Musicians show enhanced perception, but not production, of native lexical tones

Article

Accepted Version

Ong, J. H., Wong, P. C. M. and Liu, F. ORCID: https://orcid.org/0000-0002-7776-0222 (2020) Musicians show enhanced perception, but not production, of native lexical tones. Journal of the Acoustical Society of America, 148 (3443). ISSN 0001-4966 doi: https://doi.org/10.1121/10.0002776 Available at http://centaur.reading.ac.uk/93814/

To link to this article DOI: http://dx.doi.org/10.1121/10.0002776

Publisher: Acoustical Society of America

Musicians show enhanced perception, but not production, of native lexical tones

Jia Hoong Ong

School of Psychology and Clinical Language Sciences, University of Reading, Earley Gate, Reading RG6 6AL, United Kingdom. Email: jiahoong.ong@reading.ac.uk

Patrick C. M. Wong

Department of Linguistics and Modern Languages and Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong, China. Email: p.wong@cuhk.edu.hk

Fang Liu[a)]

School of Psychology and Clinical Language Sciences, University of Reading, Earley Gate, Reading RG6 6AL, United Kingdom. Email: f.liu@reading.ac.uk

[a)] Corresponding author

ABSTRACT

Many studies have reported a musical advantage in perceiving lexical tones among non-native listeners, but it is unclear whether this advantage also applies to native listeners, who are likely to show ceiling-like performance and thus mask any potential musical advantage. The ongoing tone merging phenomenon in Hong Kong Cantonese provides a unique opportunity to investigate this as merging tone pairs are reported to be difficult to differentiate even among native listeners. In the present study, native Cantonese musicians and non-musicians were compared on their discrimination and identification of merging Cantonese tone pairs to determine whether a musical advantage in their perception will be observed, and if so, whether this is seen on the phonetic and/or phonological level. The tonal space of their lexical tone production was also compared. Results indicated that the musicians outperformed the non-musicians on the two perceptual tasks, as indexed by their higher accuracy and faster reaction time, particularly on the most difficult tone pair. In the production task, however, there was no group difference in various indices of their tonal space. Taken together, musical experience appears to facilitate native listeners' perception, but not production, of lexical tones, which partially supports a music-to-language transfer effect.

Keywords: lexical tone; musicians; perception; production

I. INTRODUCTION

Music and spoken language (speech) share many commonalities including the use of similar acoustic cues (e.g., pitch, duration, loudness, etc.) as well as the same mechanisms and resources to process these cues (Besson et al., 2011; Patel, 2008). One line of evidence to support this is cross-domain transfer, the phenomenon that expertise or ineptitude in one domain (e.g., music) may lead to a facilitatory or inhibitory effect in the other (e.g., speech; Alexander, Wong, & Bradlow, 2005; Liu, Patel, Fourcin, & Stewart, 2010; Tillmann, 2014).

A large part of cross-domain transfer research has focused on the transfer between musicality and linguistic pitch. For example, relative to non-musicians, English-speaking musicians were more accurate in perceiving emotions based on speech prosody (Thompson et al., 2004) and in discriminating and encoding lexical tones (Burnham et al., 2014; P. C. M. Wong et al., 2007), the building blocks of tone languages in which pitch, along with consonants and vowels, distinguishes lexical meaning. In addition to large pitch changes such as perceiving differences across lexical tone categories, musicians also have increased sensitivity to fine-grained linguistic pitch than non-musicians. For instance, relative to non-musicians, musicians were better able to detect subtle incongruous prosodic patterns in speech (Magne et al., 2006; Marques et al., 2007) and to discriminate pairs of lexical tones with small interval differences, such as those from a synthesized tone continuum (Zhao & Kuhl, 2015). Tone language listeners, too, exhibited cross-domain transfer of pitch, when perceiving non-native lexical tones (Cooper & Wang, 2010). On the other end of the musicality continuum, listeners with congenital amusia (or tone deafness), who have difficulties perceiving and producing musical pitch accurately (Ayotte et al., 2002), tend to be poorer at differentiating the prosodic patterns of statements and questions (Hutchins et al., 2010; F. Liu et al., 2010) as well as lexical tones (F. Liu et al., 2016; Tillmann et al., 2011). These examples are taken as evidence of a shared processing mechanism of pitch; though the accounts differ in the underlying source of transfer (Asaridou & McQueen, 2013). For example, transfer effects seen among musicians may be due to the enhancement of general auditory skills from extensive musical training (Kraus & Chandrasekaran, 2010) and/or the heightened attention to particular acoustic cues relevant to listeners' experience (Ong et al., 2016). On the other hand, deficits in music and speech processing in congenital amusia are likely caused by a domain-general pitch processing impairment (Vuvan et al., 2015).

While many studies have shown transfer effects with non-native stimuli, it is unclear if this would similarly be observed among listeners perceiving native stimuli. Though it is rarely investigated directly, some insight can be gained from previous studies. When explicitly ignoring the effect of musicianship, even though Mandarin listeners had larger difference limens for frequency (DLFs, or the threshold to discriminate two frequencies) than English listeners (Stagray & Downs, 1993), they nonetheless showed stronger categorical perception to Mandarin tones than English listeners (Yisheng Xu et al., 2006). While not explicitly mentioned that participants varied in their musical experience, Taiwanese Mandarin listeners also showed stronger categorical perception to Taiwanese Mandarin tones to French listeners (Hallé et al., 2004). Thus, based on the findings across these studies, it seems that tone language experience, rather than musical experience, afforded tone language listeners the advantage in perceiving native stimuli at least in comparison to non-tone language listeners. It remains unclear if musical experience will provide tone language listeners with any additional advantage. A direct comparison on the effect of musicianship among tone language listeners, particularly among native listeners, is rarely studied, presumably because ceiling-like performance are to be expected and thus mask any cross-domain transfer (Lee & Lee, 2010; Maggu, Wong, et al., 2018).

The ongoing tone merging phenomenon in Hong Kong Cantonese may circumvent this issue and provide a unique opportunity to investigate this (Maggu et al., 2016). Standard

Cantonese has six distinct tones, three of which are level tones (high-level Tone 55 (T55), mid-level Tone 33 (T33), and low-level Tone 22 (T22)) and three dynamic tones (high rising Tone 25 (T25), low falling Tone 21 (T21), and low rising Tone 23 (T23)). Certain tone pairs are said to be in the process of merging, presumably due to the acoustic similarity between them, language contact, and the growing influence of Mandarin in Hong Kong (Mok et al., 2013; Mok & Zuo, 2012). Three such pairs identified to be merging are (i) T25 and T23; (ii) T33 and T22; and (iii) T21 and T22 (Fung & Lee, 2019). These 'tone mergers' are said to be difficult for some native Cantonese speakers to differentiate in perception and production, though there are large individual variations in its manifestation. For example, some have difficulty with just perceiving or producing the tone mergers in a distinctive way and others have difficulty with both (Fung & Lee, 2019). The mechanism may also be different for different individuals: for example, some merge the two rising tones as either T25 or T23, and others merge the two as an approximate between them (Bauer et al., 2003; Fung et al., 2011; Kei et al., 2002). Most studies on tone merger have only looked at the younger population and we are unaware of any that have systematically compared the demographic details of those that merge tones ('tone mergerers') and those that do not (but see Fung & Wong (2010) on some preliminary evidence of younger adults showing less accurate T25 production than older adults). In the present study, we used tone mergers as a tool to investigate whether musical training may enhance native listeners' perception and production, in order to elucidate cross-domain transfer effects among native listeners.

A previous study examined this question by comparing native English and native Cantonese musicians and non-musicians on discriminating the merging tone pairs in both speech and non-speech contexts (Mok & Zuo, 2012). The native speakers were also compared on the tonal space of their lexical tone production. Since the tones in the merging tone pairs are said to be produced similarly, the assumption, then, is that a larger tonal space index for each merging tone pair would suggest that the two tones in that pair were more differentiated in their production than a smaller tonal space index. The authors found an effect of musicianship only among non-tone language listeners, suggesting that musical experience has little influence on native listeners' perception and production of lexical tones (i.e., similar in their discrimination performance and tonal space index, respectively). While the paper has shed light on this topic, several issues need to be addressed. Firstly, the study only had a discrimination task to index perception, which relies on sensitivity to lower-level acoustic/phonetic cues, and so it remains to be seen whether differences may be observed for higher-level perceptual tasks such as an identification task, which is more sensitive to phonological processing. Secondly, the scores in the previous study were still quite high (native Cantonese musicians and non-musicians had a group mean of approximately 98%), which suggests that the tone language listeners may be performing at ceiling and therefore mask any group differences. Finally, there were only approximately 10 participants in each comparison group, and so the study may be underpowered.

The present study addresses these issues directly to investigate whether musical experience may have an effect in the perception and production of native lexical tones that are difficult to differentiate (i.e., merging tones). Specifically, we extended Mok and Zuo (2012) by comparing a larger group of native Cantonese musicians and non-musicians (*n* = 26 in each group) on three tasks: (i) discrimination of the merging tone pairs in speech and non-speech contexts, given that differences in performance have been observed depending on the context (Burnham et al., 2014); (ii) identification of the merging tone pairs; and (iii) lexical tone production. Given previous findings, we hypothesized that musicians would outperform non-musicians in their perception and production of these merging tone pairs.

II. METHOD

A. Participants

Participants were 52 native Hong Kong Cantonese speakers recruited using advertisements through mass mail services at the Chinese University of Hong Kong. Half were musicians (defined in the present study as having at least six years of formal extracurricular musical training in the present study; 20 females and 6 males, $M_{age} = 23.65$, $SD_{age} = 6.24$; $M_{musical\ training} = 11.12$, $SD_{musical\ training} = 3.63$) whereas the other half were non-musicians (defined as having at most two years of formal extracurricular musical training in the present study; 18 females and 8 males, $M_{age}= 23.42$, $SD_{age}= 6.49$; $M_{musical\ training} = 0.46$, $SD_{musical\ training} = 0.81$). The two groups did not differ in their age ($t(50) = 0.13$, $p = .897$) nor gender distribution ($\chi^2(1) = 0.10$, $p = .765$). We conducted a power analysis using G*Power (Faul et al., 2007) and determined that our sample size is above that required (i.e., $n = 18$ per group) to achieve at least 80% power with alpha = .05 to detect a significant interaction between Tone and Group (of a small-to-medium effect size, f= 0.2) in an ANOVA. All had normal hearing, defined as pure-tone thresholds of 25 dB or less on each ear for frequencies 0.5, 1, 2, and 4 kHz. Participants gave their written informed consent prior to participating. The Institutional Review Board of Northwestern University and The Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee approved the study protocol.

B. Stimuli and Tasks

1. Discrimination task

We used speech and non-speech stimuli of four tone pairs, taken from previous studies (F. Liu et al., 2016; A. M. Y. Wong et al., 2009), in the discrimination task. Three of the tone pairs were the merging tone pairs (T21-T22, T22-T33, and T23-T25) whereas the fourth acted as a control tone pair (T25-T55). Within each tone pair, the tones were carried by the same syllable, which resulted in a minimal pair of real Cantonese words (/min21/ 綿 'cotton' - /min22/ 麵 'noodle'; /bei22/ 鼻 'nose' - /bei33/ 臂 'arm'; /jyu23/ 雨 'rain'- /jyu25/ 鱼, 'fish'; and /tong25/ 糖 'candy' - /tong55/ 湯 'soup'). The speech stimuli were produced in 2004 by a male native Cantonese speaker in a sound-attenuated room. The monosyllabic words were produced in a carrier phrase /ŋɔ23 wui33 tuk2 __ pei35 nei23 t^{h}ɛŋ55/ ("I will read __ for you to listen"), which were later extracted from the carrier. Each stimulus was produced five times in a random order, and the best token for each monosyllabic word was chosen by three native Cantonese speakers with four years of phonetic training. To create the non-speech tone pairs, the F0 values were first extracted from the speech stimuli, which were then used to be synthesized as hums using the pulse-pitch option on Praat. The resulting hummed sounds were then low-pass filtered at 1900 Hz. The amplitude contours from the original speech sounds were extracted and applied to the hummed sounds. Each pair was normalized for duration and amplitude, and so the only difference between the tones in each pair is their F0, fundamental frequency (see Figure 1 for the pitch contours of the stimuli used).

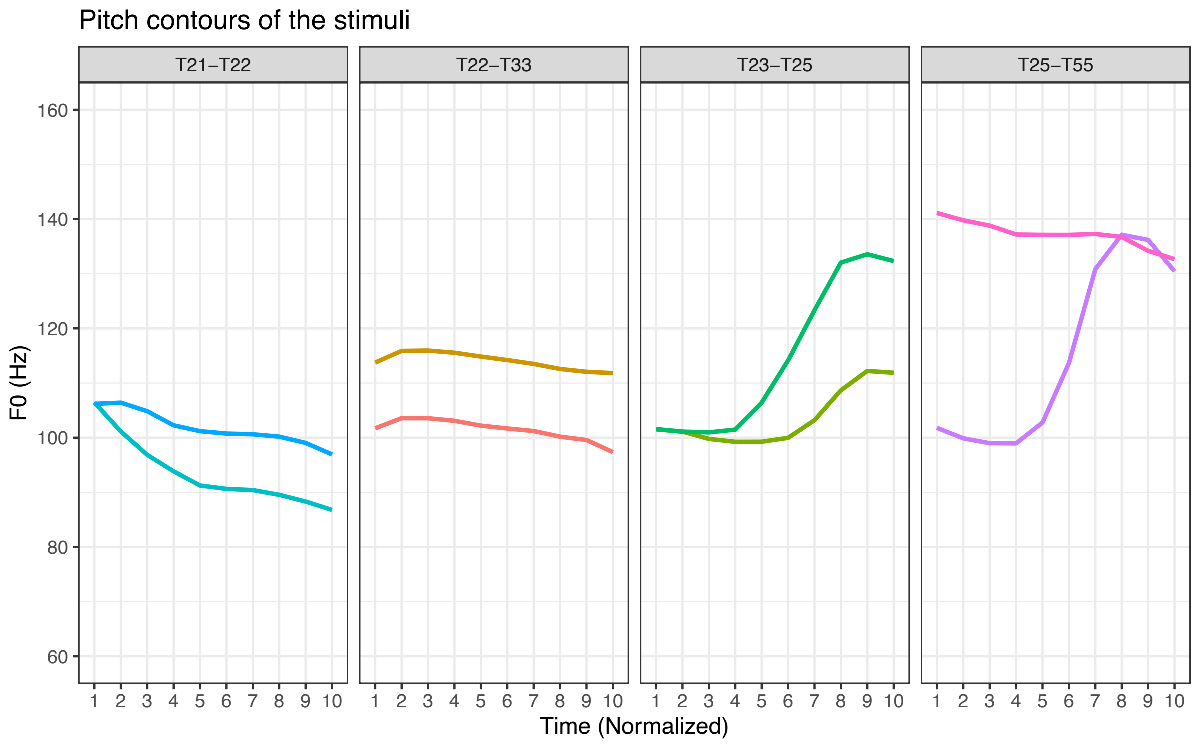

Figure 1. Time-normalized F0 contours of the stimuli used.

Participants completed an AX discrimination task, in which they had to indicate using a keyboard response whether the tone pair presented were the same or different. The task was presented in blocks by Stimuli Type, with the order of speech and non-speech stimuli counterbalanced across participants. Within each block, there were 80 stimuli pairs (4 tone pairs x [2 same + 2 different trials] x 5 repetitions), and the interstimulus interval was set at 500 ms. The trials were presented in randomised order for each participant. The task was preceded by practice trials with a different set of stimuli to familiarize participants with the task procedure.

2. *Identification task*

The speech stimuli from the discrimination task was also used as stimuli for a two-alternative forced-choice identification task. On every trial, participants were presented with a speech stimulus and they had to indicate which of two characters presented on the screen they heard without any time limit. There were eight distinct stimuli (4 tone pairs x 2 tones), each of which was repeated five times, resulting in a total of 40 trials, presented in randomised order. Prior to the actual task, participants completed several practice trials using different stimuli.

3. Production task

For the production task, following the procedure of a previous study (F. Liu et al., 2016), participants read aloud six tones on the same syllable /si/, which results in a real word for each tone: /si55/ 詩 'poem', /si25/ 史 'history', /si33/ 試 'exam', /si21/ 時 'time', /si23/ 市 'market', and /si22/ 是 'right'. Participants produced those words in a carrier sentence ("下一個字係 __" ["The next word is __"]) three times per word at a normal pace and in random order. Their production was recorded using Praat (Boersma & Weenink, 2013) with a Shure SM10A headworn microphone and a Roland UA-55 Quad-Capture audio interface at a sampling rate of 44100 Hz. Sample recordings from a musician and a non-musician can be found in the supplementary material (Supplementary Multimedia 1).

C. Procedure

Participants completed the tasks in the following fixed order: (i) tone discrimination; (ii) tone production; and (iii) tone identification in a sound-proof booth. Each task took approximately 5-10 mins to complete. For the two perceptual tasks (discrimination and identification tasks), participants' headphone volume was set at comfortable listening level.

D. Data analysis

Statistical analyses were conducted using R (R Core Team, 2019). Participants' accuracy on the discrimination task for each tone pair was scored using d', in which Hit was defined as 'different' pairs being judged as different and False Alarm was defined as 'same' pairs being judged as different. This measure thus takes response biases into account and a high d’ score would indicate higher sensitivity to the tones. We performed a linear mixed effects analysis using the *lme4* (Bates et al., 2015) and *lmerTest* packages (Kuznetsova et al., 2017) with d’ as the dependent variable and Group (effect-coded: non-musicians vs. musicians), Stimuli Type (effect-coded: speech vs. low-pass filtered hums), and Tone Pair (dummy-coded: T25-T55 as the reference level vs. each of the merging tone pairs, i.e., T21-T22, T22-T33, and T23-T25) as well as all possible interactions as fixed effects. As random effects, we included by-subject intercepts. We initially included by-subject random slopes for Tone Pair and Stimulus Type; however, due to convergence issues, these were removed. For this, and all the other models conducted in the present study, we tested statistical significance of the fixed effects in linear models using the *anova*() function from *lmerTest* and in generalized linear models using the *mixed*() function from the *afex* package (Singmann et al., 2019). Note that, whereas the estimates from the mixed models in *lme4*/*lmerTest* may be based on a specific contrast (e.g., T25-T55 vs. T23-T25), the output from the *anova*() function (or *mixed*() function) informs us of differences between any of the levels within a predictor (e.g., a statistically significant Tone Pair suggest that at least two levels within the predictor are significantly different). Subsequent post-hoc comparisons, if any, were conducted using the *emmeans* package (Lenth, 2019).

Accuracy on the identification task was scored as a binary outcome (Correct/Incorrect) and as such, we performed a binomial generalized linear mixed effects model on the accuracy data. We entered Group (effect-coded: non-musicians vs musicians), Tone Pair (dummy-coded: T25-T55 as the reference level vs. each of the merging tone pairs, i.e., T21-T22, T22-T33, and T23-T25) and the interaction between the two as fixed effects. As random effects, we included by-subject and by-item intercepts. By-subject random slopes for Tone Pair and by-item random slopes for Group were initially modelled but had to be removed due to convergence issues. Participants' reaction time (RT) on the identification task, measured from stimulus offset, was based on correct responses only. Following standard practice, RTs less than 150 ms and more than 2.5 SD of the mean of each participant for each tone pair were excluded (Ratcliff, 1993). We analyzed the RT data using a linear mixed-effects analysis with log-transformed RT data as the dependent variable and Group (effect-coded: non-musicians vs musicians), Tone Pair (dummy-coded: T25-T55 as the reference level vs. each of the merging tone pairs, i.e., T21-T22, T22-T33, and T23-T25) and the interaction between the two as fixed effects. Due to convergence issues, only by-subject intercepts were included as random effects in the final model. In addition to mixed models, chi-squared tests were used to compare group differences in response distribution of each tone within each tone pair. Due to participant unavailability, identification data from seven participants (musician, n = 2; non-musician, n = 5) were not collected.

For the production task, F0 contours for each tone were estimated using 10 time-normalized points using ProsodyPro (Yi Xu, 2013) on Praat. (Given that the primary focus of the present study is on pitch, the manuscript will only report results of the pitch analyses. For descriptive statistics and an analysis on the duration of the tones produced by participants, please see Supplementary Table 1.) We analysed the production data in several ways. To model participants' pitch contour production, we converted the F0 into log scores, which were then z-score normalized for each speaker. These z-score normalized log F0 were then subjected to a linear mixed effects model, with Group (effect coded: musicians vs. non-musicians), Tone (dummy-coded: T55 (reference level), T25, T33, T21, T23, T22), Time (continuous variable: 1-10), and their interactions as fixed effects. By-subject intercepts and by-subject random slopes for Tone, Time, and their interaction were included as random effects. To compare whether the groups differed in their tonal space, following from a previous study (Mok & Zuo, 2012), a tonal space quotient at the 9th time point was calculated for each tone pair of interest, with the larger F0 value in Hz of each tone in the tone pair being the numerator. The higher the quotient value, the larger the tonal space for that tone pair, which we take to assume that the tone pair is more distinct from each other. As another measure of tonal space, we compared the excursion size, defined as difference between the maximum and minimum F0 in semitones as estimated by ProsodyPro, for each tone. We assume that the larger the excursion size of each tone, the greater the difference in its contour. Group differences for these tonal space measures were analyzed using mixed ANOVA (quotient: Group as a between-subject factor and Time and Tone as within-subject factors; excursion size: Group as a between-subject factor and Tone Pair as a within-subject factor).

III. RESULTS

A. Discrimination task

For simplicity, only statistically significant estimates of the linear mixed effects model will be reported (the entire model output is displayed in Supplementary Table 2). As expected, compared to the non-merging tone pair (T25-T55), d' was lower for each of the merging tone pairs (T25-T55 vs. T21-T22: β = -0.35, SE = 0.08, *t*(350) = 4.54, *p* < .001; T25-T55 vs. T22-T33: β = -0.18, SE = 0.08, *t*(350) = 2.38, *p* = .018; T25-T55 vs. T23-T25: β = -0.91, SE = 0.08, *t*(350) = 11.79, *p* < .001). Moreover, a significant interaction between Group and T25-T55 vs. T23-T25 (β = 0.57, SE = 0.15, *t*(350) = 3.69, *p* < .001) suggests that the d' difference between the two tone pairs was smaller among musicians than non-musicians. A similar trend of d' difference was also seen for T21-T22, but this was not statistically significant (β = 0.28, SE = 0.15, *t*(350) = 1.81, *p* = .072). No other predictors were significant. The omnibus ANOVA conducted to examine the statistical significance of the fixed effects in the linear mixed effects model revealed that main effects of Group (*F*(1,50) = 5.81, *p* = .020) and Tone Pair (*F*(3, 350) = 51.89, *p* < .001) and a significant interaction between the two (*F*(3, 350) =

5.09, *p* = .002; see Figure 2). Pairwise comparisons on each Tone Pair revealed that musicians outperformed non-musicians on the T21-T22 (*t*(148) = 2.05, *p* = .042) and on the T23-T25 (*t*(148) = 4.10, *p* < .001) pairs.

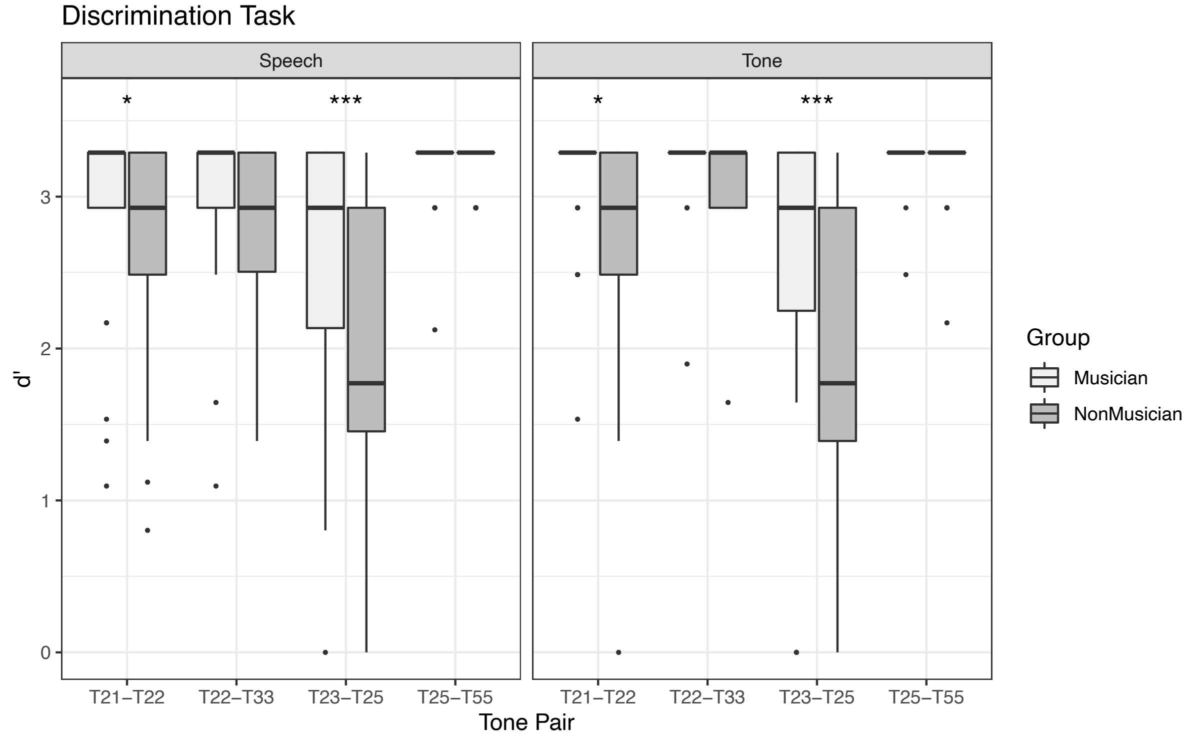

Figure 2. d' of each tone pair by stimuli tone and by group from the discrimination task.

B. Identification task

From the generalised linear mixed effects model on the accuracy data (see Supplementary Table 3 for the entire model output), performance on two of the three merging tone pairs T22-T33 and T23-T25 were significantly poorer than that of the non-merging tone pair T25-T55 (T25-T55 vs. T22-T33: β = -1.43, SE = 0.67, *z* = 2.13, *p* = .033; T25-T55 vs. T23-T25: β = -1.39, SE = 0.51, *z* = 2.74, *p* = .006). These effects interacted with Group (Group × T25-T55 vs. T22-T33: β = 2.93, SE = 0.77, *z* = 3.78, *p* < .001; Group × T25-T55 vs. T23-T25: β = 1.91, SE = 0.71, *z* = 2.70, *p* = .007), which suggests that difference in performance between the merging and non-merging tone pairs was smaller among musicians than non-musicians. Using the *mixed*() function, there were main effects of Group ($\chi^2(1) = 4.31$, $p = .040$) and Tone Pair ($\chi^2(3) = 9.72$, $p = .020$), and a significant interaction between the two ($\chi^2(3) = 24.78$, $p < .001$). Pairwise comparisons on each Tone Pair revealed that musicians outperformed non-musicians on the T22-T33 ($z = 4.20$, $p < .001$) and on the T23-T25 ($z = 2.93$, $p = .003$) pairs (see Figure 3).

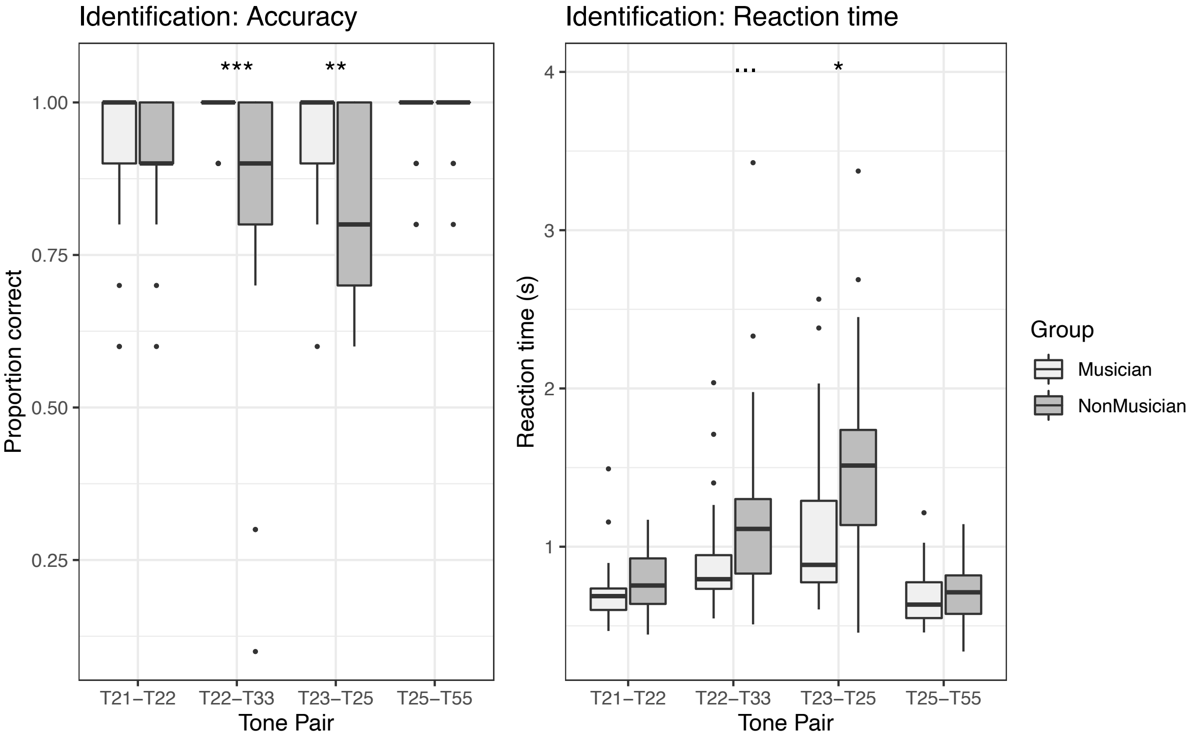

Figure 3. Proportion correct (left) and reaction time (right) of each tone pair by group from the identification task. Note that the reaction time plot is displayed using an untransformed scale for easier interpretation.

The linear mixed effects analysis on log-transformed RT data revealed complementary findings to the accuracy data (see Supplementary Table 4 for the entire model output): RT on two of the three merging tone pairs T22-T33 and T23-T25 were significantly longer than on the non-merging tone pair T25-T55 (T25-T55 vs. T22-T33: β = 0.33, SE =

0.03, *t*(1613.32) = 10.18, *p* < .001; T25-T55 vs. T23-T25: β = 0.53, SE = 0.03, *t*(1611.57) = 16.55, *p* < .001), and these interacted with Group (Group × T25-T55 vs. T22-T33: β = -0.14, SE = 0.06, *t*(1613.32) = 2.16, *p* = .031; Group × T25-T55 vs. T23-T25: β = -0.20, SE = 0.06, *t*(1611.57) = 3.14, *p* = .002), which suggests that the difference in RT between the merging and non-merging tone pairs were smaller among musicians. In addition, the model revealed that the RT was marginally longer on the merging tone pair T21-T22 compared to the non-merging tone pair (β = 0.06, SE = 0.03, *t*(1611.35) = 1.95, *p* = .051). The omnibus *anova*() function on the model revealed a main effect of Tone Pair (*F*(3, 1612.49) = 115.36, *p* < .001) and an interaction between Group and Tone Pair (*F*(3, 1612.49) = 4.01, *p* = .007). Pairwise comparisons on each Tone Pair revealed that musicians were significantly faster on the T23-T25 pair (*t*(69.1) = 2.60, *p* = .012) and marginally faster on the T22-T33 pair (*t*(69.3) = 1.88, *p* = .065) than non-musicians (see Figure 3).

We also compared participants' response distribution for the tone pairs using chi-squared tests (see Figure 4). No group differences in response distribution was observed for tone pairs T25-T55 (T25, $\chi^2(1) = 1.10$, *p* = .294; T55, $\chi^2(1) = 0.00$, *p* = 1) and T21-T22 (T21, $\chi^2(1) = 0.05$, *p* = .828; T22, $\chi^2(1) = 0.00$, *p* = 1). In contrast, for the T22-T33 tone pair, musicians correctly identified both target tones more often than non-musicians (T22, $\chi^2(1) = 14.95$, *p* < .001; T33, $\chi^2(1) = 13.46$, *p* < .001). For the T23-T25 tone pair, musicians had more correct responses than non-musicians for T23 only (T23, $\chi^2(1) = 17.58$, *p* < .001; T25, $\chi^2(1) = 0.06$, *p* = .804).

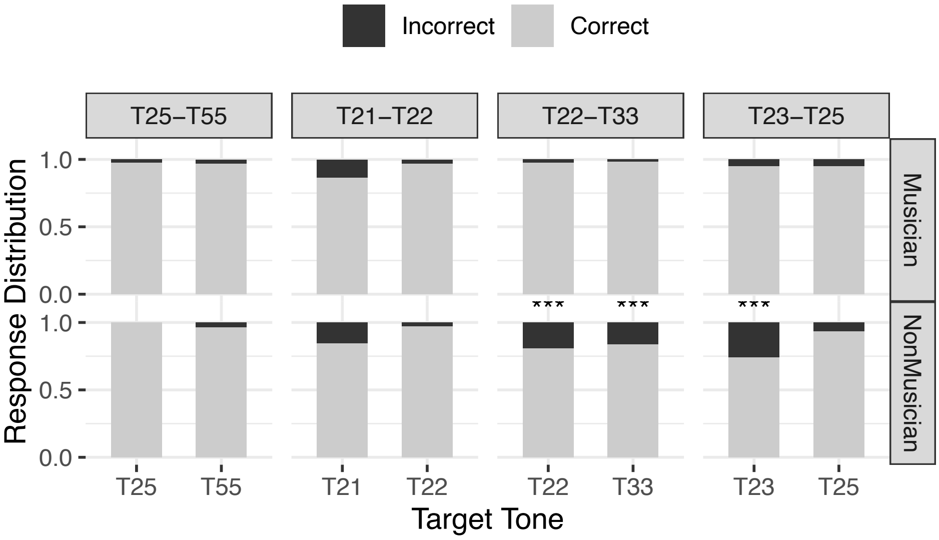

Figure 4. Response distribution of each tone pair by group from the identification task.

C. Production task

Figure 5 displays the mean time-normalized pitch contour (in z-score normalized log F0) for the six tones by group. The linear mixed effects model on z-score normalized log F0 is in reference to Tone 55 (the output of which may be found in Supplementary Table 5), which may not be useful for our current purpose, given that we want to determine whether musicians may differ from non-musicians in their tone realisations in general. As such we report only the findings on the omnibus ANOVA here. Unsurprisingly, there were main effects of Time ($F(1, 51.67) = 142.37, p < .001$) and Tone ($F(5, 59.58) = 161.10, p < .001$), and a significant interaction between the two ($F(5, 59.57) = 194.26, p < .001$), which suggests that the contour shape is different between the different tones. Importantly, the three-way interaction between Time, Tone, and Group was not significant ($F(5,59.57) = 0.88, p = .499$), suggesting that the change in contour over time was not significantly different between groups.

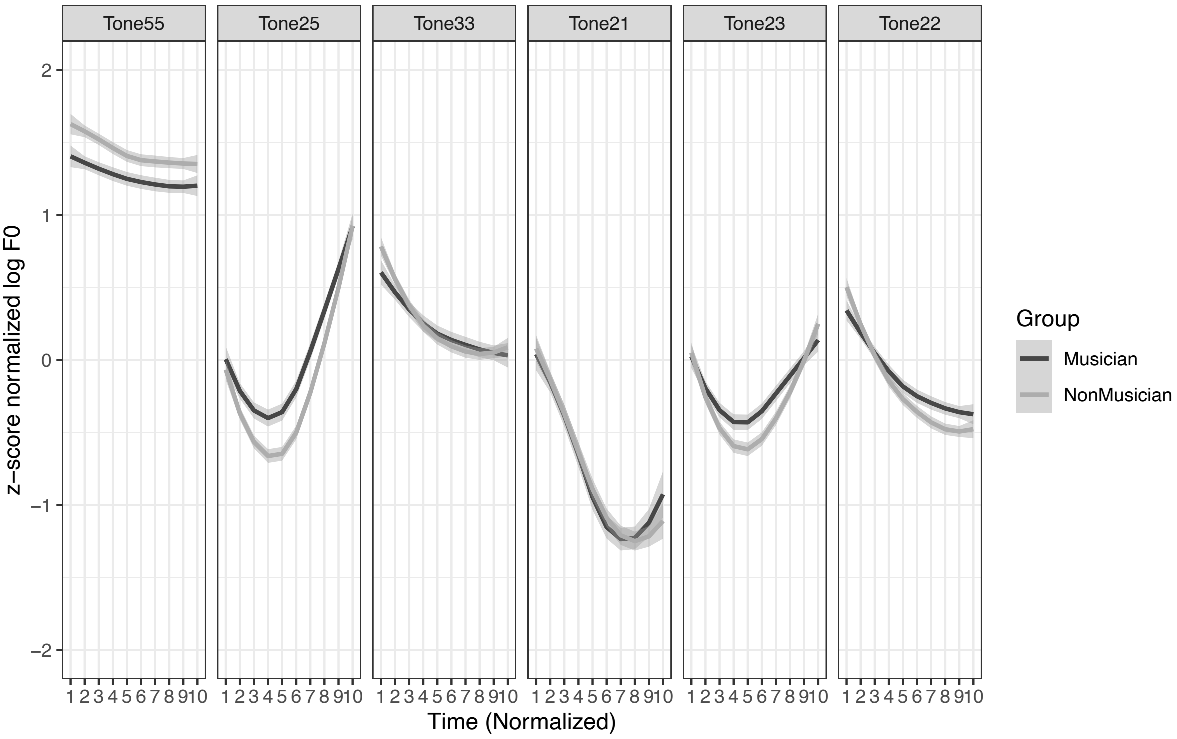

Figure 5. Mean time-normalized F0 contours (in z-score normalized log F0) of the six Cantonese tones by group. Shading represents 95% confidence intervals.

To determine if there were any group differences in their tonal space, an ANOVA on tonal space quotient with Group (musicians, non-musicians) as a between-subject factor and Tone Pair (T21-T22, T22-T33, T23-T25, T25-T55) as a within-subject factor was conducted. There was only a main effect of Tone Pair (*F*(3,150) = 26.31, *p* < .001), with the T21-T22 pair having a higher tonal space quotient than the other pairs (T21-T22 vs T25-T55, *t*(150) = 7.22, *p* < .001; T21-T22 vs. T23-T25, *t*(150) = 7.10, *p* < .001; T21-T22 vs. T22-33, *t*(150) =7.43, *p* < .001). No effects involving Group were observed (see Figure 6).

An ANOVA on excursion size with Group (musicians, non-musicians) as a between-subject factor and Tone (T55, T25, T33, T21, T23, T22) as a within-subject factor similarly revealed only a main effect of Tone. Not surprisingly, the level tones had smaller excursion size than the dynamic tones (T55 vs T25, *t*(250) = 7.26, *p* < .001; T55 vs T21, *t*(250) = 15.54, *p* < .001; T25 vs T33, *t*(250) = 6.12, *p* < .001; T25 vs. T22, *t*(250) = 5.16, *p* < .001; T21 vs

T22, *t*(250) = 13.44, *p* < .001) and within the dynamic tones, some differences reflecting the degree of glide change were also observed (T25 vs T21, *t*(250) = 8.28, *p* < .001; T25 vs. T23, *t*(250) = 4.94, *p* < .001; T21 vs T23, *t*(250) = 13.22, *p* < .001). Importantly, there were no effects involving Group (see Figure 6).

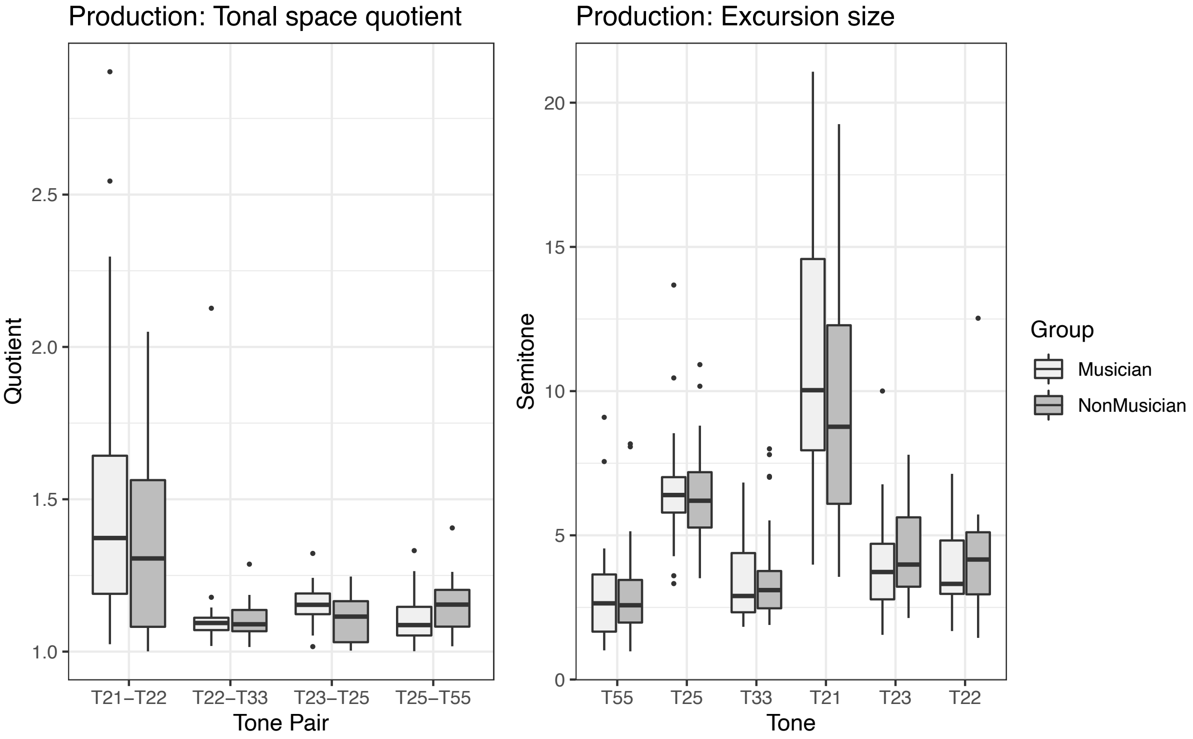

Figure 6. Tonal space quotient (left) of each tone pair and excursion size (right) of each tone by group.

IV. DISCUSSION

The present study investigated whether musical training may benefit native listeners' perception and production of lexical tones. To overcome the possibility of linguistic influence and/or ceiling effects, we examined this using merging lexical tone pairs, which are said to be relatively more difficult to differentiate than non-merging lexical tone pairs even for native listeners. Our findings suggest that a musical advantage was seen among native listeners in their perception but not in their production of these merging tone pairs.

We measured listeners' perceptual ability using a discrimination and an identification task in the present study since the tasks may partially rely on different levels of sensitivity (e.g., the former measures lower-level acoustic sensitivity whereas the latter measures higher-level phonological distinctions; Pisoni & Lazarus, 1974) and/or may be subjected to different encoding variance (e.g., discrimination involves processing and integrating two stimuli, the first of which would be subjected to decay, whereas identification only involves one stimulus; Yisheng Xu, Gandour, & Francis, 2006). Moreover, for the discrimination task, we also differentiated the discrimination of speech and non-speech stimuli, since non-speech stimuli are generally easier to discriminate (Burnham et al., 2014). We found that regardless of task and stimuli, musicians were more accurate in their perception, particularly for what appears to be the most difficult tone pair (T23-T25) based on participants' performance in both the perceptual tasks. Musicians also identified the difficult tone pair more quickly than non-musicians, suggesting that there was no speed-accuracy trade-off. Based on their confusion matrices, musicians made fewer errors than non-musicians in identifying similar-contour tone pairs, that is, the rising tone pair (T23-T25) and level tone pair (T22-T33). The fact that musicians were more accurate at both rising and level tone pairs suggests that they were sensitive to subtle changes in both pitch height and pitch direction. Our findings also revealed that non-musicians were more biased to identify T23 as T25 when presented in isolation, which may reflect the higher frequency of occurrence of T25 in Hong Kong Cantonese than T23 (Leung et al., 2004). Non-musicians may need more context (e.g., additional speech signal) or other acoustic cues (e.g., duration, amplitude, creaky voice, etc.) to help disambiguate the two rising tones. Indeed, previous studies have demonstrated the importance of other acoustic cues in the perception of lexical tones (e.g., Whalen & Xu, 1992; K. M. Yu & Lam, 2014), though it should be noted that F0 is the dominant cue used in tone perception and the contribution of the other cues is only secondary and may only be helpful for specific tones (e.g., Lin & Repp, 1989; S. Liu & Samuel, 2004; Tong et al., 2015). Nonetheless, it may be that a perceptual advantage among musicians may only be evident when pitch information is the only cue available.

While previous studies have consistently demonstrated a positive effect of musicality on lexical tone perception among *non-native* listeners (with and without tone language experience), little has been done to investigate the effect of musicality on lexical tone perception among *native* listeners. One previous study that is directly relevant to the work reported herein reported no musical advantage among native listeners on their ability to discriminate merging tone pairs (Mok & Zuo, 2012), which contradicts our perceptual findings. The discrepancy may be due to several factors. First, it may be related to when the data was collected, which would reflect different stages of the tone merger process. It is not known precisely when the data for Mok and Zuo (2012) was collected but ours was collected after theirs, that is, in 2013 and 2014. Thus, it may be that the difference between musicians and non-musicians may be more apparent when the merging process is more advanced. Alternatively, the discrepancy may be due to a larger sample size in the present study (n = 26 vs. n = 10 in each group), leading to a greater statistical power to detect a difference between groups. The sample in Mok and Zuo (2012) also had a slightly higher overall discrimination accuracy (mean ranging between 96-99%) than the present study (mean ranging between 91-97%), which is likely due to idiosyncratic differences in the stimuli, and so the ceiling-like performance in Mok and Zuo (2012) may have masked any subtle effect of musicianship on native lexical tone perception. Indeed, as we have found, the musical advantage is particularly evident for tone pairs that are difficult to differentiate.

Our results of a positive transfer effect of musical experience to lexical tone perception is in line with previous studies that have mostly investigated this with non-native listeners (Alexander et al., 2005; Burnham et al., 2014; Cooper & Wang, 2010), which adds to the growing evidence of the possibility of a shared pitch processing mechanism between lexical tone and musical pitch (Besson et al., 2011; Patel, 2008). Our study has further demonstrated that given lexical tone pairs that are difficult to perceive, musical experience may provide a boost in their perceptual ability above and beyond their native linguistic experience. This suggests that while musical experience and tone language experience may not have an additive effect, as suggested in previous studies (Cooper & Wang, 2012; Maggu, Wong, et al., 2018), musical experience may compensate where tone language experience fails to facilitate listeners' perception.

In contrast to perception, we did not observe any effect of musical experience on their production of lexical tones, at least in terms of their pitch realization. Musicians' and non-musicians' pitch contour and their excursion size (the difference between the local minimum and maximum pitch) for each tone and their tonal space for each tone pair (as measured using a quotient of the final portion for each tone pair) were similar. Thus, it appears that despite being able to hear the difference between the merging tone pairs better, musicians produced those tone pairs similarly as non-musicians, at least as indexed by our measures. We propose several possibilities on the divergent results of perception and production below.

Firstly, the positive effect of musical experience may be limited to the domain on which the musicians were trained. Whereas it is likely that most, if not all, musicians would have extensive training in *perceiving* subtle pitch differences, not all musicians would receive the same degree of training in vocal pitch *production*. To test this proposal, future research should compare instrumental musicians, trained vocalists, and non-musicians on their perception and production of lexical tones. If this proposal is true, then a musical benefit may be observed in perception for the instrumental musicians and trained vocalists whereas only trained vocalists will show an advantage in production. Note, however, that a previous study that compared English instrumentalists and vocalists failed to find any significant difference in their Mandarin lexical tone production as judged by two native listeners (Kirkham et al., 2011). Given the relatively small sample size in that comparison (*n* = 7 per group), care should be taken in interpreting the null finding pending further studies to confirm with a larger sample.

Secondly, the production task itself may have masked any observable group differences. Though the production task used in the present study is commonly used in the field, it is still relatively artificial in nature, which may lead participants to speak more carefully and produce clear and unambiguous tones. This is in contrast to more natural conversational speech, which is likely to be less precise in their production (Lindblom, 1990).

A musical advantage in perception but not in production may reflect that the perception-production link for lexical tones may not be as tight as that suggested for segments such as consonants and vowels (Diehl et al., 2004). Indeed, a growing body of research seems to suggest a dissociation between the two abilities for lexical tones. For instance, native Cantonese-learning children show a weak relationship in their lexical tone perception and production ability (P. Wong & Leung, 2018). Among non-native adults, training their tone production does not seem to improve their tone perception above and beyond tone perception training alone (Lu et al., 2015). Our results are parallel to that found among tone language listeners with congenital amusia ('tone deafness') who show typical tone production despite impaired tone perception (F. Liu et al., 2016; Nan et al., 2010). If this proposal is true, then lexical tones may indeed have different characteristics than consonants and vowels (Burnham et al., 2011), which would merit further research on lexical tones.

Similar to our divergent findings between perception and production, several studies on sound change, including those on lexical tones, have also reported a dissociation between the two abilities (e.g., Fung & Lee, 2019; Yu, 2007). For example, Mok, Zuo, and Wong (2013) classified native Hong Kong Cantonese speakers as likely to be 'tone mergerers' and 'non-tone mergerers' impressionistically (i.e., based on screening their production of the six Cantonese tones by the authors). The two groups, who were similar in age and gender distribution (though the latter was not compared statistically), produced distinctive Cantonese tones but the 'tone mergerers' were slower than the 'non-tone mergerers' at discriminating Cantonese tones. Law, Fung, and Kung (2013) classified their participants as 'tone mergerers' and 'non-tone mergerers' based on their performance on a perceptual task, and the two groups were similar in their age and gender distribution (though this was not compared statistically). They found that the 'tone mergerers' did not have a significant mismatch negativity (MMN) response to Cantonese T21-T22 contrast that was seen among 'non-tone mergerers', suggesting that the former could not discriminate the contrast, despite both groups being able to produce all the Hong Kong Cantonese tones distinctively. It is still unclear how these so-called 'near mergers' (Labov et al., 1972) only pose perceptual but not production difficulty but their existence implies that perception and production abilities are dissociable to a certain extent.

Drawing on the studies on lexical tone near mergers, and from our own findings, we propose that the dissociation between perception and production ability may in part be modulated by cognitive factors. Most, if not all, perceptual tasks (e.g., discrimination, identification, etc.) involve cognitive processes to some extent (e.g., comparing two or more memory traces to determine if they are similar or different; Heald & Nusbaum, 2014), or at the least more so than production tasks (Loui et al., 2008). So, any difference in performance between perception and production may be partly explained by individual differences in cognitive abilities. While we do not have any direct evidence for this, findings from previous literature does seem to corroborate this claim. For example, tone mergerers and non-tone mergerers in Mok et al. (2013) had similar performance on a perceptual task, but the former was significantly slower in their response, which may reflect a more conscious, effortful processing that might imply the use of top-down strategies and/or cognitive abilities to compensate for performance. In Law et al. (2013), the tone mergerers had a weaker P3a component, which is said to measure attentional switch, than non-tone mergerers when perceiving lexical tone preattentively. These findings suggest that tone mergerers' perception may be constrained by their cognitive abilities. In terms of our own findings, musicians have been reported to have enhanced cognitive abilities including abilities that are likely to be important in perceptual tasks such as verbal memory, general intelligence, and executive functions (see Schellenberg & Weiss, 2013, for a review). So if our proposal is true, then this may explain their superior performance in perceiving lexical tones than non-musicians. Another possibility for the dissociation between perception and production abilities may be related to differences in the range of interindividual variations in pitch and non-pitch cues (e.g., voice onset time (VOT), place of articulation, etc.). That is, whereas listeners will hear various pitch ranges, even ones outside of their own pitch range, they are likely to hear non-pitch cues that are closer to their own range. To be sure, there are talker differences in the production of non-pitch cues (e.g., voice onset time: Allen et al., 2003; Oh, 2011), but it may be that these differences are smaller than those of pitch cues. Indeed, a previous study using a corpus of spontaneous speech of American English found gender differences for pitch but not for VOT (Syrdal, 1996), suggesting that at least in the gender dimension, the variation in pitch is larger than in VOT between genders. These possibilities remain speculative and need to be examined systematically. Prior to that, however, future research must first determine conclusively that there is a dissociation between the two abilities for lexical tones.

Sound change may be partly due to factors relating to speakers (Ohala, 1989, 1993), such as variation in speech production and misperception, which may be due to acoustic factors (e.g., failing to hear a difference between two similar sounds due to native language interference) or sociolinguistic factors (e.g., to adopt a particular sound change to elevate one's status). While it is beyond the scope of the present study to pinpoint the reason(s) for sound change in Hong Kong Cantonese, it is possible that genuine perceptual difficulties partly contribute to the change, as native listeners' behavioural performance on merging tones correlated with the fidelity of their brainstem representations of merging tones (Maggu et al., 2016). The source for this perceptual difficulty may be partly due to language contact (Maggu, Zong, et al., 2018) and/or a genetic basis, given the direct association between a genetic variant, APSM (rs41310927), and lexical tone perception, even after taking into consideration of confounding factors such as musical experience and IQ (P. C. M. Wong et al., 2020). Regardless, assuming that the sound change in Hong Kong Cantonese results from genuine misperception/mispronunciation, our study provides an intriguing idea that musicianship may help resist sound change, at least for the perception of difficult contrasts. That is, extensive musical experience may provide veridical perception of lexical tones, which may limit the merging of ambiguous tone pairs. Further work is necessary to determine whether the tone merging phenomenon is indeed less likely among those with musical experience, and if so, whether the effect is causative or correlational in nature.

In conclusion, we found that musical training provides native listeners with an advantage in perceiving lexical tone contrasts that are undergoing sound change, particularly those that are the most difficult to differentiate. This suggests that musical experience may provide a boost in perception where linguistic experience may fail. Our findings raise the possibility that musical experience may provide a buffer to help resist sound change, at least in cases of phonemic mergers where the merging occurs due to a loss of distinction between sounds. Conversely, there was no musical advantage seen in the production of these merging tones. Though this may be due to methodological reasons such as the type of musicians examined and the production task, it may also be due to a weaker perception-production relationship for lexical tones relative to consonants and vowels and/or due to individual differences in cognitive abilities, which are likely to be more involved in perceptual rather than production tasks. Future work is necessary to further understand this relationship and how musical experience may modulate both abilities to deepen our knowledge of the interaction between music and language.

ACKNOWLEDGEMENTS

This work was supported by a National Science Foundation Grant (BCS-1125144) and a Research Grants Council (RGC) of Hong Kong grant (34000118) to P.C.M.W. and a European Research Council (ERC) Starting Grant to F.L. (CAASD, 678733). We thank Hilda Chan for assistance with data collection.

REFERENCES

Alexander, J. A., Wong, P. C. M., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. *Interspeech 2005*, 397–400.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. https://doi.org/10.1121/1.1528172

Asaridou, S. S., & McQueen, J. M. (2013). Speech and music shape the listening brain: Evidence for shared domain-general mechanisms. *Frontiers in Psychology*, *4*(June), 1–14. https://doi.org/10.3389/fpsyg.2013.00321

Ayotte, J., Peretz, I., & Hyde, K. L. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain: A Journal of Neurology*, *125*(Pt 2), 238–251.

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

Bauer, R. S., Kwan-hin, C., & Pak-man, C. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change*, *15*(02). https://doi.org/10.1017/S0954394503152039

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of training between music and speech: Common processing, attention, and memory. *Frontiers in Psychology*, *2*(May), 94. https://doi.org/10.3389/fpsyg.2011.00094

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer*. http://www.praat.org

Burnham, D., Brooker, R., & Reid, A. (2014). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychology of Music*, 1–17. https://doi.org/10.1177/0305735614546359

Burnham, D., Kim, J., Davis, C., Ciocca, V., Schoknecht, C., Kasisopa, B., & Luksaneeyanawin, S. (2011). Are tones phones? *Journal of Experimental Child Psychology*, *108*(4), 693–712. https://doi.org/10.1016/j.jecp.2010.07.008

Cooper, A., & Wang, Y. (2010). The role of musical experience in Cantonese lexical tone perception by native speakers of Thai. *Speech Prosody 2010*.

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*, 4756.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

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

Fung, R. S. Y., & Lee, C. K. C. (2019). Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *The Journal of the Acoustical Society of America*, *146*(5), EL424–EL430. https://doi.org/10.1121/1.5133661

Fung, R. S. Y., & Wong, C. S. P. (2010). *Mergers and near-mergers in Hong Kong Cantonese Tones*. The Fourth European Conference on Tone and Intonation (TIE4), Stockholm, Sweden, Stockholm, Sweden. http://www.su.se/polopoly_fs/1.30143.1320939963!/HKC_tone_mergers_Roxana_Cathy.pdf

Fung, R. S. Y., Wong, C. S. P., & Law, S. P. (2011). *The mechanism of rising tone merger in Hong Kong Cantonese: An acoustic approach*. Phonetics & Phonology in Iberia (PaPI), Tarragona, Spain. http://hdl.handle.net/10722/136295

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. https://doi.org/10.1016/S0095-4470(03)00016-0

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00035

Hutchins, S., Gosselin, N., & Peretz, I. (2010). Identification of changes along a continuum of speech intonation is impaired in congenital amusia. *Frontiers in Psychology*, *1*(DEC), 1–8. https://doi.org/10.3389/fpsyg.2010.00236

Kei, J., Smyth, V., So, L. K. H., Lau, C. C., & Capell, K. (2002). Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure. *Asia Pacific Journal of Speech, Language and Hearing*, *7*(1), 25–38. https://doi.org/10.1179/136132802805576535

Kirkham, J., Lu, S., Wayland, R., & Kaan, E. (2011). Comparison of vocalists and instrumentalists on lexical tone perception and production tasks. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, *August*, 1098–1101.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Labov, W., Yaeger, M., & Steiner, R. (1972). *A quantitative study of sound change in progress*. U.S. Regional Survey.

Law, S. P., Fung, R., & Kung, C. (2013). An ERP study of good production vis-à-vis poor perception of tones in Cantonese: Implications for top-down speech processing. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0054396

Lee, C.-Y., & Lee, Y.-F. (2010). Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *The Journal of the Acoustical Society of America*, *127*, 481.

Lenth, R. V. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://cran.r-project.org/package=emmeans

Leung, M.-T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 500–505. https://doi.org/10.3758/BF03195596

Lin, H.-B., & Repp, B. H. (1989). Cues to the perception of Taiwanese tones. *Language and Speech*, *32*(1), 25–44.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-2037-8

Liu, F., Chan, A. H. D., Ciocca, V., Roquet, C., Peretz, I., & Wong, P. C. M. (2016). Pitch perception and production in congenital amusia: Evidence from Cantonese speakers. *The Journal of the Acoustical Society of America*, *140*(1), 563–575. https://doi.org/10.1121/1.4955182

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain*, *133*(6), 1682–1693. https://doi.org/10.1093/brain/awq089

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*, *47*(2), 109–138. https://doi.org/10.1177/00238309040470020101

Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action-perception mismatch in tone-deafness. *Current Biology : CB*, *18*(8), R331–R332. https://doi.org/10.1016/j.cub.2008.02.045

Lu, S., Wayland, R., & Kaan, E. (2015). Effects of production training and perception training on lexical tone perception—A behavioral and ERP study. *Brain Research*, *1624*, 28–44. https://doi.org/10.1016/j.brainres.2015.07.014

Maggu, A. R., Liu, F., Antoniou, M., & Wong, P. C. M. (2016). Neural Correlates of Indicators of Sound Change in Cantonese: Evidence from Cortical and Subcortical Processes. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00652

Maggu, A. R., Wong, P. C. M., Liu, H., & Wong, F. C. K. (2018). Experience-dependent influence of music and language on lexical pitch learning is not additive. In B. Yegnanarayana, C. Chandra Sekhar, S. Narayanan, S. Umesh, S. R. M. Prasanna, H. A. Murthy, P. Rao, P. Alku, & P. K. Ghosh (Eds.), *Interspeech 2018* (pp. 3791–3794). International Speech Communication Association (ISCA). https://doi.org/10.21437/Interspeech.2018-2104

Maggu, A. R., Zong, W., Law, V., & Wong, P. C. M. (2018). Learning two tone languages enhances the brainstem encoding of lexical tones. *Interspeech 2018*, 1437–1441. https://doi.org/10.21437/Interspeech.2018-2130

Magne, C., Schön, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neuroscience*, *18*(2), 199–211. https://doi.org/10.1162/089892906775783660

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

Mok, P. P. K., & Zuo, D. (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. https://doi.org/10.1121/1.4747010

Mok, P. P. K., Zuo, D., & Wong, P. W. Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change*, *25*(3), 341–370. https://doi.org/10.1017/S0954394513000161

Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain*, *133*(9), 2635–2642. https://doi.org/10.1093/brain/awq178

Oh, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics*, *39*(1), 59–67. https://doi.org/10.1016/j.wocn.2010.11.002

Ohala, J. (1989). Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (Eds.), *Language chane: Contributions to the study of its causes* (pp. 173–198). Mouton de Gruyter.

Ohala, J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical linguistics: Problems and perspectives* (pp. 237–278). Longman.

Ong, J. H., Burnham, D., Stevens, C. J., & Escudero, P. (2016). Naïve learners show cross-domain transfer after distributional learning: The case of lexical and musical pitch. *Frontiers in Psychology*, *7*(August), 1–10. https://doi.org/10.3389/fpsyg.2016.01189

Patel, A. D. (2008). *Music, language, and the brain*. Oxford University Press, USA.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America*, *55*(2), 328–333. https://doi.org/10.1121/1.1914506

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532.

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 499–550). Elsevier. http://dx.doi.org/10.1016/B978-0-12-381460-9.00012-2

Singmann, H., Bolker, B., Westfall, J., & Frederik, A. (2019). *afex: Analysis of Factorial Experiments.* (R package 0.25-1) [Computer software]. https://CRAN.R-project.org/package=afex

Stagray, J. R., & Downs, D. (1993). Differential sensitivity for frequency among speakers of a tone and a nontone language. *Journal of Chinese Linguistics*, *21*(1), 143–163.

Syrdal, A. K. (1996). Acoustic variability in spontaneous conversational speech of American English talkers. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *1*, 438–441. https://doi.org/10.1109/ICSLP.1996.607148

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, *4*(1), 46–64. https://doi.org/10.1037/1528-3542.4.1.46

Tillmann, B. (2014). Pitch processing in music and speech. *Acoustics Australia*, *42*(2), 124–130.

Tillmann, B., Burnham, D., Nguyen, S., Grimault, N., Gosselin, N., & Peretz, I. (2011). Congenital amusia (or tone-deafness) interferes with pitch processing in tone languages. *Frontiers in Psychology*, *2*.

Tong, X., Lee, S. M. K., Lee, M. M. L., & Burnham, D. (2015). A tale of two features: Perception of Cantonese lexical tone and English lexical stress in Cantonese-English bilinguals. *PLOS ONE*, 33.

Vuvan, D. T., Nunes-Silva, M., & Peretz, I. (2015). Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia. *Cortex*, *69*, 186–200. https://doi.org/10.1016/j.cortex.2015.05.002

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*, *49*(1), 25–47. https://doi.org/10.1159/000261901

Wong, A. M. Y., Ciocca, V., & Yung, S. (2009). The perception of lexical tone contrasts in Cantonese children with and without Specific Language Impairment (SLI). *Journal of*

Speech, Language, and Hearing Research, *52*(6), 1493–1509. https://doi.org/10.1044/1092-4388(2009/08-0170)

Wong, P. C. M., Kang, X., Wong, K. H. Y., So, H.-C., Choy, K. W., & Geng, X. (2020). ASPM-lexical tone association in speakers of a tone language: Direct evidence for the genetic-biasing hypothesis of language evolution. *Science Advances*, *6*(22), eaba5090. https://doi.org/10.1126/sciadv.aba5090

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T. M., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420–422.

Wong, P., & Leung, C. T. T. (2018). Suprasegmental features are not acquired early: Perception and production of monosyllabic cantonese lexical tones in 4- to 6-year-old preschool children. *Journal of Speech, Language, and Hearing Research*, *61*(5), 1070–1085. https://doi.org/10.1044/2018_JSLHR-S-17-0288

Xu, Yi. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. *Proceedings of Tools Resource Analysis Speech Prosody*, 7–10.

Xu, Yisheng, Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*, 1063.

Yu, A. C. L. (2007). Understanding near mergers: The case of morphological tone in Cantonese. *Phonology*, *24*(1), 187–214. https://doi.org/10.1017/S0952675707001157

Yu, K. M., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *The Journal of the Acoustical Society of America*, *136*(3), 1320–1333. https://doi.org/10.1121/1.4887462

Zhao, T. C., & Kuhl, P. K. (2015). Effect of musical experience on learning lexical tone categories. *The Journal of the Acoustical Society of America*, *137*(3), 1452–1463. https://doi.org/10.1121/1.4913457

FIGURE CAPTIONS

Figure 1. Time-normalized F0 contours of the stimuli used.

Figure 2. d' of each tone pair by stimuli tone and by group from the discrimination task.

Figure 3. Proportion correct (left) and reaction time (right) of each tone pair by group from the identification task. Note that the reaction time plot is displayed using an untransformed scale for easier interpretation.

Figure 4. Response distribution of each tone pair by group from the identification task.

Figure 5. Mean time-normalized F0 contours (in z-score normalized log F0) of the six Cantonese tones by group. Shading represents 95% confidence intervals.

Figure 6. Tonal space quotient (left) of each tone pair and excursion size (right) of each tone by group.