Effects of Stimulus Duration and Vowel Quality in Tone Perception by English Musicians and Non-musicians

Si Chen
Department of Chinese and Bilingual Studies,
The Hong Kong Polytechnic University
Kowloon, Hong Kong SAR, China
sarah.chen@polyu.edu.hk

Yiqing Zhu
Department of Linguistics,
University of Florida
Florida, U. S.
yiqingzhu@ufl.edu

Ratree Wayland
Department of Linguistics,
University of Florida
Florida, U. S.
ratree@ufl.edu

Yike Yang
Department of Chinese and Bilingual Studies,
The Hong Kong Polytechnic University
Kowloon, Hong Kong SAR, China
yi-ke.yang@connect.polyu.hk

Abstract

The link between music and language has been a subject of great interest, and evidence suggesting a connection between musical abilities and prosodic processing skills in language is growing. Acoustic fundamental frequency (F0), perceived as pitch, differentiates notes in music and word meaning in lexical tone languages. This study examines categorical perception of pitch stimuli among 14 English musicians and 15 English non-musicians, both groups having no exposure to tonal languages. The stimuli consist of continua of falling and rising F0 contours produced on high and low vowels with 9 different durations. The results revealed that musicians were more sensitive to variation in stimulus duration than non-musicians were, and music experience enhanced the sharpness of category boundaries. Significant main effects of vowel quality and pitch directions as well as two-way interactions between vowel and pitch direction, vowel and duration, group and duration, and pitch direction and duration on identification rate were also found. Formulae for minimum duration required for English musicians and non-English musicians to perceive rising and falling F0 were derived, revealing that musicians require less time to perceive a pitch fall and rise if the change is less than 12semitones.

Index Terms: categorical perception, Mandarin tones, vowel quality, stimulus duration

1 Introduction

Categorical perception (CP) refers to the phenomenon wherein listeners are better at distinguishing stimuli from different categories than distinguishing stimuli from the same category (Goldstone and Hendrickson, 2010). Enhanced perceptual sensitivity to cross-category differences requires the mapping of acoustically variable phonetic units into finite phonological categories, a skill crucial for speech perception (Kuhl, 2004). CP has been investigated at both segmental and suprasegmental levels (Flege et al., 1994; Changet al., 2016; Boersma and Chládková, 2013) and is commonly characterized by a sharp category boundary, a corresponding discrimination peak, and the predictability of discrimination from identification (Treisman et al., 1995). For lexical tones, a few studies showed that Mandarin tones are perceived categorically by native listeners, but not by non-native tone language listeners (Peng et al., 2010; Xu et al., 2006).

Acoustic fundamental frequency (F0) or its auditory impression of pitch is a common perceptual object in music and language. It differentiates notes in music and word meaning in lexical tone languages such as Mandarin.

Associations between musical ability and accuracy at perceiving lexical tone contrasts have been reported in previous investigations (Wayland et al., 2010; Zhao and Kuhl, 2015; Lee and Hung, 2008). For example, English-speaking musicians and non-musicians with no prior tone language experience were tested in identification and discrimination tasks of four Mandarin tones, and the musicians significantly outperformed the non-musicians in both tasks, exhibiting higher accuracy rate and shorter reaction times, indicating an overlap in the processing of pitch in music and speech (Alexander et al., 2005; Lee and Hung, 2008). CP of pitch stimuli among musicians is less well studied. It is reported that categorization of synthetic speech triads is more prominent and discrimination pattern is better predicted by categorization among musicians than among non-musicians (Locke and Kellar, 1973). Similarly, musical training was found to sharpen categorical boundaries and improve discrimination of three-pure-tone sequences (Raz and Brandi, 1977). In addition, musicians' processing of pitch (sine-tone) intervals is reported to be categorical (Siegel and Siegel, 1977). More recently, comparable identification functions with similar steepness and locations of categorical boundary were found among Chinese musicians and non-musicians in their processing of Mandarin Tone 1-Tone 4 continuum. However, musicians exhibit superior discrimination performance on within-category stimuli over non-musicians (Wu et al., 2015).

Stimulus duration also plays a critical role in categorical perception, where shorter vowels are perceived more categorically than longer vowels (Fujisaki and Kawashima, 1970). In addition, the interaction between perceived vowel duration and the tone it bears has also been found. In general, dynamic tones such as falling and rising tones lead to longer perception of vowels than level tones (e.g. Pisoni, 1976; Yu et al., 2014). In comparison to pitch production, it was found that a relative shorter duration is required to perceive than to produce pitch contours, with non-tonal listeners needing longer duration to detect a change in pitch direction (Chen et al., 2017). More importantly, the study found that native tone listeners are more sensitive to pitch contour than non-tone listeners, and their perception is also more categorical. However, for both groups, their perception is more categorical, with sharper category boundary, for the falling contours than the rising ones. Category boundary becomes sharper with increasing duration but at a faster pace for tone listeners. The intrinsic F0 (IF0) effects also play a role in pitch perception. IF0 refers to the fact that high vowels are usually correlated with higher F0 values and low vowels with low F0 values in speech production cross-linguistically (Whalen & Levitt, 1995). The opposite of this relationship has been observed in speech perception. It has been shown that intrinsic F0 effects significantly contributed to CP among non-musicians, though the effect was relatively limited (Chen et al., 2017). Lastly, duration asserts a stronger effect on between- and within-category discrimination patterns among tonal listeners. The current study examines effects of musical experience on CP of tonal continua by listeners without tone language background while considering the factor of vowel quality and stimulus duration.

2 Methodology

2.1 Participants

Fourteen English musicians (7 males; 7 females; mean age ± SD: 23.38 ± 5.06 years) and fifteen English non-musicians (7males; 8 females; mean age ± SD: 20.4 ± 1.96 years) participated in the study. All of the English musicians received professional training (mean years ± SD: 14.23 ± 4.60 years). Partial results of the English non-musicians were reported in (Chen et al., 2017). The experiment procedure was approved by the University of Florida Internal Review Board (IRB), and all participants were paid for their participation. None of the participants reported history of speaking, hearing or language impairments, and none received formal instruction in Mandarin Chinese or other tone languages.

2.2 Stimuli

We manipulated tones on both low and high vowels ([a] and [i]) produced by a male native Mandarin speaker with no reported speaking or hearing problems as described in Chen et al. (2017). The pitch synchronous overlap add (PSOLA) method (Moulines and Laroche, 1995) was used to generate 7-step, level-to-falling and level-to-rising pitch continua on [a] and [i] vowels with 9 different durations (200ms, 180ms, 160ms, 140ms, 120ms, 100ms, 80ms, 60ms and 40ms). Thus, a total of 36 continua [2 (pitch directions/contours) x 2 (vowels) x 9 (durations)] were generated and each continuum contains 7 steps.

All contour stimuli were linear with slope and intercept values manipulated based on the estimation of underlying pitch targets (slope: 93.4; intercept: -2.2 semitones (st) obtained from a corpus of real speech in Mandarin (Prom-On et al., 2009). For example, to generate a rising continuum, we first calculated the onset and offset for each duration value. If the stimulus duration is 200ms (0.2s), based on the underlying pitch target formula as in Eq. (1), where *t* is duration (s) *X(t)* is the F0 value (st) at duration *t*,

$$X(t) = 93.4 * t - 2.2 \qquad (1)$$

the onset was calculated by setting t = 0s, which is -2.2st. Using 130Hz as a baseline as chosen in Lehnert-Lehouillier (2013) and Chen et al. (2017), all the st can be transformed to Hz. In this case, -2.2st equals 123.02 Hz. We then calculate the offset value by setting t = 0.2s since the stimulus duration is 0.2s, and the F0 value is 16.48st (196.56Hz). Thus, there is an onset-to-offset distance of 18.68st. Next, we set the estimated intercept of the underlying pitch target [-2.2st] as the cut-off point for the onset value, then various onsets and offsets for different steps for a particular stimulus duration were calculated. We aim to find the lowest (LO < 2.2st) and highest (HO > 2.2st) onset values where the value -2.2st serves as a cut-off point. The distance between LO and -2.2st is one third between HO and -2.2st. The value -2.2st was not chosen as the median because we were more interested in perception of shallower slopes. HO is the same as offset since the tone becomes a level tone when the onset is the highest, equal to offset. LO can be calculated as the point with a distance of one third below -2.2st, which is -8.43st or 105.23Hz (-2.2st-1/3*18.68st). After determining LO and HO for each stimulus duration, seven perceptually equal steps were created based on the ERB scale, which reflects natural perception (Xu et al., 2006). The specific values for all rising pitch directions are presented in Table 1. For falling pitch directions, the reversed order of onset values in rising pitch directions were used as offset values, and the highest offset value was equal to the onset values of all steps (e.g. 196.56 Hz when stimulus duration is 200ms).

All stimuli were peak normalized to the same intensity level for presentation. Details on stimulus generation procedure can be found in Chen et al. (2017).

Duration	**0.2**	**0.18**	**0.16**	**0.14**	**0.12**
Onset 1	105.23	106.89	108.57	110.28	112.01
Onset 2	118.67	118.92	119.21	119.55	119.92
Onset 3	132.77	131.49	130.27	129.13	128.05
Onset 4	147.58	144.61	141.76	139.03	136.41
Onset 5	163.12	158.31	153.69	149.26	145.02
Onset 6	179.43	172.62	166.09	159.84	153.86
Onset 7	196.56	187.56	178.97	170.78	162.96
Duration	**0.1**	**0.08**	**0.06**	**0.04**	
Onset 1	113.78	115.57	117.39	119.24	
Onset 2	120.33	120.79	121.28	121.82	
Onset 3	127.04	126.11	125.24	124.43	
Onset 4	133.91	131.52	129.24	127.06	
Onset 5	140.94	137.04	133.30	129.72	
Onset 6	148.14	142.66	137.42	132.40	
Onset 7	155.50	148.38	141.59	135.11	

Table 1: Onset values for each step of stimuli created based om stimulus duration for linear rising pitch directions

All stimuli were presented in two blocks for an identification task, one for falling pitch directions and the other for rising pitch directions. In total, 1260 stimuli were presented randomly (5 repetitions * 7 steps * 9 duration * 2 vowels * 2 pitch directions). The order of blocks was also randomized across participants.

2.3 Data analysis

First, to examine factors that significantly contributed to identification of tone stimuli, we fitted a generalized linear mixed model with a random effect of subjects using the “lme4” R package (Bates et al., 2015). We divided all stimuli into eight subgroups according to musical experience, pitch direction and vowel quality (FEMA, FEMI, REMA, REMI, FENA, FENI, RENA, RENI), where F and R stand for falling and rising pitch directions, EM and EN stand for English musicians and English non-musicians, and A and I stand for [a] and [i] vowels. Thus, FEMA stands for a subset of data consisting of falling pitch directions (F) on the vowel [a] (A) perceived by English Musicians (EM).

Second, to obtain category boundary sharpness and location, we fit a generalized linear mixed model again with identification scores (0 or 1) as the response variable and step number (x = 0–6) as a factor within each subgroup. Following (Xu et al., 2006), we treated the step number as a continuous variable. When only the fixed effects are considered, the model is similar to a logistic regression model in Eq. (2)

$$\log_e\left(\frac{p_1}{1 - p_1}\right) = b_0 + b_1 x \tag{2}$$

We were able to extract the coefficients b_1 and b_0 from the fixed effects of the generalized linear mixed model. From Xu et al. (2006), the coefficient b_1 represents the sharpness of category boundary. A post-hoc analysis was conducted to examine the effects of stimulus duration on the sharpness of category boundary by comparing identification between different duration values for each subgroup. A likelihood ratio test was employed for the post-hoc analyses, where each pair of stimuli with different duration was compared using two models. The first model treated the coefficient b_1 from each pair as the same, and the second model treated them as different. Significant differences obtained between the two models indicated different coefficient b_1 and hence sharpness of category boundary between the pair. Similarly, we conducted likelihood ratio tests for a post-hoc analysis on effects of music experience, pitch direction and vowel quality. We further modeled the relationship between the sharpness of category boundary and duration for native English musicians and non-musicians using regression models.

The step number at category boundary (*xcb*) can be estimated using the proportion of $-b_0/b_1$ after obtaining the estimates for b_0 and b_1 (Xu et al., 2006). A post-hoc analysis was also conducted to test for the effects of musical experience, duration and vowel quality in determining category location after obtaining individual category boundary for subjects in each subgroup (e.g. FEMA). Regression models were then fitted to capture the relationship between duration and category boundary for listeners with and without musical experience.

Finally, we obtained formulae for minimum duration required for pitch perception to compare with those proposed for pitch production (Prom-On et al., 2009). The form of the formulae is as shown in Eq. (3).

$$t = b_0 + b_1 d \tag{3}$$

where *t* stands for the stimulus duration needed for perceiving *d* st differences (rising or falling) from level tones. To obtain the formula, we first estimated the step where the identification rate was 50% and designated it as a cut-off point because a step greater than this cut-off point meant that it was less likely to be perceived as a level tone. In other words, the cut-off step is the smallest number of steps where a rising or falling pitch direction was perceived as different from a level tone. The calculated step numbers were then transformed back to st values, and the baseline was set to the pitch value of the level tone for each stimulus duration value. Then, linear mixed effects models were fitted to the cut-off st values and duration. Formulae for rising and falling pitch directions were obtained for both English musicians and non-musicians combined as well as for each group separately. We further obtained formulae for each subgroup considering the effect of vowel quality. To test for the differences among those formulae, we conducted likelihood ratio tests.

3 Results

3.1 Sharpness of category boundary

A generalized linear mixed model was fitted to the identification rate of both English speakers with and without professional musical training. The results revealed significant main effects of vowel quality ($\chi^2(1) = 43.463$, $p < 0.001$) and duration ($\chi^2(1) = 6341.9$, $p < 0.001$) and a marginally significant main effect of pitch direction ($\chi^2(1) = 3.52$, $p = 0.06$). The two-way interaction between vowel and pitch directions ($\chi^2(1) = 12.97$, $p < 0.001$), vowel and duration ($\chi^2(1) = 9.35$, $p = 0.002$), music experience and duration ($\chi^2(1) = 462.39$, $p < 0.001$), and pitch directions and duration ($\chi^2(1) = 135.35$, $p < 0.001$) also reached significance. Neither of the three-way interactions reached significance.

Due to significant contribution of stimulus duration in identification of pitch directions, we further tested how sensitive musicians were to a change in duration. We conducted pair-wise comparisons of duration within each subgroup to test whether the category sharpness significantly varied as the duration changed, and most pairs reached significance, except for a few pairs listed in Table 2.The English musician group showed no significant differences for approximately 6% of all the pairs in comparison to 13% for the English non-musicians. English musicians were extremely sensitive to falling (0 non-significant pair) and rising pitch contours (2 non-significant pairs) on the vowel [a]. Both English non-musicians and musicians showed an increasing trend of sharpness of category boundary as stimulus duration increased. However, English musicians exhibited sharper category boundary than English non-musicians in most cases as shown in Figure 1.

Sub group	Duration pair	χ^2 (1)	P value
FEMI	0.18_0.2	1.87	0.17
	0.16_0.18	1.18	0.28
	0.04_0.06	0.05	0.82
REMA	0.18_0.2	1.29	0.26
	0.1_0.12	0.65	0.42
	0.04_0.06	2.46	0.12
REMI	0.16_0.18	0.23	0.63
	0.04_0.06	0.17	0.68

Table 2: *Pairs of stimulus duration for comparison in English musicians*

In addition, likelihood ratio tests suggested a significant effect of musical experience on the sharpness of category boundary ($\chi 2(1) = 386.64$, $p < 0.001$). Pitch direction ($\chi 2(1) = 23.70$, $p < 0.001$) and vowel quality ($\chi 2(1) = 44.36$, $p < 0.001$) also significantly affected category boundary sharpness. Post-hoc analyses were performed to further examine group and pitch direction effects. English musicians and English non-musicians differed across conditions (FEMA vs. FENA, FEMI vs. FENI, REMA vs. RENA and REMI vs. RENI) for most duration values. However, the significant effects of pitch direction were found for most duration values, but only among pairs involving the vowel [i] (FEMI vs. REMI and FENI vs. RENI). Other pairs only showed significant differences for about three duration values.

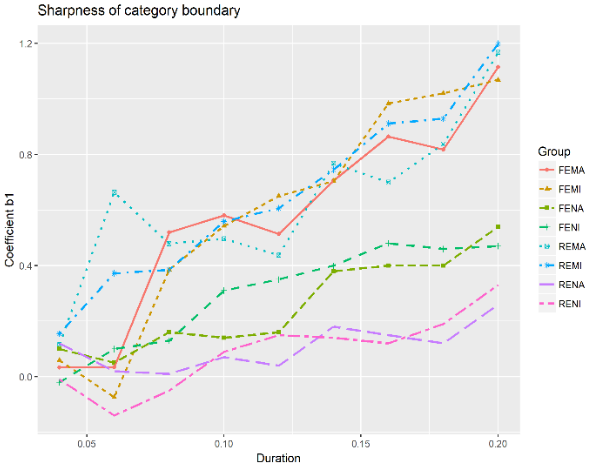

Figure 1: *Sharpness of category boundary for musicians and non-musicians*

According to likelihood ratio tests, the group effect with and without musical training reached significance; therefore, two regression models were fitted for each group separately to examine the relationship between duration and the sharpness of category boundary. For English musicians, a regression model with an extra quadratic term did not further improve the model as suggested by likelihood ratio tests ($\chi^2(1) = 0.05$, $p = 0.82$). The regression model predicted the sharpness well (English musicians: $F(1, 34) = 210.3$, $p < 0.001$; adjusted $R2 = 0.85$). For English non-musicians, a linear regression without a quadratic term could be used to model the sharpness of category boundary with respect to duration ($\chi 2(1) = 0.12$, $p = 0.73$) and yielded good predictions ($F(1, 34) = 38.26$, $p < 0.001$; adjusted $R^2 = 0.52$) (Chen et al., 2017). Specifically, the formula for the relationship between the sharpness of category boundary (b1) and duration (d) for English musicians is presented in Eq. (4).

$$b_1 = 5.98 * d - 0.10 \qquad (4)$$

For comparison, the proposed formula for the English non-musicians is presented in Eq. (5) (Chen et al., 2017).

$$b_1 = 2.36 * d - 0.09 \qquad (5)$$

From the formulae, category boundary sharpness for both English musicians and non-musicians increased as stimulus duration increased, but English musicians showed a steeper slope, indicating a faster increment of sharpness of category boundary with increased duration.

3.2 Category boundary location

For English musicians, a regression model with an extra quadratic term turned out to be significantly different than the model with only the slope and intercept terms, as suggested by a likelihood ratio test (χ2(1) = 8.75, p = 0.003). Specifically, the formula for the category boundary (cb) and duration (d) is listed in Eq. (6). The chosen regression model was significant (F(2, 33) = 17.97, p < 0.001; adjusted $R^2 = 0.52$).

$$cb = 2158.59 * d^2 - 735.9 * d + 64.34 \tag{6}$$

However, for native English non-musicians, a model with a slope and intercept did not vary from a model with an extra quadratic form (χ2(1) = 0.05, p = 0.82). For comparison, the formula is listed in Eq. (7), and the regression model was also significant (F(1, 30) = 33.47, p < 0.001; adjusted $R^2 = 0.51$).

$$cb = -258.66 * d + 50.53 \tag{7}$$

English musicians showed earlier category boundary locations than English non-musicians for various durations, and their location of category boundary decreased faster than non-musicians in most cases.

3.3 Formulae

Formulae for speech production of rising and falling pitch directions have been proposed (Prom-on et al., 2009). The formulae were obtained by fitting simple linear regression models to production time and pitch changes (excursion size). Our formulae also show a linear relationship between stimulus duration required for perception and pitch changes (rising or falling). However, our model yielded a bigger R^2 than the value reported for the production results (Prom-on et al., 2009) (rising pitch perception: $R^2 = 0.26$; rising pitch production $R^2 = 0.3533$; falling pitch perception: $R^2 = 0.4797$; falling pitch production $R^2 = 0.10$). The following formulae were proposed for perception time required by both English musicians and non-musicians combined, where *t* is the duration (ms) required to perceive *d* st differences from level tones.

Rising, English listeners:

$$t = 122.67 + 3.79 * d \tag{8}$$

Falling, English listeners:

$$t = 109.04 + 4.75 * d \tag{9}$$

Compared with formulae proposed for pitch production, the time required to effectively perceive pitch directions was shorter than the time to produce them when the rise is greater than 6 st and the fall is greater than 8 st. In addition, across English musicians and non-musicians, a shorter duration was required to perceive a falling pitch than a rising pitch direction when the rise and fall were not greater than 14st. However, for one st increment, a longer time is needed to perceive a falling pitch direction than a rising pitch direction.

To explore the effects of music experience, we compared formulae obtained for each pitch direction for musicians to those obtained for non-musicians in Chen et al. (2017). Again, *t* is the duration needed to perceive *d* st differences from level tones.

Rising, English musicians:

$$t = 108.84 + 4.64 * d \tag{10}$$

Falling, English musicians:

$$t = 101.97 + 5.16 * d \tag{11}$$

Rising, English non-musicians:

$$t = 137.33 + 1.89 * d \tag{12}$$

Falling, English non-musicians:

$$t = 144.6 + 1.5 * d \tag{13}$$

For rising pitch directions, musicians needed less time than non-musicians when the rise is less than 11st. For falling pitch directions, musicians needed less time than non-musicians when the fall is less than 12st. Moreover, we obtained formulae for each subgroup. When we submitted the intercept and slope values to the likelihood ratio tests, we found that English musicians and non-musicians had significantly different intercept and slope values ($\chi^2(2) = 12.18$, p = 0.002) with smaller intercepts and sharper slopes for English musicians. The intercept and slope values for falling and rising pitch also differed significantly ($\chi^2(2) = 10.81$, p = 0.045). Both intercept and slope values reached marginal significance for low and high vowels ($\chi^2(2) = 5.15$, p = 0.076). High vowels tended to have smaller slopes and intercepts, and thus required less time to be effectively perceived in comparison to low vowels.

4 Discussion

This study is the first to examine effects of stimulus duration and vowel quality on pitch perception by musicians. We found that musical experience significantly increases the sharpness of category boundary for both falling and rising pitch directions on low and high vowels across various duration values. The results were inconsistent with the findings reported in Zhao and Kuhl (2015), where no significant differences were found between musicians and non-musicians. However, category boundary was less sharp for English musicians than native Mandarin speakers as reported in Chen et al. (2017).

It is known that musical training enhances auditory processing of (fundamental) frequency and the enhancement can be transferred to speech processing, especially pitch processing in lexical tones (e.g. Kraus and Chandrasekaran, 2010). Musicians usually have higher precision in processing of pitch. However, it is yet unknown whether the precision found musicians require more time in perception, and whether musicians are less affected by intrinsic effects of vowels in perceiving tones. Our results show that increasing rate of category sharpness differs between musicians and non-musicians. Specifically, although category boundary sharpness increased with duration among both English musicians and non-musicians, English musicians showed a faster increment than non-musicians, indicating that musical experience draws greater benefits from extra stimulus duration. Moreover, English musicians showed earlier category boundary than English non-musicians for many duration values, and the location of category boundary decreased faster than non-musicians in most cases. Similar to studies done by Alexander et al. (2005) and Burnham et al. (2014), our results indicate that musical training can in fact promote pitch processing and categorical perception in tone languages. More importantly, superior pitch processing ability shortens the time needed for its categorization, leading to less time required for tone perception in the language domain.

As for minimal duration required to produce and perceive pitch direction, our results showed that the minimum time required to perceive pitch perception was usually shorter than what is needed to produce pitch direction, similar to the findings reported for Chinese and English listeners (Chen et al., 2017), lending further support to the claim that physical constraints affect speech production than perceptual constraints do on speech perception (Janse, 2003). However, the relationship holds only when the rise is greater than 6st and the fall is greater than 8 st, showing asymmetric thresholds for pitch rise and fall.

For rising pitch perception, musicians need less duration than non-musicians when the rise is less than 11 st. For falling pitch perception, musicians need shorter duration than non-musicians when the fall is less than 12 st. English musicians and non-musicians differed significantly in intercept and slope values, and English musicians tended to have smaller intercepts and sharper slopes, suggesting that musical ability heightens musicians' sensitivity to a rise and fall in pitch within 12 st range.

In addition, vowel quality significantly contributed to identification of tones and sharpness of category boundary in musicians and non-musicians. Vowel quality also plays a role in determining the time required for tone perception by musicians, where pitch directions on high vowels require less time to be accurately perceived. We found that onset of the rising pitch is perceived to be lower on a high vowel than on a low vowel, but the offsets are perceived to be similar. Thus, the perceived slope of a rising pitch on a high vowel is sharper and more different from a level pitch than on a lower vowel, leading to a shorter duration for its perception. We also found that a falling pitch on a high vowel requires less time to identify that on a low vowel because its perceived slope is shaper on a high vowel than on a low vowel. Specifically, while the onset values of a falling pitch on a low and on a high vowel is perceived to be equal, its offset values are perceived to be relatively lower on a high vowel than on a low vowel.

For non-musicians, vowel quality plays a less consistent role where rising pitch on a low vowel requires less time for perception (rise < 16st), but falling pitch on a low vowel requires more time.

Acknowledgements

This work was supported by the Early Career Scheme of Research Grants Council of the Hong Kong Special Administrative Region (grant number: 25602816) and Fund for ECS Project Rated 3.5 CRG (grant number: G-YBGK).

References

Alexander, Jennifer A., Patrick CM Wong, and Ann R. Bradlow. (2005). Lexical tone perception in musicians and non-musicians. In *Ninth European Conference on Speech Communication and Technology*, 397-400

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Boersma, Paul, and Kateřina Chládková. (2013). Detecting categorical perception in continuous discrimination data. *Speech Communication*, *55*(1), 33-39.

Burnham, Denis, Ron Brooker, and Amanda Reid. (2015). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychology of Music*, *43*(6), 881-897.

Chang, Daniel, Nancy Hedberg, and Yue Wang. (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.

Chen, Si, Yiqing Zhu, and Ratree Wayland. (2017). Effects of stimulus duration and vowel quality in cross-linguistic categorical perception of pitch directions. *PloS one*, *12*(7), e0180656.

Flege, James Emil, Murray J. Munro, and Robert Allen Fox. (1994). Auditory and categorical effects on cross-language vowel perception. *The Journal of the Acoustical Society of America*, *95*(6), 3623-3641.

Fujisaki, Hiroya, Takako Kawashima. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute*, *29*: 207-214.

Goldstone, Robert L., & Andrew T. Hendrickson. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69-78.

Janse, Esther. (2003). *Production and perception of fast speech* (Doctoral dissertation), Utrecht University.

Kraus, Nina, and Bharath Chandrasekaran. (2010). Music training for the development of auditory skills. *Nature reviews neuroscience*, 11(8), 599–605.

Kuhl, Patricia K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831.

Lee, Chao-Yang, and Tsun-Hui Hung. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustical Society of America*, *124*(5), 3235-3248.

Lehnert-Lehouillier, Heike. (2013). From long to short and from short to long: Perceptual motivations for changes in vocalic length. In Alan C. L. Yu (Ed.) *Origins of Sound Change: Approaches to Phonologization* (pp. 98–111). Oxford: Oxford University Press.

Locke, Simeon, and Lucia Kellar. (1973). Categorical perception in a non-linguistic mode. *Cortex*, *9*(4), 355-369.

Moulines, Eric, and Jean Laroche. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech communication*, *16*(2), 175-205.

Peng, Gang, Hong-Ying Zheng, Tao Gong, Ruo-Xiao Yang, Jiang-Ping Kong, and William S-Y. Wang. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, *38*(4), 616-624.

Pisoni, David B. (1976). Fundamental frequency and perceived vowel duration. *The Journal of the Acoustical Society of America*, 59(S1): S39–S39.

Prom-On, Santitham, Yi Xu, and Bundit Thipakorn. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, *125*(1), 405-424.

Raz, Israel, and J. F. Brandi, (1977). Categorical perception of nonspeech stimuli by musicians and nonmusicians. *The Journal of the Acoustical Society of America*, *62*(S1), S60-S60.

Siegel, Jane A., and William Siegel. (1977). Categorical perception of tonal intervals: musicians can't tell sharp from flat. *Perception & Psychophysics*, *21*(5), 399-407.

Treisman, Michel, Andrew Faulkner, Peter LN Naish, and Burton S. Rosner. (1995). Voice-onset time and tone-onset time: the role of criterion-setting mechanisms in categorical perception. *The Quarterly Journal of Experimental Psychology*, *48*(2), 334-366.

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

Whalen, Douglas H., and Andrea G. Levitt. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *23*(3), 349-366.

Wu, Han, Xiaohui Ma, Linjun Zhang, Youyi Liu, Yang Zhang, and Hua Shu. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Frontiers in psychology*, *6*, 436.

Xu, Yisheng, Jackson T. Gandour, and Alexander L. Francis. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063-1074.

Yu, Alan CL, Hyunjung Lee, and Jackson Lee. (2014). Variability in perceived duration: pitch dynamics and vowel quality. In *Proceedings of TAL-2014*, 41-44.

Zhao, T. Christina, and Patricia K. Kuhl. (2015). Effect of musical experience on learning lexical tone categories. *The Journal of the Acoustical Society of America*, *137*(3), 1452-1463.