Music and speech are distinct in lexical tone normalization processing

Ran Tao
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
rantao@polyu.edu.hk

Gang Peng
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
gpeng@polyu.edu.hk

Abstract

This paper presents the results of a Cantonese lexical tone normalization experiment. Two non-linguistic (i.e., music and nonspeech) context conditions were used in addition to a speech context condition to elicit listeners' normalization of level tones following contexts with fundamental frequency alterations. Participants showed clear tone normalization in the speech context condition. Participants' normalization performances in the non-linguistic conditions were not significant but comparable to each other. The findings indicated that the non-linguistic music context is not sufficient to support lexical tone normalization.

1 Introduction

In daily life, people speak with huge inter- and intra-talker variance (Peng et al., 2012). The task of normalizing variable vocal streams to identical linguistic codes seems challenging. When it comes to the tone languages such as Cantonese, the situation is even more complicated (Wong and Diehl, 2003). Cantonese has three level tones conveying distinct meanings when superimposing to identical syllables, leaving the pitch height the only clue for listeners' successful judgments. The level tones' pitch heights frequently overlap with each other among speakers and within a speaker, such as in different emotional states. However, the task of identifying words in a vocal stream, i.e., lexical tone normalization, can be accomplished by typical Cantonese without conscious awareness (Wong and Diehl, 2003; Francis et al., 2006; Zhang et al., 2016; Zhang et al., 2017).

Research on lexical tone normalization has revealed the significant role of immediate context which provides a reference to normalize the target pitch height. Researchers debated on whether such processes depend on a general-purpose auditory mechanism or a speech-specific mechanism. There are contradictory findings in the literature, providing evidence to support both sides. Huang and Holt (2009; 2011) found both nonspeech and speech context is sufficient to facilitate the listener's tone normalization, with a restricted effect in the nonspeech context. However, other researchers failed to replicate the facilitation effect of nonspeech context, and mainly support the speech specific mechanism (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2016; Zhang et al., 2017).

While the nonspeech context can replicate the pitch pattern in speech context, they differ in aspects other than whether or not containing a linguistic meaning. The nonspeech context is novel to listeners because such stimuli are rare in daily lives. It is plausible that listeners cannot take advantage of the immediate context of nonspeech as they are not familiar with such auditory patterns.

Music, as non-linguistic stimuli, frequently appears in daily lives. More interestingly, previous research suggested that music could benefit linguistic abilities. Nan et al., (2018) found that piano training can enhance the neural processing of pitch and thus improves speech perception in Mandarin-speaking children. Other researchers have found that music training can benefit the lexical tone perception of

non-tone language speakers (Wayland et al., 2010), suggesting that lexical tone and music may share a similar processing mechanism. However, a closer study on Cantonese speakers failed to find such a facilitating effect (Mok and Zuo, 2012). Thus, it is intriguing whether music can serve as an efficient context for lexical tone normalization. An evident tone normalization following the music context may support the general-purpose auditory mechanism. If the music context does not evoke tone normalization, the result may favor the speech-specific mechanism.

In addition to the investigation into the musical context, it is also interesting to compare the function of different non-linguistic contexts. One limitation in previous studies is that researchers tended to include only one contrastive condition to the speech context condition (Zhang et al., 2013; Zhang et al., 2017), which may limit the interpretation and generalization of the findings.

In this study, we inspect the context's role in native Cantonese speakers' lexical tone normalization by including two non-linguistic conditions to compare with the speech context condition, e.g., a nonspeech context condition and a music context condition. Our primary question is whether music context can suffice the normalization of Cantonese level tones. We are also interested in the profile of the contextual effect between the two non-linguistic contexts.

2 Methodology

We adopted similar stimuli and experiment design as our teams' previous studies. The stimuli preparation and experiment procedure are reported briefly as follows, but see (Zhang et al., 2013; Zhang et al., 2017) for detailed descriptions.

Participants

28 native Cantonese speakers (15 female, mean age = 21.9 yrs, SD = 2.95) were recruited in the current study, all without hearing impairment. Two female participants were left-handed. None of the participants worked as professional musicians. All participants signed written consent before the experiment. Experiment protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

Stimuli

Stimuli of this study consisted of contexts and targets in the four context conditions, e.g., no context, speech context, and two non-linguistic contexts. speech context and all targets were produced by four native Cantonese speakers, who were a female speaker with a high pitch range, a female speaker with a low pitch range, a male speaker with a high pitch range, and a male speaker with a low pitch range (coded as FH, FL, MH, and ML respectively in the following text). Speech context was a four-syllable meaningful sentence, i.e., 呢個字係 (/li55 ko33 tsi22 hɐi22/, "This word is meaning"). After recording the natural production of the sentence from the four talkers, the F0 trajectories of the sentences were then lowered and raised three semitones. In sum, three kinds of speech contexts were formed: an F0 lowered context, an F0 unshifted context, and an F0 raised context. All targets from three context conditions were the natural production of the Chinese character 意 (e.g., /ji33/ mid-level tone, "meaning").

The nonspeech contexts were produced by applying the F0 trajectory and intensity profile from speech contexts to triangle waves. The music contexts were piano notes that had the closest pitch height to each of the syllables in the speech context. All the targets were adjusted to 55dB in intensity and 450 ms in duration. All speech contexts were adjusted to 55 dB in intensity. The non-linguistic contexts were adjusted to 65dB in intensity to match the hearing loudness of speech contexts. The duration of non-linguistic contexts were the same as their corresponding speech contexts.

There were also fillers in the tasks. In the speech context condition, the filling context was a four-syllable sentence, i.e., 我以家讀 (/ŋo23 ji21 ka55 tuk2/, "Now I will read"). All the targets were Chinese characters 醫 (e.g., /ji55/ high level tone, "a doctor") or 意. The nonspeech contexts and music contexts of the fillers were produced with the same procedure above.

Experiment Procedure

Participants attended two practice blocks and four experiment blocks. The four experiment blocks con-

sisted of four context conditions respectively, e.g., the no context condition, the speech context condition, the nonspeech context condition, and the music context condition. In the no context condition, the participants heard the targets without preceding contexts. The no context condition is coded as isolated in the following text.

The task was a word identification task that asked participants to make a judgment on the target syllable after listening to the preceding context attentively. In each experiment trial, participants first heard a context, and after a jittering silence (range: 300 - 500 ms), a target syllable was presented. In the isolated condition, Participants heard the target without a context. Participants then made a judgment on the target syllable, whether it is 醫, 意, or 二 (e.g., /ji22/low-level tone, "two") by pressing corresponding keys on the keyboard when they saw a cue on the screen. The cue was presented 800 ms after the onset of the target. Such a manipulation of response time widow was because the experiment was part of a large project in which participants were tested with EEG recording. This restriction minimized the artifacts of EEG signal due to muscle movement. In this kind of setting, reaction times were not a meaningful index of participants' psycholinguistic properties and thus not analyzed in this study. We focused on the judgments of the targets from the participants, which was also the standard procedure in previous research.

The isolated condition consisted of 16 repetitions of each target. The three context conditions each consisted of nine repetitions of three F0 shifts of four talkers, making 27 trials for each talker in each context condition. The four experiment blocks were counterbalanced to prevent order effects.

Analysis

Following previous research (Wong and Diehl, 2003; Zhang et al., 2012; Zhang et al., 2017), perceptual height (PH) and identification rate (IR) were analyzed to investigate participants' lexical tone normalization performance. For the PH analysis, a response of high-level tone was coded as 6, middle-level tone as 3, and low-level tone as 1. The mean Perceptual Height close to 6 indicated that participants generally perceived the targets as high-level tones. In a lowered F0 condition, this could serve as an evidence of evoking participants' tone normalization. Perceptual height close to 1 indicated that participants generally perceived the targets as low-level tone. In a raised F0 condition, this could serve as an evidence of evoking participants' tone normalization. The identification rate was the percentage of expected responses in each condition. The expected responses were the judgments that participants should make when successfully evoked tone normalization, e.g., low-level tone response in the raised F0 condition, middle-level tone response in the unshifted F0 condition, and high-level tone response in the lowered F0 condition.

We conducted one-way repeated measures ANOVAs on PH and IR, with Context as the main factor. Then, we conducted three-way repeated measures ANOVAs on PH and IR. The isolated condition was excluded from this analysis because it did not match the design matrix of other context conditions, e.g., there was no context and thus no F0 Shift manipulations. Three main factors were Context (music, nonspeech, speech), F0 shift (lowered, unshifted, raised), and talker (FH, FL, MH, ML). Greenhouse–Geisser correction was applied when the data violated the Sphericity hypothesis. Tukey method for comparing families of multiple estimates were applied for necessary post-hoc analysis.

3 Results

Perceptual Height

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(2.78, 75.18) = 4.91, p = 0.004, ges = 0.054$), indicating that the PH is influenced by the targets' preceding contexts (see Figure 1). Post-hoc analysis revealed that the speech context condition (3.23) had higher PH than nonspeech (2.81) and music (2.80) conditions ($ps < 0.01$), but not isolated (2.94) condition ($p = 0.117$). All other comparisons were not significant. Next, we report the results of the three-way repeated measures ANOVA on PH.

There was a significant main effect of Context ($F(1.83, 49.44) = 7.20, p = 0.002, ges = 0.031$), replicating that participants perceived the same set of targets as different lexical tones across three context conditions. Post-hoc analysis showed that the

PH is highest in the speech context (3.23, $ps < 0.01$ when compared with the two non-linguistic contexts) and showed no difference between music context (2.80) and nonspeech context (2.81). The main effect of F0 Shift was also significant ($F(1.13, 30.50) = 162.77, p < 0.001, ges = 0.173$), indicating participants perceived targets as different lexical tones with contexts' F0 manipulated. As expected, the PH was highest in the lowered F0 conditions (3.62) and lowest in the raised F0 conditions (2.34). The PH in the unshifted F0 condition was 2.87 and the comparisons among the three PHs were all significant (all $ps < 0.001$) which indicated participants' tone normalization evoked in general. The main effect of talker was also significant ($F(2.04, 55.14) = 8.80, p < 0.001, ges = 0.100$). Participants perceived targets produced by FH with the highest (3.41) picth height, significantly higher than that of FL (2.35), MH (3.00), and ML (3.01). There was no difference between FL, MH, and ML.

The interaction between Context and Shift was significant ($F(1.33, 35.92) = 142.88, p < 0.001, ges = 0.286$). The large suggests the participants perception of the same set of targets was strongly influenced by the preceding context with different F0 Shift manipulation (see Figure 1). Post-hoc analysis revealed that only speech context elicited successful lexical tone normalization: the PH was significantly different among the three F0 Shift conditions (lowered: 5.21, unshifted: 3.02, raised: 1.46, all $ps < 0.001$). The PHs were not different among the three F0 Shift conditions of the two non-linguistic contexts (PH range: 2.77 - 2.85, all $ps > 0.7$).

The interaction between Context and talker was also significant ($F(2.96, 80.03) = 11.59, p < 0.001, ges = 0.046$), indicating participants' perception of tones was modulated by the talkers in the three Contexts. However, such modulation was not significant in speech contexts. Participants perceived the four talkers as the same PH (all $ps > 0.7$). In the two non-linguistic contexts, the targets produced by FL (music: 1.95, nonspeech: 1.92) were always perceived lowest (all $ps < 0.01$). No other comparisons were significant except that the targets produced by FH (3.46) were perceived higher than that of ML (2.77) in the music context ($p < 0.05$).

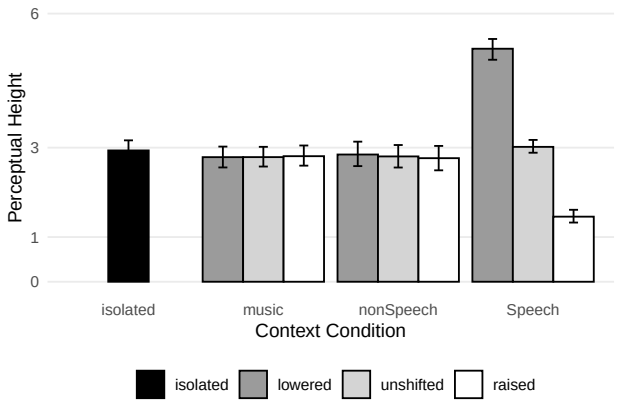

Figure 1: Average perceptual heights (PH) of the word identification task. Black bar respresents the PH in isolated condition. Gray bars represent the PH in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

Identification Rate

For the one-way repeated measures ANOVA, there was a significant main effect of Context ($F(1.75, 47.37) = 95.47, p < 0.001, ges = 0.705$, see Figure 2). The large effect size indicated that participants' IR is strongly influenced by the targets' preceding contexts. Post-hoc analysis revealed that the speech context condition (78.4%) had higher IR than nonspeech (34.1%), music (33.3%) and isolated (45.2%) conditions (all $ps < 0.01$). The isolated condition had higher IR than the two non-linguistic context conditions (all $ps < 0.01$). There was no IR difference between the two non-linguistic conditions. Next, we report the results of the three-way repeated measures ANOVA on IR.

There was a significant main effect of Context ($F(1.10, 29.75) = 135.93, p < 0.001, ges = 0.359$). This replicated that participants' identification rate is strongly influenced by targets' attaching context. The post-hoc analysis also revealed the same result as in the one-way repeat measures ANOVA. The main effect of Shift ($F(1.52, 41.06) = 8.83, p = 0.002, ges = 0.058$) was also significant, indicating participants' performance was different under the three F0 Shift conditions. Specifically, the lowered F0 condition (38.8%) yielded lower IR than unshifted (54.0%) and raised (53.1%) F0 conditions. There was no

IR difference between the unshifted and raised F0 conditions. However, the main effect of talker was not significant ($F(2.84, 76.79) = 1.37, p = 0.259$). Although the PHs was different as percieved from targets produced by the four talkers, this perceptual difference did not influence participants' judgment: their performance was overall the same on the four talkers.

There were two significant two-way interactions. The interaction between Context and Shift was significant ($F(3.08, 83.09) = 8.55, p < 0.001, ges = 0.035$), indicating participants' performance was modulated by F0 shifts in different contexts. Post-hoc analysis showed that there was no difference in IRs among F0 Shifts of speech context (range: $75.7 - 81.7\%$, all comparisons' $ps > 0.79$). However, the patterns of IR were highly similar in music and nonspeech conditions. In both non-linguistic context conditions, lowered F0 context (music: (17.4%), nonspeech: (21.1%)) elicited lower identification rate than unshifted ((45.8%), (40.4%)) and raised ((36.8%), (40.7%)) F0 context (all $ps < 0.001$, see Figure 2).

The interaction between Shift and talker was also significant ($F(3.18, 85.93) = 6.22, p = 0.002, ges = 0.055$), indicating participants' performance was also modulated by F0 Shift in different talkers. Interestingly, for targets produced by the male talkers, the IRs showed no difference among three F0 shifts (range: $40.3 - 54.9\%$, all $ps > 0.1$). For targets produced by female talkers, The IR was lower in the lowered F0 condition compared with unshifted F0 condition (in FL: 27.7% and $50.9\%, p < 0.01$; in FH: 44.8% and $61.6\%, p < 0.05$). In addition, the IRs were higher in the raised F0 condition (64.6%) than lowered F0 condition in FL ($p < 0.001$), and higher in the raised F0 condition (40.2%) than unshifted F0 condition in FH ($p < 0.05$).

Finally, the three-way interaction among Context, Shift and talker was significant ($F(6.74, 182.05) = 6.22, p < 0.001, ges = 0.039$). To decode this interaction, we conducted two-way ANOVAs on the IRs of targets produced by each of the four talkers (see Figure 3). The Context main effect were significant across talkers (all $ps < 0.001$). Post-hoc analysis revealed that IRs of the speech contexts were always highest (all $ps < 0.001$), with no difference between

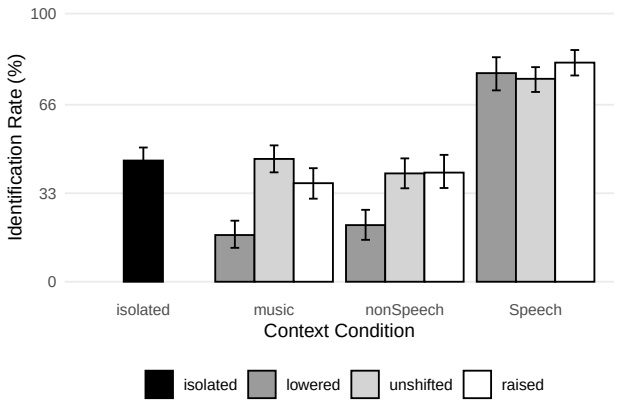

Figure 2: Identification rates (IR) of the word identification task. Black bar respresents the IR in isolated condition. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of the other three Context conditions. Error bars represent 95% confidence intervals.

the two non-linguistic contexts (all $ps > 0.8$). The Shift main effect was only significant in the analysis of female talkers (all $ps < 0.05$) and the lowered F0 condition always produced the lowest IRs. The interaction between Context and Shift was significant in the analysis of FH, FL, and ML (all $ps < 0.001$). The post-hoc analysis revealed that in music conditions, lowered F0 contexts had consistently low IRs (all $ps < 0.01$), while in nonspeech conditions, lowered F0 context had low IRs only in the analysis of FL ($ps < 0.001$).

4 Discussion

In this study, we included two non-linguistic context conditions in addition to a speech context condition to investigate native Cantonese's lexical tone normalization. Specifically, we were interested in participants' performance in the music context condition. The results revealed successful tone normalization only in speech contexts, suggesting that the music and language abilities were not mutually transferrable in the scenario of lexical tone normalization. Besides, the two non-linguistic conditions' performance were comparable. In general, the finding favors the speech-specific mechanism in lexcial tone normalization.

Our results on the perceptual heights and identification rates under different conditions largely

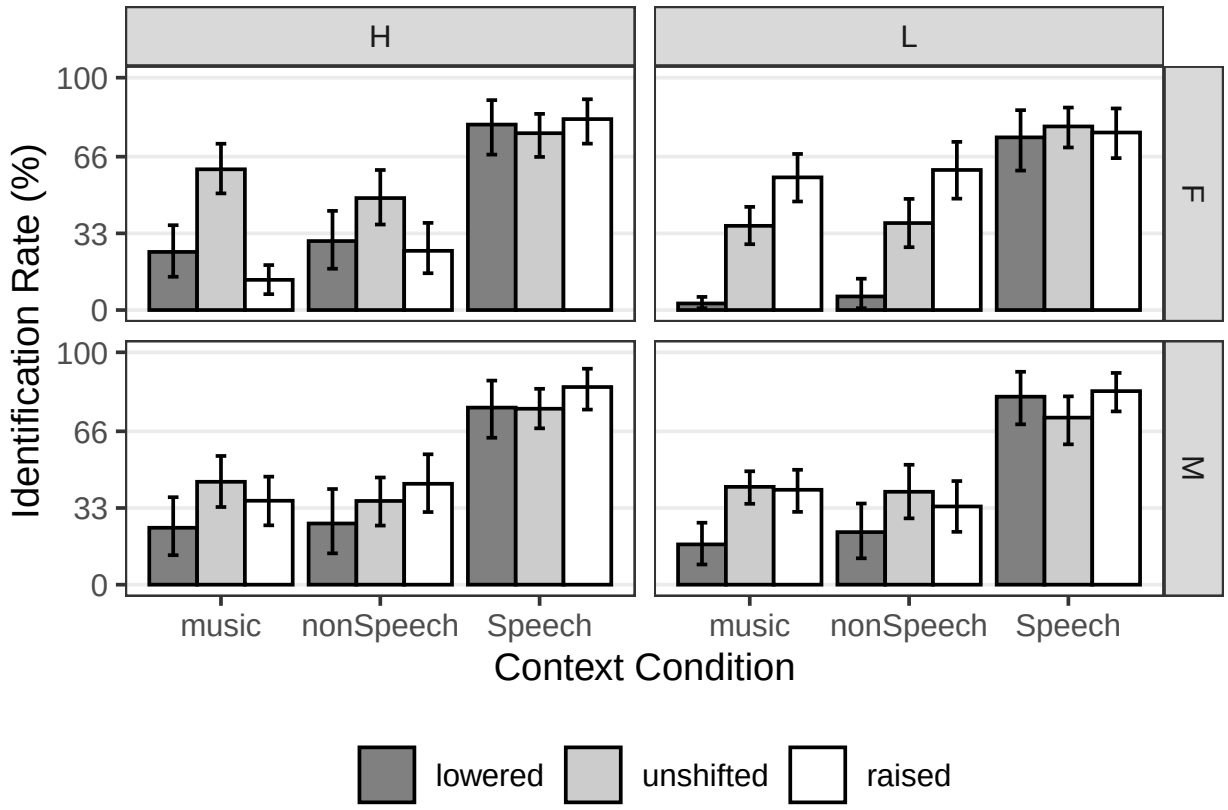

Figure 3: Identification rates (IR) of the targets produced by four talkers in the word identification task. Gray bars represent the IR in the lowered, unshifted and raised F0 Shift conditions of three Context conditions. Error bars represent 95% confidence intervals. Rows of the panels are divided by the talkers gender: Femal and Male; columns of the panels are divided by the talkers' pitch range: High and Low.

replicated our previous studies (Zhang et al., 2012; Zhang et al., 2013; Zhang et al., 2017). In accordance with our previous research, participants only showed evident lexical tone normalization effect under the speech context conditions. The perceptual height under music and nonspeech conditions were close to 3, indicating that participants perceived the middle-level tone as it was in these non-linguistic conditions.

The participants' perceived perceptual heights were influenced by the talker who produced the targets under the non-linguistic conditions. In both music and nonspeech context conditions, targets produced by FL was lowest. Targets produced by FH were perceived highest, however, the perceptual height differences between targets produced by FH and two male talkers were not significant. In contrast, the identification rates among the targets produced by four talkers were not different from each other. This suggests that the perceived height does not influence the ability to correctly normalize lexical tones.

Interestingly, the identification rate was significantly lower in the lowered F0 conditions, compared with unshifted and raised F0 conditions. Such a pattern was consistent across the non-linguistic contexts but not evident in the speech context condition. In both music and nonspeech context conditions, the identification rates in the lowered F0 contexts were significantly lower than unshifted and raised F0 contexts, while the identification rate under unshifted and raised F0 contexts showed no difference.

Here we offer two possible explanations. Firstly, the normalization of level tones under the lowered F0 context conditions was probably harder than that of unshifted and raised. The pitch difference

between high-level tone and middle-level tone is greater than that between low-level tone and middle-level tone (Wong and Diehl, 2003). To elicit successful tone normalization, the F0 shift has a higher requirement for high-level tones, i.e., in the lowered F0 condition. In our manipulation, the lowered and raised F0 condition both altered 3 semitones. While this manipulation was sufficient to elicit tone normalization in speech contexts, it might bring bias in the non-linguistics context, resulting in a unbalanced IR in the three F0 alternation conditions. A future study with an unbalanced alteration (larger F0 shift in lowered F0 condition) may explicitly examine this explanation.

However, such a bias was not consistent across talkers: the Shift main effect was only significant in the analysis of female talkers. Thus, a second explanation is that the low IR in the lowered F0 condition was influenced by the response tendency driven by the natural pitch range of different talkers. Indeed, the interaction between Context and Shift had variable patterns among the analysis of four talkers.

Both of the two explanations seemed hard to be compatible with the speech-specific mechanism which would predict a null effect of F0 Shift in non-linguisitic conditions. In future studies, a closer examination on the low IR in the lowered F0 Shift conditions of non-linguistic contexts will be meaningful to unravel the possible role of general auditory mechanism in lexcial tone normalization. For example, the F0 alterations can be exagerated to increase the effect of non-linguistic contexts.

5 Conclusion

We revisited the lexical tone normalization process by examining the possible function of music as an untested non-linguistic context. Participants successfully normalized target lexical tones following the speech context but not the music or nonspeech context. The behaivoral pattern in the two non-linguistic conditions were similar. Our findings mainly supported the idea that lexical tone normalization relies on the speech-specific mechanism.

Acknowledgments

This work was partially supported by the General Research Fund (No. 15607518) from Research Grants Council (RGC) of Hong Kong and a PolyU internal fund (PolyU 156002/17H).

References

Alexander L. Francis, Valter Ciocca, Natalie King Yu Wong, Wilson Ho Yin Leung, and Phoebe Cheuk Yan Chu. 2006. Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America*, 119(3):1712–1726.

Jingyuan Huang and Lori L. Holt. 2009. General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, 125(6):3983–3994.

Jingyuan Huang and Lori L. Holt. 2011. Evidence for the central origin of lexical tone normalization (l). *The Journal of the Acoustical Society of America*, 129(3):1145–1148.

P. K. Peggy Mok and Donghui Zuo. 2012. The separation between music and speech: evidence from the perception of cantonese tones. *The Journal of the Acoustical Society of America*, 132(4):2711–2720.

Yun Nan, Li Liu, Eveline Geiser, Hua Shu, Chen Chen Gong, Qi Dong, John D. E. Gabrieli, and Robert Desimone. 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.

Gang Peng, Caicai Zhang, Hong-Ying Zheng, James W. Minett, and William S.-Y. Wang. 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.

Ratree Wayland, Elizabeth Herrera, and Edith Kaan. 2010. Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, 38(4):654–662.

Patrick C. M. Wong and Randy L. Diehl. 2003. Perceptual normalization for inter- and intratalker variation in cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2):413–421.

Caicai Zhang, Gang Peng, and William S-Y Wang. 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.

Caicai Zhang, Gang Peng, and William S-Y Wang. 2013. Achieving constancy in spoken word identification: time course of talker normalization. *Brain and language*, 126(2):193–202.

Caicai Zhang, Kenneth R. Pugh, W. Einar Mencl, Peter J. Molfese, Stephen J. Frost, James S. Magnuson, Gang Peng, and William S-Y Wang. 2016. Functionally integrated neural processing of linguistic and talker information: An event-related fmri and erp study. *NeuroImage*, 124(Pt A):536–549.

Kaile Zhang, Xiao Wang, and Gang Peng. 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.