Interaction of long-term acoustic experience and local context information on the perceptual accommodation of talker variability

Caicai Zhang[a, b], *Gang Peng*[a], *and William S-Y. Wang*[a]

[a] Language and Cognition Laboratory, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong
[b] Haskins Laboratories, Yale University, USA

Abstract:
How do listeners recover speech content from acoustic signals, given the immense variability between talkers? In this study, two experiments were conducted on Cantonese level tones, comparing the perception of multi-talker speech stimuli in isolation and within a speech context. Without prior knowledge of a talker's pitch range, listeners resort to the population-average pitch range as a default reference for perception. This effect is attested by the significant correlation between the distance from population-average pitch range and identification accuracy in the isolation condition ($r=-.24$, $p<0.01$). The closer a talker's pitch range is to the population-average, the higher the identification accuracy is. The population-average reference is gender-specific, showing separate accommodation scales for female and male talkers. Such default reference is presumably built from one's long-term acoustic experience, reflecting the dense distribution of talkers in a community whose pitch is close to the population-average. Above the effect of long-term experience, the presence of a speech context allows listeners to tune to talker-specific pitch range, boosting the identification accuracy from 43% (in isolation) to 86%. Our findings demonstrate that listeners have built-in knowledge of population-average pitch and can shift from the default reference to talker-specific reference with the facilitation of context information.

INTRODUCTION

In a speech community, there is an enormous amount of differences in the way that people talk. Due to physiological differences in the vocal folds and vocal tract length, the same word spoken by different talkers carries different acoustic characteristics as shaped by the vocal apparatus (Liberman et al., 1967, Johnson, 2005). The acoustic realizations of speech is further modulated by the speaking rate (Kessinger and Blumstein, 1998) and emotional status of the talker (Protopapas, 1997). Above the individual speaker level, a spoken language often has dialects and sub-varieties that are spoken by different groups of talkers depending on their socio-economic status.

In tone languages, fundamental frequency (F0) is used to determine the lexical meaning (Wang, 1967). Talkers differ physiologically in the vocal folds, which gives rise to variation in tone production (Rose, 1996). For example, female and male speakers show different F0 realizations of the same tone. Speaking F0 also varies between talkers of the same sex (Rose, 1996) and within the same talker across the day (Garrett and Healey, 1987).

Despite the talker variability in speech production, the listeners can understand the speech of different talkers without much difficulty. How the listeners perceptually accommodate talker variability is a fundamental question in speech perception (Johnson, 2005, Kuhl, 2011). For lexical tones, the important determinant of perception is not the absolute values of F0, but the relative F0 with reference to a particular talker's F0 range (Wong and Diehl, 2003, Francis et al., 2006, Peng et al., 2012, Zhang et al., 2012).

Estimation of a talker's F0 range can be obtained from the acoustic characteristics of a talker's voice and the external context. A previous study suggested that English listeners may be able to estimate the F0 range based on the voice quality of a stranger's voice (Honorof and Whalen, 2005). However, a recent study places some limitation on the role of voice quality, and suggests that listeners rely more on the absolute values of F0 in judging the location of the F0 in an unfamiliar talker's range (Bishop and Keating, 2012). These authors also found that listeners have different expectations of the average F0 of each sex. This finding highlights the importance of long-term auditory experience with female and male talkers in a speech community in speech perception.

Another important factor for F0 range estimation is the external context (i.e. what a talker said earlier). The listeners likely build a talker's F0 reference from the context. The context enables the listeners to adapt to a talker's F0 range, which serves as the talker reference for mapping the variable acoustic signals onto the invariant phonological category. Previous studies have found that the same word tended to be perceived as having a high tone when embedded in the context with lowered F0, and as having a low tone when embedded in the same context with raised F0 (Leather, 1983, Moore and Jongman, 1997, Wong and Diehl, 2003, Francis et al., 2006, Huang and Holt, 2009, Huang and Holt, 2011, Zhang et al., 2012). When the overall F0 of the context is raised or lowered, it requires the listeners update the talker F0 reference. As a result, the same speech signals are mapped onto different tone categories.

In this study, we examine the effects of the average F0 range of the population and the external context on lexical tone perception. The knowledge of the population-average F0 range is probably important for F0 judgment in both tone and non-tone languages. But in tone languages, F0 determines the acoustic form of lexical tones. The listeners' expectations of the likely F0 form of lexical tones could have been shaped by the long-term auditory experience with the F0 range of female and male speakers in a speech community. For example, a flat F0 at 160 Hz is likely to be a high level tone from a male speaker, but not very likely from a female speaker. Above the effect of the population-average F0 range, the external context provides a talker-specific F0 reference for estimating the relative F0 height of a word, which determines the tone category and the lexical meaning of a word. For example, a flat F0 at 160 Hz could be the high level tone or the mid level tone from a male talker, which depends on the talker's F0 range as indicated by the F0 of the context.

Cantonese level tones are ideal for studying this question. There are three level tones in Cantonese, high level tone, mid level tone, and low level tone, which mainly differ in F0 height (Peng et al., 2012). A word with a flat F0 is ambiguous and can be mapped to any of these three tones (Peng et al., 2012, Zhang et al., 2012). We examine the perception of Cantonese level tones produced by talkers with different F0 ranges. If the population-average F0 range plays a role, it means that the closer a particular talker's F0 range is to the population-average reference, the more likely the words produced by this talker can be correctly identified. In particular, if a talker's F0 range is higher than the population-average reference, the words from this talker are likely to be misperceived as having the high level tone; if a talker's F0 range is lower than the population-average reference, the words from this talker are likely to be misperceived as having the low level tone. The external context that explicitly provides cues of a talker's F0 range is expected to boost up the identification accuracy. With the facilitation of the external context, the listeners are expected to correctly identify the words from all the talkers, no matter the talker's F0 range is higher or lower than the population-average reference.

Two perception experiments were conducted in this study. We compared the identification accuracy of multi-talker word stimuli presented in isolation and in the context. In Experiment 1, the target word occurred in the middle of the context; in Experiment 2, the target word occurred at the end of the context. Different positions of the target word in two experiments provide additional information about the effect of the external context. The population-average F0 range for female and male Cantonese speakers was separately obtained from a speech database (Lee et al., 2002). We calculated the correlation between the distance of the talkers' F0 range from the population-average reference and the identification performance.

METHODS

Participants

16 native speakers of Hong Kong Cantonese (8 female, 8 male; mean age = 22.6, s.d. = 2.6) participated in Experiment 1, and 18 native speakers of Hong Kong Cantonese (10 female, 8 male; mean age = 21.0, s.d. = 1.2) participated in Experiment 2. No subjects had hearing impairment or music training. All subjects gave informed consent in compliance with a protocol approved by the Survey and Behavioral Research Ethics Committee of The Chinese University of Hong Kong.

Stimuli

In Experiment 1, four native Cantonese speakers (2 female, 2 male) all in their early twenties were recruited to record the stimuli. These four talkers had different but overlapping F0 ranges (F01: 212.41~331.37 Hz; F02: 197.76~279.81 Hz; M01: 128.05~189.78 Hz; M02: 91.04~121.79 Hz; see Figure 1(a)). The F0 range of these four talkers was estimated from their production of two words in isolation, 醫 (/ji55/ 'a doctor') which carries the highest tone in Cantonese, and 兒 (/ji21/ 'a son') which carries the lowest tone. The upper F0 range was measured from the average F0 of six repetitions of 醫 /ji55/, and the lower F0 range was obtained from the average F0 of six repetitions of 兒 /ji21/.

Each of the four talkers was asked to read aloud a Cantonese sentence for six times, i.e., 請你讀意嚟聽下 /tsʰiŋ25 lei23 tuk22 ji33 lei21 tʰiŋ55 ha23/ 'please read /ji33/ for me'. One clear sentence was selected for each speaker and used in the perception experiment.

In Experiment 2, another four Cantonese speakers (2 female, 2 male) were recruited to make the recording (F03: 183.38 ~ 284.18 Hz; F04: 169.96 ~ 280.55 Hz; M03: 119.77 ~ 199.86 Hz; M04: 98.14 ~ 150.54 Hz; see Figure 1(b)). A meaningful sentence, i.e., 呢個字係意 /li55 ko33 tsi22 hɐi22 ji33/ 'This word is 'meaning'', was recorded from four talkers and one clear sentence was selected for each talker for the perception experiment.

In both experiments, 意 /ji33/ 'meaning' (mid level tone) was the target word, and the remaining part served as the context. In Experiment 1, the context appeared both before and after the target word. In Experiment 2, the context only occurred before the target word.

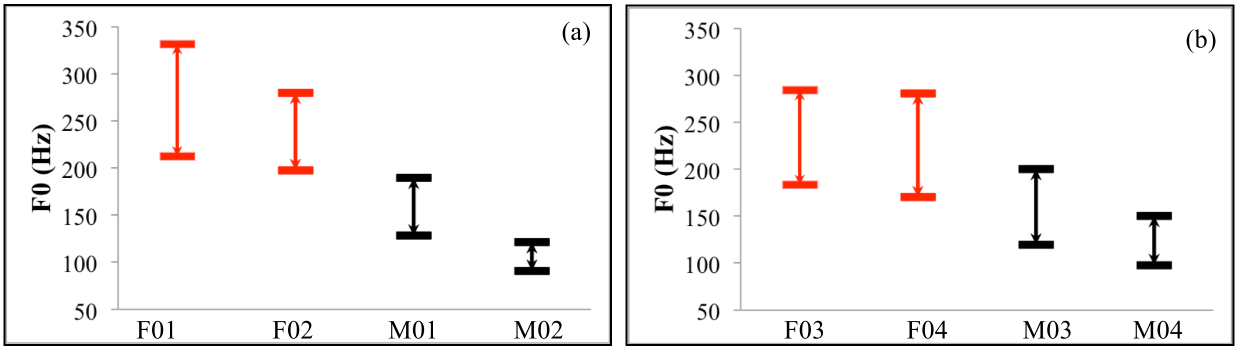

FIGURE 1. F0 range of four talkers in Experiment 1 (a) and Experiment 2 (b).

Procedure

In both experiments, the stimuli from four talkers were presented in two blocks, the isolation block and the context block. In the isolation block, the target word (i.e. /ji33/ produced by all four talkers) was extracted out of the sentence and presented in isolation. In the context block, the target word was presented with the context.

The task was three-alternative forced choice identification. Subjects were instructed to identify the target word as any of the three Cantonese words, 醫 (/ji55/ 'a doctor'), 意 (/ji33/ 'meaning'), and 二 (/ji22/ 'the second') by pressing labeled buttons on a computer keyboard as soon as possible. These three words correspond to high level tone, mid level tone and low level tone respectively, and differ exclusively in tones.

Data Analysis

The average F0 range of female and male Cantonese speakers in the population was measured from CUSENT, a database which includes read speech materials from 68 native Cantonese speakers (34 female, 34 male) in the training set (Lee et al., 2002). The upper and lower F0 range was measured from words carrying the highest and lowest tones read by all female and male speakers. Only sentence-initial words were included in order to minimize the effect of intonation. Average F0 range of female speakers was approximately 220~290 Hz, and average F0 range of male speakers was approximately 110~160 Hz. The eight talkers recruited in this study were compared with the average F0 range of female and male talkers separately and the distance was calculated in semitones. The calculation was done in semitones rather than in Hz to reflect the perceptual distance.

The behavioral performance in the two experiments was analyzed in terms of identification accuracy and perceptual height scores. For identification accuracy, the percentage that the target word was correctly identified was calculated for the isolation and the context condition respectively. The second analysis was to calculate the average perceptual height of all responses (i.e. not only the correct responses) in a condition. Following (Wong and Diehl, 2003), each identification response was coded according to the perceptual height of the selected tone. According to the phonological description of Cantonese tones, the high level tone is three semitones higher than the mid level tone, which is in turn two semitones higher than the low level tone (Chao, 1947). Following this perceptual scale, each high level tone response was coded as '6', each mid level tone response as '3', and each low level tone response as '1', with '6', '3' and '1' referring to the perceptual height of a tone (see Figure 2). The average perceptual height score was then obtained from all the responses in the isolation condition. If the average perceptual height was close to ‘1’, it indicated that the target word was mostly identified as having the low level tone. If it was close to ‘6’, it indicated that the target word was mostly identified as having the high level tone.

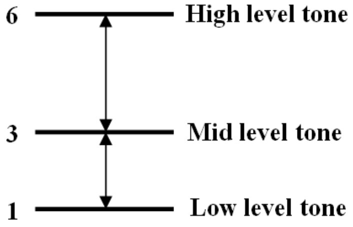

FIGURE 2. Coding scheme of the perceptual height of Cantonese level tones.

RESULTS

Experiment 1

Figure 3 shows the identification accuracy for each of the four talkers in the isolation and context condition in Experiment 1. Two-way repeated measures ANOVA was conducted on the identification accuracy by indicating *context* (isolation and context) and *talker* (F01, F02, M01, and M02) as two within-subjects factors. Greenhouse-Geisser correction was applied where the sphericity was violated. There were significant main effects of *context* (F(1, 15) = 91.212, p < 0.001), *talker* (F(1.844, 27.657) = 6.628, p < 0.01), and significant interaction effect of *context* by *talker* (F(3,45) = 6.385, p < 0.01). Post-hoc tests were conducted to analyze the interaction effect. One-way ANOVAs were conducted with *talker* as the factor in the isolation and context conditions separately. There was a significant main effect of *talker* in the isolation condition (F(3, 60) = 7.861, p < 0.001), but not in the context condition (F(3, 60) = 1.909, p = 0.138). It means that the identification accuracy varied among the four talkers in the isolation condition, but it was similar in the context condition. Without the context, the perception was interfered with the talker difference in F0 range, which gives rise to the variability in identification accuracy. Such interference was controlled via the facilitation of the external context, which elicited equally high identification accuracy.

In order to examine whether the variability in accuracy in the isolation condition is related to the expectation of population-average F0 ranges, bivariate correlation analysis was carried out between the accuracy and the distance of the four talkers’ F0 range from the population-average F0 range. Absolute values of the distance (i.e. no matter a talker’s F0 range is higher or lower than the population-average) were used as an indicator of the general deviance from the population-average reference. The correlation analysis was conducted for the lower F0 range and upper F0 range respectively. There was a significant negative correlation between the identification accuracy and the distance from the population-average for the lower F0 range (r = -0.327, p < 0.01). There was a trend of negative correlation for the upper F0 range, but it did not reach significance (r = -0.152, p = 0.232). It means that the smaller the distance of a talker’s lower F0 range was from the population-average, the higher the identification accuracy was.

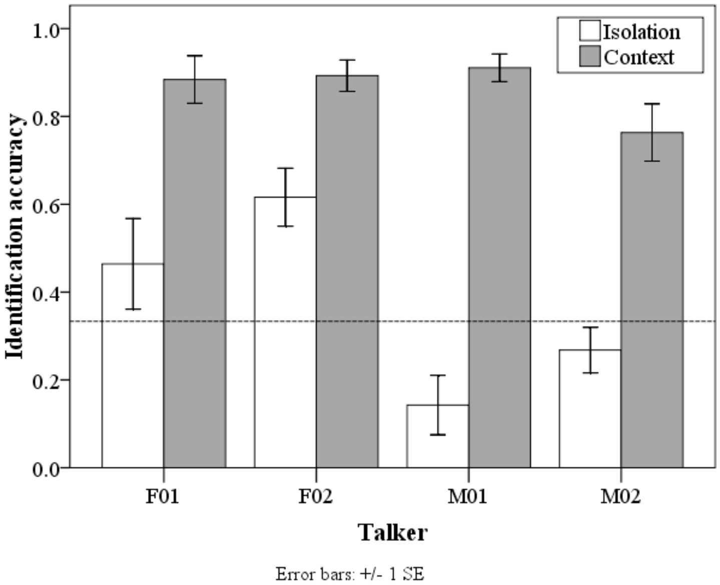

FIGURE 3. Identification accuracy for the four talkers in the isolation and the context condition in Experiment 1. The dotted line indicates the chance-level accuracy (0.33).

Experiment 2

Figure 4 shows the identification accuracy for the four talkers in the isolation and the context condition in Experiment 2. Two-way repeated measures ANOVA with *context* (isolation and context) and *talker* (F03, F04, M3, and M04) as two within-subjects factors obtained similar results as in Experiment 1. There were significant main effects of *context* (F(1, 17) = 145.626, $p < 0.001$), *talker* (F(3, 51) = 6.915, $p < 0.01$), and significant interaction effect of *context* by *talker* (F(3, 51) = 6.601, $p < 0.01$). Post-hoc one-way ANOVAs found that there was a significant effect of *talker* in the isolation condition (F(3, 68) = 8.46, $p < 0.001$), but not in the context condition (F(3, 68) = 1.055, p = 0.374).

Similar to Experiment 1, bivariate correlation analysis found that there was a significant negative correlation between the identification accuracy and the distance from population-average for the lower F0 range (r = -0.303, $p < 0.01$). The correlation was not significant for the upper F0 range (r = -0.002, p = 0.988). It means that the variability in identification accuracy among talkers in the isolation condition is related to the expectation of population-average F0 range. The closer the talker's lower F0 range was to the population-average, the more accurate the identification was.

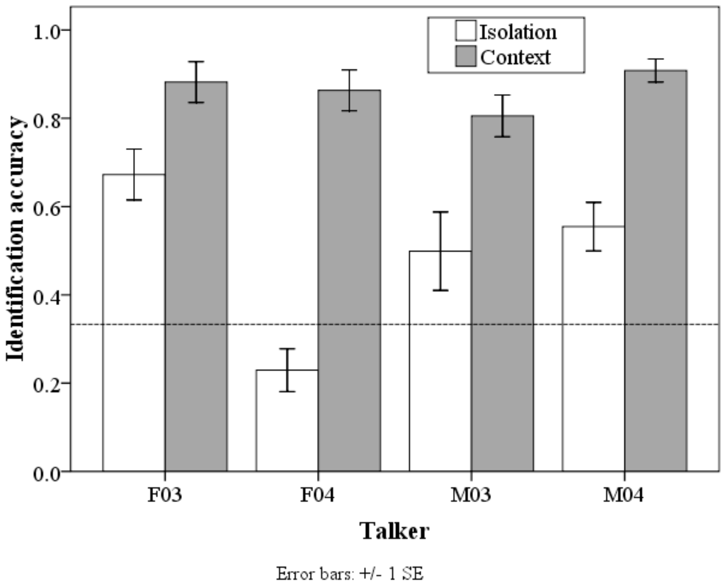

FIGURE 4. Identification accuracy for the four talkers in the isolation and the context condition in Experiment 2. The dotted line indicates the chance-level accuracy (0.33).

Combined Analysis of Experiment 1 and 2

Similar results obtained in two experiments corroborated to show that the listeners have expectations of the population-average F0 ranges. As a result, if a talker's F0 range is close to the population-average, the target word from that talker can be correctly identified without the facilitation of the external context. When an external context is present (no matter the context occurs both before and after the target word as in Experiment 1, or only before the target word as in Experiment 2), it explicitly provides cues of a particular talker's F0 range, boosting the mean identification accuracy to 86%.

In this subsection, the results of Experiment 1 and 2 were grouped together and input to the linear regression analysis. The purpose is to analyze the contribution of the population-average F0 range to the perceptual performance in the isolation condition. The dependent variable was the perceptual height scores. The independent variables were the distance of the eight talkers' lower and upper F0 range from the population-average. Figure 5 shows the trend of the perceptual height scores as a function of the distance if the eight talkers' lower and upper F0 ranges from the population-average reference.

If the listeners resort to the population-average F0 ranges for the perception of tones from unfamiliar talkers, the higher a talker's F0 range is compared to the population-average reference, the more likely the target word from that talker would be identified as having the *high level tone*. In that case, the perceptual height score would be close to '6'. If a talker's F0 range is lower than the population-average reference, the target word from that talker is like to be identified as having the *low level tone*. In that case, the perceptual height score would be close to '1'.

The regression model with the distance from the population-average reference as the predictors reached significance and accounted for 53.9% of the variance in the perceptual height scores ($r^2 = 0.539$, $p < 0.001$). Moreover, if the distance from the lower and upper F0 range was input to the model in a stepwise manner, the lower F0 range alone accounted for 50% of the variance in the data ($r^2 = 0.5$, $p < 0.001$), and adding the upper F0 range as a second predictor only accounted for an additional 3.9% of the variance.

In summary, the linear regression analysis further confirmed that lexical tone perception was influenced by the listeners' expectation of population-average F0 ranges. Moreover, the lower F0 range contributes more to the perceptual performance.

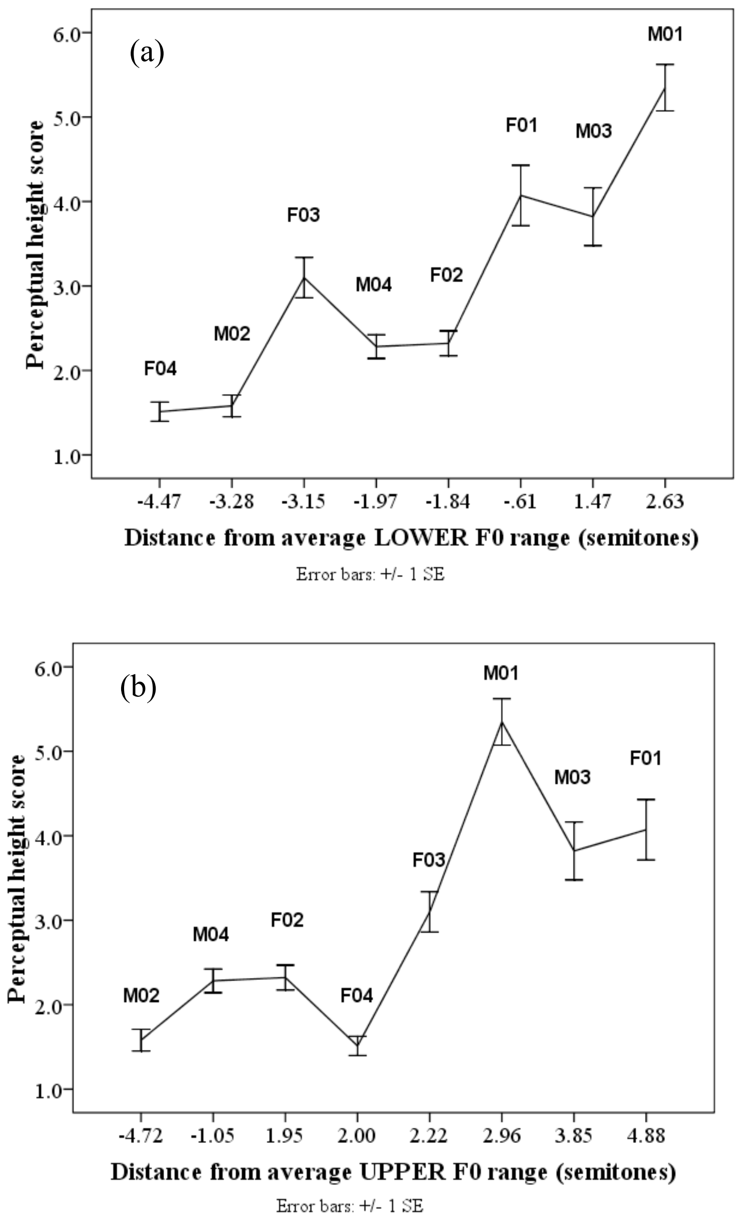

FIGURE 5. Perceptual height score in the isolation condition for eight talkers according to the distance from the population-average lower F0 range (a) and lower F0 range (b).

DISCUSSION

In this study, we have found that Cantonese listeners have built-in knowledge of population-average F0 ranges for female and male speakers, which shape the expectation of the F0 forms of lexical tones. When a particular talker's F0 range is unknown (as in the isolation condition), the population-average F0 range serves as the default reference. Therefore the identification accuracy is higher for those talkers whose F0 ranges are closer to the population-average. Moreover, for the talkers whose F0 range is higher than the population-average, the target word tends to be misperceived as having the high level tone; for the talkers whose F0 range is lower, the target word tends to be misperceived as having the low level tone. When the external context is present, the listeners can build the talker-specific reference from the context, thereby eliciting equally high identification accuracy for any talker. It indicates that the listeners can switch from the default reference to the talker-specific reference when the talker cues are available.

Our findings also imply that the mental representation of lexical tones is not abstract or deprived of phonetic details. Rather, the representation of tones encodes episodic traces of previous encountered examples of the tones. Long-term auditory experience with the speech of many different speakers in a community likely accumulates and adds to the listeners' internal knowledge of the population-average F0 range. It is possible that more talkers' F0 ranges are close to the population-average than those far away. As a result, the global distribution of F0 in a community sets the mental representation of tones to the F0 which most frequently occurs. Such episodic representation is formed separately for female and male speakers, who have different speaking F0 (Honorof and Whalen, 2005, Smith and Patterson, 2005, Bishop and Keating, 2012). This finding is consistent with the exemplar theory which suggests that the representation of phonological categories are episodic (Johnson, 1997).

It is intriguing that the lower F0 range is a better predictor of the perceptual performance than the upper F0 range. This result seems unexpected at first glance because the between-talker variability is usually greater at the upper F0 range than at the lower F0 range (Bishop and Keating, 2012, Keating and Kuo, 2012). It may mean that the upper F0 range is more indicative of the between-talker difference. However, the upper F0 range may also vary more within a talker. For example, how high the upper F0 range can reach is probably influenced by many factors such as emotional status. Therefore, the lower F0 range may be a more reliable indicator of a particular speaker's F0. The unequal effects of lower and upper F0 range on lexical tone perception are worth more investigation in future studies.

ACKNOWLEDGMENTS

We thank Dr. Patricia Keating for thoughtful discussions which inspired this study. This study is supported in part by grants from the National Basic Research Program of the Ministry of Science and Technology of China (973 Grant: 2012CB720700), from National Science Foundation of China (NSFC: 61135003, 11074267), and from Research Grant Council of Hong Kong (GRF: 455911).

REFERENCES

Bishop J., and Keating, P. (**2012**). "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," J. Acoust. Soc. Am. **131**, 1-13.

Chao, Y.-R. (**1947**). Cantonese Primer (Harvard University Press).

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., and Chu, P. C. Y. (**2006**). "Extrinsic context affects perceptual normalization of lexical tone," J. Acoust. Soc. Am. **119**, 1712-1726.

Garrett, K. L., and Healey, E. C. (**1987**). "An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day," J. Acoust. Soc. Am. **82**, 58-62.

Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**, 2193-2200.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**, 3983-3994.

Huang, J., and Holt, L. L. (**2011**). "Evidence for the central origin of lexical tone normalization (L)," J. Acoust. Soc. Am. **129**, 1145-1148.

Johnson, K. (**1997**). "Speech percepiton without speaker normalization: An exemplar model," In *Talker Variability in Speech Processing*, edited by K. Johnson, and J. W., Mullennix (Academic Press, San Diego), pp. 145-166.

Johnson, K. (**2005**). "Speaker normalization in speech perception," In *The handbook of speech perception* edited by D. B. Pisoni, and R. E. Remez (Blackwell Publishing), pp. 363-389.

Keating, P., and Kuo, G. (**2012**). "Comparison of speaking fundamental frequency in English and Mandarin," J. Acoust. Soc. Am. **132**, 1050-1060.

Kessinger, R. H., and Blumstein, S. E. (**1998**). "Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies," J. Phon. **26**, 117-128.

Kuhl, P. K. (**2011**). "Who's talking?" Science **333**, 529-530.

Leather, J. (**1983**). "Speaker normalization in perception of lexical tone," J. Phon. **11**, 373-382.

Lee, T., Lo, W. K., Ching, P. C., and Meng, H. (**2002**). "Spoken language resources for Cantonese speech processing," Speech Commun. **36**, 327-342.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**, 431-461.

Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864-1877.

Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., and Wang, W. S-Y. (**2012**). "The effect of inter-talker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems," J. Speech Lang. Hear. Res. **55**, 579-595.

Protopapas, A, and Lieberman, P. (**1997**). "Fundamental frequency of phonation and perceived emotional stress," J. Acoust. Soc. Am. **101**, 2267-2277.

Rose, P. (**1996**). "Cantonese citation tones," In *Vocal Fold Physiology: Controlling Complexity and Chaos,* edited by P. J. Davis, and N. H. Fletcher (Singular Pub. Group), pp. 307-324.

Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177-3186.

Wang, W. S. Y. (**1967**). "Phonological Features of Tone," Int. J. Am. Linguist. **33**, 93-105.

Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual Normalization for Inter- and Intratalker Variation in Cantonese Level Tones," J. Speech Lang. Hear. Res. **46**, 413-421.

Zhang, C., Peng, G., and Wang, W. S-Y. (**2012**). "Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones," J. Acoust. Soc. Am. **132**, 1088-1099.