https://dx.doi.org/10.1016/j.bandl.2023.105348

The advantage of the music-enabled brain in accommodating lexical tone variabilities

Kaile Zhang[1#], Ran Tao[2#], Gang Peng[2*]

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

[2]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

*__Corresponding author__: Gang Peng

E-mail: gpeng@polyu.edu.hk

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

# The authors contributed equally to this work.

Abstract:

The perception of multiple-speaker speech is challenging. People with music training generally show more robust and faster tone perception. The present study investigated whether music training experience can facilitate tonal-language speakers to accommodate speech variability in lexical tones. Native Cantonese musicians and nonmusicians were asked to identify Cantonese level tones from multiple speakers. Two groups were equally well in using context cues to normalize lexical tone variability at behavioral level. However, the advantage of music training was observed at cortical level. The time-domain ERP analysis suggested that musicians normalized lexical tone variability much earlier than nonmusicians (N1: 70-175 ms vs. P2: 175-280 ms). An exploratory source analysis further revealed that two groups probably relied on different cortical regions to normalize lexical tones. Left BA41 showed stronger involvement in musicians in accommodating tone variability, but right auditory cortex (including BA 41, 42 and 22) activated to a greater extend in nonmusicians.

Keywords:

Music training experience; speech normalization; lexical tone; time course; source analysis

1. Introduction

As two fundamental forms to convey information and emotion, both speech and music make use of pitch and heavily depend on auditory learning (Patel, 2013; Zatorre et al., 2007). Neuroimage studies revealed shared neural circuits for pitch processing across music and linguistic domain, including the pars triangularies of Broca's area and the superior temporal gyrus (Nan & Friederici, 2013). According to the expanded OPERA hypothesis (Patel, 2014), music places higher demands on the sensory and cognitive processing mechanisms that are shared by music and speech, and thus the frequent music training with emotional reward and focused attention can shape the neural plasticity in processing sound signals, for instance, the enhanced synaptic strength (Sanju & Kumar, 2016; Tremblay et al., 2001) and increased cortical thickness (Schneider et al., 2002). The enhanced neural plasticity by music training in turn facilitates the pitch processing in language domain.

1.1. The music training experience affects the linguistic pitch processing

Although pitch movement shows relatively different forms in music and language (the continuous and curvilinear contours vs. the discrete and stair-stepped notes), the advantage of music training experience on linguistic pitch processing is widely observed (Bidelman et al., 2011). French musicians can detect the weak pitch incongruity in perceiving sentence prosody (Schön et al., 2004; Marques et al., 2007) and can detect lexical tone variations more accurately in identifying Mandarin four-word sequences (Marie et al., 2011). English musicians' tone identification approximates the perceptual pattern of native Mandarin speakers in the categorical perception of Mandarin tones, being more categorical than English nonmusicians (Chang et al., 2016). Even for native tonal language speakers with long-term experience of tone processing, music training also modulates their linguistic pitch processing. Mandarin-speaking musicians outperformed nonmusicians in discriminating within-category Mandarin tone pairs, indicating that musical training enhanced Mandarin speakers' sensitivity to subtle pitch differences which may contribute to robust mental representations of tone categories (Wu et al., 2015). Musicians also attended more to acoustic details such as intrinsic fundamental frequency (F0) and pitch types in lexical tone processing (Chen et al., 2020). Mandarin-speaking musicians benefited more from increased stimulus duration in identifying rising and falling tone continua, perhaps due to their greater sensitivity to temporal information (Chen et al., 2020).

At both subcortical and cortical levels, musicians showed more robust linguistic pitch processing. Frequency following response measured at the rostral brainstem suggested that amateur musicians without tonal language experience showed more faithful and robust pitch tracking to Mandarin tones than nonmusicians, and that the brainstem pitch tracking was significantly correlated with both the years of and the onset age of musical training (Wong et al., 2007). French musicians showed enhanced P3b components in discriminating Mandarin tone variations (Marie et al., 2011). Mandarin-speaking children who received piano training demonstrated enhanced positive mismatch responses (pMMRs) to lexical tone changes (Nan et al., 2018). Enlarged cortical response to changes in Mandarin lexical tones was also observed in Mandarin adult musicians, as indexed by the increased MMN amplitudes (Tang et al., 2016). Even Mandarin amateur musicians who received less than five-year music training showed significantly larger MMN for the within-category deviants than nonmusicians in the categorical perception of Mandarin tones (Zhu et al., 2021).

The enhanced sensitivity to fine pitch differences and the robust pitch encoding at cortical and subcortical levels result in a faster lexical tone perception in musicians. In perceiving sentences with congruent or incongruent pitches at the final word position, the positive potential triggered by strong pitch incongruencies has its onset 50 ms earlier for French musicians than nonmusicians, and the positive potential triggered by weak pitch incongruities has its onset latency 100 ms earlier in musicians (Schön et al., 2004). In discriminating Mandarin tones, tone variations elicited an increased N2/N3 that developed 100 ms earlier in French musicians than in nonmusicians (Marie et al., 2011). Mandarin musicians also discriminated Mandarin tones with shorter response times than nonmusicians (Tang et al. 2016).

1.2. Speech variabilities and lexical tone normalization

F0 is the primary acoustic cue for lexical tone perception. However, due to the anatomical difference of vocal folds, speech signals produced by different speakers vary a lot. A female speaker's low-level tone may have a similar F0 as a male speaker's production of a high-level tone. Therefore, speech variability is a big challenge for listeners to categorize two tones merely based on their intrinsic F0s (Wong & Diehl, 2003; Peng et al., 2012). This is especially difficult for Cantonese speakers who must deal with three level tones in their language which have similar pitch contours and were primarily differentiated by pitch heights. The base syllable /ji/ means doctor (/ji55/) with a high-level tone (T55), 'meaning' with a mid-level tone (T33), and 'two' with a low-level tone (T22). The inter- and intra-talker variability makes pitch height a less reliable perceptual cue. In such a condition, listeners tend to rely on contextual cues to interpret ambiguous target words and thus to some extent reduce the ambiguity caused by the inter- and intra-talker variability, a process known as *extrinsic* *normalization* (Nearey, 1989). Wong & Diehl, (2003) and Peng et al. (2012) showed that Cantonese speakers' perception of three level tones improved a lot with the help of speech contexts.

In most cases, contextual cues affect speech perception in a contrastive way. That is, a lexical tone is more frequently perceived as a high tone if its preceding context has low pitch and as a low tone if its preceding context has high pitch, which is also known as contrastive context effect in the speech normalization process (Campbell & Tyler, 2018; Francis et al., 2006; Ladefoged & Broadbent, 1957; K. Zhang et al., 2017; K. Zhang & Peng, 2021). The contrastive context effect suggests that the spectro-temporal contrast between context and target is a prerequisite for the normalization process (Holt, 2006). In addition to the spectro-temporal contrast, the reliable normalization process also requires phonetic and phonological information. Contexts composed of temporally reversed speech that had phonetic information and contexts composed of meaningless word sequences that had phonological information triggered significantly larger normalization effect during the perception of Cantonese tones than nonspeech contexts that only had spectro-temporal contrast (C. Zhang et al., 2015). It was proposed that listeners used the acoustic, phonetic, and phonological information provided by contexts to construct a talker-specific acoustic-phonemic mapping, and then they used this mapping to recalibrate the ambiguous target cues (Nusbaum and Morin,1992). For example, two tones in a Cantonese speaker's greeting 早晨 (Good morning, /tʃou25 ʃɐn21/) could roughly outline the speaker's tonal space since the ending point of T25 was produced with the highest F0 and the ending point of T21 with the lowest F0. The incoming acoustic signals can be categorized by referring to the talker-specific tonal space (Wong and Dielt, 2003; Francis et al., 2006).

Consistent with the acoustic-phonemic mapping explanation, most neurological studies suggest that the normalization process probably occurs at the early stage of speech perception process, for examples, the N1 and/or P2 time windows. Sjerps et al. (2011) found that listeners' perception of ambiguous target words triggered different N1 amplitudes in high contexts vs. in low contexts. Their research suggested that listeners started to integrate the contextual cues with target phonemes at around 80 – 160 ms after the stimulus onset. Sjerps et al. (2019) followed up to investigate the neural basis of speech normalization by recording the electroencephalography (EEG) signals with high-density intracranial electrode arrays placed on the perisylvian region. They asked listeners to identify ambiguous syllables in either the high-F1 context or low-F1 context and found that a subset of electrodes on the auditory cortex demonstrated a contrastive activation pattern of the target speech cues relative to the context F1s. The context-dependent activation of the neurons occurred at around 60 – 190 ms after the stimulus onset (see Figure 2C in that paper), which was largely consistent with their previous findings (Sjerps et al. 2011). K. Zhang and Peng (2021) compared word identification in speech and nonspeech contexts. They only observed the context-dependent perception of the target words in speech contexts but not in nonspeech contexts. By comparing the ERPs in speech- and nonspeech-context conditions, they found that target word perception in speech and nonspeech contexts elicited different P2 (130 – 250 ms) and N400 (350 – 470 ms) amplitudes. Since N400 was related to the word retrieval (Kutas & Federmeier, 2011), K. Zhang and Peng (2021) suggested that the normalization process should largely occur at the P2 time window and was related to the phonetic and phonological processing (Crowley & Colrain, 2004). Although most studies about online speech normalization process focused on vowels (e.g., Sjerps et al., 2011; Sjerps et al., 2019; and K. Zhang & Peng, 2021), lexical tone normalization probably also occurs at the similar speech processing stages (i.e., acoustic-phonemic processing) and triggers similar ERP components (i.e., N1 and/or P2) since vowels are the primary carriers of lexical tones. Indeed, with the same speech-nonspeech comparison paradigm, C. Zhang et al. (2013) observed N1 (100 – 200 ms) component in Cantonese tone normalization process. Besides, they also observed N400 (250 – 500 ms) in the lexical tone normalization process. Considering that N1 only showed in the mid-pitch context condition, but N400 consistently showed in high, mid, and low contexts, Zhang et al. (2013) suggested that the normalization process occurred in the N400 time window. However, as stated in K. Zhang and Peng (2021), the normalization process in which the ambiguous acoustic cue is recalibrated and mapped to an abstract phonological category should be finished before word retrieval. Therefore, N1 rather than N400 might be a more reliable index of the normalization process in C. Zhang et al. (2013).

1.3. The present study: the musicianship and the lexical tone normalization

So far, it remains largely unknown whether music training helps tonal-language speakers to tackle speech variability, although the carrier-over advantage from music training to lexical tone perception and learning has been widely reported. The studies about congenital amusia, a neurodevelopmental disorder of pitch processing (Ayotte et al., 2002), suggest that the impairment in music aptitude impedes the extrinsic normalization processing. People with congenital amusia cannot use contextual cues effectively in accommodating Cantonese (Shao & Zhang, 2019) and Mandarin (Liu et al., 2021) tone variabilities. However, no empirical studies tested if musicians, people with higher music aptitude, can perform significantly better in perceptual normalization of lexical tones. The present study aimed to address this question by comparing the Cantonese tone normalization of musicians and nonmusicians. Native Cantonese speakers with either intensive music training or not were recruited to perceive Cantonese level tones produced by multiple speakers in speech and nonspeech contexts. The pitch heights of contexts were manipulated to be either high or low to introduce speech variability. If listeners rely on contextual cues to perceive the target tone token, they will show a contrastive context effect. Specifically, the same Cantonese tone token will be perceived as T22 in contexts of low pitch and as T55 in contexts of high pitch. Therefore, the contrastive context effect enables the present study to observe if listeners show the extrinsic normalization of Cantonese tones at the behavioral level, and by comparing the performance of musicians and nonmusicians, the present study shall observe if music training facilitates the normalization process of lexical tones.

The electrophysiological data was also collected in the present study to see how the neural plasticity induced by music training interacted with the normalization process. The normalization process at the neurological level was detected by comparing tone perception in speech and nonspeech contexts, instead of comparing tone perception in contexts of different pitch heights. The cognitive process of normalization occurs in all speech contexts of different pitch heights, and thus the comparison between them at the cortical level may only show the different word retrieval processes rather than the normalization process. On the contrary, the contrastive context effect was observed only in speech contexts but not in nonspeech contexts, indicating that listeners might not do the normalization process in nonspeech contexts but did so in speech contexts (K. Zhang & Peng, 2021). The unequal effect between speech and nonspeech context in the perceptual normalization has been reported in many studies (e.g., Francis et al., 2006; C. Zhang et al., 2013; K. Zhang et al., 2017), suggesting that the speech-nonspeech comparison is an effective and reliable method to measure the normalization process at the cortical level.

The auditory evoked potentials, N1 and P2, would be the focus of the present study. On the one hand, as reviewed above, the normalization process has been observed in the N1 (Sjerps et al., 2011) and P2 (K. Zhang and Peng, 2021) time windows. On the other hand, N1 and P2 are sensitive to remodeling of the auditory cortex by music training (Shahin et al., 2003). Musicians showed enhanced N1m (the N1 in MEG) to piano notes compared with nonmusicians (Pantev et al., 1998). Their N1 amplitudes to frequency changes were positively related to behavioral thresholds for frequency discrimination (Lee et al., 2020). Significantly higher amplitudes in P2 were observed in musicians than nonmusicians when perceiving consonants /m/, /t/, and /g/ (Polat & Ataş, 2014). Musicians also showed increased P2 amplitude in perceiving the /u/-/a/ vowel continuum which was coupled with steeper identification functions and shorter response time (Bidelman et al., 2014). The increase in amplitudes of the cortical auditory event potentials (i.e., N1 and P2) indicates an increase in neural synchrony and strengthened neural connections in processing sound signals that are introduced by music training (Sanju & Kumar, 2016; Tremblay et al., 2001). Therefore, N1 and P2 which are sensitive to both the normalization process and the neural plasticity induced by music training are ideal neural markers to evaluate the effect of music training on the lexical tone normalization process. It was hypothesized that native Cantonese speakers with extensive music training would show a larger contrastive context effect at the behavioral level and that ERPs related to the normalization process would show larger amplitudes and/or shorter latencies in the musician group.

2. Methods

2.1. Participants

Twenty-four nonmusicians (12 female, M_{age} = 23.1, SD_{age} = 3.5) and 24 musicians (13 female, M_{age} = 24.9, SD_{age} = 5.6) were recruited to participate in this study. Nonmusicians had less than three years of musical training, while musicians experienced at least seven years of professional musical training and still actively engaged in music-related activities (Cooper & Wang, 2012; Wayland et al., 2010; Wong et al., 2007), such as regularly practicing, studying in music-related major, or having a music-related occupation (e.g., band members, private music tutors, or music teachers in schools), when they participated in this study (see Supplementary Table S1 for a detailed description of their music background). Two groups were matched in age [*Welch's* $t(38.5) = 1.36$, $p = 0.182$]. All participants, but one ambidextrous, were right-handed according to the Edinburgh handedness scale (Oldfield, 1971). All participants signed consent forms before the experiment and received a small remuneration after the experiment. The study was approved by Human Subjects Ethical Committee of The Hong Kong Polytechnic University.

2.2. Stimuli

The auditory stimuli were adapted from C. Zhang et al. (2013). Stimuli consisted of speech contexts, nonspeech contexts, and speech targets. Speech contexts were four-syllable

Cantonese phrase “呢個字係” (/li55 ko33 tsi22 hɐi22/, “This word is”), and target stimuli were Cantonese character “意” with mid-level tone (e.g., /ji33/, “meaning”). Speech contexts and targets were recorded from two female and two male native Cantonese speakers who varied in their natural pitch heights and these speakers were denoted as FH (Female speaker with High pitch range), FL, MH, and ML, respectively. The F0 trajectories of the contexts were raised or lowered three semitones to introduce the intra-talker variability and also to elicit contrastive context effect. As predicted by the contrastive context effect, the target tone following a high-F0 context would more likely be perceived as a low-level tone (e.g., character “二”, /ji22/, “son”), and the target following a low-F0 context would more likely be perceived as a high-level tone (e.g., character “醫”, /ji55/, “doctor”). The durations of contexts were kept as their original durations (FH:1005 ms, FL: 888 ms, MH: 811 ms, ML: 821 ms) to maximize the naturalness, while the duration of the target was normalized to 450 ms for the precise timing in the evaluation of participants’ EEG responses during the target tone perception. The intensities of all speech stimuli were adjusted to 55 dB. Nonspeech contexts are composed of triangle waves. They matched the speech context stimuli in F0 trajectories and duration. The intensities of nonspeech contexts were adjusted to 75dB to match the perceived loudness which was rated by native Cantonese listeners. Finally, the manipulation results in six types of contexts for each speaker: high-F0 speech context, mid-F0 speech context, low-F0 speech context, high-F0 nonspeech context, mid-F0 nonspeech context, and low-F0 nonspeech context. In each trial, the targets and contexts are congruent in speaker, i.e., the target always followed the context produced by the same speaker or its nonspeech counterpart. The experiment also included speech and nonspeech fillers prepared with the same procedure. The speech-context fillers were four-syllable Cantonese phrases “我而家讀” (/ŋo23 ji21 ka55 tuk2/, “Now I will read”) from FL and MH and “請留心聽” (/ts^{h}iŋ25 ləu21 sɐm55 t^{h}iŋ55/, “Please listen carefully to”) from FH and ML. The targets in filler trials were Cantonese Characters “意” (/ji33/) from FL and MH or “二” (/ji22/) from FH and ML.

2.3. Experiment procedure

All participants performed a Cantonese word identification task in a sound-proof booth while their EEG signals were recorded. The experiment consisted of two sessions: one session with speech contexts and one with nonspeech contexts. The session order was counterbalanced across participants. Each session had nine blocks and each block had twelve experiment trials (4 speakers × 3 *Pitch Shift*s) and four filler trials.

In each trial (Figure 1), participants first saw a 500 ms fixation at the center of the screen followed by the context stimulus played bilaterally through insert earphones. After a jittering silence (300 – 500 ms), a target syllable was played. A question mark which delayed 800 – 1000 ms from the onset of the target, appeared at the center of the screen. Participants were instructed that in each trial, they would hear a Cantonese sentence, and they needed to identify whether the last syllable was "醫" (/ji55/), "意" (/ji33/), or "二" (/ji22/) by pressing the designated keys on the keyboard when they saw the question mark. The maximum response time was 1500 ms.

All trials within each block were randomly presented. Short rests were provided between blocks to prevent fatigue. Participants got familiar with the trial procedure with a short practice session before the experiment. The stimuli used in the practice session were recorded from two speakers (one female and one male) different from the formal task. Speech contexts were all four-syllable Cantonese phrase "呢個字係", while the target syllables were low-, mid-, and high-level tone characters (i.e., "二", "意", and "醫") in the high-, mid-, and low-F0 speech context trials respectively, to reduce ambiguity during the practice.

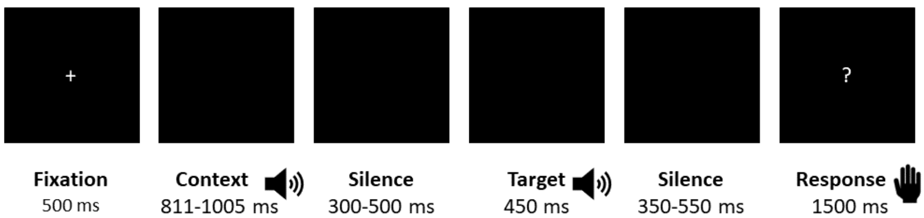

Figure 1: The trial procedure of the Cantonese word identification task.

2.4. EEG signal recording and preprocessing

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

The preprocessing of EEG signals was conducted using self-written scripts with functions from EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) in the MATLAB environment. For the time-domain ERP analysis, the EEG signal was filtered offline with a 0.1 Hz high-pass and a 30 Hz low-pass filters (both slopes = 12 dB/Oct) and re-referenced offline to the average of the two mastoid recordings. Epochs ranging from -100 to 800 ms (time locked to the onset of target stimulus) were extracted. Baseline correction was performed according to the -100 – 0 ms pre-target stimulus activity.

EEG epochs containing horizontal or vertical eye movements or exceeding an absolute threshold of ±100 μV at any scalp channels were excluded from the analysis. Eye blinks were detected automatically by a moving window peak-to-peak threshold criterion on the VEOG data with the threshold of 100 μV, the window size of 200 ms, and the widow step of 50 ms. Horizontal eye movements were detected automatically by a step-like threshold criterion on the HEOG data with the threshold of 40 μV, the window size of 400 ms, and the window step of 10 ms. All participants' EEG epochs in each experimental condition have reasonably good acceptance rate (> 75%) and thus were included in the following analysis. The acceptance rates did not differ between nonmusicians (Mean = 90.06%, SD = 7.02%) and musicians (Mean = 93.20%, SD = 6.53%; p = 0.207).

3. Results

3.1. Behavioral results: Speech superiority in lexical tone normalization

The present study adopted two indices: the perceptual height (PH) and the expected identification rate (IR) to evaluate participants' Cantonese level tone normalization in each experimental condition (Wong & Diehl, 2003; C. Zhang et al., 2012). PH was defined by assigning a number to the response, e.g., one for “二” (/ji22/), three for “意” (/ji33/), and six for “醫” (/ji55/). This coding scheme reflected the relative pitch distance among three Cantonese level tones (Wong & Diehl, 2003). IR was defined by whether the response was in accordance with the contrastive context effect. Specifically, in the high-F0 context condition, “二” (/ji22/) responses were coded as 1 and other responses [“意” (/ji33/) or “醫” (/ji55/) responses] as 0; in the mid-F0 context condition, “意” (/ji33/) responses were coded as 1 and other responses [“二” (/ji22/) or “醫” (/ji55/) responses] as 0; in the low-F0 context condition, “醫” (/ji55/) responses were coded as 1 and other responses [“二” (/ji22/) or “意” (/ji33/) responses] as 0. These two indices provided different views to demonstrate the normalization results. PH can intuitively visualize how the pitch information of target tone was encoded according to the preceding contexts, while IR can better quantify how frequently listeners did the normalization process. These two indices were selected for better comparison with previous speech normalization studies which adopted similar paradigm and analysis strategies (e.g., Wong & Diehl, 2003; C. Zhang et al., 2012, K. Zhang et al., 2017).

Three-way ANOVAs were performed on PH and IR respectively to reveal whether participants used the context cues to normalize the lexical tone variability. The three factors of interest were the between-subject factor *Group* (two levels: nonmusicians and musicians), the within-subject factor *Sound Type* (two levels: speech context and nonspeech context), and the within-subject factor *Pitch Shift* (three levels: high-F0 context, mid-F0 context, and low-F0 context). Greenhouse-Geisser corrections were used to correct violations of sphericity assumption whenever necessary in all analyses.

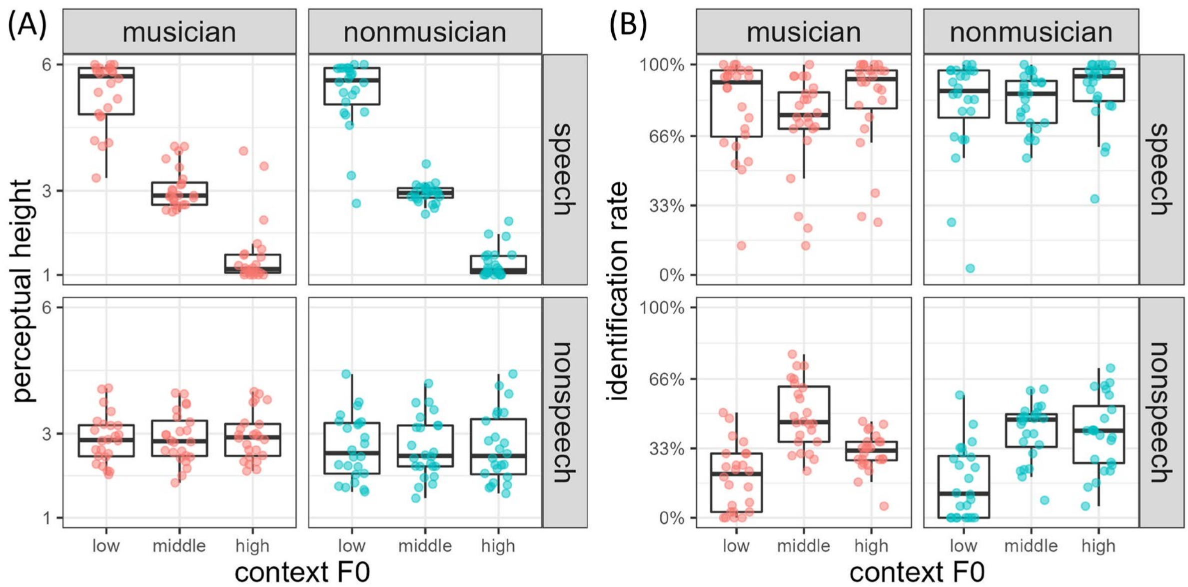

Figure 2: Boxplot of behavioral performance summarized as (A) the perceptual height and (B) the identification rate in each context condition. Each dot represents one subject's result.

The analysis on PH showed significant main effect of *Sound Type* [$F(1,46) = 21.95$, $\eta^2 = 0.105$, $p < 0.001$]. The speech context in general elicited a higher PH than nonspeech context [3.23 (0.068) vs. 2.79 (0.137)]. There was also a significant main effect of *Pitch Shift* for PH analysis [$F(1.24, 57.24) = 465.48$, $\eta^2 = 0.614$, $p < 0.001$]. Post-hoc analysis showed that the low-F0 context elicited the largest PH and high-F0 context elicited the lowest PH [low-F0: Mean (SE) = 4.05 (0.116), mid-F0: 2.89 (0.086), and high-F0: 2.08 (0.099); all *ps* < 0.001]. The interaction between *Sound Type* and *Pitch Shift* [$F(1.30, 60.01) = 395.14$, $\eta^2 = 0.621$, $p <$

0.001] was also significant. Post-hoc analysis revealed that only in speech-context conditions, listeners showed diverged PHs in different *Pitch Shift* conditions (see the upper panels of Figure 2A), indicating a successful normalization process of lexical tones in speech contexts. The low-F0 speech contexts elicited highest PH [5.32 (0.114)] and the mid-F0 speech contexts elicited a higher PH than high-F0 speech context condition [2.99 (0.055) vs. 1.36 (0.087), all ps < 0.001], while the PHs in low-, mid-, and high-F0 nonspeech contexts were similar [2.78 (0.099), 2.78 (0.096), and 2.80 (0.104), respectively; all ps > 0.7]. Critical to the interest of the present study, neither the main effect of *Group* nor its interactions with other factors was significant in the analysis of PH (all ps > 0.5).

The analysis on IR showed similar result patterns. There were significant main effects of *Sound Type* [$F(1,46) = 376.98$, $\eta^2 = 0.651$, $p < 0.001$] and *Pitch Shift* [$F(1.53, 70.32) = 13.33$, $\eta^2 = 0.091$, $p < 0.001$], and a significant *Sound Type* by *Pitch Shift* interaction [$F(1.96, 90.03) = 22.32$, $\eta^2 = 0.105$, $p < 0.001$]. The IR was higher in speech contexts than nonspeech contexts [81.1% (3.29%) vs. 32.8% (0.71%)]. For the levels of *Pitch Shift*, the mid- and high-F0 conditions elicited similar IRs [61.2% (2.74%) vs. 60.6% (2.19%), $p = 0.935$], while the low-F0 condition elicited lowest IR [49.0% (3.08%), all ps < 0.01]. The post-hoc analysis on the *Sound Type* by *Pitch Shift* interaction showed that the three speech-context conditions elicited higher IRs than their nonspeech counterparts: in low-F0 conditions, speech contexts elicited higher IR than nonspeech contexts [80.2% (3.31%) vs. 17.9% (2.33%)], and similarly in mid- [77.6% (2.78%) vs. 44.8% (2.19%)] and high-F0 conditions [85.5% (2.84%) vs. 35.6% (2.11%), all ps < 0.0001]. In addition, a significant *Group* by *Sound Type* by *Pitch Shift* interaction was observed in the analysis of IR [$F(1.96, 90.30) = 3.31$, $\eta^2 = 0.017$, $p = 0.042$]. Post-hoc analysis showed that, only in the high-F0 nonspeech-context condition, the musician group showed a lower IR than nonmusician group [31.1% (1.85%) vs. 40.1% (3.60%), $p = 0.032$], and no other significant differences were found in the pairwise comparison (all ps > 0.05). However, in high-F0 nonspeech contexts, both IR of the musician group [$t(23) = -1.21$, $p = 0.238$] and that of the nonmusician group [$t(23) = 1.87$, $p = 0.075$] were not significantly different from the chance level (33.3%), suggesting no reliable normalization process in such conditions. Therefore, higher IR in the nonmusician group can hardly be interpretated as a better normalization process.

In summary, both PH and IR analyses replicated previous results that native Cantonese speakers could normalize Cantonese level tones only if speech contexts were provided (C. Zhang et al., 2013; K. Zhang et al., 2017). In addition, our results indicated that both groups have very similar talker normalization performance, and that neither musicians nor nonmusicians showed noticeable advantage in accommodating lexical tone variability.

3.2. The results of time-domain ERP analysis: Tone normalization was early for musicians

Figure 3A showed the global field power (GFP) which was computed as the root mean square of the ERP voltage and then was averaged across the scalp electrodes, different contexts, and two groups. The ERPs at the nine representative electrodes were plotted in Figure 3B. As stated in Section 1.3, the present study planned to examine the early ERP components: N1 and P2. The GFP and the ERPs suggested that N1 and P2 did emerge during participants' perception of the Cantonese level tones in the present study. The time-windows chosen for N1 were 70 to 175 ms and 175 to 280 ms for P2 based on the visual inspection of the GFP. Only ERPs from -100 ms to 450 ms were plotted in Figure 3B to match the duration of the target stimuli, which was also enough to cover the N1 and P2 time windows. According to the topographic images (Figure 3C), electrodes where the ERP components were expected to peak were chosen to quantify the corresponding ERP components. The electrodes selected for N1 and P2 are listed in Table 1.

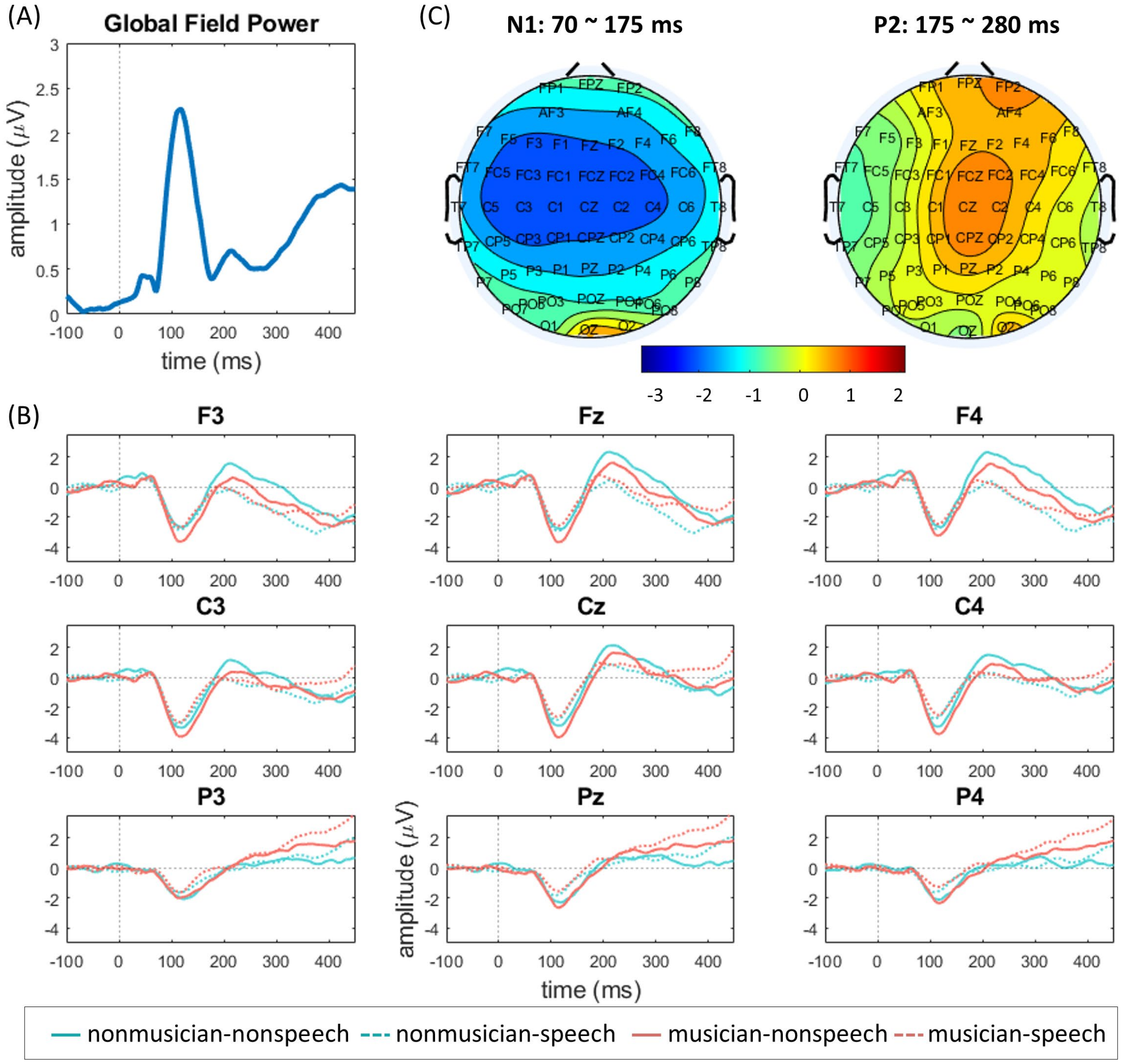

Figure 3: (A) Global field power; (B) ERPs at nine representative electrodes in speech- and nonspeech-context conditions for two groups; and (C) Topographies of N1 (left) and P2 (right).

Table 1: Summary of the time windows and electrodes chosen for N1 and P2.

Components	Time Windows (ms)	Electrodes
N1	70~175	F3, F1, Fz, FC5, FC3, FC1, FCz, FC2, C5, C3, C1, Cz, C2, C4
P2	175~280	Fz, F2, F4, FC1, FCz, FC2, FC4, C1, Cz, C2, C4, CP1, CPz, CP2

The ERP in each experimental condition was obtained by averaging the pre-processed EEG epochs, by which the trial information was lost. Therefore, instead of using linear mixed-effect model, three-way repeated-measure ANOVAs were conducted on the mean amplitude and the peak latency of each ERP component (N1 and P2) with *Group* as the between-subject factor and *Sound Type* and *Pitch Shift* as the within-subject factors. Since the present study was not interested in how participants responded to a particular talker, the ERPs were averaged across four talkers to increase the signal-to-noise ration.

3.2.1. ERP amplitude

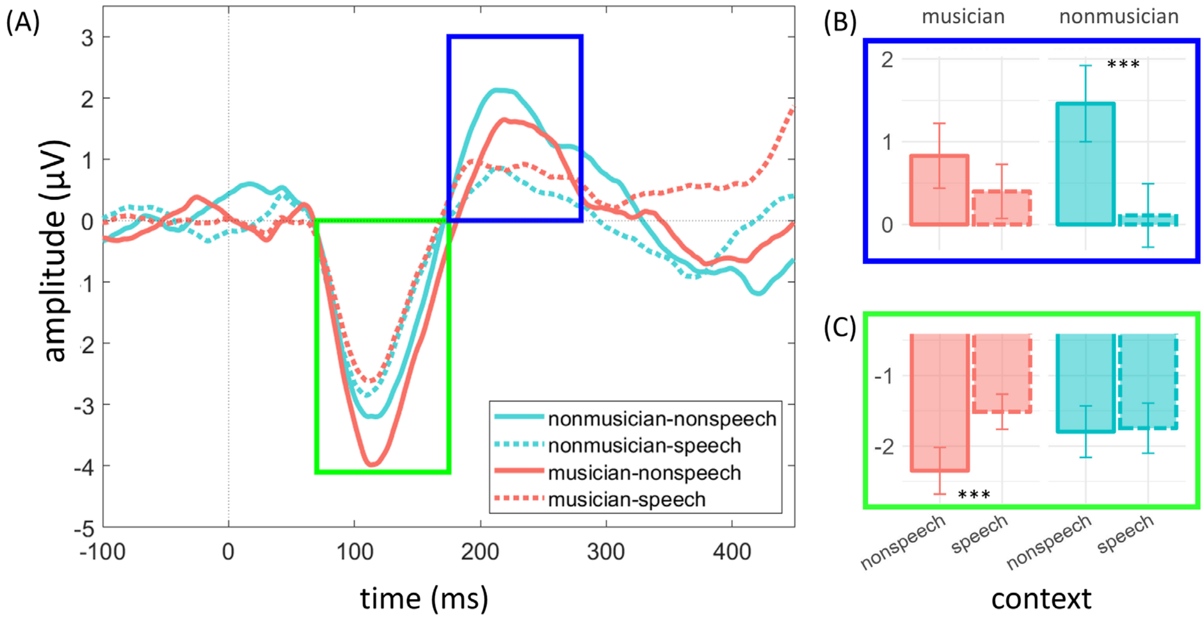

Figure 4: (A) The ERPs at Cz for each group in the nonspeech- and speech-context conditions. Bar plots show the mean amplitude of (C) N1and (B) P2 for each group in the nonspeech- and speech-context conditions; error bars represent standard error. ***: $p < 0.001$.

N1: Musicians showed the speech-nonspeech context difference

Mean amplitudes of ERP were first averaged across selected electrodes and then entered the three-way ANOVAs. The analysis on N1 revealed a main effect of *Sound Type* [$F(1, 46) = 7.6$, $\eta^2 = 0.015$, $p = 0.008$], where nonspeech contexts [Mean (SE) = -2.07 (0.160)] elicited a larger amplitude than speech contexts [-1.63 (0.147)]. Most importantly, the factor *Group* was involved in a significant *Sound Type* by *Group* interaction [$F(1, 46) = 5.99$, $\eta^2 = 0.012$, $p = 0.018$]. Simple main effect analysis showed that, for the musician group, nonspeech contexts elicited larger N1 amplitude than speech contexts [-2.35 (0.332) vs. -1.51 (0.249), $p < 0.001$; see Figure 4C], but nonspeech- and speech-context conditions elicited similar N1 amplitude for the nonmusician group [1.80 (0.365) vs. 1.75 (0.355), $p = 0.828$]. A significant main effect of *Pitch Shift* was also observed [$F(1.98, 91.21) = 5.00$, $\eta^2 = 0.011$, $p = 0.009$].

Low-F0 condition elicited larger N1 amplitude than high-F0 condition [-2.11 (0.176) vs. -1.66 (0.196), $p = 0.007$]. N1 amplitudes in mid-F0 context condition [-1.79 (0.194)] was statistically comparable to the other two conditions (*ps* > 0.05). The target tone in the present study was always T33 whose pitch height was much closer to T22 than T55 (Peng, 2006), but the pitch shift in low- and high-F0 contexts was the same (i.e., three semitones). Therefore, normalization in low-F0 contexts (i.e., perceiving T33 as T55) was more difficult than normalization in high-F0 contexts (perceiving T33 as T22) (IR: 49.0% vs. 60.6%), which probably resulted in a larger N1 amplitude in low-F0 contexts than in high-F0 context.

P2: Nonmusicians showed the speech-nonspeech context difference

Within the P2 time-window, a main effect of *Sound Type* was observed [$F(1, 46) = 15.99$, $\eta^2 = 0.040$, $p < 0.001$], where nonspeech context [Mean (SE) = 1.15 (0.302)] elicited significantly larger P2 amplitude than speech context [0.25 (0.250)]. There was also an interaction between *Sound Type* and *Group* [$F(1, 46) = 4.20$, $\eta^2 = 0.011$, $p = 0.046$]. Simple main effect analysis showed that, for nonmusician group, nonspeech contexts elicited larger P2 amplitude than speech contexts [1.46 (0.460) vs. 0.11 (0.383), $p < 0.001$; see Figure 4B]. In contrast, the musicians had statistically similar P2 amplitudes across the nonspeech- and speech-context conditions [0.83 (0.391) vs. 0.40 (0.328), $p = 0.175$]. Other factors or their interactions were not significant.

3.2.2. ERP latency

Peak latencies were averaged across selected electrodes and then entered the three-way ANOVAs. No significant main effects were observed in the N1 and P2 latency analyses. There was a significant *Pitch Shift* by *Sound Type* interaction in the N1 time window [$F(1.87,86.21) = 3.37$, $\eta^2 = 0.013$, $p = 0.042$]. Simple main effect analysis showed that the three *Pitch Shift*s elicited similar N1 latencies in nonspeech contexts (*ps* > 0.05), while in speech contexts, low-F0 context elicited later N1 than mid-F0 [Mean (SE): 121 (3.8) ms vs. 114 (3.4) ms, $p = 0.023$] and high-F0 [121 (3.8) ms vs. 114 (3.6) ms, $p = 0.033$] contexts did. Considering relatively larger pitch difference between T33 and T55, perceiving T33 as T55 in low-F0 context was more difficult, which might account for the later N1 in low-F0 contexts than other two contexts.

3.3 The correlation between ERP and behavioral results

To investigate if the neural responses could explain the behavioural results, the Spearman correlation between ERP (N1 and P2) amplitudes and participants' IR were calculated on all participants. Only the speech-context condition was included in the correlation analysis since behavioural results suggested that significant normalization process was observed only in the speech-context condition. Only the musician group showed a strong correlation between N1 amplitude and IR ($r = 0.42$, $p = 0.042$), indicating that musicians with a better behavioural performance showed smaller N1 (note that N1 is negative). It is possible that successful normalization process eases the acoustic-phonemic encoding of target speech signals, and thus resulting in a smaller N1. No other significant correlations were found (all *p*s > 0.16, see supplementary analysis).

3.4. An exploratory ERP source analysis: partially different brain regions are involved in different groups.

The ERP source analysis was inspired by the results of the time-domain EEG analysis, which suggested that lexical tone normalization triggered different ERP components in musicians and nonmusicians, with N1 for musicians and P2 for nonmusicians (see Section 4.1 for more detailed discussions). The topographies of N1 and P2 (Figure 3C) illustrated that the contributing electrodes for N1 and P2 were different, with N1 more left-lateralized and P2 slightly right-lateralized, indicating the spatial differences in lexical tone normalization process between two groups. Besides, previous studies also reported that N1 and P2 have different (albeit partially overlapped) cortical sources, with N1 most likely in the primary auditory cortex (Näätänen & Picton, 1987; Woods, 1995) and P2 in the planum temporale and Brodmann's Areas (BA) 22 (Crowley & Colrain, 2004; Godey et al., 2001; Verkindt et al., 1994).

To further evaluate if musicians and nonmusicians adopted different cortical regions for lexical tone normalization, a follow-up exploratory source localization was performed using the standard Low Resolution Electromagnetic Tomography (sLORETA, see Pascual-Marqui, 2002) implemented in LORETA-KEY software (version: v20210701). The detailed methods for the ERP source analysis can be seen in the Supplementary materials. Strong activations relative to pre-target stimulus baseline were indeed observed in the bilateral auditory cortices within the N1 and P2 time windows in both groups. To explore whether musicians and nonmusicians adopted the auditory cortex differently for lexical tone normalization, the current source densities were examined more closely in six ROIs: the primary auditory cortex (left and right BA 41), secondary auditory cortex (left and right BA 42), and the Wernicke's area (left and right BA 22). The six ROIs were chosen because they were reported to be responsible for normalization process by previous fMRI (von Kriegstein et al., 2010; C. Zhang et al., 2016) and ECoG (Sjerps et al., 2019) studies which had high spatial resolution. The six ROIs were operationally defined with the ROI maker module in LORETA-KEY.

For each participant, the current source densities were converted from individual ERPs averaged across three *Pitch Shifts* (high-F0 context, mid-F0 context, and low-F0 context) for each *Sound Type* (speech context and nonspeech context). For each time-point of the N1 and P2 time windows, the standardized current source densities of each ROI were averaged and entered two-way ANOVAs with a between-subject factor *Group* (two levels: musician and nonmusician) and a within-subject factor *Sound Type* (two levels: speech context and nonspeech context) to determine if there was any significant effect. To evaluate the effect directions, values of significant time points within each time-window were averaged for post-hoc analysis.

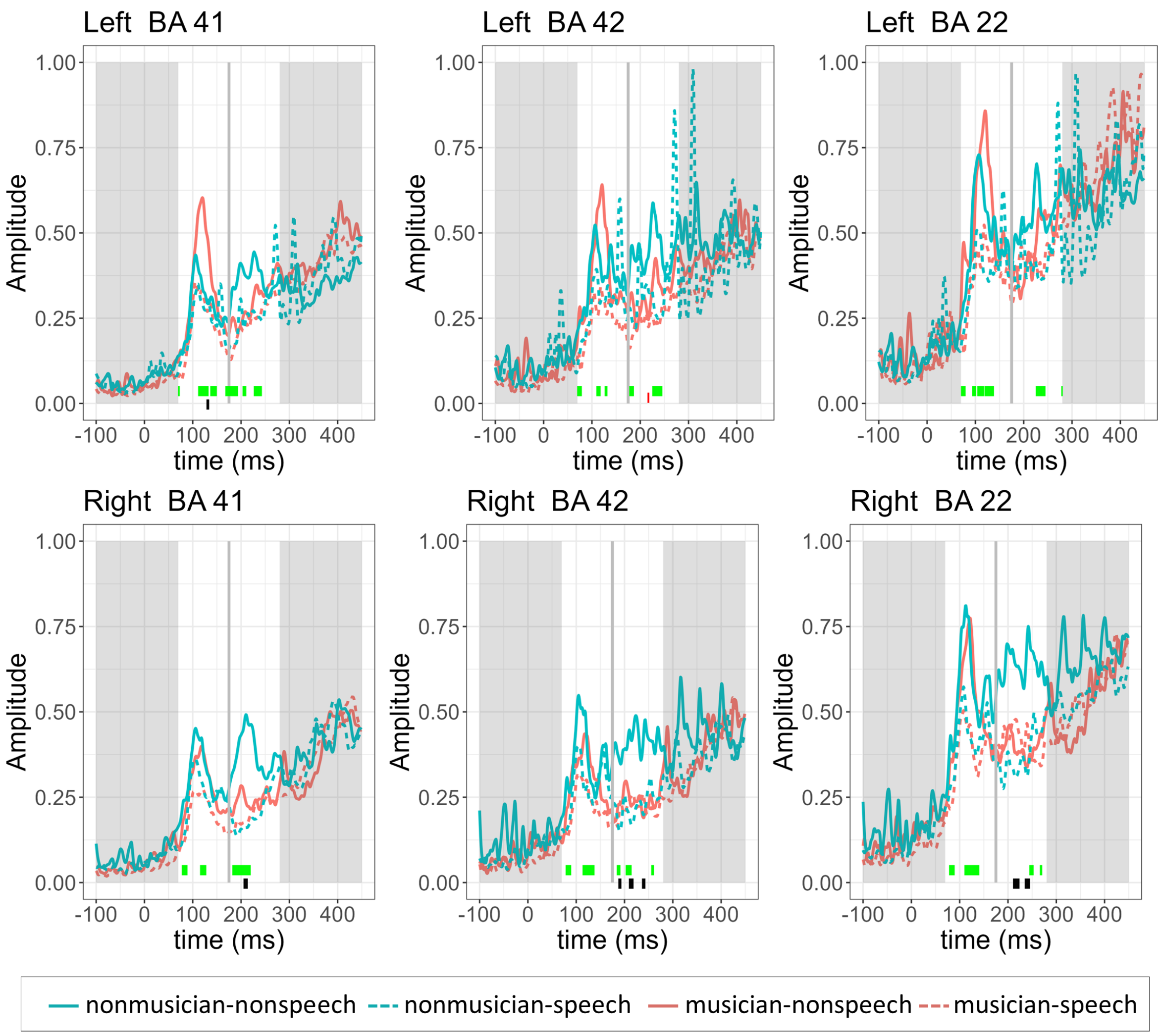

Figure 5: The time series of standardized current source densities (arbitrary unit) in the six ROIs. Green bars on the bottom of each figure refer to the significant main effect of *Sound*

Type; red bars refer to the significant main effect of *Group*; black bars refer to the significant interactions between *Group* and *Sound Type*.

Most interestingly, significant *Sound Type* by *Group* interactions were observed in 129-133 ms within the N1 time window in left BA 41, and in several periods within the P2 time window in right BA 41, 42 and 22 (*p*s < 0.05, as indicated by black bars in Figure 5). Post hoc analysis on the time points with significant *Sound Type* by *Group* interactions showed that, in 129-133 ms within the N1 time-window, only the musician group exhibited larger current source density in the nonspeech-context condition than the speech-context condition in left BA 41 ($p < 0.05$). In several periods within the P2 time-window (see black bars in Figure 5), however, only the nonmusician group exhibited a contextual difference in the right BA 41, right BA 42 and right BA 22, with larger current source density shown in the nonspeech context condition than in the speech context condition (*p*s < 0.05). There was no interaction between *Sound Type* and *Group* in the N1 or P2 time window in left BA 42 and 22. These results indicated that auditory cortex activated differently in two groups for the lexical tone normalization process. Specifically, the musician group showed slightly stronger normalization effect on the left primary auditory cortex (left BA 41) in an early time-window, while the nonmusician group may rely more on the right auditory cortex (right BA 41, 42 and 22) in a later time-window, and left BA42 and 22 were equally important for the two groups in lexical tone normalization. Besides, *Sound Type* main effect was observed in six ROIs (*p*s < 0.05, as indicated by green bars in Figure 5). Nonspeech contexts consistently elicited larger current source density than speech contexts (*p*s < 0.05). *Group* main effect was observed in left BA 42 (*p*s < 0.05, as indicated by red bars in Figure 5) where nonmusicians briefly showed higher current source densities than musicians. The detailed time ranges showing significant main effects or interactions can be seen in supplementary table S2.

4. Discussion

4.1 Musicians' earlier accommodation of lexical tone variability

The present study tested the effect of music training experience on Cantonese speakers' lexical tone normalization process. Two groups of participants with different music backgrounds were recruited to finish a Cantonese word identification task. They were instructed to listen to a Cantonese sentence and identify the last syllable (i.e., the target syllable). Critically, the preceding contexts were either speech or nonspeech and their F0s also differ to elicit normalization process. The analysis on subjects' PH revealed a significant *Sound Type* by *Pitch Shift* interaction, indicating difference perceptual patterns in speech- and nonspeech-context conditions. Indeed, the post hoc analysis suggested that the PHs varied in speech contexts of different F0s, with the highest PH in the low-F0 speech contexts and the lowest PH in the high-F0 speech contexts, showing a typical contrastive context effect. However, the PHs in nonspeech contexts of different F0s were comparable, showing no context-dependent perception of target tones. Similar conclusion can also be driven from the analysis of IR. The significant main effect of *Sound Type* in IR analysis suggested that subjects did the normalization process more frequently in the speech context condition (81.1%) than in the nonspeech context condition (32.8%) which was almost around the chance level (33.3%). The behavioral data reduplicated the unequal effect of speech and nonspeech contexts: significant normalization process emerged in speech but not in nonspeech contexts (e.g., Francis et al., 2006; C. Zhang et al., 2012). The unequal effects suggested that the comparison of tone perception in speech and nonspeech contexts could effectively detect the normalization process at the cortical level.

At the cortical level, the present study did observe a main effect of *Sound Type* (i.e., the speech-nonspeech context difference) in N1 and P2 times windows. Since the context difference reflected the normalization process, the ERP results suggested that lexical tone normalization occurred in the N1 and P2 time windows when two groups were pooled together. Nonspeech contexts triggered significantly larger ERP amplitude than speech contexts did in both the N1 and P2 time window. It was possible that lexical tone normalization process eased the acoustic-phonemic encoding of target lexical tones and thus the N1 and P2 amplitudes were smaller in the speech-context condition than in the nonspeech-context condition. However, the interpretation of the main effects of *Sound Type* should be careful as there were significant *Sound Type* by *Group* interactions in both the N1 and P2 time windows, suggesting that speech-nonspeech context difference differed in the musician and nonmusician groups. The simply main effect analyses on the *Sound Type* by *Group* interactions revealed that musicians showed speech-nonspeech context difference in the N1 time window (70 – 175 ms) but nonmusicians showed such a difference in the P2 time window (175 – 280 ms), indicating that lexical tone normalization triggered N1 in musicians, but P2 in nonmusicians.

Both N1 and P2 were reported to index the normalization process (Sjerps et al., 2011; K. Zhang & Peng, 2021). Nonmusicians in the present study showed the normalization effect in the P2 time window (175-280 ms). This was consistent with the findings in K. Zhang and Peng (2021) which recruited nonmusicians as their participants and adopted the similar speech-nonspeech comparison paradigm to investigate the time course of Cantonese vowel normalization. Although the speech cues varied in these two studies (i.e., F0 for lexical tones and F1 for vowels), both studies observed that normalization process triggered P2 component with overlapping time windows (175 - 280 ms for lexical tones vs. 130 - 250 ms for vowels). These studies together suggested that P2 might be a reliable ERP component in normalization process especially for nonmusicians. The musicians in the present study showed the normalization effect in the N1 time window. C. Zhang et al. (2013) with similar experiment designs also found a relatively weak context effect in the N1 time window probably due to the mixed music backgrounds of their participants. Sjerps et al. (2011) embedded the Dutch vowel normalization task into an active multiple-deviant oddball paradigm. They found a context modulation on the target vowel perception in the N1 time window (80 – 160 ms), which largely overlaps the N1 in the present study (70 – 175 ms). These results suggested that normalization can also be observed reliably in the N1 time window, especially when there was no strict control on participants' music backgrounds. In summary, the time-domain ERP analysis suggested that the lexical tone normalization occurs no later than P2 time window. In addition, musicians may begin the normalization process as early as in N1 time window.

The time-domain ERP analysis showed that musicians started to use contextual cues to guide target tone perception as early as 70 – 175 ms after the stimulus onset, but nonmusicians did so 175 – 280 ms after the stimulus onset, indicating a music-training advantage in the time course of the online lexical tone normalization process. The earlier normalization process of Cantonese level tones in the musician group is consistent with previous studies which showed that music training enhances the neural plasticity and advances the sound signal processing (Schön et al., 2004; Marques et al., 2007; Marie et al., 2011). Therefore, the present study extends previous research by showing that musicians not only process the intrinsic cues of the target tone earlier but also integrate the context and target cues earlier than nonmusicians do.

Musicians' faster context-target cue integration might be partially due to the relative pitch practice in music domain. Most musicians are trained to perceive the scale degrees and the tonal functions of musical tones within an established tonal context (Miyazaki et al., 2018). According to the acoustic-phonemic mapping explanation, lexical tone normalization is essentially a relative pitch processing, i.e., the recalibration of the target cue with the contextual reference (Nusbaum and Morin,1992). Skills in a perceptual expertise domain can influence the processes involved in other domains depending on their similarities (Liao et al., 2022). Moreover, relative pitch processing in music is more fine-grained than that in the linguistic domain, since there are twelve music notes within an octave in western music, but almost no languages have more than five level tones in their tonal systems (Tsai et al., 2018). Compared with tonal language speakers, musicians were better at discriminating different pitches and more consistent in their assessments of the direction and magnitude of relative pitch change (Ngo et al., 2016). Therefore, the finer-grained relative pitch in music probably enables musicians in the present study to have a faster online process of relative pitch in the linguistic domain.

4.2. Musicians and nonmusicians rely on partially different brain regions to normalize tone variability

Lexical tone normalization triggered N1 in musicians but P2 in nonmusicians. The supporting electrodes for N1 and P2 also varied a lot, indicating a spatial difference in the normalization process at cortical level between two groups. Besides, considering that N1 and P2 have partially different brain sources, a follow-up exploratory source location analysis was carried out to test if musicians and nonmusicians used different brain regions to accommodate lexical tone variability. The significant main effect of *Sound Type* was observed in all six ROIs, suggesting that all these regions were responsible for the speech normalization process when two groups were pooled together, which was consistent with the findings in previous fMRI (von Kriegstein et al., 2010; C. Zhang et al., 2016) and ECoG studies (Sjerps et al., 2019). The current source density was larger in the nonspeech contexts than in the speech contexts, which showed the similar pattern as the ERP amplitude analysis. The larger current source density in the nonspeech-context condition suggested that without effective speech normalization, the target tone processing was more demanding for cortical resources.

Additionally, the source for the normalization process varied in the musician and nonmusician groups, as suggested by significant *Sound Type* by *Group* interactions. In left BA41, only the musician group showed a significant speech-nonspeech context difference from 129 ms to 133 ms (a period within the N1 time window), but in several periods within the P2 time window (see Supplementary Table S2), the speech-nonspeech context difference was observed in right BA41, right BA42, and right BA22 only for the nonmusician group. The source analysis echoed the time-domain ERP analysis by similar *Sound Type* by *Group* interactions. The time-domain and source-domain analyses together suggested that musicians and nonmusicians not only accommodated tone variability in different time windows but also relied on partially different brain regions. The left primary auditory cortex was involved to a greater extent in the lexical tone normalization process in the musician group, but the primary, secondary, and associate auditory cortex in the right hemisphere showed stronger involvement in the nonmusician group, implying that musicians used less cortical resources to achieve the same normalization results. Schneider et al. (2002) reported that due to music practice, musicians showed enlarged gray matter volume of Heschl's gyrus. This might be the structural basis for the stronger involvement of the left primary auditory cortex in musicians' lexical tone normalization process.

It is worth noting that the exploratory source analysis in the present study has noticeable limitations. First of all, the present study was not designed specifically for identifying the brain regions of the normalization process. The participants' precise EEG electrode coordinates were not obtained, nor were structural images of their brains. Thus, templates from LORETA software were applied. Future studies may increase the source estimation accuracy by deliberately obtaining the precise electrode coordinates and corresponding brain structural information. Second, the spatial resolution of sLORETA was low (e.g., voxel size of 5 mm^3). Large voxels may not well estimate the generator from folded structures such as the Heschl's gyrus. Therefore, the source analysis cannot give an exclusive conclusion about the brain regions that are responsible for the normalization process in groups with different music training experience. Studies with the fMRI technique which has much higher spatial resolution need to be carried out to clarify this question in the future.

5. Conclusions

The present study tested if musicians of tonal language speakers showed superiority in accommodating lexical tone variability compared with nonmusicians. Although at the behavioral level, musicians' performance was similar to that of the nonmusicians in normalizing Cantonese level tones, the music-training advantage was observed at the cortical level. Specifically, the lexical tone normalization occurred in the N1 time window (70 – 170 ms) in the musician group, but in the P2 time window (170 – 300 ms) in the nonmusician group, indicating an earlier normalization process of musicians. Moreover, the source analysis revealed that musicians probably relied on left BA 41 more heavily, but nonmusician on right BA 41, 42 and 22, to normalize lexical tone variabilities.

Acknowledgments

This study was supported by a grant from the Research Grants Council of Hong Kong (GRF: 15607518).

References:

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia. A group study of adults afflicted with a music-specific disorder. *Brain*, *125*(2), 238–251. https://doi.org/10.1093/brain/awf028

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience*, *23*(2), 425–434. https://doi.org/10.1162/jocn.2009.21362

Bidelman, G. M., Weiss, M. W., Moreno, S., & Alain, C. (2014). Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *The European Journal of Neuroscience*, *40*(4), 2662–2673. https://doi.org/10.1111/ejn.12627

Campbell, K. L., & Tyler, L. K. (2018). Language-related domain-specific and domain-general systems in the human brain. *Current Opinion in Behavioral Sciences*, *21*, 132–137. https://doi.org/10.1016/j.cobeha.2018.04.008

Chang, D., Hedberg, N., & Wang, Y. (2016). Effects of musical and linguistic experience on categorization of lexical and melodic tones. *The Journal of the Acoustical Society of America*, *139*(5), 2432–2447. https://doi.org/10.1121/1.4947497

Chen, S., Zhu, Y., Wayland, R., & Yang, Y. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? In *PLoS ONE* (Vol. 15, Issue 5). https://doi.org/10.1371/journal.pone.0232514

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America*, *131*(6), 4756–4769. https://doi.org/10.1121/1.4714355

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

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

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

Godey, B., Schwartz, D., De Graaf, J. B., Chauvel, P., & Liégeois-Chauvel, C. (2001).

Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: A comparison of data in the same patients. *Clinical Neurophysiology*, *112*(10), 1850–1859. https://doi.org/10.1016/S1388-2457(01)00636-8

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

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

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

Lee, J., Han, J. H., & Lee, H. J. (2020). Long-Term Musical Training Alters Auditory Cortical Activity to the Frequency Change. *Frontiers in Human Neuroscience*, *14*(August), 1–13. https://doi.org/10.3389/fnhum.2020.00329

Liao, W., Li, S. T. K., & Hsiao, J. H. (2022). Music reading experience modulates eye movement pattern in English reading but not in Chinese reading. *Scientific Reports*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-12978-9

Liu, F., Yin, Y., Chan, A. H. D., Yip, V., & Wong, P. C. M. (2021). Individuals with congenital amusia do not show context-dependent perception of tonal categories. *Brain and Language*, *215*(February), 104908. https://doi.org/10.1016/j.bandl.2021.104908

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(1 APR), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., & Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *Journal of Cognitive Neuroscience*, *23*(10), 2701–2715. https://doi.org/10.1162/jocn.2010.21585

Marques, C., Moreno, S., Castro, S. L., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: Behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, *19*(9), 1453–1463. https://doi.org/10.1162/jocn.2007.19.9.1453

Miyazaki, K., Rakowski, A., Makomaska, S., Jiang, C., Tsuzaki, M., Oxenham, A. J., Ellis, G., & Lipscomb, S. D. (2018). Absolute pitch and relative pitch in music students in the east and the west: Implications for aural-skills education. *Music Perception*, *36*(2), 135–155. https://doi.org/10.1525/mp.2018.36.2.135

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

Nan, Y., & Friederici, A. D. (2013). Differential roles of right temporal cortex and broca's area in pitch processing: Evidence from music and mandarin. *Human Brain Mapping*, *34*(9), 2045–2054. https://doi.org/10.1002/hbm.22046

Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C. C., Dong, Q., Gabrieli, J. D. E., & Desimone, R. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(28), E6630–E6639. https://doi.org/10.1073/pnas.1808412115

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

Ngo, M. K., Vu, K. P. L., & Strybel, T. Z. (2016). Effects of music and tonal language experience on relative pitch performance. *American Journal of Psychology*, *129*(2), 125–134. https://doi.org/10.5406/amerjpsyc.129.2.0125

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

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

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature*, *392*(6678), 811–814. https://doi.org/10.1038/33918

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods and Findings in Experimental and Clinical Pharmacology*, *24 Suppl D*, 5–12.

Patel, A. D. (2013). Sharing and Nonsharing of Brain Resources for Language and Music. In *Language, Music, and the Brain*. The MIT Press. https://doi.org/10.7551/mitpress/9780262018104.003.0014

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research*, *308*, 98–108.

https://doi.org/10.1016/j.heares.2013.08.011

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

Polat, Z., & Ataş, A. (2014). The investigation of cortical auditory evoked potentials responses in young adults having musical education. *Balkan Medical Journal*, *31*(4), 328–334. https://doi.org/10.5152/balkanmedj.2014.14171

Sanju, H. K., & Kumar, P. (2016). Enhanced auditory evoked potentials in musicians: A review of recent findings. *Journal of Otology*, *11*(2), 63–72. https://doi.org/10.1016/j.joto.2016.04.002

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience*, *5*(7), 688–694. https://doi.org/10.1038/nn871

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341–349. https://doi.org/10.1111/1469-8986.00172.x

Shahin, A., Bosnyak, D. J., Trainor, L. J., & Roberts, L. E. (2003). Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *Journal of Neuroscience*, *23*(13), 5545–5552. https://doi.org/10.1523/jneurosci.23-13-05545.2003

Shao, J., & Zhang, C. (2019). Talker normalization in typical Cantonese-speaking listeners and congenital amusics: Evidence from event-related potentials. *NeuroImage: Clinical*, *23*(April), 101814. https://doi.org/10.1016/j.nicl.2019.101814

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, *10:2465*. https://doi.org/10.1038/s41467-019-10365-z

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (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

Tang, W., Xiong, W., Zhang, Y. xuan, Dong, Q., & Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: Behavioral and neural evidence. *Neuropsychologia*, *91*, 247–253. https://doi.org/10.1016/j.neuropsychologia.2016.08.003

Tremblay, K., Kraus, N., McGee, T., Ponton, C., & Otis, A. B. (2001). Central auditory plasticity: Changes in the N1-P2 complex after speech-sound training. *Ear and Hearing*, *22*(2), 79–90. https://doi.org/10.1097/00003446-200104000-00001

Tsai, C. G., Chou, T. L., & Li, C. W. (2018). Roles of posterior parietal and dorsal premotor cortices in relative pitch processing: Comparing musical intervals to lexical tones. *Neuropsychologia*, *119*(March), 118–127. https://doi.org/10.1016/j.neuropsychologia.2018.07.028

Verkindt, C., Bertrand, O., Thevenet, M., & Pernier, J. (1994). Two auditory components in the 130-230 ms range disclosed by their stimulus frequency dependence. In *NeuroReport* (Vol. 5, Issue 10, pp. 1189–1192). https://doi.org/10.1097/00001756-199406020-00007

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience*, *30*(2), 629–638. https://doi.org/10.1523/JNEUROSCI.2742-09.2010

Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, *38*(4), 654–662. https://doi.org/10.1016/j.wocn.2010.10.001

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., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420–422. https://doi.org/10.1038/nn1872

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

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., & Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Frontiers in Psychology*, *6*(APR), 1–7. https://doi.org/10.3389/fpsyg.2015.00436

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nat Rev Neurosci*, *8*(7), 547–

558. https://doi.org/10.1038/nrn2152

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, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & Wang, W. S. Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*, *124*. https://doi.org/10.1016/j.neuroimage.2015.08.064

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., Wang, X., & Peng, G. (2017). Normalization of lexical tones and nonlinguistic pitch contours: Implications for speech-specific processing mechanism. *The Journal of the Acoustical Society of America*, *141*(1), 38–49. https://doi.org/10.1121/1.4973414

Zhu, J., Chen, X., & Yang, Y. (2021). Effects of Amateur Musical Experience on Categorical Perception of Lexical Tones by Native Chinese Adults: An ERP Study. *Frontiers in Psychology*, *12*(March), 1–17. https://doi.org/10.3389/fpsyg.2021.611189

Supplementary materials

Table S1: Demographic information and music training experience of musician group.

Participant	Gender	Age	Years of Training	Age of Onset	Instrument
1	female	19.08	13	6	Not reported
2	male	21.84	7	14	Piano, Vocal
3	female	18.91	10	6	Piano
4	male	29.68	13	4	Piano, Contrabass
5	male	23.94	16	4	Piano
6	female	33.98	30	3	Piano, Vocal, Chorus, Erhu
7	male	34.03	10	8	Erhu, Piano, Guitar
8	male	31.07	20	11	Piano, Drum
9	female	19.1	13	4	Piano, Flute, Vocal, Chorus
10	female	18.81	9	5	Piano
11	female	30.86	24	6	Huqin, Piano
12	male	33.98	28	6	Not reported
13	female	27.14	8	12	Cello, Double Bass,
14	female	20.2	12	9	Zhongruan, Guitar
15	male	20.6	7	14	Guitar
16	male	19.35	12	6	Erhu
17	female	20.98	14	6	Piano
18	female	20.08	8	13	Erhu
19	male	20.61	7	14	Guitar
20	male	26.74	14	12	Sanxian, Pipa
21	male	28.89	18	10	Violin, Viola
22	female	27.55	19	8	Percussion, Violin, Piano
23	female	31.34	10	9	Piano, Zheng
24	female	19.52	15	3	Piano

Supplementary analysis for the relationship between musicians' music training experience and their behavioral and cortical responses

In Section 3.3 of the main text, a significant correlation between N1 amplitude and IR was found in the musician group in the speech-context condition. To further explore if the musician's music background influences their normalization process at behavioral and cortical levels, linear regression models were fitted to the IR and the amplitude of N1 in the speech contexts with years of musical training and age of musical training onset as predictors (see Table S1).

The results showed that neither years of musical training [β= -0.004, SE = 0.006, t = -0.673, p = 0.508] nor age of musical training onset [β= -0.018, SE = 0.011, t = -1.606, p = 0.123] significantly affected the IR in the speech contexts. Similarly, neither years of musical training [β= -0.003, SE = 0.048, t = -0.067, p = 0.948] nor age of musical training onset [β= -0.015, SE = 0.083, t = -0.177, p = 0.861] significantly affected the N1 amplitude. It is possible that other music-related indices, rather than self-reported years of musical training and age of musical training onset, contributed to the normalization process, which can be an interesting topic for the future studies.

Supplementary methods for the exploratory ERP source analysis

The ERPs for source analysis were obtained with preprocessing steps largely the same as procedures in Section 3.2.1, except that the bad channels were replaced with interpolation before filtering. After preprocessing, the ERPs were averaged across *Pitch Shift*s and *Sound Type*s on the individual level for further processing in the LORETA-KEY software. Modules in LORETA-KEY determined the electrode coordinates with built-in template and calculated a sLORETA transformation matrix with the head model of MNI152 template. The sLORETA transformation matrix converted participants' scalp electric potentials to standardized current source densities of 6,239 voxels in the brain, which were used to determine if bilateral auditory cortices were activated during the N1 and P2 time windows.

The current source density maps were then entered into group-level, voxel-wise randomization tests with 5000 permutations. The voxel-wise tests compared the current source densities of N1 and P2 time-windows with baseline (-100 to 0 ms) densities, based on statistical nonparametric mapping (SnPM) on the collapsed group of musicians and nonmusicians. Voxels with significant results (for corrected $p < 0.01$) were labeled with specific brain regions, Brodmann Areas (BAs), and MNI coordinates. Strong source activations were indeed observed in the bilateral auditory cortices within the N1 and P2 time windows among a widely distributed network.

Table S2: Summary of time points showing significant main effects and interactions within the N1 and P2 time windows.

	N1 (70-175 ms)	P2 (175-280 ms)
***Group* main effect**		
left BA 42	216-218	-
***Sound Type* main effect**		
left BA 41	70-72, 112-132, 137-149, 168-175	176-193, 204-210, 227-242
right BA 41	78-88, 116-127	183-219
left BA 42	70-78, 110-117, 127-131,	178-186, 226-245,
right BA 42	79-89, 114-137,	185-191, 203-214, 256-260,
left BA 22	71-79, 94-101, 105-117, 120-138	226 -244, 279-280
right BA 22	79-89, 111-140	245-252, 267-270
Interaction between *Group* and *Sound Type*		
left BA 41	129-133	-
right BA 41	-	206-213
left BA 42	-	-

right BA 42	-	188-193, 210-218, 237-242,
left BA 22	-	-
right BA 22	-	211-223, 236-245