Development of extrinsic normalization of lexical tones in Cantonese-speaking children

Yicheng Rong[1, 2#], Yi Weng[3#], Gang Peng[3]*

1. School of Foreign Languages, Shanghai Jiao Tong University, China

2. National Research Centre for Language and Well-Being, Shanghai Jiao Tong University, China

3. Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China

# Y.R. and Y.W. contributed equally to this study

*** 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: This study aimed to investigate the age by which Cantonese-speaking children reach adult level in using contextual cues to adjust for speech variability in identifying level tones. Another aim of this study was to explore the external and internal factors on the level tone normalization, that is the influence of context type and individual attributes including linguistic skill and musical pitch sensitivity.

Methods: The study involved 62 Cantonese-speaking children aged seven to 10 (31 boys, 31 girls) and 24 young adults (12 men, 12 women). Participants were asked to identify Cantonese level tones in different conditions: condition without context and condition with contexts: speech, music, or pure tone. Child participants' linguistic skill was assessed using a subtest of the standardized language test, and their sensitivity to musical pitch changes was assessing using three subtests related to pitch perception of Montreal Battery of Evaluation of Musical Abilities.

Results: Children aged eight and above showed comparable performance with adults in the condition with speech context, and performed significantly better than younger children. Non-speech contexts (music and pure tone) did not elicit contrastive context effect in participants across all age groups. The children with better linguistic skill or higher musical pitch sensitivity performed better in using speech contextual cues to identify level tones.

Conclusions: Cantonese-speaking children matured in their ability to normalize level tones at age of eight. This ability was positively associated with linguistic skill and musical pitch sensitivity. In addition, Cantonese level tone normalization is a speech-

specific perceptual process.

Keywords: Development; Extrinsic normalization; Lexical tones

1. Introduction

A pivotal inquiry in speech perception revolves around understanding how listeners, confronted with the inherent variability of human speech, achieve constancy in mapping speech signals to linguistic categories. The variability, seen both between and within speakers, intensifies diversity within the same linguistic category and generates overlap across different categories, thus complicating the distinction between similar phonemes. This phenomenon is evident in Cantonese tone system, where the pitch heights of the high-level tones produced by male speakers are comparable to those of the Cantonese mid-level tones produced by female speakers (Peng et al., 2012). Research has shown that adults are able to achieve perceptual constancy despite variability in the acoustics of level tones (e.g., Wong & Diehl, 2003; Zhang et al., 2015). While perceptual constancy is well-documented in adults, the developmental trajectory of this skill in children remains unexplored.

1.1 Normalizing acoustic variability with context cues

Despite the extensive variability in human speech, interlocutors are able to understand one another. One potential reason for this is that, in actual communication situations, words frequently appear with context, which furnishes insights for mitigating speech variability and achieving perceptual constancy. Ladefoged and

Broadbent's seminal work in 1957 revealed that the ambiguous "bVt" syllable tends to be perceived as "bit" in sentences characterized by a high first formant (F1) and as "bet" in those with a low F1. This research underscores how contextual acoustic information influences the perception of speech sounds, thereby reducing ambiguity stemming from both inter- and intra-talker variability through a process termed extrinsic normalization (Nearey, 1989). Typically, context influences the perception of target sounds in a contrastive manner as demonstrated in Ladefoged and Broadbent (1957), where the speech sound with a vowel midway between /ɛ/ and /ɪ/ is more likely to be perceived as /ɪ/, which has a lower F1 compared to /ɛ/, when the preceding sentence has a relatively high F1. Conversely, it is more likely to be perceived as /ɛ/ in the context of a preceding sentence with a relatively low F1. This phenomenon has also been observed in the perception of consonants. For instance, previous studies have revealed that the speaking rate of a context sentence will affect the perception of stop consonants regarding their voice onset time (VOT). Typically, voiced stop consonants (like /b/, /d/, /g/) have shorter VOTs than their voiceless counterparts (like /p/, /t/, /k/) in English. In faster speech, listeners might perceive these consonants with larger VOTs, categorizing them closer to the voiceless end of the spectrum. Conversely, in slower speech, the same consonants might be perceived with shorter VOTs, leading to a voiced interpretation (Miller & Volaitis, 1989; Summerfield, 1981).

In addition to segments (like consonants and vowels), the perception of suprasegmental features (such as lexical tones) is also contrastively affected by contextual cues. The fundamental frequency (F0) is a critical acoustic component in

realizing lexical tones, with its height or slope being essential for differentiating between lexical tones (Gandour, 1983). For instance, Cantonese features three level tones—high-level, mid-level, and low-level—which are solely distinguished by pitch height (Peng, 2006). A Cantonese speech sound with an ambiguous level tone is more frequently perceived as a word with high-level tone when following a context of low F0, and as one with low-level tone after a context of high F0, demonstrating a contrastive context effect (Tao et al., 2021; Wong & Diehl, 2003; Zhang et al., 2015; Zhang et al., 2012; Zhang et al., 2018; Zhang et al., 2017). However, since these studies have focused on adults, there is a dearth of understanding regarding how children who are native speakers of Cantonese develop the capacity to use contextual cues to normalize the acoustic variability of level tones. Therefore, the current study aimed to explore the developmental trajectory of this ability in Cantonese-speaking children.

1.2 Development of extrinsic normalization in speech perception

Previous studies suggest that the development of the ability to use contextual cues for normalizing lexical tones lags behind the development of the similar ability for consonants (Campbell et al., 2018; Chen et al., 2023; Miller & Eimas, 1983). The study by Miller and Eimas (1983) demonstrated that even 3- to 4-month-old infants are capable of categorizing consonants, /d/ and /th/, based on the duration of formant transitions in relation to the overall syllable duration, indicating an early sensitivity to contextual cues in speech when processing consonants. Campbell et al. (2018) extended this research by examining the consonant normalization of school-age children,

112 comparing them to adults. Their study revealed that both children and adults adjust their
113 identification of voicing contrasts in VOT continua based on the speaking rate of the
114 surrounding context. In addition, for children aged 5 to 10, the magnitude of this
115 speaking rate effect does not significantly differ across ages, suggesting a relatively
116 stable period in the development of this perceptual adjustment ability. Chen et al. (2023)
117 further contributed to this line of research by focusing on Mandarin-speaking children
118 and their ability to use contextual cues to normalize speech variability in perceiving
119 lexical tones. Their findings showed that Mandarin-speaking children did not
120 effectively utilize acoustic-phonemic cues in contexts to accommodate lexical tone
121 variability until they reach the age of 6 years. It remains unclear whether these findings
122 can be extended to Cantonese-speaking children, particularly regarding the perception
123 of level tones in Cantonese. Furthermore, Chen et al. (2023) did not compare the
124 performance of children and adults, leaving the age at which children reach adult-level
125 performance in this area undetermined. In our study, we aimed to explore for the first
126 time when children matured in their ability to normalize lexical tones.

127 Although Mandarin and Cantonese share similarities as tone languages, they also
128 have distinct features in their tone systems. For example, Cantonese has three level
129 tones: high, mid, and low, which are distinguished by pitch height alone, whereas
130 Mandarin features a single high-level tone. In Chen et al. (2023), participants were
131 required to identify two Mandarin lexical tones: high-level and mid-rising tones. Since
132 these two lexical tones are distinguished not only by pitch height but also by slope, they
133 can be easily identified by young native Mandarin speakers without context. Wong et

al. (2005) demonstrated that 3-year-old Mandarin-speaking children can accurately identify the high-level and mid-rising tones in isolation, even without segmental cues, as seen in minimal pairs. However, the identification of Cantonese level tones proves to be more challenging in the absence of context, with adults' accuracy levels merely hovering around chance. Interestingly, when provided with context, Cantonese-speaking adults demonstrate a significantly higher accuracy rate in perceiving level tones, exceeding 90% (Wong & Diehl, 2003). This improvement underscores the crucial role of contextual cues in the perception of Cantonese level tones. Given the distinctions between Mandarin and Cantonese, and the importance of utilizing contextual cues, this study sought to explore the developmental path of Cantonese-speaking children's ability to leverage context in perceiving level tones. This exploration was conducted in relation to the findings reported by Chen et al. (2023) on Mandarin-speaking children, offering a point of reference for understanding how children process lexical tone normalization in different tonal systems.

1.3 Factors influencing extrinsic normalization of lexical tones

The factors that may affect the extrinsic normalization of lexical tones can be categorized into external and internal ones. The external factors encompass the type of context, while the internal factors include linguistic skill, musical pitch sensitivity, and other individual attributes. In the exploration of external factors, previous studies have primarily concentrated on the distinction between speech and non-speech contexts. For instance, Francis et al. (2006) investigated the normalization of level tones in

Cantonese-speaking adults, comparing them in speech and non-speech settings. The speech context involved comprehensible Cantonese sentences, while the non-speech context was created by using Praat to apply the pitch contour from the speech context to a hummed vocal sound /ə/. Despite the non-speech context including essential cues for pitch range assessment, akin to those in the speech context, Cantonese native speakers exhibited minimal normalization effects when presented with the non-speech context. Conversely, a distinctive perceptual contrast was observed in the speech context, suggesting a disparity in the impact of speech versus non-speech contexts on the process of lexical tone normalization. Zhang et al. (2015) delved deeper into the role of speech information at various levels in adults. They tasked native Cantonese speakers with recognizing ambiguous Cantonese level tones in several contexts: meaningful speech, meaningless speech (sequences of Cantonese monosyllables), reversed speech (normal speech played backward), and non-speech (triangle waves). Their findings revealed that meaningful speech had the most significant impact on normalization, followed by meaningless speech. Reversed speech also showed some benefit for the normalization process, but the effect was minimal in non-speech contexts. Tao et al. (2021) explored a specific non-speech context: music. They found that, like other non-speech contexts, the music context did not effectively elicit lexical tone normalization in Cantonese-speaking adults, regardless of their musical training or lack thereof. However, Huang and Holt (2009) has shown that non-speech contexts can influence lexical tone perception in Mandarin-speaking adults. In their research, non-speech contexts were created using sine-wave harmonics or pure tones. They observed

that the contrastive context effect in both non-speech contexts, though quantitatively smaller, was statistically comparable to that in speech contexts. In addition, young native Mandarin speakers, particularly 6- and 7-year-olds, have been found to possess the ability to utilize non-speech cues for normalizing lexical tone, as evidenced by a significant context effect observed in the non-speech context in the study by Chen et al. (2023). However, it is noteworthy that older children and adults in the same study did not exhibit the use of non-speech contextual cues for lexical tone perception, suggesting a gradual decline or unlearning of this ability with age. Given the limited understanding of how the ability to use various contexts for normalizing Cantonese level tones develops in Cantonese-speaking children, the current study aimed to investigate this normalization process at different ages in both speech and non-speech contexts.

Regarding the internal factors that may affect the extrinsic normalization of lexical tones, previous research has suggested that musical pitch sensitivity plays a role. For example, Zhang et al. (2018) compared the normalization of Cantonese level tones between two groups of native speakers, one of which scored significantly lower in tests designed to assess musical pitch sensitivity. This group was also found to perform significantly worse in normalizing level tones within speech contexts compared to the other group, indicating a potential link between musical pitch sensitivity and the ability to normalize tones in Cantonese. However, increased music training does not appear to further enhance level tone normalization significantly, as Tao et al. (2021) found that Cantonese-speaking musicians and non-musicians performed similarly in using speech contextual cues to normalize level tones. This similarity in performance might be

attributed to a ceiling effect. To address this, the current study aimed to examine the influence of musical pitch sensitivity on level tone normalization in Cantonese-speaking children, whose ability in this area is still developing.

Linguistic skill represents another internal factor that potentially influences the process of extrinsic normalization in speech perception. Speech perception involves understanding spoken language, which is challenging due to the variability in how something can be expressed. Insufficient exposure to a language may hinder one's ability to effectively accommodate this variability, as evidenced in second language (L2) learners (Tamati & Pisoni, 2014). Furthermore, the presence of a meaningful speech context has been shown to enhance the normalization effect by 11%, compared to contexts consisting of meaningless word sequences (Zhang et al., 2015). This suggests that a lack of linguistic knowledge could contribute to subpar performance in the normalization process. While direct research on this factor in children remains limited, insights can be drawn from studies on L2 learners. For instance, Zhang et al. (2024) explored how linguistic skill affects Cantonese level tone normalization in adults with different levels of proficiency in Cantonese. By comparing Mandarin-speaking learners to native Cantonese speakers, they found that normalization effects were significantly weaker in the learner group, particularly among those with lower proficiency. This suggests that linguistic skill plays a role in the normalization process. However, such findings cannot be directly extended to children. Unlike adult L2 learners, who build on a fully developed first language, children are still acquiring their native language while undergoing cognitive development. By examining children at different

222 developmental stages, the present study provides a window into how tone normalization
223 gradually emerges and stabilizes during first language acquisition, offering a
224 perspective on the role of language experience in early lexical tone processing.
225
226 1.4 The current study
227 Overall, the present study aimed to examine (1) the developmental trajectory of
228 Cantonese-speaking children's ability to utilize contextual cues—such as speech, music,
229 or pure tones—in level tone perception, (2) in the contexts where a general context
230 effect has been observed, the development of the ability to normalize lexical tones—
231 reflected in contrastive perceptual shifts in response to contextual F0 variation, and (3)
232 the influence of individual attributes including linguistic skill and sensitivity to musical
233 pitch on children's ability to normalize level tones. Previous studies indicate that the
234 ability to use contextual cues for normalizing lexical tones matures at a later stage.
235 Specifically, 5-year-olds were not yet adept at using contextual cues for lexical tone
236 normalization (Chen et al., 2023), in contrast to the normalization of consonants, which
237 reached adult-like levels by the age of 5 (Campbell et al., 2018). It is important to note
238 that Chen et al. (2023) focused on the perception of Mandarin lexical tones. The
239 developmental timeline for normalizing Cantonese level tones, however, is less clear.
240 Cantonese, with its six tones including three level tones distinguished solely by pitch
241 height, presents a more complex tonal system than Mandarin, which has four tones—
242 one level and three contour tones. Mandarin tones can be recognized more easily
243 without context, suggesting they might be less influenced by context than Cantonese

tones. Given the challenge of differentiating Cantonese level tones by pitch alone, Cantonese-speaking children likely learn to rely on contextual cues at a young age, developing proficiency in using these cues for lexical tone perception earlier than Mandarin-speaking children. However, it is crucial not to overlook the impact of tonal complexity of Cantonese, which may complicate the mastery of each tone category. Consequently, Cantonese-speaking children may require a more extended period to fully develop their ability to utilize contextual cues effectively.

2. Methods

2.1 Participants

Sixty-eight Hong Kong Cantonese-speaking children and 24 native adults participated in this study. The sample size was mainly determined by availability, which was further confirmed using the G*Power 3 (Faul et al., 2007). When opting for a moderate sample size (η_p^2 = .06), .80 power, an alpha of .05, and .50 as correlation among repeated measure to pursue the interaction among Group (7-, 8-, 9-year-olds, adults), and Condition (isolated, music context, pure tone context, speech context), the desired total sample size turned out to be 52. Participants were recruited through multiple channels. Recruitment posters were displayed on university campuses and published on the faculty website. In addition, flyers were distributed in primary school communities, both online via group chats and in person at school entrances. A snowball sampling approach was also employed, where previously tested participants referred new participants. Inclusion criteria required that all participants be native Cantonese

speakers born and raised in Hong Kong, with typical language, speech, and hearing abilities and no reported history of developmental, neurological, or psychological disorders. Six children were excluded for failing to pass the post-training test (see subsection 2.3 Procedure). These included one five-year-old, three six-year-olds, and two seven-year-olds. The final data analysis included 20 seven-year-olds, 22 eight-year-olds, and 20 children aged between 9 and 10. Initially, we attempted to include younger children but found that those below age seven often failed to pass the post-training test assessing level tone identification in meaningful speech contexts. These children struggled to reliably distinguish mid- and low-level tones. This difficulty is consistent with previous findings on Cantonese tone acquisition, which show that children aged six identified T3 and T6 with less than 65% accuracy (e.g., Mok et al., 2019). Consequently, it would be futile to proceed with tasks requiring level tone identification, as the inconsistent performance in level tone identification would raise concerns about whether any observed lack of contextual effect was due to an insufficient understanding of Cantonese level tones or an inability to effectively use contextual cues. Therefore, to ensure that participants had a foundational understanding of Cantonese tones and could potentially benefit from contextual cues, we focused on children aged seven and above. In addition, 24 native adults were recruited in the current study. Both child and adult participants were recruited from Hong Kong, China, and they were native speakers of Hong Kong Cantonese. Table 1 presents the information of child and adult participants. According to self-reports or reports from their guardians, none exhibited any language, speech, or hearing impairments. Written consent was obtained from adult participants

and the caregivers of child participants before the experiment commenced. They were compensated for their participation after finishing all tasks. The research was conducted with the approval of the Human Subjects Ethics Subcommittee of The Hong Kong Polytechnic University.

[Table 1 about here]

2.2 Stimuli

Stimuli were prepared with reference to Tao et al. (2021). A trial was composed of two parts: a context and a syllable as the target. There were three types of contexts: speech, pure tone, and music. The speech contexts and all target stimuli were produced by two female native speakers of Hong Kong Cantonese, one with a high pitch range and the other with a low pitch range. Figure 1a and Figure 1b present their F0 contours across the stimuli, illustrating the pitch differences between the two speakers. Female voices were chosen for two reasons: their pitch range better matches that of children, and including male voices would have made the experiment too long for children to stay focused, aligning with prior studies on child speech perception (Chen et al., 2023). The speech context consisted of a four-syllable semantically coherent phrase, specifically /li55 kɔ33 tsi22 hai22/, which translates to “This word is”. Following the recording of the phrase’s natural delivery by the two speakers, the F0 contours of the speech were adjusted downwards and upwards by three semitones, resulting in two additional versions of the speech context: a low-pitch and a high-pitch version, alongside the original mid-pitch version. This manipulation followed the design of previous studies on adult listeners (Tao et al., 2020; Zhang et al., 2015; 2024), which

allows for comparability.

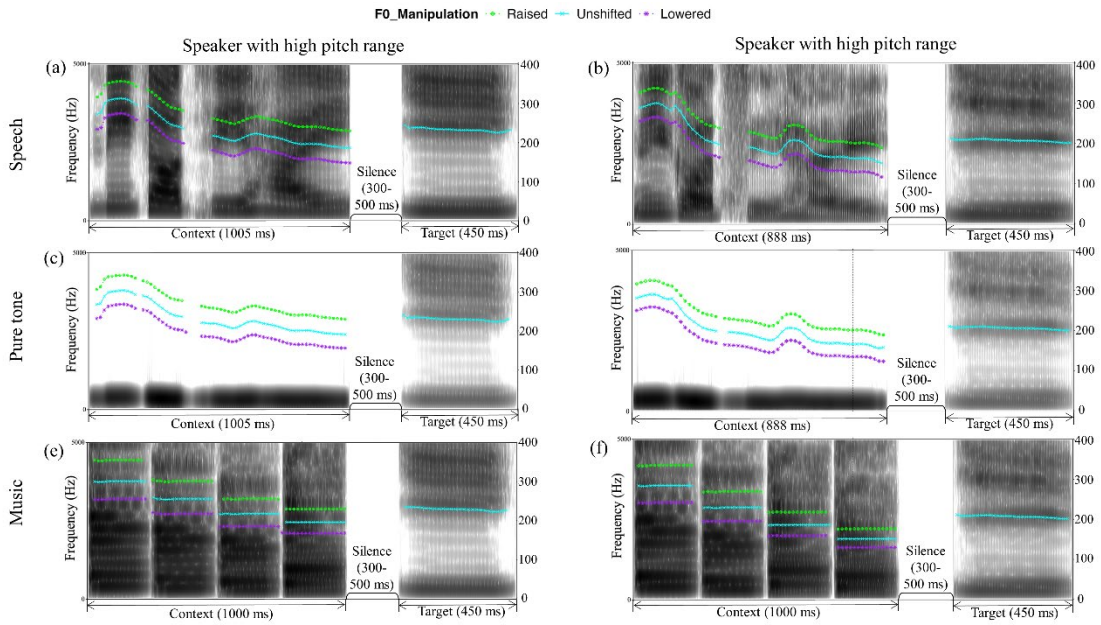

Figure 1. F0 contours of the three F0-manipulated versions for each context type and each speaker: (a) speech context with high-pitch speaker; (b) speech context with low-pitch speaker; (c) pure tone context with high-pitch speaker; (d) pure tone context with low-pitch speaker; (e) music context with high-pitch speaker; (f) music context with low-pitch speaker.

The creation of pure tone contexts involved the application of pitch contours and intensity patterns from the speech contexts to pure tones. The musical contexts were generated using piano notes that closely matched the pitch of each syllable in the speech context, produced through a Kurzweil K2000 synthesizer set to the standard A4 tuning of 440 Hz (Peng et al., 2013). This approach of selecting the nearest piano notes, rather than synthesizing a piano sound with the average pitch of each syllable, was chosen to ensure that participants would experience a more natural sound. The target stimulus in

each trial was the Cantonese syllable /ji33/, produced naturally by the same speakers who provided the preceding contexts. This syllable carries a mid-level tone in Cantonese and means “meaning”. Across all trials, the targets syllable remained acoustically identical. Figure 1 shows the F0 contours of the contexts and targets across conditions. In addition, there is another condition where the targets were presented in the absence of context.

The speech contexts and targets were normalized to 55 dB in intensity. Following previous studies (Tao et al., 2021; Zhang et al., 2018), the average acoustic intensity of the nonspeech contexts was set to 75 dB, as nonspeech stimuli were perceived as softer compared to speech. This adjustment was made to match the perceived loudness of the speech stimuli, as evaluated by native Cantonese speakers. The duration of the speech contexts was preserved to ensure naturalness: 1005 ms for the speaker with a high pitch range and 888 ms for the speaker with a low pitch range. Pure tone contexts matched the duration of their corresponding speech contexts, while the music contexts were standardized to 1000 ms, with each note lasting 250 ms.

Filler items were introduced to maintain participant engagement and to reduce predictability effects caused by the repetitive structure of the speech contexts. Each filler item for contextual conditions, like an experimental trial, was structured into two components: a four-syllable phrase acting as the context and a target word. In the speech context condition, two fillers were used. The first one, /ŋɔ23 ji21 ka55 tuk2/, which translates to “Now I will read”, was followed by the target word /ji33/. This was recorded by the speaker with a lower pitch range. The second one, /tsʰɪŋ25 lɐu21 sɐm55

tʰɛŋ55/, meaning “Please listen carefully to”, was followed by the target word /ji22/ with a low-level tone, meaning “two”. This was recorded by the speaker with a higher pitch range. The creation of filler items for pure tone and music contexts was identical to the method employed for generating the experimental trials of those two contexts. For the isolation condition, the two fillers were a different /ji33/ produced by the speaker with a lower pitch range and a /jj22/, meaning “two”, produced by the speaker with a higher pitch range.

2.3 Procedure

2.3.1 Level tone identification task

To ensure child participants understood the concept of the three Cantonese level tones, a training session was conducted. The experimenter displayed three slides, each representing one of the level tones carried by the syllable /ji/. Each slide included a rectangular pattern with a unique background color, Jyutping with tone label displayed alongside the target character, and an image related to the character’s meaning (Jyutping, developed by the Linguistics Society of Hong Kong in 1993, see Figure 2a). Although Jyutping with tone numbers was displayed alongside the target syllables to increase the transparency of tonal contrasts, children were not instructed to read or rely on them. After showing a slide, the corresponding recording of the target syllable, preceded by the phrase /li55 kɔ33 tsi22 hai22/, meaning “This word is”, would be played three times. For example, when presenting the slide showing the syllable with the high-level tone /ji55/, the experimenter played the recording of /li55 kɔ33 tsi22 hai22 ji55/ three times.

The children were then asked to mimic the targets they heard.

To ensure that participants could reliably identify the level tones in clear, unambiguous contexts, we included a post-training test. Specifically, after hearing a phrase, participants were required to point to the pattern that matched the targets they heard at the end of the phrase (see Figure 2b). The post-training test included nine trials (three per level tone). Only after correctly identifying all patterns were they considered ready for the formal experiment. While this criterion may appear demanding for children, especially given that previous research has shown even adults sometimes struggle to distinguish mid- from low-level tones (Mok et al., 2019), identification accuracy for level tone pairs in that study still exceeded 90%. In our study, participants who did not pass on the first attempt received repeated training, up to five rounds if necessary. Moreover, the items used in the post-training test were identical to those used during training. Taken together, the criterion was achievable. The sound materials employed in the training session differed from those used in the formal experiment; all of them originated from a different female native speaker of Hong Kong Cantonese.

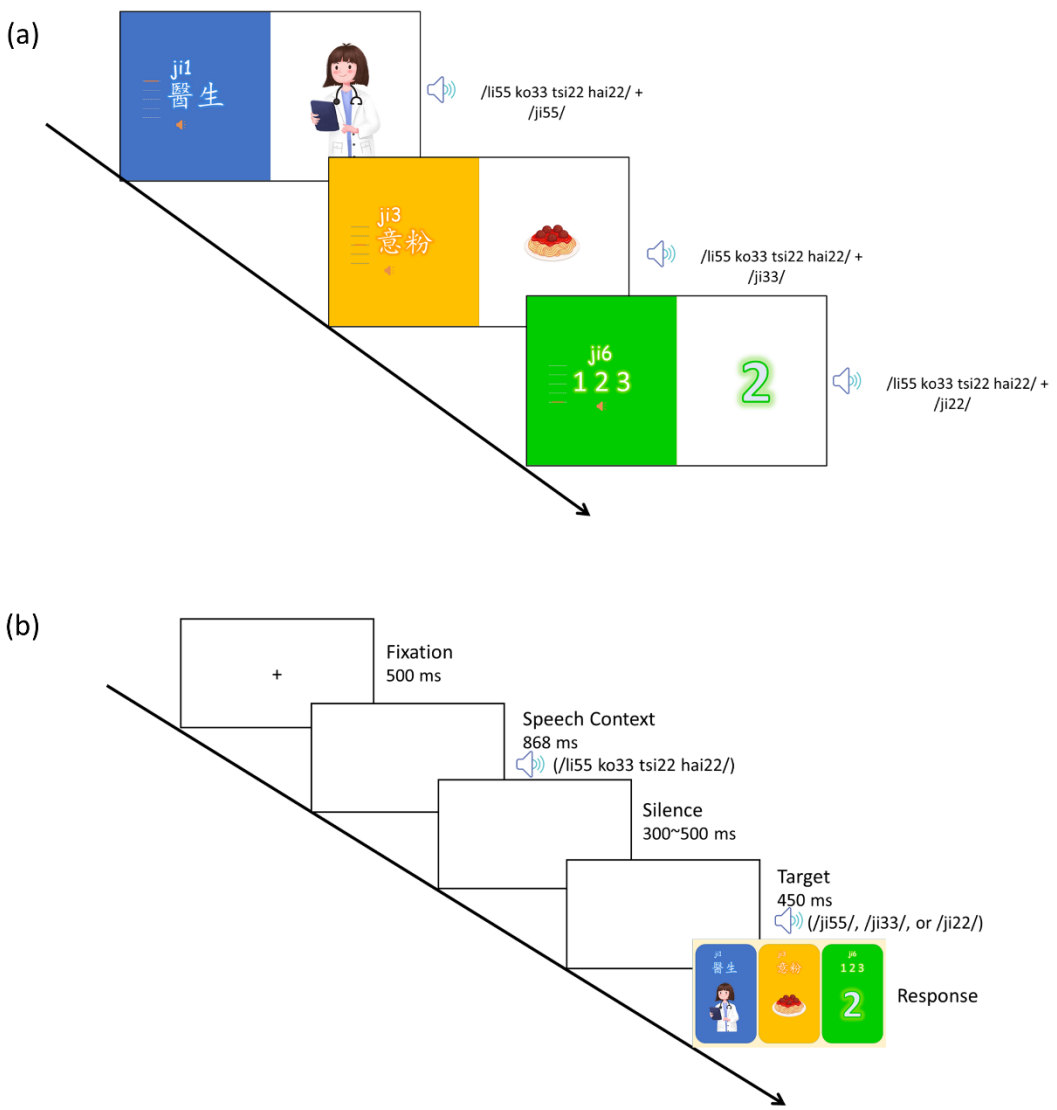

Figure 2. (a) Procedure of the training session for child participants. (b) Procedure of a sample trial used to examine the ability to identify /ji55/, /ji33/, and /ji22/.

In the formal experiment, participants were required to make judgments on the target syllable under four conditions: without context, speech context, pure tone context, and music context. The no-context condition involved five iterations, with each iteration consisting of one target syllable and one filler, produced by two speakers—one using a high pitch range and the other a low pitch range. These iterations were presented in random order, resulting in a total of 20 trials. Only the 5 target syllables per speaker

(10 in total) were analyzed, while the filler items were not scored. For each contextual condition, a total of 40 trials were conducted, featuring three versions of contexts (F0 lowered, unshifted, and raised), along with one filler, all produced by two speakers. These trials were also presented in random order, with each context type repeated five times. Among these, only the 30 trials containing target syllables were analyzed. All iterations of a given target syllable were the same recorded token to maintain acoustic consistency. Likewise, the context phrases within each F0 condition (lowered, unshifted, raised) were also repeated as the same recorded tokens. For the three contextual conditions, each trial followed a structured sequence: a fixation screen (500 ms), an auditory presentation of the audio context, a jittering silence (ranging from 300–500 ms), an auditory presentation of the target syllable, and a response screen (see Figure 2b). In the isolated condition, absent of context, the fixation screen was immediately followed by the jittering silence. These conditions were counterbalanced across participants to mitigate order effects.

Adult participants responded independently, whereas child participants indicated their judgments by pointing to one of three rectangular patterns representing their choices on the response screen, with the experimenters positioned behind and to the right of the child, recording responses via keypress. To ensure neutrality, the experimenters only provided minimal interaction, such as acknowledging responses with phrases like “okay” or “got it”, without giving any feedback or cues regarding correctness.

Responses were scored in two ways. First, we calculated the percentage of mid-level

tone responses. Second, we assessed expectation alignment based on the expected perceptual shifts induced by the manipulated F0 of the preceding context. Specifically, when the context had a lowered F0, listeners were expected to perceive the target as a high-level tone, /ji55/; when the context had a raised F0, the target was expected to be perceived as a low-level tone, /ji22/. The F0-unshifted context served as a baseline, in which the target was expected to be identified as /ji33/.

2.3.2 Linguistic skill test

Test of Hong Kong Cantonese Grammar, a component of the Hong Kong Cantonese Oral Language Assessment Scale (T'Sou et al., 2006) was employed to assess the linguistic skill of child participants. This standardized assessment, with a total possible score of 83, is widely used by speech therapists in Hong Kong to evaluate both receptive and expressive grammar skills in children aged between five and 12. It comprises four distinct tasks: sentence-picture mapping, question answering, sentence judgment, and picture description. These tasks collectively assess the child's comprehension and application of a broad spectrum of Cantonese forms and structures. Raw language ability scores were used instead of standardized scores to maintain consistency with musical pitch sensitivity measures, which do not have standardized scores (see below). In addition, age would be included as a covariate, and using age-corrected scores would have resulted in redundant control for age.

2.3.3 Musical pitch sensitivity task

Three subtests related to pitch perception of Montreal Battery of Evaluation of Musical Abilities (MBEMA) were used to assess the child participants' sensitivity to musical pitch: Scale test, Contour test, and Interval test (Peretz et al., 2013). These were conducted by trained experimenters who were familiar with the test procedures and requirements following instructions provided by the test developers. In each trial, after hearing two melodies separated by a short silence, children were required to judge whether the two melodies were the same or different. It has been proven to be a valid tool for evaluating children's musical abilities across ages and cultures. There are 20 trials for each subtest. One point was awarded for every correct response. Full marks for 100% correct responses totaled 60.

All participants first completed the level tone identification task to ensure they understood the concept of the three Cantonese level tones. The Linguistic skill test and the musical pitch sensitivity task were then administered in a counterbalanced order. The entire session, including training, breaks, and all experimental tasks, lasted approximately 1.5 hours.

2.4 Testing environment and equipment

All tasks were conducted in a sound-treated booth. Auditory stimuli were presented in free-field through two loudspeakers positioned bilaterally at approximately ±45° azimuths and 50 cm distance from the participant. The volume was adjusted to a comfortable level for each child participant.

The level tone identification task (including the post-training test and formal experiment) was administered using E-Prime 3. The initial training phase employed

PowerPoint presentations. The linguistic skill test used Praat to present stimuli, with children responding while viewing a digital response booklet in PDF format. The musical pitch sensitivity task was delivered via PowerPoint.

2.5 Data analysis

To address our research questions concerning the development of lexical tone normalization, we conducted two sets of analyses. The overarching goal was to examine at what age children begin to utilize contextual cues in tone perception like adults do, and whether such context effects reflect adult-like normalization processes. First, we tested whether the presence of a preceding context (speech, music, or pure tone) influenced children's perception of mid-level tones compared to an isolation baseline. This addressed the question of whether children of different ages show sensitivity to contextual information in lexical tone perception, and whether this sensitivity varies by context type. The dependent variable was the rate of mid-level tone responses. Because the dependent variable was a continuous proportion bounded between 0 and 1, we used a generalized linear mixed model (GLMM) with a beta distribution, implemented via the glmmTMB package (Brooks et al., 2017). A smoothing transformation was applied to ensure values fell within the (0,1) range required for beta regression. The model included two fixed effects: Condition (isolation, music, pure tone, speech) and Age Group (7-, 8-, 9-year-olds, adults), as well as their interaction. Random intercepts for participants were included to account for subject-level variability. For this analysis, we only considered the F0 unshifted version of each contextual condition. This allowed us

to isolate the influence of contextual information on tone perception without introducing additional variability from F0 shifts. The second analysis focused on contrastive context effects within the contexts where a general context effect had been observed in the first analysis. Previous studies have found that when the F0 of the preceding context is artificially shifted, listeners systematically adjust their perception of the target tone in a contrastive manner (Zhang et al., 2012; Wong & Diehl, 2003; Zhang et al., 2015; Tao et al., 2021). This contrastive effect is a key indicator of normalization. To quantify this effect, we adopted the measure of expectation alignment, defined as the percentage of responses that followed the expected contrastive pattern, i.e., choosing low-level tone in F0-raised contexts, mid-level tone in F0-unshifted contexts, and high-level tone in F0-lowered contexts. GLMM with a beta distribution was carried out on expectation alignment. Fixed effects included F0 manipulation (raised, unshifted, lowered), Age Group (7-, 8-, 9-year-olds, adults), and their interaction. Random intercepts for participants were included to account for repeated measures. Context was included as an additional within-subject factor only if multiple contexts showed a general effect in the first analysis. In both sets of analyses, post-hoc pairwise comparisons were performed using the "emmeans" package with Bonferroni adjustment when necessary.

Next, to examine how linguistic skill and sensitivity to musical pitch affected the ability to use contextual cues to normalize Cantonese level tones, we first conducted Pearson correlation analyses among tone normalization performance, age, language ability, and musical pitch sensitivity with Holm adjustment. If tone normalization was

significantly correlated with age, we proceeded with hierarchical regression analyses to examine whether either language ability or musical pitch sensitivity accounted for additional variance beyond what could be explained by age. For each factor (language ability and musical pitch sensitivity), we only conducted hierarchical regression if that factor was significantly correlated with tone normalization. In these cases, we first entered age as a predictor in the model, followed by the variable of interest in a second step. This allowed us to determine whether the added variable improved the model significantly after accounting for age. We focused exclusively on contextual conditions where a contrastive context effect was evident, as these were the most pertinent to our study. Conditions where expectation alignment was near chance levels were not included, as analyzing contributing factors in these instances would not yield meaningful insights.

3. Results

3.1 Identification of Cantonese level tones in conditions with different contexts

Table 2 shows the percentages of mid-level tone responses in the isolated condition and the F0 unshifted version of the three contextual conditions in different age groups. Statistical analysis revealed main effects of Age Group ($\chi^2(3) = 10.95$, $p = .012$) and Condition ($\chi^2(3) = 214.85$, $p < .001$), and a significant Age Group × Condition interaction ($\chi^2(9) = 31.90$, $p < .001$). The interaction was analyzed under groups to assess whether the presence of a preceding context influenced the perception of mid-level tones in children of different ages, in comparison to adults.

[Table 2 about here]

In the adult group, only the speech context significantly influenced the proportion of mid-level tone responses. They chose the mid-level tone more often in the speech context than in the isolated ($z = 8.00$, $p < .001$), pure tone context ($z = 7.90$, $p < .001$), and music context conditions ($z = 8.24$, $p < .001$). No significant differences were found among the isolated, pure tone context, and music context conditions (ps > .05). The pattern of results for both 8- and 9-year-old groups mirrored that of the adults. In both groups, children were significantly more likely to choose the mid-level tone in the speech context compared to the isolated, pure tone context, and music context conditions (ps < .001). No significant differences were found among isolated, pure tone context, and music context conditions (ps > .05). Unlike older children and adults, 7-year-olds did not show a significant difference between the isolated condition and any of the contextual conditions (ps > .05).

Based on the results of the first analysis, a general context effect was observed only in the speech context, but not in the music or pure tone contexts. Moreover, expectation alignment in the music and pure tone contexts did not significantly exceed chance level (33.33%) across age groups (all ps > .45; see Table 3). Therefore, the second set of analyses focused specifically on the speech context to examine whether children showed adult-like contrastive context effects indicative of lexical tone normalization.

[Table 3 about here]

Table 4 presents the expectation alignment for the speech context with their F0 raised, unshifted, or lowered, across different age groups. A significant main effect was

found for Age Group ($\chi^2(3) = 21.75$, $p < .001$), but not for F0 Manipulation. Post-hoc comparisons revealed that 7-year-olds showed significantly lower alignment with expected responses than 8-year-olds, 9-year-olds, and adults ($ps < .05$), while no significant differences were found among the 8-year-olds, 9-year-olds, and adults ($ps > .05$). In addition, the interaction between Age Group and F0 Manipulation was also significant ($\chi^2(6) = 13.09$, $p = .042$). Post hoc comparisons revealed that the group differences (i.e., 7 < 8/9/adult) emerged only in the lowered F0 condition ($ps < .05$), while no significant group differences were found in the raised or unshifted F0 conditions. When analyzed by group, we found that 7-year-olds showed significantly higher expectation alignment in the raised F0 condition compared to the lowered F0 condition ($t = -3.36$, $p = .003$), while no significant differences were observed across the three F0-shifted conditions in any of the older age groups or adults.

[Table 4 about here]

3.2 Influence of linguistic skill and musical pitch sensitivity on tone identification performance

Given that the expectation alignment was around chance level in conditions with non-speech contexts (see table 3), indicating the limited utility of non-speech cues for identifying ambiguous level tones, we focused our examination on the influence of linguistic skill and musical pitch sensitivity within the speech context condition, where contextual cues could more effectively aid identification. Table 5 presents expectation alignment in speech context, scores of linguistic skill test and musical pitch sensitivity task in children by age group.

570 [Table 5 about here]

571 Pearson correlation analyses revealed that tone normalization performance,
572 language ability, musical pitch sensitivity, and age positively correlated with each other
573 (see Table 6). To further examine the individual contributions of language ability and
574 musical pitch sensitivity beyond age, we conducted two hierarchical regression
575 analyses. In the first analysis, age was entered in Step 1 and significantly predicted tone
576 normalization performance, accounting for 23.85% of the variance ($F(1, 59) = 18.48$,
577 $p < .001$). When language ability was added in Step 2, the model significantly improved
578 ($F(1, 58) = 19.24$, $p < .001$), with the total variance explained increasing to 42.82%. In
579 this model, language ability was a significant predictor ($p < .001$), while age was no
580 longer significant. In the second analysis, musical pitch sensitivity was added in Step 2
581 instead of language ability. The addition did not significantly improve the model (p
582 $= .197$), as musical pitch sensitivity was not a significant predictor, while age remained
583 significant.

584 [Table 6 about here]

585 **4. Discussion**

586 To explore how Cantonese-speaking children develop the ability to use context
587 when perceiving level tones, we conducted a study involving 7- to 10-year-old children
588 and young adults. Participants were asked to perceive ambiguous Cantonese level tones
589 in various contexts, including speech and non-speech contexts (music and pure tones)
590 and in isolation. Neither children nor adults showed significant influence posed by the
591 non-speech contexts; their perception in these conditions was similar to that in the

context-free condition, and expected response alignment did not significantly exceed chance level. In the speech context condition, a significant context effect was observed. Children aged eight and above demonstrated the ability to use speech context to identify level tones with accuracy similar to that of adults, whereas 7-year-olds did not perform at a comparable level. Furthermore, our findings indicate that expectation alignment is associated with both linguistic skill and musical pitch sensitivity among children. Even when age is controlled for, linguistic skill remains a significant factor influencing tone identification performance.

4.1 Late development of level tone normalization

This study reveals that Cantonese-speaking children attain an adult-like proficiency in utilizing speech context cues to adjust for lexical tone variability by the age of eight. This finding aligns with Chen et al. (2023), which suggests that the development of the capacity to employ contextual cues for lexical tone normalization lags behind the development of a similar ability for consonants. Specifically, extrinsic normalization of consonants reaches an adult-like level at around 5 years of age, whereas 5-year-old Mandarin-speaking children have not yet mastered the context-dependent interpretation of lexical tones. The findings on the delayed development of level tone normalization are in line with what previous research has shown about Cantonese tone acquisition. While early infant perception studies (e.g., Mattock & Burnham, 2006; Mattock et al., 2008) and production studies based on transcription (e.g., To et al., 2013) suggest that tone categories emerge quite early, recent research has found that tone acquisition

614 continues well into later childhood. In particular, studies have shown that even by age
615 six, children are not yet fully adult-like in their perception and production of certain
616 tones, especially the level ones (e.g., Mok et al., 2019; 2020). These findings support
617 the view that level tone normalization follows a gradual developmental course,
618 consistent with our results. The delayed developmental trajectory of lexical tone in
619 speech contexts might be attributed to the multifaceted roles of pitch movements.
620 Throughout children's speech and language development, the various functions of pitch
621 movement are gradually realized, including expressing emotions, signaling lexical
622 stress, and conveying intonation patterns. Four- to 5-year-olds showed the ability to
623 recognize words based on their lexical tones, regardless of the intonation used, whereas
624 younger children's ability to recognize tones is significantly affected by intonational
625 variations (Singh & Chee, 2016). It is not until around the age of 4 that children start to
626 use pitch to express emotions effectively (Quam & Swingley, 2012). Even at 5 years
627 old, children still face challenges in using pitch to emphasize words correctly (Quam &
628 Swingley, 2014). Consequently, it may require a longer period to establish a robust
629 phonological representation of tonal categories. Without a stable, abstract mental
630 representation of each tonal category, it becomes challenging to form a reliable
631 perceptual reference point for recalibrating ambiguous targets. Moreover, native
632 speakers perceived Cantonese level tones in a gradient manner instead of a categorical
633 one, as there lacks a sharp identification boundary and obvious discrimination peaks
634 across boundaries among the three level tones (Francis et al., 2003). As a result, native
635 speakers may easily confuse tokens from one level tone category with another, making

successful decoding level tones in isolation require high acoustic sensitivity, which is challenging even for adults. This would make level tones harder for children to perceive and normalize, leading to a later developmental trajectory. In addition, lexical tones offer less information compared to consonants in distinguishing phonological neighbors, as demonstrated by (Tong et al., 2008). Thus, it is plausible that children might not allocate as much attention to lexical tones as they do to consonants. This lack of attention would lead to a higher tolerance for lexical tone variation (Wewalaarachchi & Singh, 2020), which in turn results in less distinct category boundaries. This distinction helps to explain why the normalization process for level tones has been observed to develop at a slower rate compared to that of consonants.

Despite the similarities, our findings diverge from those of Chen et al. (2023). Their study concentrated on children aged 5 to 8 and revealed that, although 5-year-old Mandarin speakers have not fully grasped the context-dependent interpretation of lexical tones, there was no significant influence of age, indicating no developmental aspect in using context for Mandarin tone perception. In contrast, we found a significant age effect in the identification of Cantonese level tones. Given that even 3-year-old speakers of Mandarin can identify lexical tones in their native language without contextual support (Wong et al., 2005), whereas Cantonese level tones are more reliant on sufficient context for accurate identification, it is not surprising that the skill to utilize contextual cues in lexical tone perception develops specifically in Cantonese but not in Mandarin.

In addition, 7-year-olds' poor performance was primarily observed in the F0

lowered condition and they showed significantly lower expectation alignment in this condition compared to the raised F0 condition, whereas no significant differences were observed across the three F0-shifted conditions in the older age groups or adults. This uneven pattern in the 7-year-olds might reflect their relatively immature use of contextual cues. Instead of integrating contextual information, these younger children might have relied more heavily on the absolute pitch of the target tone itself. Given that the target tones were produced by adults, their pitch might have sounded relatively low to younger listeners. Therefore, they performed poorly when the expected response was the high-level tone (i.e., in the F0-lowered condition). While this explanation is speculative, further investigation would be needed to test this hypothesis, which is beyond the scope of the current study.

4.2 No context-dependent level tone perception in non-speech contexts

Unlike Chen et al. (2023), our study did not find a context effect in non-speech contexts among children, suggesting that speech-specific information is necessary for the normalization of level tones in Cantonese (Francis et al., 2006; Tao et al., 2021; Zhang et al., 2015). It has been believed that listeners will use context to create a speaker-specific link between sounds and linguistic units, resolving ambiguous speech by referencing this link (Joos, 1948). Thus, a context with cues to estimate a specific talker's acoustic-phonological space is necessary for normalization, like the contexts carrying information of extremes of a phonological space (e.g., high and low tone of a tonal space, or /a/, /i/, /u/ of a vowel space), indicating a speech-specific process (Joos,

1948; Zhang et al., 2015).

The absence of context effects in non-speech conditions may reflect processing differences between speech and non-speech stimuli. Prior neuroimaging studies have shown that speech and non-speech are processed in distinct cortical pathways (Belin et al., 2000; Norman-Haignere et al., 2015), with voice-selective areas specifically responsive to human vocal signals. In addition, neurophysiological studies have demonstrated that speech stimuli elicit mismatch negativity (MMN) responses that are influenced by the type of deviant—specifically, whether it is phonologically distinct from the standard—even when acoustic distances are controlled (Xi et al., 2010). In contrast, non-speech stimuli generate MMNs that scale with raw acoustic distance (Rong et al., 2024), reflecting neural responses that are more directly tied to the absolute acoustic differences between stimuli. This indicates that speech perception operates in a more non-linear manner, whereas non-speech perception relies more heavily on absolute acoustic cues. This distinction likely underlies the lack of context-dependent normalization in our non-speech conditions.

However, Chen et al. (2023) reported that 6- and 7-year-old Mandarin-speaking children demonstrated significant context effects in non-speech conditions. Chen et al. attributed the emergence and subsequent disappearance of this effect in older children to the instability of the lower-level acoustic normalization mechanism. While some sensitivity to spectrotemporal contrast may emerge around age 6, it appears to be inconsistent and susceptible to both task demands and stimulus design. Importantly, the discrepancy may also stem from differences in stimulus properties. Chen et al. used

synthetic stimuli with ambiguous F0 contours between high-level and high-rising tones, which are more likely to elicit context-dependent perception based on non-speech acoustic cues. These stimuli lack strong phonemic anchors and thus may prompt listeners to rely more heavily on adjacent non-speech contexts to resolve the ambiguity. In contrast, our study used naturally produced level tones with phonetically clear F0 cues and required three-way identification, reducing perceptual ambiguity and potentially diminishing the impact of non-speech context. Furthermore, previous studies with Cantonese-speaking adults have consistently failed to find robust normalization effects from non-speech contexts (e.g., Francis et al., 2006; Tao et al., 2021; Zhang et al., 2012, 2017, 2018, 2024), suggesting that this mechanism may not be reliably available even in mature listeners. Together, our findings suggest that lower-level acoustic normalization might be constrained by language background and task design.

4.3 Relation between linguistic skill and lexical tone normalization

Our study reveals a link between children's linguistic proficiency and their ability to effectively normalize level tones within a meaningful speech context. This connection is particularly evident in the case of Mandarin speakers acquiring Cantonese as a second language (Zhang et al., 2024). The enhancement of this skill over time has been attributed to the increased perceptual practice that comes with prolonged exposure to the Cantonese language. This linguistic experience not only facilitates the learners' ability to discern and interpret these variations accurately but also enhances their skill

in utilizing contextual cues to mitigate the effects of such variability. While the research on adult L2 learners highlights the role of language experience in shaping lexical tone normalization, the relationship between linguistic ability and normalization in children likely reflects a different set of underlying mechanisms. Unlike L2 learners, who possess a fully developed first language system and established cognitive-linguistic infrastructure, children are still acquiring their first language, with ongoing development in auditory perception and general cognitive functions. Thus, improvements in tone normalization ability during childhood likely stem not merely from increased exposure to language. Thus, findings from L2 studies are used here to illustrate the role of language experience, not to conflate the two populations.

Furthermore, research indicates that to effectively establish an acoustic-phonemic mapping for the normalization of lexical tones, listeners must cultivate a robust phonological representation for each tonal category. A substantial body of research has demonstrated that the development of such phonological representations is closely associated with linguistic skill, both in children and adults (Chen & Peng, 2021; Rong et al., 2024; Stewart et al., 2018). Consequently, it stands to reason that inadequate linguistic skill could hinder the formation of a stable, abstract mental representation for each tonal category. Without these mental representations, individuals may struggle to establish a reliable perceptual reference point, which is crucial for recalibrating ambiguous targets during speech perception. This suggests a pivotal role for linguistic skill in the development of effective lexical tone normalization abilities.

4.4 Relation between musical pitch sensitivity and lexical tone normalization

This study also examined whether the ability to use contextual cues for lexical tone normalization was influenced by musical pitch sensitivity. Our findings revealed that children with diminished sensitivity to musical pitch changes exhibited less expected responses in identifying lexical tones when presented with speech context. This reduced performance could be attributed to a poor ability to discern the pitch distance between the target word and the context. Specifically, in the current study, the F0 of the context was shifted upwards or downwards by three semitones, which was sufficient to change the perception of the target tone significantly for children with adequate sensitivity to musical pitch changes. For children with reduced sensitivity to musical pitch changes, a more pronounced contextual F0 shift might be necessary for them to perceive the F0 shift to a similar extent as those with heightened sensitivity to musical pitch changes.

Our observation of positive relation between musical pitch sensitivity and level tone normalization seemingly contrasts with the findings of Tao et al. (2021), who reported that musical training did not enhance the normalization of level tones among adults. One plausible explanation for this discrepancy is that the adults in their study had already reached full development in their ability to use contextual cues, leaving little room for further improvement. This hypothesis aligns with the study by Zhang et al. (2018) on amusia, which showed that individuals with poor musical abilities also performed less well in utilizing contextual cues for lexical tone normalization. This suggests a potential link between musical pitch sensitivity and the effectiveness of lexical tone normalization, particularly during developmental stages when cognitive

768 and perceptual skills are still maturing.

769 However, after controlling for age in the hierarchical regression analysis, musical
770 pitch sensitivity no longer significantly predicted lexical tone normalization
771 performance. This suggests that the observed correlation between pitch sensitivity and
772 tone normalization may be attributable to their shared association with age. Age appears
773 to be a common developmental factor influencing both musical pitch perception and
774 lexical tone processing, likely due to increased exposure to language and music as
775 children grow older.

776 Although the pitch-related subtests of the MBEMA were well-suited to our
777 research focus, we acknowledge the possibility of ceiling effects in this typically
778 developing child sample, which may have reduced the sensitivity to detect meaningful
779 individual differences. Future studies may explore alternative assessments with greater
780 variability. Another limitation of this study is that we did not measure working memory
781 (WM), which has been proposed as a factor influencing speech normalization
782 (Nusbaum & Morin, 1992). In addition, participants' language, speech, and hearing
783 status were based on self-reports or guardian reports, without formal clinical screening.
784 Future studies may benefit from incorporating standardized assessments to confirm
785 participants' auditory and linguistic profiles.

786 Theoretically, our results suggest that lexical tone normalization in children follows
787 a protracted developmental trajectory, shaped by linguistic acquisition and perceptual
788 mechanisms. The emergence of context effects in speech but not non-speech contexts
789 supports the view that speech-specific processing, beyond acoustic contrast, plays a

critical role and matures over time. Practically, these findings may inform language assessment and early intervention strategies for children with atypical tone perception or delayed phonological development. In particular, our study highlights the importance of evaluating contextual processing abilities, not just isolated tone identification, in clinical or educational settings.

5. **Conclusion**

This study explored the development of level tone normalization in Cantonese-speaking children aged between seven and ten, and its link to linguistic skill and musical pitch sensitivity. The results suggest that the children reached adult level in using speech contextual cues to normalize level tones at the age of eight. In addition, children across all age groups did not exhibit contrastive context effect in non-speech contexts, providing further evidence that the normalization of level tones is specific to speech. Children's performance in level tone normalization was significantly associated with their linguistic skills, even after controlling for age, suggesting a robust role of language ability. The correlation with musical pitch sensitivity, however, was no longer significant after age was accounted for, indicating that its influence may be age-related rather than independent.

Data Availability Statement: Please contact the corresponding author for information about accessing the data presented in this study.

Ethics Approval Statement: Approval of the research was granted by the Human Subjects Ethics Sub-committee at The Hong Kong Polytechnic University (Project No. PolyU/RFS2122-5H01).

Funding Statement: This research was partly supported by a fellowship award from

814 the Research Grants Council of the Hong Kong SAR, China (Project No.
815 PolyU/RFS2122-5H01) and Shanghai Magnolia Talent Plan Pujiang Project (No.
816 24PJC055).
817

References:

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309-312.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated Generalized Linear Mixed Modeling. *The R Journal, 9*(2), 378–400.

Campbell, J. A., McSherry, H. L., & Theodore, R. M. (2018). Contextual influences on phonetic categorization in school-aged children. *Frontiers in Communication*, 3 http://doi.org/10.3389/fcomm.2018.00035

Chen, F., Zhang, K., Guo, Q., & Lv, J. (2023). Development of achieving constancy in lexical tone identification with contextual cues. *Journal of Speech Language and Hearing Research*, 66(4), 1148-1164. http://doi.org/10.1044/2022_JSLHR-22-00257

Chen, F., & Peng, G. (2021). Categorical perception of pitch contours and voice onset time in Mandarin-speaking adolescents with autism spectrum disorders. *Journal of Speech Language and Hearing Research*, 64(11), 4468-4484. http://doi.org/10.1044/2021_JSLHR-20-00725

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Francis, A. L., Ciocca, V., & Ng, B. K. (2003). On the (non)categorical perception of lexical tones. *Percept Psychophys*, 65(7), 1029-1044. http://doi.org/10.3758/bf03194832

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11(2), 149-175. http://doi.org/DOI 10.1016/S0095-4470(19)30813-7

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

Joos, M. (1948). Acoustic phonetics. *Language*, 2(24), 5-136. http://doi.org/https://doi.org/10.2307/522229

Linguistic Society of Hong Kong. (1993). *The Linguistic Society of Hong Kong Cantonese Romanization Scheme (Jyutping)*. Retrieved April 3, 2025, from https://lshk.org/jyutping-scheme/

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*(3), 241-265.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. Cognition, 106(3), 1367-1381.

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, 13(2), 135-165. http://doi.org/10.1016/0010-0277(83)90020-3

862 Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure
863 of a phonetic category. *Percept Psychophys*, 46(6), 505-512.
864 http://doi.org/10.3758/bf03208147
865 Mok, P. P. K., Fung, H. S. H., & Li, V. G. (2019). Assessing the link between perception
866 and production in Cantonese tone acquisition. *Journal of Speech, Language, and*
867 *Hearing Research, 62*(5), 1243-1257.
868 Mok, P. P. K., Li, V. G., & Fung, H. S. H. (2020). Development of phonetic contrasts
869 in Cantonese tone acquisition. *Journal of Speech, Language, and Hearing*
870 *Research, 63*(1), 95-108.
871 Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception.
872 *Journal of the Acoustical Society of America*, 85(5), 2088-2113.
873 http://doi.org/10.1121/1.397861
874 Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical
875 pathways for music and speech revealed by hypothesis-free voxel decomposition.
876 *Neuron, 88*(6), 1281-1296.
877 Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to difference among talkers.
878 In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception,*
879 *speech production, and linguistic structure* (pp. 113–134). IOS Press.
880 Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based
881 comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics*,
882 34(1), 134-154.
883 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 and Mandarin tone systems. *Journal of Speech, Language and Hearing Research*, 55(2), 579-595

Peretz, I., Gosselin, N., Nan, Y., Caron-Caplette, E., Trehub, S. E., & Beland, R. (2013). A novel tool for evaluating children's musical abilities across age and culture. *Frontiers in Systems Neuroscience*, 7, 30. http://doi.org/10.3389/fnsys.2013.00030

Quam, C., & Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Development*, 83(1), 236-250. http://doi.org/10.1111/j.1467-8624.2011.01700.x

Quam, C., & Swingley, D. (2014). Processing of lexical stress cues by young children. *Journal of Experimental Child Psychology*, 123, 73-89. http://doi.org/10.1016/j.jecp.2014.01.010

Rong, Y., Weng, Y., Chen, F., & Peng, G. (2023). Categorical perception of Mandarin lexical tones in language-delayed autistic children. *Autism, 27*(5), 1426-1437.

Rong, Y., Weng, Y., & Peng, G. (2024). Processing of acoustic and phonological information of lexical tones at pre-attentive and attentive stages. *Language, Cognition and Neuroscience, 39*(2), 215-231.

Singh, L., & Chee, M. (2016). Rise and fall: Effects of tone and intonation on spoken word recognition in early childhood. *Journal of Phonetics*, 55, 109-118. http://doi.org/10.1016/j.wocn.2015.12.005

Stewart, M. E., Petrou, A. M., & Ota, M. (2018). Categorical speech perception in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*, 48(1), 72-82. http://doi.org/10.1007/s10803-017-3284-0

906 Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic
907 perception. *Journal of Experimental Psychology-Human Perception and*
908 *Performance*, 7(5), 1074-1095. http://doi.org/10.1037//0096-1523.7.5.1074
909 Tamati, T. N., & Pisoni, D. B. (2014). Non-native listeners' recognition of high-
910 variability speech using PRESTO. *Journal of the American Academy of Audiology*,
911 25(9), 869-892. http://doi.org/10.3766/jaaa.25.9.9
912 Tao, R., Zhang, K., & Peng, G. (2021). Music does not facilitate lexical tone
913 normalization: A speech-specific perceptual process. *Frontiers in Psychology*, 12,
914 717110. http://doi.org/10.3389/fpsyg.2021.717110
915 To, C. K., Cheung, P. S., & McLeod, S. (2013). A population study of children's
916 acquisition of Hong Kong Cantonese consonants, vowels, and tones. *Journal of*
917 *Speech, Language, and Hearing Research, 56*(1), 103-122.
918 Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between
919 segmental and suprasegmental features in Mandarin Chinese. *Language and*
920 *Cognitive Processes*, 23(5), 689-708. http://doi.org/10.1080/01690960701728261
921 T'Sou, B. K. Y., Lee, T., Tung, P., Chan, A., Man, Y., & To, C. K. S. (2006). *Hong*
922 *Kong Cantonese oral language assessment scale*. Language Information Sciences
923 Research Centre, City University of Hong Kong.
924 Wewalaarachchi, T. D., & Singh, L. (2020). Vowel, consonant, and tone variation exert
925 asymmetrical effects on spoken word recognition: Evidence from 6-year-old
926 monolingual and bilingual learners of Mandarin. *Journal of Experimental Child*
927 *Psychology*, 189 http://doi.org/10.1016/j.jecp.2019.104698

928 Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and
929 intratalker variation in Cantonese level tones. *Journal of Speech, Language, and*
930 *Hearing Research*, 46(2), 413-421. http://doi.org/10.1044/1092-4388(2003/034)
931 Wong, P., Schwartz, R. G., & Jenkins, J. J. (2005). Perception and production of lexical
932 tones by 3-year-old, Mandarin-speaking children. *Journal of Speech Language and*
933 *Hearing Research*, 48(5), 1065-1079. http://doi.org/10.1044/1092-4388(2005/074)
934 Xi, J., Zhang, L., Shu, H., Zhang, Y., & Li, P. (2010). Categorical perception of lexical
935 tones in Chinese revealed by mismatch negativity. *Neuroscience, 170*(1), 223-231.
936 Zhang, C., Peng, G., Wang, X., & Wang, W. S. (2015). *Cumulative effects of phonetic*
937 *context on speech perception*. 18th International Congress of Phonetic Sciences,
938 Glasgow, UK.
939 Zhang, C., Peng, G., & Wang, W. S. (2012). Unequal effects of speech and nonspeech
940 contexts on the perceptual normalization of Cantonese level tones. *Journal of the*
941 *Acoustical Society of America*, 132(2), 1088-1099.
942 http://doi.org/10.1121/1.4731470
943 Zhang, C., Shao, J., & Chen, S. (2018). Impaired perceptual normalization of lexical
944 tones in Cantonese-speaking congenital amusics. *The Journal of the Acoustical*
945 *Society of America*, 144(2), 634-647. http://doi.org/10.1121/1.5049147
946 Zhang, K., Li, D., & Peng, G. (2024). Achieving perceptual constancy with context
947 cues in second language speech perception. *Journal of Phonetics*, 103
948 http://doi.org/10.1016/j.wocn.2024.101299
949 Zhang, K., Wang, X., & Peng, G. (2017). Normalization of lexical tones and

950 nonlinguistic pitch contours: Implications for speech-specific processing
951 mechanism. *Journal of the Acoustical Society of America*, 141(1), 38.
952 http://doi.org/10.1121/1.4973414
953

954 Table 1. Characteristics of participants across groups.

Group	N	Age in year (SD)
7-year-olds	20 (9 girls)	7.38 (0.22)
8-year-olds	22 (11 girls)	8.47 (0.29)
9-year-olds	20 (11 girls)	9.72 (0.47)
Adults	24 (12 females)	23.02 (3.07)

955

956

957 Table 2. Percentages of mid-level tone responses in the isolated condition and the F0

958 unshifted version of the three contextual conditions in different age groups.

	7-year-olds (SD)	8-year-olds (SD)	9-year-olds (SD)	Adults (SD)
Isolation	42.00% (14.36%)	37.27% (14.53%)	40.00% (14.87%)	48.50% (15.73%)
Pure tone	40.00% (19.19%)	44.55% (15.35%)	49.50% (20.38%)	48.27% (12.06%)
Music	39.50% (22.82%)	41.36% (16.99%)	50.00% (22.24%)	46.52% (14.95%)
Speech	55.00% (20.39%)	75.45% (20.41%)	71.08% (20.96%)	82.24% (18.16%)

959

960

961 Table 3. Expectation alignment for the music, pure tone and speech contexts across age

962 group.

Context	7-year-olds (SD)	8-year-olds (SD)	9-year-olds (SD)	Adults (SD)
Music	32.83% (27.00%)	33.64% (21.67%)	35.67% (23.60%)	32.64% (20.64%)
Pure tone	33.67% (24.21%)	34.39% (19.31%)	36.17% (23.73%)	33.02% (22.00%)
Speech	58.00% (27.05%)	74.55% (22.75%)	75.01% (25.96%)	81.25% (23.67%)

963

964

Table 4. Expectation alignment under speech context, with F0 manipulations, categorized by age group.

F0 Manipulation	7-year-olds (SD)	8-year-olds (SD)	9-year-olds (SD)	Adults (SD)
Raised	71.00% (14.83%)	76.82% (17.29%)	72.58% (27.37%)	85.42% (23.39%)
Unshifted	55.00% (20.39%)	75.45% (20.41%)	71.46% (20.85%)	84.26% (15.34%)
Lowered	48.00% (36.65%)	71.36% (29.49%)	81.00% (29.18%)	74.07% (29.43%)

969 Table 5. Children's performance in tone identification task in speech context, linguistic

970 skill test, and musical pitch sensitivity task across age group.

Group	Age in year (SD)	Expectation alignment (SD)	Raw score of linguistic skill test (SD)	Total score of musical ability task (SD)
7-year-olds	7.38 (0.22)	58.00% (27.05%)	63.65 (8.68)	47.3 (7.77)
8-year-olds	8.47 (0.29)	74.55% (22.75%)	73.77 (4.33)	53.29 (2.69)[1]
9-year-olds	9.72 (0.47)	75.01% (25.96%)	73.80 (4.80)	51.6 (3.08)

971

972

[1] As one 8-year-old child did not complete the tasks assessing musical abilities, the performance of 21 children in the 8-year-old group is shown here.

973 Table 6. Correlation among coefficients among tone normalization performance,
974 language skill, musical pitch sensitivity, and age.

	Expectation alignment	Linguistic skill	Musical pitch sensitivity	Age
Expectation alignment	-	0.63 ($p < .001$)	0.29 ($p = .027$)	0.49 ($p < .001$)
Linguistic skill	0.63 ($p < .001$)	-	0.53 ($p < .001$)	0.55 ($p < .001$)
Musical pitch sensitivity	0.29 ($p = .022$)	0.53 ($p < .001$)	-	0.31 ($p = .027$)
Age	0.49 ($p < .001$)	0.55 ($p < .001$)	0.31 ($p = .014$)	-

975 Note. Values below the diagonal are two-tailed Pearson correlation *p*-values. Values
976 above the diagonal are *p*-values adjusted for multiple comparisons using the Holm
977 method.
978