**low**	**/ji55/**	**0.27**	**0.02**	
	low	**/ji33/**	**0.36**	**0.02**	**.38**		**low**	**/ji33/**	**0.44**	**0.02**	**< .001**

	low	**/ji22/**	**0.19**	**0.01**	**< .001**		**low**	**/ji22/**	**0.29**	**0.02**	**1**
BMN	**high**	**/ji55/**	**0.22**	**0.02**	**< .001**	BN	**high**	**/ji55/**	**0.23**	**0.02**	**0.01**
-10 dB	**high**	**/ji33/**	**0.4**	**0.02**	**1**	-10 dB	**high**	**/ji33/**	**0.43**	**0.02**	**.2**
	high	**/ji22/**	**0.38**	**0.02**			**high**	**/ji22/**	**0.34**	**0.02**	
	mid	/ji55/	0.23	0.02	< .001		**mid**	**/ji55/**	**0.22**	**0.02**	**< .001**
	mid	/ji33/	0.46	0.02			**mid**	**/ji33/**	**0.42**	**0.02**	
	mid	/ji22/	0.31	0.02	< .001		**mid**	**/ji22/**	**0.36**	**0.02**	**1**
	low	**/ji55/**	**0.3**	**0.02**			**low**	**/ji55/**	**0.22**	**0.02**	
	low	**/ji33/**	**0.43**	**0.02**	**< .01**		**low**	**/ji33/**	**0.44**	**0.02**	**< .001**
	low	**/ji22/**	**0.27**	**0.02**	**1**		**low**	**/ji22/**	**0.35**	**0.02**	**< .001**
SP in quiet	high	/ji55/	0.06	0.01	< .001						
	high	/ji33/	0.31	0.02	< .001						
	high	/ji22/	0.63	0.02							
	mid	/ji55/	0.23	0.02	< .001						
	mid	/ji33/	0.54	0.02							
	mid	/ji22/	0.23	0.02	< .001						
	low	/ji55/	0.62	0.02							
	low	/ji33/	0.29	0.02	< .001						
	low	/ji22/	0.1	0.01	< .001						

**Notes: PH refers to pitch height and Resp. for response. Conditions in bold did not follow the typical contrastive context effect. P values were calculated by comparing each response with the expected response in each condition (i.e., the row without a p value).*

As shown in Table 1 which summarizes the results of the post-hoc analysis, when context was presented in quiet, in BMN at 10 dB SNR, in BMN at 5 dB SNR, in BMN at 0 dB SNR, in BN at 10 dB SNR, or in BN at 5 dB SNR, participants provided significantly more expected tone responses than the other two alternatives, demonstrating the typical context-dependent interpretation of target tones (*ps* < .05; see Table 1 for specific *prob* and *p*-values in each condition). However, when context was presented in BN at -5 dB SNR, in BN at -10 dB SNR, or in BMN at -10 dB SNR, participants were more likely to perceive the target tone token as /ji33/, regardless of context pitch heights, indicating no effective extrinsic normalization processes. The results for contexts with BN at 0 dB SNR and BMN at -5 dB SNR were complex.

Participants in mid- and high-F0 contexts gave more expected responses than two alternative choices, exhibiting a typical context-dependent interpretation of target tones. However, in the low-F0 context, they gave comparable /ji33/ and /ji55/ responses (*p*s > .3).

In summary, the multinomial logistic regression analysis demonstrated that participants can effectively use context cues to perceive lexical tones when context cues were presented in quiet, in BMN at 10, 5, and 0 dB SNR, and in BN at 10 and 5 dB SNR. However, the context cues were almost useless for accommodating Cantonese tone variability if they were presented in BMN at -10 dB SNR and BN at -5 and -10 dB SNR. When the context cues were presented in BN at 0 dB SNR and BMN at -5 dB SNR, extrinsic normalization of Cantonese tones was evident in the mid and high- F0 contexts, but not in the low-F0 contexts.

3.2. The effects of noise level, noise type, and attentional control on the extrinsic normalization of Cantonese tones

The accuracy rate for each noise condition was illustrated in Figure 3 (a). The Stroop interference of each subject was illustrated in Figure 3 (b). The final output of the binomial logistic regression model which was to evaluate the effects of noise level, noise type, and attentional control on normalization process was summarized in Supplementary Material (Table 2S). The analysis revealed that all main effects and interactions in the model were statistically significant (see Table 2). To further understand the significant main effects of SNR, noise type, and pitch height**,** post hoc pairwise comparisons were conducted using the *emmeans* package (Lenth, 2019) in R with Tukey's adjustment for multiple comparisons. All results are reported in log odds ratio unless otherwise specified. Post hoc comparisons revealed significant differences across almost all SNR conditions (*ps* < .01), except the comparison between SNR at 10 dB and 5 dB (*p* = .07). Low SNR generally led to lower normalization accuracy. BMN yielded significantly higher accuracy compared to BN (β = 0.173, *SE* = 0.028, *z* = 6.213, *p* < .001). Mid-F0 led to significantly higher accuracy than both the low F0 (β = 0.372, *SE* = 0.034, *z* = 10.881, *p* < .001) and high F0 (β = 0.089, *SE* = 0.034, z = 2.637, *p* < .023). Low F0 led to significantly lower accuracy than high F0 (β = -0.283, *SE* = 0.034, *z* = -8.221, *p* < .001).

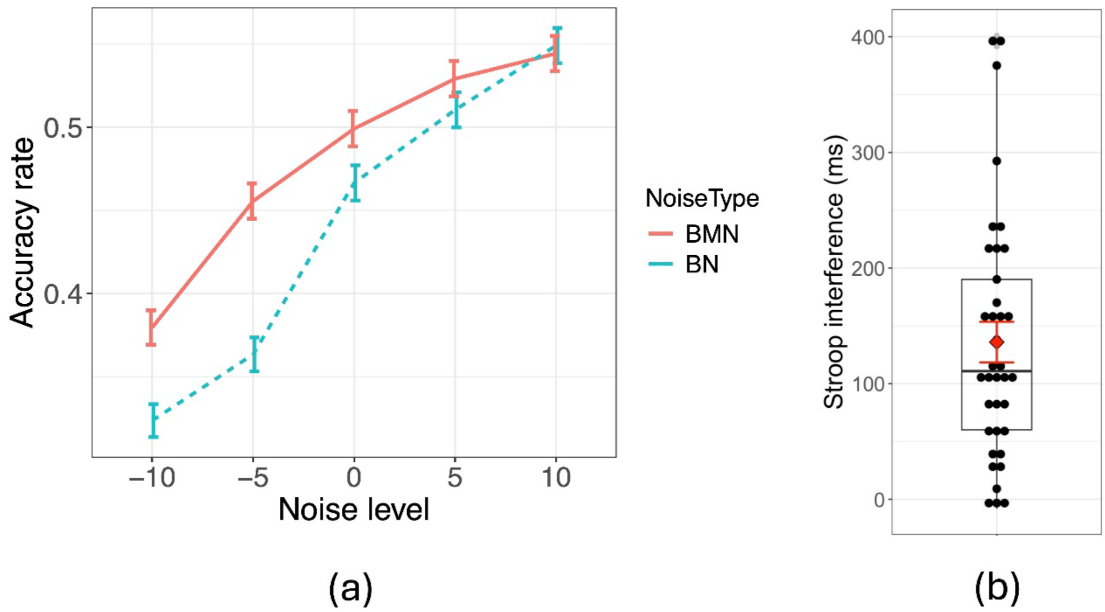

Figure 3. (a) The normalization accuracy rate of the Cantonese word identification task in contexts with noise and (b) the Stroop interference for each subject in the Word-Color Stroop task.

Table 2. The significance of each predictor in the generalized linear mixed-effects model

Predictor	χ^2	df	*p*-value
SNR	454.64	4	< .001
Noise type	35.54	1	< .001
Stroop interference	3.94	1	.047
Pitch height	116.11	2	< .001
SNR : Noise type	27.80	4	< .001
SNR : Stroop interference	16.31	4	.003
Noise type : Stroop interference	6.37	1	.012
Pitch height : Stroop interference	43.06	2	< .001
Pitch height : SNR	72.37	8	< .001
Pitch height : Noise type	19.00	2	< .001

Similar analyses were conducted on the significant SNR by noise type, noise type by pitch height, and SNR by pitch height interactions. Post hoc comparison for the significant SNR by noise type interaction revealed significant differences between BMN and BN for SNR at -10 dB (β = 0.272, *SE* = 0.07, *z* = 4.19, *p* = .001) and -5 dB (β = 0.398, *SE* = 0.06, *z* = 6.37, *p* < .001), but not for SNR at 0 dB, 5 dB, and 10 dB (*ps* > .4), indicating that BN had a greater negative impact on normalization accuracy at lower SNRs. Post hoc comparison for the significant noise type by pitch height interaction revealed that BMN led to significantly higher accuracy than BN in the low- (β = 0.327 *SE* = 0.049, *z* = 6.66, *p* < .001) and high- F0 conditions (β = 0.163, *SE* = 0.048, *z* = 3.38, *p* = .009), but not in the mid-F0 condition (*p* = .989). Post hoc comparison for the significant SNR by pitch height interaction revealed greater differences among SNRs in the low- and high- compared to mid- F0 conditions. Specifically, in the low-F0 condition, normalization accuracy significantly differed across most SNR levels (*ps* < .05), expect for the comparisons between 0 dB and 5 dB (*p* = .712) and between 5 dB and 10 dB (*p* = .995). Similarly, in the high F0 condition, normalization accuracy significantly differed across most SNR levels (*p* < .05), except for the comparisons between -10 dB and -5 dB (*p* = .681), 0 dB and 5 dB (*p* = .182), and 5 dB and 10 dB (*p* = .74). In contrast, in the mid-F0 condition, significant differences were observed only between pairs with relatively large SNR differences: -10 dB and 0 dB (*p* = .028), -10 dB and 5 dB (*p* < .001), -10 dB and 10 dB (*p* < .001), and -5 dB and 10 dB (*p* = .004).

To explore the effect of Stroop interference (a continuous variable) on normalization accuracy, post hoc simple slope analyses were conducted at each level of the categorical factors that significantly interacted with Stroop interference (i.e., SNR, noise type, and pitch height). Estimated marginal trends were computed using the emtrends() function in R, and significance was assessed using z-tests with Tukey’s adjustment for multiple comparisons. The analysis of the Stroop interference by SNR interaction revealed a simple main effect of Stroop interference at 0 dB (β = -1.354, SE = 0.527, z = -2.57, p = .01) and 10 dB (β = -1.589, SE = 0.527, z = -3.02, p = .00), but not at -10 dB, -5 dB, and 5 dB (ps >.1), indicating that greater Stroop interference led to worse normalization at moderate and low noise levels. The analysis on the Stroop interference by noise type interaction revealed that Stroop interference significantly reduced normalization accuracy in BMN (β = -1.21, SE = 0.477, z = -2.54, p = .011), but not in BN (p = .255). The analysis on the Stroop interference by pitch height interaction revealed a significant negative effect of Stroop interference on normalization accuracy in the mid- (β = -1.564, SE = 0.494, z = -3.17, p = .002) and high- (β = -1.416, SE = 0.495, z = -2.86, p = .004) F0 conditions but not in the low-F0 condition (p = .48).

In summary, the normalization accuracy rate increased as the SNR improved from -10 to 5 dB, but did not continue to improve with further increase in the SNR. The normalization accuracy rate was higher in contexts presented with BMN compared to those with BN when SNR was below 0 dB, but comparable in the two noise types when SNR was equal to or above 0 dB. The influence of Stroop interference on normalization accuracy was modulated by both SNR and noise types. Subjects with shorter Stroop interferences showed higher normalization accuracy rates in BMN, but Stroop interference did not significantly affect lexical tone normalization in BN. Meanwhile, subjects with shorter Stroop interference also had significantly better extrinsic normalization of lexical tones at 10 and 0 dB SNRs, but not at other SNRs. The normalization process in the present study is more difficult in the low-F0 condition than in the mid- and high- F0 conditions, probably due to the stimuli manipulation (see section 4.1 for more discussion).

4.Discussion

Perceiving high-variability speech is challenging in noise. The present study aims to examine whether noise hinders listeners' use of context cues to overcome talker variability by having them perceive Cantonese level tones with context cues, either in quiet or noisy conditions. The findings are generally consistent with our hypothesis: the effect of noise on listeners’ ability to use context cues to accommodate talker variability depends on noise levels, noise types, and attentional control. Specifically, participants were able to use context cues to normalize talker variability in Cantonese tones effectively when the SNRs were relatively high (e.g., ≥ 0 dB), but the extrinsic normalization process became difficult in lower SNRs (e.g., < -5 dB). This aligns closely with our prediction that listeners can still use context cues to accommodate Cantonese tone variability at SNRs of 0 dB or higher but struggle to do so when the SNR falls below 0 dB. Furthermore, our prediction that tone normalization is more difficult in BN (babble-noise) was only partially supported. The results revealed that SNR exhibited an interaction with noise types. BN demonstrated a stronger negative impact on the extrinsic normalization of Cantonese tones only at low SNRs but not at moderate to high SNRs (i.e., 0, 5 and 10 dB). Similarly, the prediction regarding attentional control was also partially supported. Attentional control indeed facilitated listeners’ use of context cues to accommodate talker variability in Cantonese tones, but this effect was limited to specific noise conditions — namely, in BMN (babble-modulated speech-shaped noise) and at SNRs of 10 dB and 0 dB.

4.1 Noise level matters in the extrinsic normalization of Cantonese tones

The present study revealed a significant main effect of SNR in the Cantonese tone normalization accuracy, suggesting that noise level plays an important role in the Cantonese tone normalization in noise. In general, subjects can more effectively use the context cues to accommodate Cantonese tone variability when the SNR is higher. More importantly, the present study also revealed that the turning point for successful extrinsic normalization in noise may be around 0 dB SNR. The analysis on each context condition in section 3.1 revealed that listeners could use context cues to accommodate the Cantonese tone variability when the SNR was at or above 0 dB (with the exception of low-F0 context in BN at 0 dB SNR, which would be discussed later), but that context cues became ineffective for the normalization process at -10 dB SNR and only partially effective at -5 dB SNR depending on noise types.

Prior studies have shown that listeners' speech perception at 0 dB SNR is comparable to that in quiet conditions (Phatak & Allen, 2007; Qi et al., 2017; P. C. M. Wong et al., 2009), suggesting that the intelligibility of speech signal is acceptable when SNR is at or above 0 dB. Similar results were also revealed in Cantonese tone perception (Shao et al., 2016). Under such conditions, listeners may effectively extract the low-level spectro-temporal information and the high-level phonemic and semantic information in contexts, and then they could use such information to estimate the talker-specific acoustic-phonemic mapping to accommodate talker variability in target tones. This finding is also partially in alignment with Lee et al. (2013), which reported that listeners can cope with talker variability in Mandarin tones in quiet and in speech-shaped noise at 0 dB SNR, but that the perception of high-variability Mandarin tones significantly deteriorated compared to the low-variability conditions in speech-shaped noise when SNRs lowered to -10 dB and -15 dB. These results also elucidate the absence of noise effects on the perception of high-variability speech in Hazan et al. (2013), in which the noise level was at 0 dB SNR.

It was worth noting that the extrinsic normalization of Cantonese tones in BN at 0 dB and in BMN at -5 dB was successful in high-F0 contexts but not in low-F0 contexts. We also observed that normalization accuracy was significantly lower in the low-F0 condition compared with the mid- and high- F0 conditions. The unequal impact of noise in these situations could be attributed to the pitch manipulation method in the present study. The context pitch was shifted equally up and down (by three semitones). However, the pitch height difference between T55 and T33 was considerably greater than that between T33 and T22 (Peng, 2006). Therefore, interpreting T33 as T55 in low-F0 contexts was more difficult than interpreting T33 as T22 in high-F0 contexts. Successful extrinsic normalization processes might be observed in low-F0 contexts with BN at 0 dB and BMN at -5 dB if the context pitch had been lowered further. Similarly, the main effect of pitch height and interactions involving pitch height may no longer be significant if the context pitch had been further lowered.

Listeners failed to use context cues to accommodate lexical tone variability at lower SNRs, which could be caused by either the inaudibility of context cues due to noise masking or the inability to utilize the perceived context cues. L. L. N. Wong et al. (2012) observed that the intelligibility of Cantonese sentences was approximately 10% in a Chinese restaurant setting and 20% in speech-shaped noise at -10 dB. Given that BN is somewhat similar to the Chinese restaurant setting and BMN is close to the speech-shaped noise in L. L. N. Wong et al. (2012), it is plausible that Cantonese contexts in the present study were inaudible at -10 dB SNR, which resulted in the absence of the normalization process in these noise conditions. L. L. N. Wong et al. (2012) further reported over 50% intelligibility of Cantonese sentences in a Chinese restaurant setting at -5 dB SNR. Therefore, it is reasonable to assume that participants in the present study perceived some useful contextual cues in BN at an SNR of -5 dB. The absence of the normalization process in -5 dB BN may be attributed to participants' inability to effectively utilize the perceived context cues. The present study observed a significant effect of attentional control on normalization accuracy in certain noisy conditions (see 4.3 for more discussions). Listeners with poorer attentional control may leave fewer attentional resources for the normalization process, resulting in a failure to utilize the perceived context cues. These findings suggest that extrinsic normalization is an actively controlled process that relies on attentional resources (Nusbaum & Magnuson, 1997). Future studies could ask participants to report the content of the context (phonemic and semantic information) as well as the pitch height of the context (acoustic/phonetic information) to better quantify the intelligibility of context cues at each SNR. This approach would help to further investigate how two factors—context cue availability and the attentional resources required to utilize these cues—interact.

4.2. The effect of noise type on the extrinsic normalization of Cantonese tones.

The present study utilized two different types of noise: BMN and six-talker BN, and the analysis revealed a significant main effect of noise type, indicating that listeners' utilization of context cues to normalize Cantonese tones varies across noise types. In general, the utilization of context cues in Cantonese tone normalization was less affected by BMN than BN. The finding was in line with previous reports that BN compromised the Mandarin word identification (Wang et al., 2023) and English word identification (Kilman et al., 2014) more than BMN. The different effects of BN and BMN on the Cantonese tone normalization in the present study might be due to EM and IM they possessed (Kilman et al., 2014). Noise primarily affects speech perception through EM and IM (Wang & Xu, 2021). BN and speech-shaped noise are two types of noise that are frequently used to test the effect of EM and IM on speech perception (Wang et al. 2023). BN and BMN in the present study were matched in temporal envelope and LTAS, resulting in nearly comparable EM on the perception of context cues at the same SNR. However, BN, being composed of speech, was somewhat intelligible, particularly when its energy exceeded that of context speech (i.e., at low SNRs). Listeners could somewhat pick up a few prominent words from the BN. Intelligible speech maskers (i.e., BN) impose additional IM on the perception of context cues, leading to a more severe negative impact of BN on speech normalization. Meanwhile, it is also worth noting that the effect of noise type was further modulated by SNR. BN exhibited greater interference on the extrinsic normalization of Cantonese tones than BMN in low SNRs (i.e., -10 and -5 dB), but the normalization accuracies in two types of noise were similar when speech signals were equal to or exceeded noise (i.e., 0, 5, and 10 dB). The results indicate that when the energy of context cues was equal to or exceeded that of noise (i.e., audible), the noise type is no longer a modulator on the talker normalization process.

Two types of noise also revealed the impact of different context cues on the extrinsic normalization process. Context information at either the acoustic level (e.g., Holt, 2001) or language-specific level (i.e., the phonemic and semantic information) (e.g., K. Zhang et al., 2021; C. Zhang et al., 2015) has been found to contribute to extrinsic normalization. BMN mainly affected the processing of context cues by EM since its spectral and temporal characteristics overlap with those of co-occurrent context speech. BMN in the present study hindered the extrinsic normalization of Cantonese tones at -10 dB SNR and partially at -5 dB SNR, indicating that spectro-temporal information in the context speech is essential for the extrinsic normalization of Cantonese tones. Concurrently, IM alone in BN also obscured the extrinsic normalization of Cantonese tones as evidenced by the poorer extrinsic normalization of Cantonese tones compared to BMN at the same SNRs (i.e., -10 dB and -5 dB). Part of IM results from processing linguistic information in competing maskers (i.e., phonemic and semantic), which in turn disrupts the processing of linguistic information in the context speech. The interference from IM in BN in the present study suggested that the phonemic and semantic information in the context speech contributed to the extrinsic normalization of Cantonese tones. Taken together, the effects of EM and IM on the processing of context cues suggested that extrinsic normalization was more likely a multi-stage process that incorporates both acoustically contrastive encoding at the general auditory level and acoustic-phonemic mapping at the language-specific processing level, which was consist with the results from the computational modeling study (Xie et al., 2023).

4.3. Attentional control and extrinsic normalization in noise

The present study also attempted to examine how the general cognitive ability contributes to the extrinsic normalization of Cantonese tones in noise. We chose the attentional control, one of the most frequently reported factors affecting speech perception in noise (e.g., Porto et al., 2023), and found that participants with poorer attentional control, as indicated by larger Stroop interference, demonstrated inferior lexical tone normalization in BMN and in 10 and 0 dB SNRs. The findings revealed an important role of attentional control on the extrinsic normalization of Cantonese tones, which was consistent with the results from the prior research that listeners with better attentional control performed more effectively in identifying words or phonemes in noisy conditions, probably because their attentional control abilities allowed them to focus on target speech signals while disregarding background noise (Dryden et al., 2017; Stenbäck et al., 2021).

The results of the present study suggested the influence of attentional control on the extrinsic normalization of Cantonese tones in noise was modulated by SNR, as attentional control facilitated lexical tone normalization at 10 dB SNR and 0 dB SNR, but not at SNRs lower than 0 dB. The intensity of context signals was relatively more prominent compared to background noise in SNR higher than 0 dB but weaker in -10 dB and -5 dB SNRs. It is plausible that, as the intensity of noise becomes more dominant, even individuals with good attentional control struggle to effectively segregate context cues from the overwhelming background noise. In such challenging listening conditions, the negative impact of noise on the context cue processing might overshadow the potential benefits of attentional control, resulting in a lack of observable effect. The absence of a positive effect of attentional control in 5 dB conditions is difficult to explain. Stroop interference was observed in the significant interaction with noise type and pitch height at 5 dB SNR, but the post-hoc analysis showed that it was not significant at any condition at 5 dB SNR. This might indicate that the sample size was not large enough to detect the significant simple main effect of attentional control on lexical tone normalization at 5 dB SNR. It is also possible that intermediate noise levels (i.e., 5 dB SNR) impose specific demands on cognitive or auditory processing. Several other cognitive and auditory factors, such as working memory capacity (Ingvalson et al., 2015) and pitch sensitivity (Maggu et al., 2021), may also affect lexical tone normalization in noise. The relative importance of attentional control might be diminished due to the influence of these other factors in this particular listening condition. Future research should include more participants and incorporate more measurements of cognitive and auditory abilities to explore the interplay between attentional control and other factors, aiming to gain a deeper understanding of extrinsic normalization in noise.

Meanwhile, the influence of attentional control on the extrinsic normalization of Cantonese tones in noise was modulated by noise type, since the Stroop interference was significant on the analysis of normalization accuracy in BMN but not in BN. Although the BMN matched the BN in terms of LTAS and temporal envelope, the finer spectral details of BMN might be more uniformly distributed compared to BN due to the original white Gaussian noise's flat spectrum. This slight homogeneity enabled the attentional control system to inhibit BMN from disturbing the processing of context cues. Thus, we observed a significant positive effect of attentional control on lexical tone normalization in BMN. However, the BN used in each trial was randomly extracted from a 10-second six-talker BN, which is highly variable, creating a complex and dynamic auditory environment. This low predictability requires more cognitive effort to manage, which probably overwhelmed the attentional control system. As a result, listeners' attentional control ability can hardly predict their lexical tone normalization performance in BN. The modulation of noise type on how attentional control affected the extrinsic normalization process might be explained from the perspective of EM and IM exerted by BMN and BN as well. The BMN only exerted EM on the processing of context cues, which might be manageable for the attentional control system. However, the BN exerted both EM and IM on the processing of context cues, which might overwhelm the attentional control system. In sum, the findings highlight that attentional control may not always work properly in facilitating Cantonese tone normalization in noise. When noise is relatively simple and predictable, attentional control can more effectively suppress the background noise, enhancing the listener's ability to process context cues. In contrast, listening conditions with high variability and cognitive demands might limit the effectiveness of attentional control in facilitating Cantonese tone normalization.

It is noteworthy that the current study employed the Stroop Color-Word Test to assess participants' attentional control capabilities. The Stroop Color-Word Test primarily engages visual-semantic processing, whereas the talker normalization process examined in this study pertains to auditory speech processing. Although Roberts & Hall (2008) demonstrated substantial overlap in the brain regions activated during visual (color-word) and auditory (pitch-word) Stroop tasks, suggesting a partially shared neural basis for attentional control across two sensory modalities, the visual Stroop Color-Word test may not fully capture the attentional control ability in the auditory tasks. This limitation might be one of the reasons why the Stroop interference scores in the present study did not consistently predict variations in the talker normalization process at SNR of 0 dB and above. Future research should incorporate auditory Stroop tasks to further elucidate the influence of attentional control on talker normalization in noise.

4.4. Implications

This study represents the first investigation of the extrinsic normalization process in noisy environments. Using Cantonese level tones, which are particularly sensitive to talker variability, we found that intelligible noise, such as BN, and noise levels with SNR lower than 0 dB, challenge the effective utilization of context cues to mitigate talker variability. This highlights the importance of considering environmental factors, such as noise, when discussing extrinsic normalization process. Additionally, this study is the first to examine the perception of high-variability speech in noise using Cantonese tones. Previous research demonstrated that Mandarin speakers could achieve over 80% accuracy in identifying high-variability Mandarin tones in -10 dB speech-shaped noise (Lee et al., 2013). However, unlike Lee et al. (2013), who investigated Mandarin tones with distinct pitch contours, this study focuses on the impact of noise on Cantonese tone perception. In our experiment, noise was paired with contexts consisting of words with level tones (e.g., /li55 ko33 tsi22 hɐi22/). Notably, there were no significant normalization effects observed at an SNR of -10 dB, indicating that listeners were unable to effectively perceive the pitch of the Cantonese contexts under these noisy conditions. The findings suggest that the processing of finer pitch heights is more vulnerable to noise than the perception of pitch contours in lexical tone recognition. Therefore, when considering the effects of noise on tone perception, it is necessary to account for the specific tonal features of different languages.

5. Conclusion

By testing how native Cantonese speakers utilize context cues to overcome talker variability in Cantonese tones under different noise conditions (i.e., the extrinsic normalization process), the present study revealed a complex interplay between the extrinsic normalization of lexical tones, noise, and attentional control. Listeners' ability to use context cues to accommodate lexical tone variability declined with lower SNRs, and even no significant normalization process was observed at -10 dB SNR with BMN and BN and at -5 dB with BN, suggesting that SNR is critical for the extrinsic normalization in noise and that 0 dB might be a pivotal turning point. The extrinsic normalization of lexical tones in BN was worse than in BMN especially at low SNRs, indicating a more detrimental effect of BN on extrinsic normalization. Both EM and IM inhibited listeners' ability to use context cues effectively, implying that the extrinsic normalization process relies on both the general acoustic and high-level linguistic information. Another key finding was that participants with poorer attentional control exhibited inferior extrinsic normalization performance across various noisy conditions, underlining the crucial role of attention control in successful speech perception under adverse listening conditions. These insights advance our understanding of the robustness of human speech perception amidst both talker variability and noise, informing the design of cognitive training paradigms to improve speech perception in challenging conditions.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics Approval Statement: The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Macau (project number: 12304526). Written informed consent to participate in this study was provided by all participants.

Funding Statement: This study was supported by the National Natural Science Foundation of China (NSFC: 12304526) and Department General Research Funds from the Hong Kong Polytechnic University (P0052546)

References

Anderson, B. A. (2021). An Adaptive View of Attentional Control. *American Psychologist*, *76*(9), 1410–1422. https://doi.org/10.1037/amp0000917

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

Audibert, N., & Fougeron, C. (2022). Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-Septe*(September), 4755–4759. https://doi.org/10.21437/Interspeech.2022-10965

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

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/.*

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*(1), 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and Measurement of Attention Control. *Journal of Experimental Psychology: General*, *152*(8), 2369–2402. https://doi.org/10.1037/xge0001408

Calandruccio, L., Buss, E., & Bowdrie, K. (2017). Effectiveness of Two-Talker Maskers That Differ in Talker Congruity and Perceptual Similarity to the Target Speech. *Trends in Hearing*, *21*, 1–14. https://doi.org/10.1177/2331216517709385

Chen, F., Li, J., Wong, L. L. N., & Yan, Y. (2013). Effect of linguistic masker on the intelligibility of mandarin sentences. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *August*, 2099–2102. https://doi.org/10.21437/interspeech.2013-498

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

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427. https://doi.org/10.1121/1.2804952

Corbin, N. E., Bonino, A. Y., Buss, E., & Leibold, L. J. (2016). Development of Open-Set Word Recognition in Children. *Ear & Hearing*, *37*(1), 55–63. https://doi.org/10.1097/AUD.0000000000000201

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*, *21*, 1–21. https://doi.org/10.1177/2331216517744675

Hazan, V., Messaoud-Galusi, S., & Rosen, S. (2013). The effect of talker and intonation variability on speech perception in noise in children with dyslexia. *Journal of Speech,* *Language, and Hearing Research*, *56*(1), 44–62. https://doi.org/10.1044/1092-4388(2012/10-0107)

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

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

Huang, J., & Holt, L. L. (2011). Evidence for the central origin of lexical tone normalization (L). *The Journal of the Acoustical Society of America*, *129*(3), 1145–1148. https://doi.org/10.1121/1.3543994

Ingvalson, E. M., Dhar, S., Wong, P. C. M., & Liu, H. (2015). Working memory training to improve speech perception in noise across languages. *The Journal of the Acoustical Society of America*, *137*(6), 3477–3486. https://doi.org/10.1121/1.4921601

Jin, S.-H., & Liu, C. (2012). English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners. *The Journal of the Acoustical Society of America*, *132*(5), EL391–EL397. https://doi.org/10.1121/1.4757730

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

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). *Informational Masking* (pp. 143–189). https://doi.org/10.1007/978-0-387-71305-2_6

Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology*, *5*(JUL), 1–9. https://doi.org/10.3389/fpsyg.2014.00651

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

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

Lee, C. Y., Tao, L., & Bond, Z. S. (2010). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *Speech Communication*, *52*(11–12), 900–910. https://doi.org/10.1016/j.specom.2010.01.004

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

Lew, E., Hallot, S., Byers-Heinlein, K., & Deroche, M. (2024). Navigating the bilingual cocktail party: a critical role for listeners' L1 in the linguistic aspect of informational masking. *Bilingualism*, 1–9. https://doi.org/10.1017/S1366728924000944

Liu, C., Xu, C., Wang, Y., Xu, L., Zhang, H., & Yang, X. (2021). Aging effect on mandarin chinese vowel and tone identification in six-talker babble. *American Journal of Audiology*, *30*(3), 616–630. https://doi.org/10.1044/2021_AJA-20-00139

Maggu, A. R., Lau, J. C. Y., Waye, M. M. Y., & Wong, P. C. M. (2021). Combination of absolute pitch and tone language experience enhances lexical tone perception. *Scientific Reports*, *11*(1), 1–10. https://doi.org/10.1038/s41598-020-80260-x

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

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

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

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

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

4388(2011/11-0025)language

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

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, *121*(4), 2312–2326. https://doi.org/10.1121/1.2642397

Porto, L., Wouters, J., & van Wieringen, A. (2023). Speech perception in noise, working memory, and attention in children: A scoping review. *Hearing Research*, *439*(September), 108883. https://doi.org/10.1016/j.heares.2023.108883

Price, C. N., & Bidelman, G. M. (2021). Attention reinforces human corticofugal system to aid speech perception in noise. *NeuroImage*, *235*(December 2020), 118014. https://doi.org/10.1016/j.neuroimage.2021.118014

Qi, B., Mao, Y., Liu, J., Liu, B., & Xu, L. (2017). Relative contributions of acoustic temporal fine structure and envelope cues for lexical tone perception in noise. *The Journal of the Acoustical Society of America*, *141*(5), 3022–3029. https://doi.org/10.1121/1.4982247

Roberts, K. L., & Hall, D. A. (2008). Examining a supramodal network for conflict processing: A systematic review and novel functional magnetic resonance imaging data for related visual and auditory stroop tasks. *Journal of Cognitive Neuroscience*, *20*(6), 1063–1078. https://doi.org/10.1162/jocn.2008.20074

Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology*, *8*(APR), 1–8. https://doi.org/10.3389/fpsyg.2017.00557

Shao, J., Zhang, C., Peng, G., Yang, Y., & Wang, W. S. Y. (2016). Effect of noise on lexical tone perception in Cantonese-speaking amusics. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *08-12-*

Sept(September), 272–276. https://doi.org/10.21437/Interspeech.2016-891

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778. https://doi.org/10.1121/1.2062650

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

Stenbäck, V., Marsja, E., Hällgren, M., Lyxell, B., & Larsby, B. (2021). The contribution of age, working memory capacity, and inhibitory control on speech recognition in noise in young and older adult listeners. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4513–4523. https://doi.org/10.1044/2021_JSLHR-20-00251

Stevens, K. N. (1971). Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds. *Proceedings of the Seventh International Congress of Phonetic Sciences*, 206–232. https://doi.org/10.1515/9783110814750-014

Tang, W., Wang, X. jian, Li, J. qi, Liu, C., Dong, Q., & Nan, Y. (2018). Vowel and tone recognition in quiet and in noise among Mandarin-speaking amusics. *Hearing Research*, *363*, 62–69. https://doi.org/10.1016/j.heares.2018.03.004

Tierney, A., Rosen, S., & Dick, F. (2019). Speech-in-Speech Perception, Nonverbal Selective Attention, and Musical Training. *Journal of Experimental Psychology: Learning Memory and Cognition*, *46*(5), 968–979. https://doi.org/10.1037/xlm0000767

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

Viswanathan, N., Fowler, C., & Magnuson, J. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*, 74–79. https://doi.org/10.3758/PBR.16.1.74

Wang, X., Lee, C. Y., & Wiener, S. (2023). Non-native disadvantage in spoken word recognition is due to lexical knowledge and not type/level of noise. *Speech Communication*, *149*(May 2022), 29–37. https://doi.org/10.1016/j.specom.2023.03.004

Wang, X., & Xu, L. (2020). Mandarin tone perception in multiple-talker babbles and speech-shaped noise. *The Journal of the Acoustical Society of America*, *147*(4), EL307–EL313. https://doi.org/10.1121/10.0001002

Wang, X., & Xu, L. (2021). Speech perception in noise: Masking and unmasking. *Journal of Otology*, *16*(2), 109–119. https://doi.org/10.1016/j.joto.2020.12.001

Wong, L. L. N., Ng, E. H. N., & Soli, S. D. (2012). Characterization of speech understanding in various types of noise. *The Journal of the Acoustical Society of America*, *132*(4), 2642–2651. https://doi.org/10.1121/1.4751538

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., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia*, *47*(3), 693–703. https://doi.org/10.1016/j.neuropsychologia.2008.11.032

Wong, P., Cheng, S. T., & Chen, F. (2018). Cantonese Tone Identification in Three Temporal Cues in Quiet, Speech-Shaped Noise and Two-Talker Babble. *Frontiers in Psychology*, *9*(October), 1–25. https://doi.org/10.3389/fpsyg.2018.01604

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

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

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

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

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