JSLHR

Article

The Effect of Intertalker Variations on Acoustic–Perceptual Mapping in Cantonese and Mandarin Tone Systems

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

Purpose: This study investigates the impact of intertalker variations on the process of mapping acoustic variations on tone categories in two different tone languages.
Method: Pitch stimuli manipulated from four voice ranges were presented in isolation through a blocked-talker design. Listeners were instructed to identify the stimuli that they heard as lexical tones in their native language.
Results: Tone identification of Mandarin listeners exhibited relatively stable normalization regardless of the voice, whereas tone identification of Cantonese listeners was unstable and susceptible to the influence of intertalker variations. In the case of Cantonese listeners, intertalker variations had a larger effect on the perception of F0 height dimension than of F0 slope dimension.
Conclusion: The comparison between Cantonese and Mandarin listeners' performances reveals an interaction of intertalker variations and the types of tone contrasts in each language. For Cantonese tones, which depend heavily on F0 height distinctions, intertalker variations result in F0 overlapping and, consequently, ambiguities among them in isolated tone perception. For Mandarin tones, which are distinctive in terms of their F0 contours, the differences in F0 contours alone seem sufficient to elicit reliable tone identification. Intertalker variations therefore have relatively limited effect on Mandarin tone perception.

Key Words: tone perception, talker normalization, fundamental frequency, Cantonese, Mandarin

Speech sounds are characterized by great intra- and intertalker variations, yet listeners usually map acoustic variations to the appropriate linguistic categories without much difficulty. In tone categorization, discrete and invariant categories are extracted from continuous acoustic variations, where talker normalization must play an important role. Moreover, listeners with different native language backgrounds (i.e., first-language background) may attach unequal importance to each tone feature in tone categorization. Gandour (1983) addressed the question of how native language background (e.g., Cantonese, Mandarin, Taiwanese, Thai, and English) affects a listener's perception and discrimination of tones. He reported that listeners attached different importance to two underlying dimensions: height (average height of a pitch contour) and direction (direction of F0 movement within a syllable), depending on the tone contrasts in their native language. For example, Cantonese listeners placed more emphasis on the height dimension than Mandarin listeners did, possibly because three of the six tones in Cantonese contrast a relatively flat pitch contour at different pitch heights. On the other hand, Mandarin tones contrast with each other in terms of direction of F0 movement (see Table 1 for an illustration of Cantonese and Mandarin tones).

Comparing Mandarin and English listeners' performances in pitch perception tasks of speech and nonspeech sounds, Bent, Bradlow, and Wright (2006) found that the effect of native language background extended to nonspeech processing under certain stimulus and task conditions. Moreover, Mandarin listeners tended to make more errors in a pitch contour identification task, which could be related to the specific features of Mandarin tones.

[a]The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, People's Republic of China
[b]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
Correspondence to Gang Peng: gpeng@ee.cuhk.edu.hk
Editor: Anne Smith
Associate Editor: Patrick Wong
Received January 26, 2011
Revision received June 13, 2011
Accepted August 8, 2011
DOI: 10.1044/1092-4388(2011/11-0025)

Table 1. Cantonese and Mandarin tone systems.

Tone	Description	Example	Word	Gloss
		Cantonese		
Tone 1	High level	/i/ 55	醫生	doctor
Tone 2	High rising	/i/ 25	倚仗	to rely on
Tone 3	Mid level	/i/ 33	意思	meaning
Tone 4	Low falling	/i/ 21	兒科	son
Tone 5	Low rising	/i/ 23	耳窿	ear
Tone 6	Low level	/i/ 22	二胡	second
		Mandarin		
Tone 1	High level	/i/ 55	医治	doctor
Tone 2	High rising	/i/ 35	移动	to move
Tone 3	Low falling–rising	/i/ 214	倚仗	to rely on
Tone 4	High falling	/i/ 51	意外	surprised

Note. In phonological description, Cantonese contrasts six lexical tones on open syllables, and another three lexical tones on checked syllables (Bauer & Benedict, 1997), whereas Mandarin has four lexical tones. (Mandarin also has a neutral tone, but this tone does not provide lexical contrast.) Given that this study focuses on open syllables, only the six long lexical tones in Cantonese, and the four lexical tones in Mandarin, are listed here. The second column describes the F0 changes within a syllable. High, mid, and low indicate average F0; level, rising, and falling specify the direction of F0 movement. In the third column, two digits next to /i/ refer to tone transcriptions in Chao's tone letters (Chao, 1930). The fourth column contains a list of Chinese words that were used in the production test. The first Chinese character corresponds to the syllable in the third column.

In addition to the influence of native language background, sentential context has been found to affect perception of vowels and tones. An early study by Ladefoged and Broadbent (1957) investigated perception at the segmental level. Synthetic /bVt/ words were embedded at the end of a carrier sentence with various ranges of F1 and F2, and listeners were asked to identify the target vowel. They found that identification of the target vowel depended on the relative F1 distance between the target vowel and the carrier sentence. For example, the target vowel was more often identified as /ɪ/ when the carrier sentence had a relatively high F1, but it was identified as /ɛ/ when the carrier had a relatively low F1 (/ɪ/ has a lower F1 than /ɛ/)—a contrastive effect.

With regard to the process of tone perception, phonetic context, which contains information of a talker's F0 range, must also play an important role. It is highly likely that phonetic context facilitates talker adaptation (i.e., the process through which listeners can recognize the same words spoken by different talkers despite the great acoustic variations; e.g., Johnson, 2005), thereby allowing invariant categories to be extracted from talker variations.

Concerning the perception of lexical tones, Lin and Wang (1984) used both synthetic and natural speech to investigate whether identification of a target tone varies depending on the relative F0 height of the context. Different from Ladefoged and Broadbent's design, in which the last syllable of the sentence was the target, Lin and Wang asked Chinese subjects to identify the first syllable (which carried a flat pitch contour) in a disyllabic phrase. They found that, when subjects listened to cross-sliced natural speech (recombination of the first syllable and a separately produced second syllable with increased F0), identification of the target syllable shifted from a high tone (i.e., Tone 1; T1) to a low tone (i.e., Tone 3 [T3]) as the F0 height of the second syllable was increased.

Despite the relatively small number of subjects (10 Chinese subjects) and the lack of statistical analyses in this study, the general pattern reported by Lin and Wang (1984) revealed a similar phenomenon to that of Ladefoged and Broadbent's study: Subjects seemed to base their identification on the relative distance of the target from the context in terms of acoustic parameters such as F1 and F0.

Recent studies on tone perception have also found that the identification of pitch stimuli is affected by F0 information of the context in a contrastive way (i.e., identical targets tend to be perceived as low tones if the contextual F0 is high and as high tones if the contextual F0 is low [Francis, Ciocca, Wong, Leung, & Chu, 2006; Huang & Holt, 2009; Moore & Jongman, 1997; Wong & Diehl, 2003]).

Apart from native language background and sentential context, another factor related to tone categorization is talker variability. For example, Wong and Diehl (2003) looked at talker variations in the perception of Cantonese level tones. They found that different types of stimulus presentation affected the process of talker adaptation. They also found significantly higher accuracy in the blocked-talker design (only one talker in a block), in comparison with the mixed-talker design (more than one talker presented in a block), which suggests that blocked-talker presentation builds up talker expectancy, thereby assisting listeners' adaptation to a talker's voice. However, this article concentrated only on Cantonese level tones, which makes it difficult to generalize its findings to other Cantonese tones.

As discussed earlier, several factors such as native language background, sentential context, and talker variability play a role in tone identification. This study aims to investigate the influence of intertalker variations on tone identification for two groups of listeners having distinct native tone language background. Previous studies of the influence of native language background on tone perception have focused mainly on the contrast between native tone language and non-native tone language (Gandour, 1983; Hallé, Chang, & Best, 2004;

Wang, 1976; Xu, Gandour, & Francis, 2006). There is still a lack of studies that examine the impact of intertalker variation on tone identification under different configuration of tone inventories.

The present study investigates how different native language backgrounds modulate the mapping of acoustic variations on to linguistic tone categories and the impact of intertalker variations on this process. On the basis of the theory of adaptive dispersion, a phonological inventory should evolve toward a state of *maximal perceptual contrast* to reduce confusion among phonological categories (Liljencrants & Lindblom, 1972). Lindblom (1986) later developed the principle of *sufficient perceptual contrast,* which is more consistent with existing natural language systems. The four Mandarin lexical tones are distributed evenly in the tone space (Peng, 2006). Such tone distribution follows well the principle of sufficient perceptual contrast. On the other hand, the Cantonese six long lexical tones are distributed unevenly, with five of the six tones crowded into the lower part of the tone space, indicating that the Cantonese tone system violates the principle of sufficient perceptual contrast. Indeed, there are ongoing changes for Cantonese tones, mainly reflected by two tone mergers: one merger of T3 and T6 and another of T2 and T5 (Bauer, Cheung, & Cheung, 2003). Moreover, Cantonese has three level tones, making them internally ambiguous when perceived in isolation. However, Cantonese listeners are able to communicate with other talkers without difficulty. We speculate that Cantonese listeners may rely more on other information, such as sentential context and talker identity, to achieve effective tone identification. We expect that, when required to identify tones in isolation without any context, Cantonese listeners would be affected more by talker variation than Mandarin listeners.

The design of this study is inspired partly by Gandour's (1983) tone perception study and the blocked-talker design in Wong and Diehl (2003). One of the main purposes of this study is to examine whether listeners can categorize varying pitch stimuli into their corresponding linguistic categories without context or prior knowledge of the talkers' voice. In the ideal case, if listeners can estimate the upper and lower F0 bounds of a particular voice within a block, listeners should be able to map each pitch stimulus on to the corresponding tone category based on its relative position in that talker's F0 range. To address this question, acoustic stimuli are presented to listeners in isolation through a blocked-talker design. All talkers are unfamiliar to the listeners, meaning that the listeners have no prior knowledge of each talker's F0 range.

To capture intertalker variations, in this study we define four voice ranges (two male, two female) based on the survey of a speech database (CUSENT; T. Lee, Lo, Ching, & Meng, 2002). These comprise two normal voice ranges, which generally match the average F0 ranges for male and female speakers in the database, and two marginal voice ranges, one female voice that was higher than average and one male voice that was lower than average.

In summary, this study aims to answer the following questions: (a) Are native tone language listeners able to categorize acoustic variations into stable linguistic tone categories without contextual information and prior knowledge of talkers' information? (b) How do intertalker variations affect the process of tone identification in isolation? (c) Do intertalker variations have an equivalent effect on tone identification for Cantonese and Mandarin listeners?

Method

Stimuli

The pitch stimuli used in this study were naturally produced syllables with synthetic F0 contours superimposed on them. To model acoustic variations in tone perception, we designed 25 pitch stimuli following Chao's tone letters (Chao, 1930). According to Chao, a tone can be described by two (or more) digits, the first digit corresponding to the onset of the tone and the last digit to the offset. (For a three-digit tone, the intermediate digit refers to the F0 of the turning point of the pitch contour.) A digit can be any of the five integers from 1 to 5, with 1 defining the lowest pitch and 5 defining the highest. Two-digit tones capture level, rising, and falling contours, and three-digit tones are either concave or convex. For example, Mandarin Tone 4, which is a high falling tone, is transcribed as 51 in Chao's tone letters.

The test stimuli in this study include only two-digit pitch contours. Allowing each digit to be any of the five integers gives rise to an exhaustive combination of 25 stimuli ($5 \times 5 = 25$; i.e., 11, 12, 13, 14, 15, 21, 22, … 51, 52, 53, 54, and 55). Three-digit pitch contours are not considered in this study because the great majority of tones in Cantonese and Mandarin are level, rising, or falling. Although Mandarin T3 is traditionally described as a falling–rising contour 214, in continuous speech, especially on non-phrase-final positions, it is mainly realized as a low falling tone (except when it is followed by another T3 syllable, where tone sandhi changes the first T3 syllable to a rising contour; Wang & Li, 1967).

The motivation to design 25 pitch stimuli, instead of using merely the native pitch contours in Cantonese and Mandarin, is to investigate how a common set of acoustic variations is categorized into discrete classes, depending on the available tone contrasts in one's native language. To draw an analogy from color perception, it has been found that the color spectrum is categorized in a way

modulated by one's native language background (Regier, Kay, & Khetarphal, 2009). Similarly, in this study, we examine how long-term exposure to Cantonese and Mandarin affects one's categorization of acoustic variations.

Figure 1 shows the F0 trajectories of the pitch stimuli over time. Each F0 contour is a linear ramp. As can be seen from Figure 1, each stimulus lasts for 500 ms, with F0 kept constant during the first 100 ms.

Figure 2 shows the two-dimensional tone space (height and slope) in which each circle represents one pitch stimulus. The position of each stimulus in this space is determined by its height and slope, which are calculated from the formula in Equation 1. For example, for a high falling Stimulus 51, its corresponding height and slope are 3 and –1, respectively. The height dimension ranges from 1 to 5, and the slope ranges from –1 to 1. For the formula of perceptual height, what we intend to obtain is the average height of onset and offset. For the formula of perceptual slope, we intend to calculate the perceptual distance between the offset and onset, which ranges from –4 to 4 (i.e., maximally different onset and offset are 5 and 1, respectively). The aforementioned perceptual distance is then scaled down, ranging from –1 to 1, by further dividing the perceptual distance by 4.

$$\begin{aligned} Height &= (\text{onset} + \text{offset})/2 \\ Slope &= (\text{offset} - \text{onset})/4, \end{aligned} \quad (1)$$

where onset corresponds to the first digit in Chao's tone letters and offset corresponds to the last digit.

To investigate the effect of intertalker variations, we selected four voice ranges (two male, two female) on the basis of a survey of CUSENT, a database that includes read speech materials from 68 native Cantonese speakers in the training set (34 male, 34 female; T. Lee et al., 2002). First, we calculated the average F0 ranges in the database for both male and female speakers. To minimize the effect of intonation, we considered only F0 values of the first syllable of each utterance. We then selected the voices of two talkers (one male and one female), which generally matched the average voice range for each gender. Second, two marginal voice ranges (one male and one female) were defined, with F0 values higher than the female average by 2 standard deviations for the female high voice and lower than the male average by 2 standard deviations for the male low voice. The voices of two other talkers compatible with the defined marginal voice ranges were then selected.

Figure 1. A schematic representation of F0 trajectories of the designed pitch stimuli. First digit in a pitch contour can be any of the numbers from 1 to 5, and the last digit can also be any of these 5 numbers. For each stimulus, total duration of a pitch stimulus is 500 ms, and F0 is kept constant during the first 100 ms.

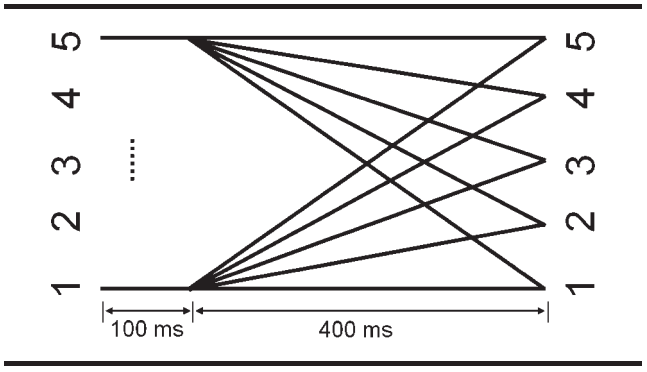
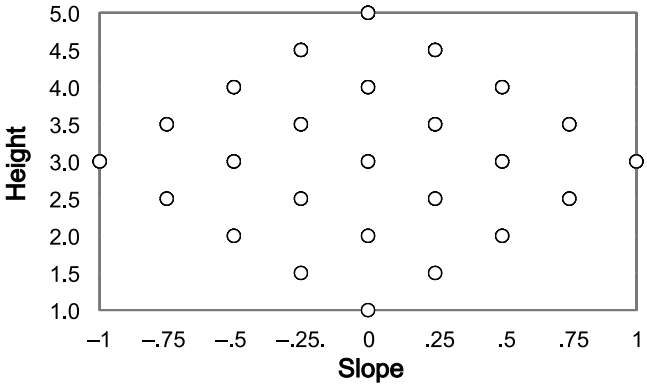


Figure 2. Two-dimensional space (height and slope) for 25 pitch stimuli. Each circle represents one stimulus.



The F0 range for each voice was as follows: female high (FH) voice, 240–350 Hz; female average (FA) voice, 200–290 Hz; male average (MA) voice, 110–160 Hz; male low (ML) voice, 85–125 Hz. For each voice, its F0 range covered approximately 0.54 octaves. These four voices form a continuum from high to low in terms of absolute F0, with some overlap between the two voices with the same gender.

For each talker, one sample of syllable /i/ with T3 was extracted from the database, and synthesized pitch contours were superimposed on this syllable (only the stable portion of syllable /i/ was used, and then time was normalized to 500 ms). Mean F1, F2, and F3 were averaged from values at the middle 30% of the whole syllable (35%–65% of the normalized duration, both ends included; Table 2). With tonal difference ignored, the syllable /i/ exists in both the Cantonese and Mandarin phonological systems.

As mentioned earlier, 25 pitch stimuli were defined in terms of tone digits without specifying F0 values. To accommodate the pitch stimuli to each voice range, we aligned the upper bound of each talker's voice to Tone Digit 5 and the lower bound to 1. After fitting pitch stimuli to the voice range, we divided the F0 range of each voice equally on a log scale into five levels (1–5) and generated 25 pitch stimuli for each voice, giving rise to

Table 2. Formant frequencies of syllable /i/ for four talkers selected in this study.

	F1		F2		F3	
Voice	Frequency	*SD*	Frequency	*SD*	Frequency	*SD*
Female high	542.11	10.55	2,778.25	29.96	3,704.27	78.70
Female average	444.10	20.10	2,927.26	51.31	3,512.37	32.08
Male average	318.85	2.43	2,245.11	6.19	3,015.33	188.04
Male low	322.99	2.37	2,086.75	6.17	2,944.03	4.48

Note. Mean F1, F2, and F3 values were obtained by averaging F1, F2, and F3 values at sampling points within the middle one third of the whole syllable. Values in the parentheses refer to SDs.

100 stimuli in total (4 voices × 25 stimuli = 100). While the pitch contour was modulated, the formant frequencies of syllable /i/ were kept intact, preserving the segmental characteristics as well as gender differences. The intensity profile was kept constant across the 100 stimuli, as our focus here is just on whether listeners can normalize intertalker variations in F0.

Subjects

Sixteen Cantonese subjects (eight male, eight female; mean age = 20.4 years, $SD = 0.78$) and 16 Mandarin subjects (eight male, eight female; mean age = 19.8 years, $SD = 1.34$) were paid to participate in the experiment. All subjects were undergraduates at The Chinese University of Hong Kong. No subject reported hearing impairment or long-term 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.

The 16 Cantonese subjects were all native speakers of Hong Kong Cantonese and were born and raised in Hong Kong. They had limited knowledge of Mandarin, which was acquired mainly through Mandarin language courses that they took during high school. The 16 Mandarin subjects were all native speakers of Mandarin, in addition to various other Chinese dialects (other than Cantonese). Most Mandarin subjects also had some exposure to Cantonese, having lived in Hong Kong for 2 to 3 years by the time of the experiment.

Tasks

There were two tasks for each subject: tone production and tone perception. It took approximately 1 hr to finish both tasks.

In the production task, we asked a subject to read aloud the first syllable of a list of meaningful disyllabic words. The list contrasted all lexical tones in the subject's native language (six words for Cantonese subjects and four words for Mandarin subjects; the word lists are shown in Table 1). For each tone, six repetitions were recorded and saved for analysis.

In the perception task, the same set of stimuli was presented to both Cantonese and Mandarin subjects.[1] A blocked-talker design was adopted to facilitate talker normalization. A practice block containing a voice that did not occur in the subsequent test blocks was presented first to familiarize subjects with the procedures. The results from the practice block were excluded from the analysis. Afterward, four test blocks were presented, each block containing stimuli from a single voice. Moreover, at the beginning of each block, one repetition of each prototypical pitch stimulus[2] in that particular voice was presented to facilitate talker normalization; this was also excluded from the analysis.

Within each block, all 25 pitch stimuli were presented in random order and repeated nine times. In other words, each subject listened to 900 test stimuli in total (25 stimuli × 9 repetitions × 4 voices = 900). Each pitch stimulus was presented in isolation; after the stimulus was presented, the subject had 3 s in which to make a choice by pressing labeled buttons on a keyboard. This study used a forced-choice identification design. Cantonese subjects were required to identify the heard stimulus as any of six labeled words, the same words as those in the production test. Similarly, Mandarin subjects were required to identify the heard

[1]Although the selection of voices is based on a database of Cantonese speakers, we believe that such a selection is generally fair for both Cantonese and Mandarin subjects. Given that Cantonese- and Mandarin-speaking populations share similar physical characteristics (e.g., body size), it is expected that the voice range extracted from the Cantonese database is, in general, compatible with that of Mandarin speakers. This point is also confirmed by the production data collected from Cantonese and Mandarin subjects in this study.

[2]As mentioned earlier, pitch stimuli were designed based on Chao's tone letters (i.e., 11, 12, 13, 14, 15, . . . 51, 52, 53, 54, and 55). Prototypical pitch stimuli in a particular voice are therefore defined as those that exactly match the phonological descriptions. Prototypical pitch stimuli in Cantonese are as follows: T1, 55; T2, 25; T3, 33; T4, 21; T5, 23; and T6, 22. Prototypical pitch stimuli in Mandarin are as follows: T1, 55; T2, 35; T3, 21; and T4, 51.

stimulus as any of four labeled words. Subjects were instructed to press the button as quickly and accurately as possible. The order of blocks was counterbalanced across subjects.

Data Analysis

The purpose of analyzing the production data is to estimate the F0 range of each subject and to make sure that the F0 ranges of the two groups of subjects were generally comparable. In this study, native Cantonese- and Mandarin-speaking subjects listened to the same set of stimuli, which comprised four voice ranges derived from a Cantonese speech database. Although we consider that these voice ranges were fair for both Cantonese and Mandarin speakers, a comparison of the voice ranges of the Cantonese and Mandarin subjects recruited in this study (as a sampling of the Cantonese- and Mandarin-speaking population) allows us to confirm the validity of the design of the intertalker variations. Moreover, when first exposed to unfamiliar voices, it is possible that the listeners may resort to their own F0 ranges as a reference. Comparing the F0 ranges of Cantonese and Mandarin subjects recruited in this study therefore allows us to determine whether the talker normalization conditions are fair for both groups of subjects.

In this study, F0 range is estimated with reference to the F0 values of the highest and lowest tones in a subject's native tone system. For the Cantonese tone system, the highest pitch was calculated from tokens of T1, whereas the lowest pitch was calculated from the lowest portion of T4; for the Mandarin tone system, the highest pitch was calculated from the highest portion of tokens of T4, whereas the lowest pitch was from the lowest portion of T3. Such a selection is based on inspection of phonetic data reported in earlier studies (Bauer et al., 2003; Xu, 1997). We report the F0 ranges measured from 16 Cantonese and 16 Mandarin subjects in Table 3.

A Language Group (Cantonese, Mandarin) × F0 Bounds (upper, lower) repeated measures analysis of variance (ANOVA) was conducted to determine whether Cantonese and Mandarin subjects differ significantly in their F0 bounds. There was neither a significant main effect of language group, $F(1, 30) = 0.40$, $p = .533$; nor a two-way interaction, $F(1, 30) = 1.50$, $p = .23$; suggesting that these two groups of subjects do not differ in their upper or lower F0 bounds and thereby share a similar F0 range.

For the analysis of identification responses, the center of gravity (CG) of each tone category was calculated as the centroid of the response distribution, following the formula in Equation 2. This formula allows us to transcribe the perceptual responses for each tone category

Table 3. Estimated fundamental frequency (F0) range (Hz) of Cantonese and Mandarin subjects according to their tone production in isolation.

Male	Lower	Upper	Average	Female	Lower	Upper	Average
			Cantonese subjects				
1	116.17	241.74	178.96	1	162.15	264.61	213.38
2	110.32	192.42	151.37	2	169.06	298.72	233.89
3	84.57	139.17	111.87	3	159.37	276.94	218.16
4	100.15	193.23	146.69	4	172.17	324.99	248.58
5	87.85	154.23	121.04	5	182.26	281.49	231.88
6	91.87	152.2	122.04	6	165.86	257.42	211.64
7	84.05	157.21	120.63	7	163.45	283.33	223.39
8	79.86	140.07	109.97	8	150.14	247.13	198.64
M	94.36	171.28	132.82		165.56	279.33	222.45
			Mandarin subjects				
1	96.24	131.16	113.70	1	168.9	271.29	220.10
2	97.56	178.43	138.00	2	155.31	267.22	211.27
3	87.87	134.77	111.32	3	160.83	257.94	209.39
4	93.38	181.58	137.48	4	146.93	238.52	192.73
5	113.64	175.07	144.36	5	151.16	284.74	217.95
6	83.62	140.41	112.02	6	143.72	224.41	184.07
7	99.81	184.94	142.38	7	161.28	277.61	219.45
8	82.36	138.06	110.21	8	163.43	253.85	208.64
M	94.31	158.05	126.18		156.45	259.45	207.95

Note. "Lower" and "upper" refer to the estimated F0 range of a subject.

with two numbers (onset and offset) comparable with that of Chao's tone letters.

$$G_{Onset} = (1 \times \sum_{j=1}^{5} R_{1,j} + 2 \times \sum_{j=1}^{5} R_{2,j} + 3 \times \sum_{j=1}^{5} R_{3,j} + 4 \times \sum_{j=1}^{5} R_{4,j} + 5 \times \sum_{j=1}^{5} R_{5,j}) / \sum_{i=1}^{5} \sum_{j=1}^{5} R_{i,j}$$

$$G_{Offset} = (1 \times \sum_{i=1}^{5} R_{i,1} + 2 \times \sum_{i=1}^{5} R_{i,2} + 3 \times \sum_{i=1}^{5} R_{i,3} + 4 \times \sum_{i=1}^{5} R_{i,4} + 5 \times \sum_{i=1}^{5} R_{i,5}) / \sum_{i=1}^{5} \sum_{j=1}^{5} R_{i,j} \quad (2)$$

CG was calculated for both onset (G_{onset}) and offset (G_{offset}) of each tone. $R_{i,j}$ refers to the number of times that a pitch stimulus is identified as a particular tone category, and i and j refer to the first digit and last digit, respectively, of the pitch contour.

For the formulae in Equation 2, the onsets and offsets are weighted according to the transcription of a stimulus in Chao's tone letters (i.e., 1, 2, 3, 4, and 5), which reflects the phonetic representation of a tone and captures a tone's perceptual property (Chao, 1930). Specifically, 1 in Chao's tone letters corresponds to low pitch, and 5 corresponds to high pitch. For example, for the Stimulus 11, which represents a low level pitch, both its onset and offset are weighted as 1.

For example, to calculate the onset CG of Mandarin Tone 1, the value for Tone Digit 1 (i.e., 1) is multiplied by the number of times that pitch stimuli with onset 1 (i.e., 11, 12, 13, 14, and 15) were identified as T1, the value for Tone Digit 2 (i.e., 2) is multiplied by the number of times that pitch stimuli with onset 2 (i.e., 21, 22, 23, 24, and 25) were identified as T1, and so on. The aforementioned five products are then summed and divided by the total number of times that T1 was identified. The offset CG of each tone category is calculated in a similar way, by multiplying a tone digit with the number of times that pitch stimuli ending with that digit were identified as a particular tone.

Having calculated the subject's perceptual responses to each tone, we calculated the perceptual height and slope, which we refer to as the *perceptual CG,* using the formula in Equation 1 given earlier. We then conducted a two-way repeated measures ANOVA for both Cantonese and Mandarin. Four such analyses were performed, with the dependent variable being either the height or the slope of the perceptual CG calculated for each tone category.[3]

[3]In the two-way repeated measures ANOVA, height or slope of the perceptual CG calculated for each tone was input to the analysis. There are cases in which one or more subjects failed to label any heard stimuli as a particular tone, which led to missing values in the input data. Before inputting the data to statistical analysis, missing values are replaced by the average value from the remaining subjects for that particular tone.

In each analysis, there were two independent variables: voice (FH, FA, MA, and ML) and tone (six tones for Cantonese and four tones for Mandarin). Results of the statistical analysis are discussed in the following section. Corrections for violations of sphericity were made, where appropriate, with the Greenhouse–Geisser method.

Results

Categorization of Cantonese and Mandarin Tones

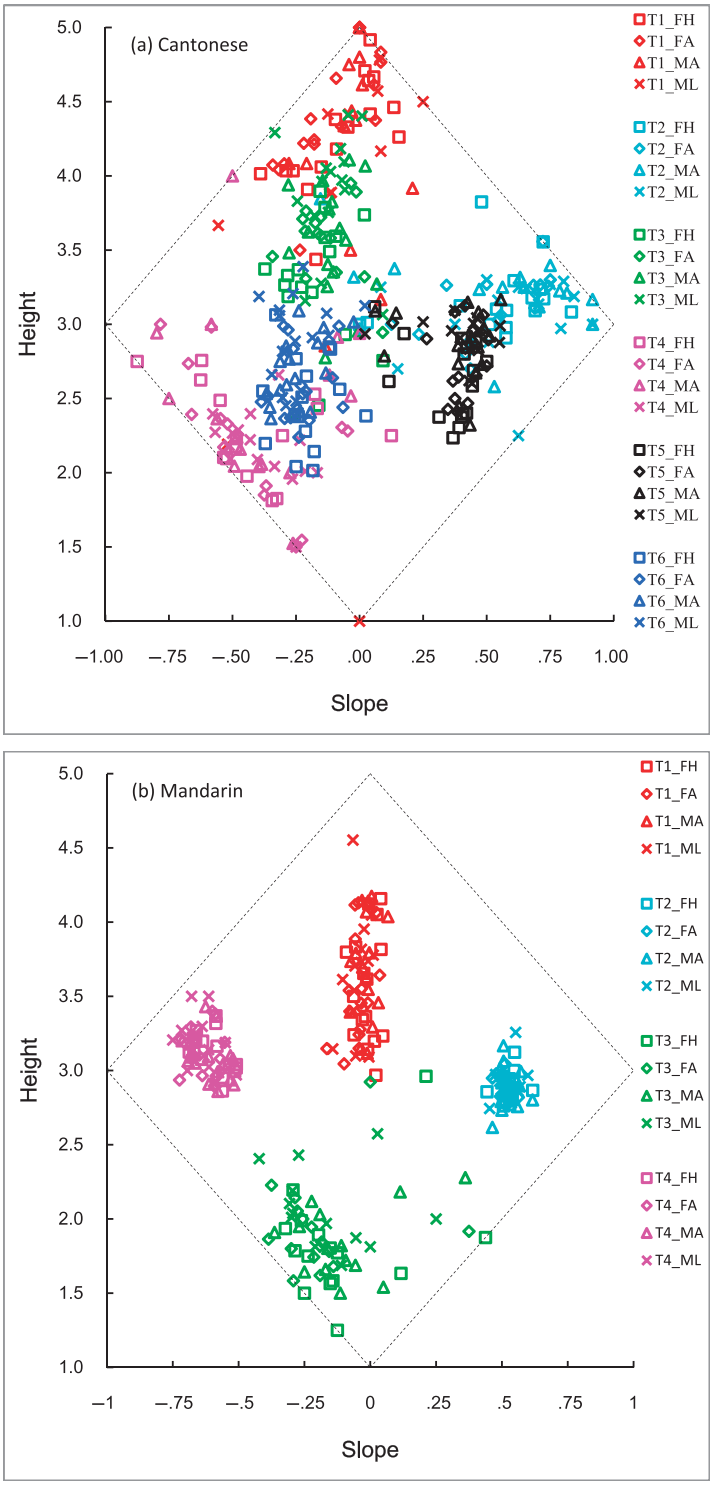
Figure 3a shows the perceptual CG of the six Cantonese tones plotted in a two-dimensional space (height and slope). The six tones are represented by different colors, and the four voices are indicated by different symbols. Each point in this space represents the perceptual CG of one tone calculated from one subject. This figure suggests that each tone category lies in an area generally separate from the other tone categories, although there is some overlap between Tones T2 and T5, T1 and T3, T3 and T6, and T4 and T6.

Inspection of Figure 3a suggests that T1, T3, T6, and T4 form a chain descending from high to low range in perceptual height. Such a distinction in perception is in line with phonological descriptions that T1 is a high level tone; T3 is a mid level tone; T6 is a low level tone; and T4, which often carries a falling contour, is the lowest tone (Bauer & Benedict, 1997).

Along the slope dimension, T1, T3, and T6 lie in an area roughly aligned to slope value 0 (indicating a level contour), although the distribution is slightly skewed toward the negative side (indicating a falling contour). This suggests that Cantonese subjects tend to perceive level or slightly falling pitch contours as these three level tones. T2 and T5, which are traditionally described as rising tones, are located mostly on the positive side of the slope dimension (indicating a rising contour). T4, often described as a low falling tone, is located on the negative side of the slope. The general match of the perceptual distribution and phonological description of Cantonese tones suggests that the perceptual space of Cantonese tones is modulated by subjects' native language background about the properties (e.g., height and slope) of Cantonese tones.

Figure 3b shows the two-dimensional perceptual space of Mandarin tones. In contrast to the Cantonese perceptual space, the Mandarin tones are located far apart from each other. The greater between-category distance in Mandarin can be attributed partly to the less dense tone system (i.e., six tones in Cantonese but only four in Mandarin). As can be seen in Figure 3b, the data points for each Mandarin tone are grouped together more compactly, especially for the two contour

Figure 3. Two-dimensional plot (height and slope) of perceptual center of gravity for (a) the Cantonese tone system and (b) the Mandarin tone system. Six Cantonese tones are represented by various colors: Tone (T) 1 is in red; T2 is in cyan; T3 is in green; T4 is in purple; T5 is in black; and T6 is in deep blue. Four Mandarin tones are represented as follows: T1 is in red; T2 is in cyan; T3 is in green; and T4 is in purple. Four voices are indicated by different symbols: squares indicate female high voice (FH); diamonds indicate female average voice (FA); triangles indicate male average voice (MA); and crosses indicate male low voice (ML). Dotted lines specify the area within which a tone can occur.



tones, T2 (rising tone) and T4 (falling tone). T1 (level tone) stretches along the height dimension, because both high and mid level pitch contour can be identified as T1. T3 shows a relatively large within-category variation compared with the other three tones. The main body of T3 lies on the negative side of the slope dimension, with some outliers on the positive side. As discussed in the *Stimuli* section, T3 has three variants conditioned according to the environment in which it occurs. Although stimuli were presented in isolation in this study, the perceptual data suggest that most subjects tend to identify a low falling pitch contour as T3.

Effect of Intertalker Variations

To compare the effects of intertalker variations on Cantonese and Mandarin tone identification, in Figure 4 we show the perceptual CG collapsed across the subjects of each language group. One outlier was removed from the data when plotting Figure 4. This outlier was caused by a particular subject, who labeled Stimulus 11 as T1 in the ML voice condition. We considered it to be an outlier because this subject identified 11 as T1 only this one time (which is also the only time that this subject labeled any stimulus as T1 in the ML condition). Moreover, Stimulus 11 was identified mainly as either T6 or T4 by this subject. It is, therefore, not unlikely that this subject intended to identify 11 as T4 (or perhaps T6) but mistakenly pressed the button for T1 (on the keyboard, the labeled buttons for T1 and T4 are adjacent to each other).

In this study, we use the term *normalization* to mean that listeners can adapt to a talker's specific characteristics. Accordingly, the same tone, despite variation in its acoustic realizations when uttered by different talkers, should be perceptually similar across different talkers. The perceptual similarity is reflected by the perceptual height and slope of a tone. Specifically, normalization of talker variations implies minimal or no shift from talker to talker within a certain tone in the perceptual space.

With perceptual similarity as the criterion for talker normalization, an obvious difference is observed between these two groups. Cantonese tones exhibit substantial fluctuation in the perceptual space as a result of intertalker variation, whereas Mandarin tones basically remain static across the voices. In the Cantonese perceptual space, the four voices cause a monotonic upward shift in perceptual CG for each of these three tones, T3, T5, and T6 (i.e., for each tone, the perceptual CG of ML is the highest, followed by that of MA, then that of FA, and the lowest CG of FH). Although there were also shifts for the other three tones, these were not monotonic. Moreover, for T3, T5, and T6, intertalker variations have an opposite effect on the perceptual CG (i.e., the higher the F0 value of a voice, the lower the perceptual height for these three tones). For instance, in the case of high voice (FH), the listeners bias their perception to higher tones. The effect of perceiving low pitch-contour stimuli of high voices as high tones implies the lack of talker normalization.

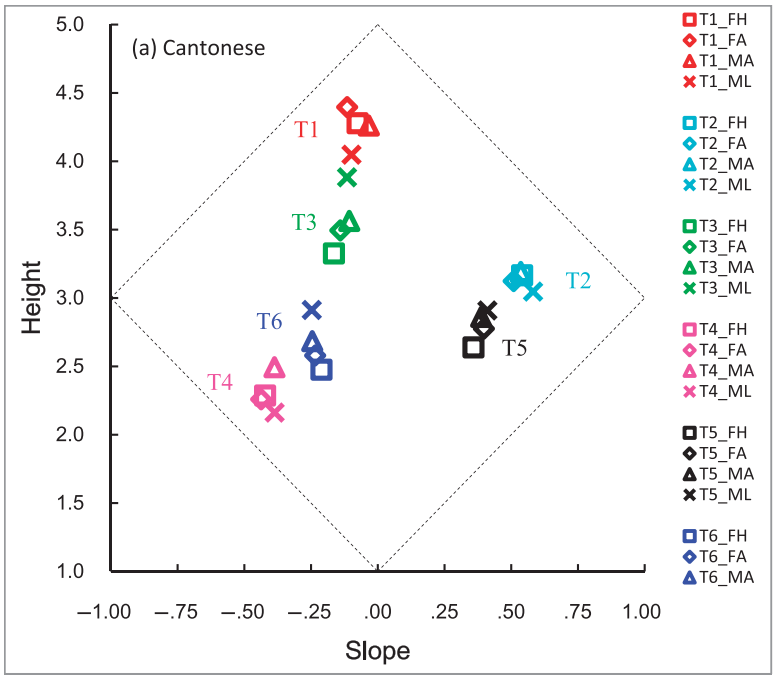
For a comparison of the degree of perceptual dispersion of the two groups of listeners due to intertalker variations, the mean absolute difference between the CG per voice and the mean CG averaged across all voices was first calculated for each tone per tonal dimension and per subject and then averaged over the tones to obtain the dispersion values. Table 4 depicts the dispersion values of the two groups of subjects. A one-way ANOVA shows that there is a significant effect of language background in both height, $F(1, 30) = 46.92$, $p < .001$, and slope, $F(1, 30) = 36.93$, $p < .001$, indicating that Cantonese tone identification exhibits significant shifts in the perceptual space as a result of intertalker variations (marginal mean dispersion for Cantonese and Mandarin in height, 0.19 and 0.08, respectively; in slope, 0.06 and 0.03, respectively).

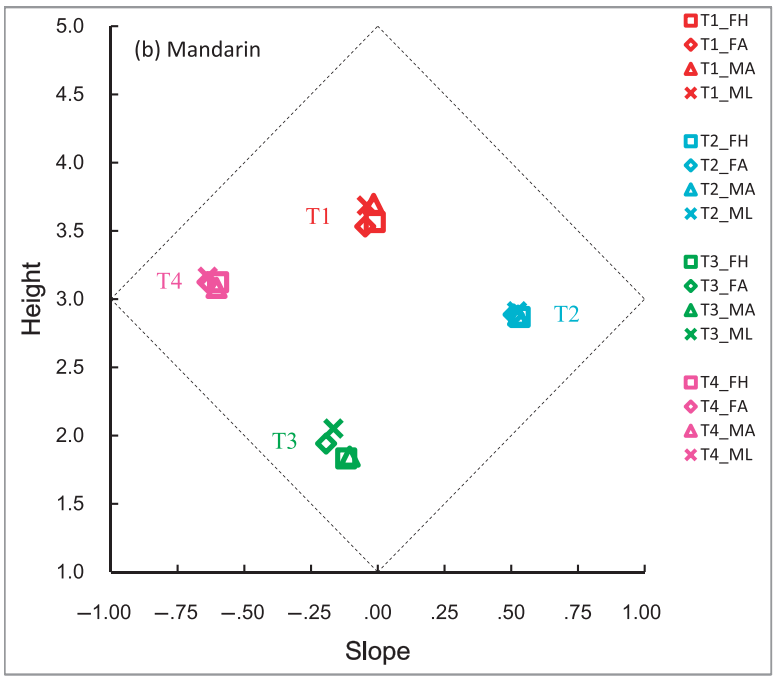
The contrast between Mandarin and Cantonese tone perception indicates that Mandarin subjects were better able to normalize intertalker variations. Cantonese subjects did not adapt to each voice very well in spite of the blocked-talker design, and their perception seems to be influenced by the absolute F0 values of each voice. In our stimulus design, the secondary cues for tone perception, such as duration and intensity profile, have been neutralized (Liu & Samuel, 2004; Zee, 1978). However, Cantonese listeners probably rely on these secondary tonal cues, as well as sentential context, more than Mandarin listeners do to achieve effective tone identification.

Given the different tone inventories of the two target languages, we conducted statistical analyses on the perceptual CG for each language group separately (Cantonese and Mandarin). For each tone category, its perceptual CG has two features, one in the height dimension and the other in the slope dimension. Within each dimension, we conducted a two-way repeated measures ANOVA, with the perceptual CG for each tone category as the dependent variable and tone and voice as two independent variables. In total, four two-way repeated measures ANOVAs were conducted, two for Cantonese (height and slope) and two for Mandarin (height and slope).

For the perceptual height of Cantonese tones, there was a significant main effect of tone, $F(3.112, 46.686) = 98.73$, $p < .001$, indicating that Cantonese tones differ from each other in terms of perceptual height. The main effect of voice failed to reach significance, $F(1.939, 29.086) = 2.23$, $p = .127$, but there was a significant Tone × Voice interaction, $F(3.572, 53.574) = 5.94$, $p < .01$. Taken together,

Figure 4. Average perceptual center of gravity collapsed across subjects for (a) Cantonese subjects and (b) Mandarin subjects. Tone categories are represented by different colors. Four voices are indicated by different symbols. Squares indicate female high voice (FH); diamonds indicate female average voice (FA); triangles indicate male average voice (MA); and crosses indicate male low voice (ML). Dotted lines specify the area within which a tone can occur.





the lack of effect of voice and significant Tone × Voice interaction can be attributed to tones shifting in different directions (some upward, some downward), perhaps canceling out the overall effect of voice.

For the perceptual slope of Cantonese tones, there were significant main effects of tone, $F(1.174, 25.705) = 95.66$, $p < .001$, and voice, $F(2.6, 39.0) = 3.04$, $p < .05$, but no significant Tone × Voice interaction, $F(5.622,$

Table 4. Dispersion according to voices.

	Cantonese			Mandarin		
	Subject	Height	Slope	Subject	Height	Slope
	1	0.223	0.045	1	0.138	0.035
	2	0.148	0.042	2	0.053	0.023
	3	0.150	0.053	3	0.085	0.023
	4	0.252	0.085	4	0.080	0.015
	5	0.193	0.092	5	0.043	0.018
	6	0.168	0.050	6	0.098	0.055
	7	0.167	0.080	7	0.105	0.023
	8	0.237	0.060	8	0.098	0.035
	9	0.172	0.070	9	0.118	0.030
	10	0.273	0.112	10	0.068	0.028
	11	0.127	0.047	11	0.068	0.030
	12	0.342	0.072	12	0.048	0.033
	13	0.203	0.062	13	0.065	0.018
	14	0.115	0.040	14	0.110	0.033
	15	0.158	0.037	15	0.073	0.023
	16	0.168	0.057	16	0.068	0.018
M		0.194	0.063		0.082	0.028

84.33) = 0.73, $p = .621$. These results suggest that the perceptual slope differs between different tones and between different talkers. The lack of interaction effect implies that intertalker variations may have a similar effect on the perceptual slope of different tones.

For the perceptual height of Mandarin tones, there were significant main effects of tone, $F(1.786, 26.789) = 166.48$, $p < .001$, and voice, $F(2.749, 41.232) = 10.11$, $p < .001$. There was also a significant Tone × Voice interaction, $F(3.6, 53.989) = 3.22$, $p < .05$. These results indicate that the perceptual height varies across tones and also across talkers. Moreover, intertalker variation may have a greater effect on some tones.

For the perceptual slope of Mandarin listeners, there were significant main effects of tone, $F(1.183, 17.743) = 476.78$, $p < .001$, and voice, $F(2.324, 34.853) = 15.08$, $p < .001$. The Tone × Voice interaction was also significant, $F(3.666, 54.993) = 3.40$, $p < .05$. Similar interpretations can be drawn here as in the case of the perceptual height of Mandarin tones.

Post Hoc Tests

This section further probes the interaction effects found earlier to determine the nature of intertalker variations on each tone for the two language groups.

Figure 5 breaks up the intertalker variations into each tone. As can be seen from Figure 5a, T3, T5, and T6 show a monotonically increasing trend in mean height from high to low voices, whereas the other three tones exhibit less systematic shifts. A follow-up one-way ANOVA comparing the four voices was conducted separately for each tone. There was a significant effect of voice for T3, T5, and T6 but not for the remaining tones: For T3, $F(3, 60) = 5.70$, $p < .01$; for T5, $F(3, 60) = 4.29$, $p < .01$; for T6, $F(3, 60) = 7.93$, $p < .001$; and for the other three tones, $p > .05$. The aforementioned monotonically increasing trends for T3 ($p < .001$, $n = 16$), T5 ($p < .01$, $n = 16$), and T6 ($p < .001$, $n = 16$) have been confirmed by Page's tests, respectively (Page, 1963).

In contrast to the intertalker variations in perceptual height, there was no significant Tone × Voice interaction effect for perceptual slope, although there was a significant main effect of voice. A one-way ANOVA also indicated that there was no significant effect of voice for any of these six tones. Therefore, intertalker variations cause only minimal variations in the slope dimension, as can be seen in Figure 5b.

Figure 6 presents the intertalker variations in perceptual height and slope for the Mandarin tone system. In terms of perceptual height, a one-way ANOVA found no significant effect of voice for all four Mandarin tones.

For slope, a one-way ANOVA revealed a significant main effect of voice for T1 only, $F(3, 60) = 3.08$, $p < .05$. However, Tukey's post hoc analysis found no significant difference between any two voice pairs on T1, suggesting that the aforementioned effect was too weak to warrant meaningful interpretation.

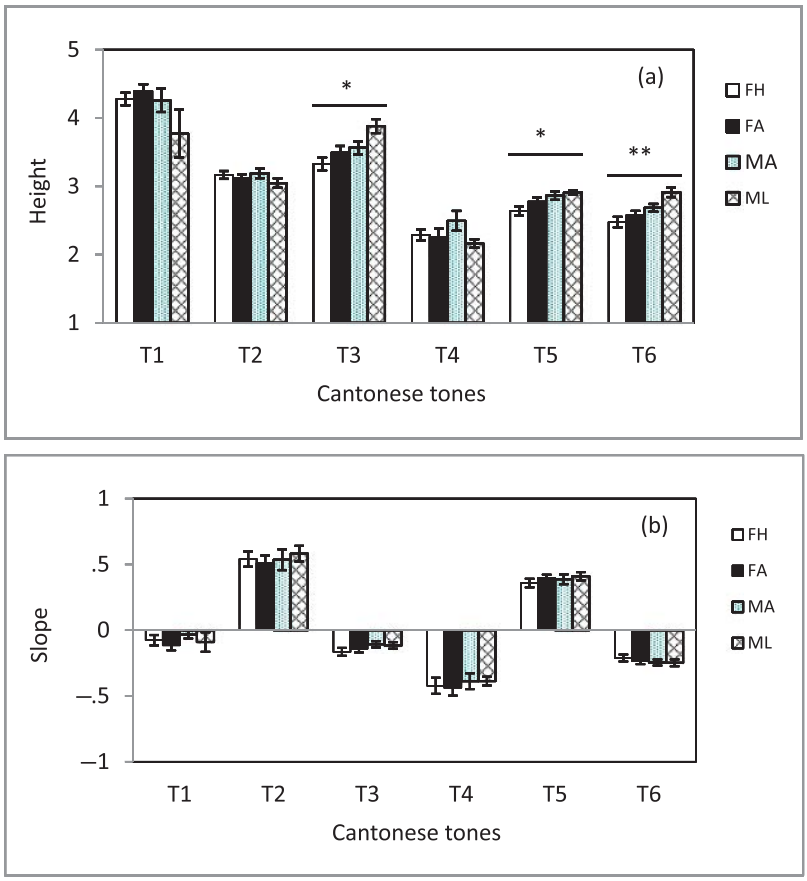
Although there were significant two-way Tone × Voice interactions for Mandarin tones for both height and slope, post hoc analyses did not reveal any significant difference between any two voice pairs for all tones.

To summarize, intertalker variations have an unequal effect on perceptual height and slope. In the case of Cantonese tone identification, intertalker variations have a larger impact on the height dimension than on the slope dimension, as indicated by the significant effects of voice for the three tones, T3, T5, and T6. For Mandarin, intertalker variations have a limited impact on both height and slope dimensions.

Discussion

The main findings of this study are as follows: (a) In terms of perceptual CG, the Mandarin tone system is organized more compactly than the Cantonese tone system, as shown by its greater between-category distances and lesser within-category dispersion. (b) Concerning talker normalization, tone identification of Mandarin listeners exhibited relatively stable normalization across the voices, whereas tone identification of Cantonese listeners was unstable and susceptible to the influence of intertalker variations. (c) In the case of Cantonese, intertalker

Figure 5. Across-voice variations in (a) perceptual height and (b) slope for six Cantonese tones. The four voices are female high (FH), female average (FA), male average (MA), and male low (ML). Asterisks indicate the results of one-way analysis of variance comparing four voices within each tone. Error bars indicate plus or minus standard error.*$p < .05$. **$p < .001$.



variations had a larger effect on the height dimension than on the slope dimension.

Between-Groups Comparison

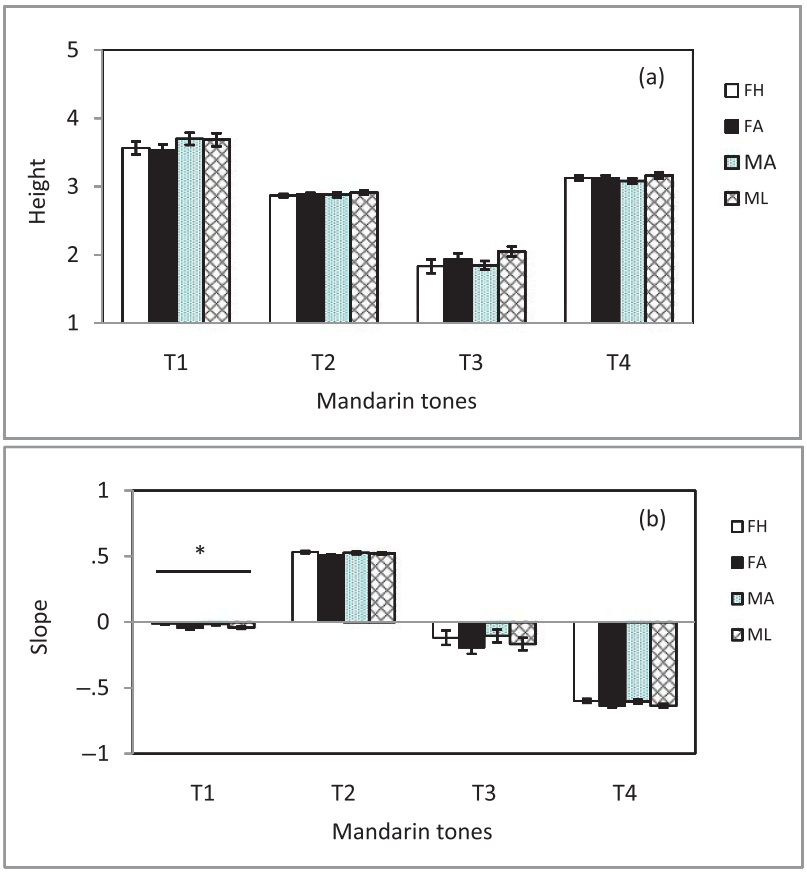
Discussion in this section focuses on the following two questions: (a) How does native tone language background shape the perceptual tone space? (b) Do intertalker variations have an equivalent effect on Cantonese and Mandarin tone categorization?

Inspection of the two-dimensional (height and slope) perceptual space reveals that the perceptual distribution of Cantonese and Mandarin tones is generally compatible with their respective phonological descriptions, providing evidence that tone perception is modulated by subjects' experience of their native language. In other words, Cantonese and Mandarin listeners are guided by the types of tone contrasts in their native language to group and map acoustic variations to corresponding linguistic tones (Gandour, 1983; Y. S. Lee, Vakoch, & Wurm, 1996).

The Mandarin and Cantonese tone spaces differ in several respects. For example, in the perceptual space, the Mandarin tone system is organized more compactly than Cantonese, as shown by the greater between-category distance and less within-category dispersion. On the other hand, the Cantonese tone system exhibits great within-category dispersion and between-category overlapping. This indicates that tone identification in isolation is less stable in Cantonese. Without access to contextual information and prior knowledge of the talkers' voice, it is not easy for Cantonese listeners to accurately map acoustic variations to corresponding tone categories. For Mandarin listeners, however, F0 differences within a syllable alone seem sufficient to elicit stable and reliable tone identification.

Lack of stable adaptation to different voices could be associated with the types of tones that are contrasted in a language. In Mandarin, each tone bears a distinctive F0 contour; therefore, F0 difference within a syllable carries sufficient information about the identity of a tone. In

Figure 6. Across-voice variations in (a) perceptual height and (b) slope for four Mandarin tones. The four voices are female high (FH), female average (FA), male average (MA), and ML male low (MA). Asterisks indicate the results of a one-way analysis of variance comparing four voices within each tone. Error bars indicate plus or minus standard error.*$p < .05$. **$p < .001$.



Cantonese, many tones share a similar F0 contour—for example, the three level tones (T1, T3, and T6) are only distinguishable in pitch height. F0 overlapping between voices may result in ambiguity for such tones, leading to less stable tone identification in Cantonese. This point is further illustrated in the next section.

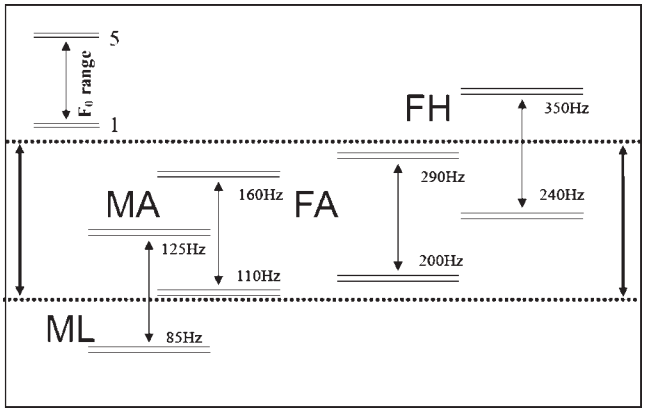
The opposite effect of intertalker variations on the perceptual height of Cantonese T3, T5, and T6 also requires an explanation (i.e., the higher the F0 value of a voice, the lower the perceptual height for a tone). To probe this question, we consider the possible strategies that listeners could have adopted in talker normalization.

One possible strategy is to figure out the F0 range of a particular voice. If listeners keep track of the heard stimuli, they may notice that one block contains one talker's voice only. On the basis of such information, they could estimate the upper and lower bounds for that particular voice and categorize a pitch stimulus by computing its relative position in that F0 range. In this case, we would expect the listeners to normalize the intertalker variations well, resulting in no dramatic shift in perceptual CG for each tone across different voices. However, such an expectation is contradicted by the finding that there is a monotonic shift in the perceptual height of certain Cantonese tones.

The other possibility is that listeners do not estimate the upper and lower bounds of a particular voice in spite of the blocked-talker design. Rather, listeners resort to an internal pitch template when listening to unfamiliar voices and categorize the pitch stimuli with reference to the internal pitch template instead of adapting to each talker's specific range (Bishop & Keating, 2010; Dolson, 1994; Honorof & Whalen, 2005; C.-Y. Lee, 2009).

The second strategy explains why perceptual CG for Cantonese tones shifts. Figure 7 is a schematic representation of this speculated internal pitch template that Cantonese listeners could have relied on in talker normalization. Let us start with the assumption that average male and female voices (MA and FA) generally fit in

Figure 7. A schematic representation of speculated internal pitch template that Cantonese listeners could have relied on in talker normalization. Area within the dotted lines indicates the speculated internal template (see the text for illustration). FH = female high voice; FA = female average voice; MA = male average voice; ML = male low voice.



this internal pitch template. When listening to a very high voice, such as FH, a listener's expected pitch height for T3 (midlevel tone in Cantonese) might correspond to the lower half of the high voice. Because perceptual CG is calculated on the basis of normalized tone digits (for each voice, highest F0 is aligned to Tone Digit 5 and lowest F0 to Tone Digit 1), identifying pitch stimuli lying in the lower half of FH as T3 would lower the perceptual height of T3. Similarly, when subjects listened to a very low voice, such as ML, a listener's expected pitch height for T3 would lie in the upper half of the ML, giving rise to relatively high perceptual CG of T3. Thus, a shift in perceptual height of Cantonese tones demonstrates that Cantonese listeners maintain an internal pitch template rather than dynamically adapt to each voice.

This discussion shows that, without information of a talker's identity, Cantonese listeners are likely to rely on an internal pitch template to identify tone categories, regardless of the specific F0 range of a talker. However, it is unclear how this internal pitch template is formed. One possibility is that it is shaped by the average F0 range of all voices that a subject has heard before. Cantonese subjects could have computed the average F0 range of Cantonese-speaking female voices and Cantonese-speaking male voices, and use such information to make tone judgments when neither the talker's identity nor contextual cues are available. This explanation seems to fit the schematic representation in Figure 7, which assumes that the average male and female voice ranges (MA and FA) estimated from a sample of 34 male and 34 female Cantonese speakers (CUSENT) generally fall within this internal template. More research is needed to further explore this question.

Within-Group Comparison

In this section, we discuss the magnitude of the influence of intertalker variations on two tone dimensions, height and slope. In the case of Cantonese tone identification, intertalker variations have a larger effect on the height dimension than on the slope dimension. For Mandarin, the overall effect of intertalker variations on tone perception is limited, regardless of height or slope dimension. Therefore, the following discussion focuses on Cantonese tone identification, especially for the height dimension.

As mentioned earlier, the three Cantonese level tones (T1, T3, and T6) bear a similar, relatively flat pitch contour. These tones are distinguishable from one another in terms of just pitch height. Two rising tones (T2 and T5) share a similar onset F0 and rising contour. Phonetic data in earlier studies show that T2 and T5 are generally distinctive from each other in terms of the magnitude of the rising slope (i.e., T2 has a steeper rising contour, ending with a higher offset F0 than T5; Bauer et al., 2003; Rose, 1996). Given that T2 and T5 share a similar onset F0, different offset F0 values cause a difference in overall pitch height as well. We suspect that tones that contrast with each other in terms of pitch height are susceptible to the influence of intertalker variations, which introduces F0 overlapping and consequently ambiguity between these tones.

This speculation predicts that the three level tones (T1, T3, and T6) and two rising tones (T2 and T5) in Cantonese are prone to the effect of intertalker variations. Such an expectation is partly confirmed by the perception patterns reported earlier (i.e., intertalker variations are found to have a significant effect on the perceptual height of T3, T5, and T6). T3 and T6 are from the level tone group, and T5 is from the rising group. It is interesting that the remaining members in level and rising groups (i.e., high level tone [T1] and high rising tone [T2]) are less influenced by intertalker variations. A possible explanation for this discrepancy is that T1 and T2 are peripheral tones that stand at the extreme ends of a tone space. In the pitch height dimension, the high level tone T1 and low falling tone T4 can be considered to be peripheral tones at two ends of the height scale. In the pitch slope dimension, the high rising tone T2, which has a larger rising slope than T5, may be considered peripheral at the upper end of the slope scale, with T4 at the lower end of the slope scale. Presumably because of their peripheral positions in the perceptual space, T1 and T2 have no extra space to shift driven by intertalker variations. On the other hand, T3, T5, and T6, which are intermediate tones, exhibit a larger shift from high to low voices.

In fact, contrary to the increasing pattern in the case of T3, T5, and T6, T1 and T2 tend to exhibit a drop in low

voices, as shown in Figure 4a. Although one-way ANOVA did not find a significant effect of voice on T1 and T2, the drop in T1 for ML voice nevertheless needs an explanation. We suspect that the drop in T1 is due to an artifact rather than indicating that T1 is perceived as so low in this condition. As discussed earlier, Cantonese listeners seem to rely on an internal template in tone identification. With reference to this range, probably the highest F0 in ML (125 Hz) is still not perceived as high enough to be labeled as T1. This can be seen from the average identification rates of T1 and T3 in ML voice. Among the total 3,600 stimuli presented in this condition (25 stimuli × 9 repetitions × 16 listeners), 18.1% of the stimuli were identified as T3, but only 2.69% of them were identified as T1. Although some subjects identified the highest pitch stimulus as T1 in this condition, the overall identification is inconsistent (e.g., one subject labeled Stimulus 11 as T1, although such cases are rare). We suspect that the low overall identification rates (only 2.69% in ML voice) compromised the accurate estimation of the perceptual height of T1. Moreover, for T1, the perceptual CG of the other three voices, FH, FA, and MA, are close to one another. Only in ML voice does T1 exhibit a substantial drop. Therefore, we conclude that perceptual CG of T1 in the ML condition reflects an artifact rather than indicates the true perceived pitch height of T1.

General Discussion

Earlier studies exploring the effect of native language background on tone perception have revealed interesting properties about Cantonese and Mandarin tones. In this section, we try to link the findings of this study to earlier studies; in particular, we focus on the role of pitch height in talker normalization.

This study found that, without access to contextual information and prior knowledge of the talkers' voice, isolated tone identification in Cantonese is unstable. Especially for those tones that rely mainly on pitch height to discriminate them, we observed a clear shift in the perceptual space due to intertalker variation. Although this study adopted a blocked-talker design to compensate for the lack of talker and contextual information, there is no clear indication of reliable talker normalization among the Cantonese listeners. This lack of consistent talker normalization suggests that contextual cues or knowledge of the talker's identity are necessary for a listener to disambiguate certain Cantonese tones (Francis et al., 2006).

On the other hand, Mandarin listeners seem to be relatively immune to the influence of different voices. For Mandarin tones, each of which carries a distinctive F0 contour (level, rising, falling, and dipping), F0 differences alone seem sufficient to elicit stable and reliable tone identification (Moore & Jongman, 1997). Talker or contextual information could be helpful, but not necessary, for Mandarin tone identification.

The contrast between Cantonese and Mandarin listeners' performances suggests that pitch height alone is not an efficient cue for tone perception, especially in isolation. For tones that are distinctive in F0 contour, such as rising and falling tones, intertalker variations have very limited influence on their identification. However, for level tones, which heavily depend on pitch height differences, intertalker variations may result in overlapping between these tones. Consequently, extra information is required to disambiguate them.

This finding—that pitch height alone is not an efficient cue for tone perception—is generally consistent with earlier findings concerning categorical perception and talker normalization of tones. For example, Francis, Ciocca, and Ng (2003) reported a lack of evidence for categorical perception of Cantonese level tones (T1, T3, and T6). On the contrary, identification crossover and discrimination peak are basically matched in the case of the high-level-to-rising tone continuum. This may suggest that the categorical boundary between level tones is vulnerable when they are presented in isolation.

In terms of talker normalization, earlier studies embedding acoustic variations in carrier sentences found that identification of a level pitch can be changed categorically depending on the F0 height of the context; therefore a level pitch can be identified as T1, T3, or T6 when the contextual F0 height varies (Wong & Diehl, 2003). On the other hand, it seems that tones that are distinctive in F0 contour are relatively more resistant to the change of context. For example, Huang and Holt (2009) embedded a continuum of level-to-rising pitch contour in a context with varying F0 height. Although they found a significant effect of context on the identification rates, it seems that F0 height of the context had relatively limited impact, compared with F0 differences of the target syllables themselves.

Although manipulation of a carrier sentence is not included in this study, our findings are generally in line with the aforementioned studies. Intertalker variations introduce difficulties in the identification of tones that rely mainly on pitch height but less so for tones having distinctive F0 contours. Therefore, contextual information is crucial in the former case, which is consistent with Wong and Diehl's (2003) finding that context changes the identification of a level pitch categorically. However, for tones that are distinctive in contour, context is not necessary in identification because F0 differences alone are sufficient to elicit reliable identification.

The last point that we draw upon here is Gandour's (1983) finding that Cantonese listeners placed more emphasis on the height dimension than Mandarin listeners

did. It seems to indicate that Cantonese listeners are more sensitive to differences in F0 height than Mandarin listeners, because Cantonese listeners have to make finer distinctions in F0 height to tell one level tone from the other.

The question that arises is whether this point contradicts our finding that Cantonese listeners did not normalize the intertalker variations very well. The answer is no. It should be pointed out that the design of this study aims to reveal the talker normalization mechanism in the process of tone identification, which is different from the phenomenon reported in Gandour's (1983) study. It is likely that Cantonese listeners do make finer distinctions than Mandarin listeners in tone perception. For example, a recent event-related potentials study (Zheng, Minett, Peng, & Wang, 2012), which compared Cantonese and Mandarin listeners' processing of rising pitch contours, revealed a categorical effect in terms of P300 amplitude only in the Cantonese group, suggesting that Cantonese listeners distinguish phonologically contrastive differences in rising pitch with greater ease than Mandarin listeners do. However, the task in this study requires a different mechanism: normalizing intertalker variations. To fulfill this task, it is necessary to figure out the F0 range of a particular voice and estimate the relative height of a pitch in that range, which we did not observe in the performance of Cantonese listeners.

In summary, the Mandarin tone system is well aligned with the principle of sufficient perceptual contrast, whose native listeners are able to normalize intertalker variations to some degree when perceiving tones in isolation. However, violation of the aforementioned principle for the Cantonese tone system makes its native listener unable to normalize intertalker variations well. The larger number of tones in Cantonese may also play a role here. Nevertheless, in practice, Cantonese listeners are able to communicate with other Cantonese talkers without difficulty. We speculate that Cantonese listeners may rely more on other information such as sentential context and talker identity, as well as secondary cues, to achieve effective tone identification. Our findings may also be applicable to other phonological inventories, such as vowel inventories. For instance, a vowel system that complies with the principle of sufficient perceptual contrast may cause less confusion among its vowels when perceived in isolation. Simulation studies on optimized phonological inventories based on the principle of sufficient perceptual contrast may shed more light on the perceptual consequence of different inventory configurations (de Boer, 2000; Ke, Ogura, & Wang, 2003).

Conclusion

Previous studies on talker normalization have usually focused on one target language (e.g., Moore & Jongman, 1997, for Mandarin; and Wong & Diehl, 2003, for Cantonese). However, the stimuli we have used here are acoustically unbiased to any particular tone system. The present study therefore serves as an initial step toward a fuller understanding of the interaction between the size and configuration of phonological inventory and the perceptual consequence of talker variability. Moreover, interest in Cantonese tones has previously focused on its level tones (e.g., Francis et al., 2006; Wong & Diehl, 2003). Therefore, our results complement previous findings on Cantonese tone perception. In summary, perception of the high level tone, T1, is less influenced by intertalker variations than the other Cantonese level tones. The other two Cantonese level tones, T3 and T6, together with the low rising tone, T5, are heavily influenced by intertalker variations, suggesting that the perception of the intermediate tones are unstable when perceived in isolation.

Acknowledgments

The work described in this article was partially supported by grants from National Science Foundation of China (NSFC: 11074267) and the Research Grant Council of Hong Kong (GRF: 455911).

References

Bauer, R., & Benedict, P. K. (1997). *Modern Cantonese phonology*. Berlin, Germany: Mouton de Gruyter.

Bauer, R., Cheung, K.-H., & Cheung, P.-M. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change, 15,* 211–225.

Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance 32,* 97–103.

Bishop, J., & Keating, P. (2010). Perception of pitch location within a speaker's own range: Fundamental frequency, voice quality and speaker sex. *UCLA Working Papers in Phonetics, 108,* 113–140.

Chao, Y.-R. (1930). A system of tone letters. *Le Maître Phonétique, 45,* 24–27.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics, 28,* 441–465.

Dolson, M. (1994). The pitch of speech as a function of linguistic community. *Music Perception, 11,* 321–331.

Francis, A. L., Ciocca, V., & Ng, B. K. (2003). On the (non)categorical perception of lexical tones. *Perception & Psychophysics, 65,* 1029–1044.

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

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11,* 149–175.

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,* 395–421.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America, 117,* 2193–2200.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America, 125,* 3983–3994.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Malden, MA: Blackwell.

Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Optimization models of sound systems using genetic algorithms. *Computational Linguistics, 29,* 1–18.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29,* 98–104.

Lee, C.-Y. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America, 125,* 1125–1137.

Lee, T., Lo, W. K., Ching, P. C., & Meng, H. (2002). Spoken language resources for Cantonese speech processing. *Speech Communication, 36,* 327–342.

Lee, Y. S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research, 25,* 527–542.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48,* 839–862.

Lin, T., & Wang, W. S.-Y. (1984). Shengdiao ganzhi wenti [Tone perception]. *Journal of Chinese Linguistics, 2,* 59–69.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando, FL: Academic Press.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47,* 109–138.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America, 102,* 1864–1877.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association, 58,* 216–230.

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A syllabus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics, 34,* 135–154.

Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language, 85,* 884–892.

Rose, P. (1996). Cantonese citation tones. In P. J. Davis & N. H. Fletcher (Eds.), *Vocal fold physiology: Controlling complexity and chaos* (pp. 307–324). San Diego, CA: Singular.

Wang, W. S.-Y. (1976). Language change. *Annals of the New York Academy of Sciences, 208,* 61–72.

Wang, W. S.-Y., & Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research, 10,* 629–636.

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46,* 413–421.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25,* 61–83.

Xu, Y., 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,* 1063–1074.

Zee, E. (1978). Duration and intensity as correlates of F0. *Journal of Phonetics, 6,* 213–220.

Zheng, H.-Y., Minett, J. W., Peng, G., & Wang, W. S.-Y. (2012). The impact of tone systems on the categorical perception of lexical tones: An event-related potentials study. *Language and Cognitive Processes, 27,* 184–209.

The Effect of Intertalker Variations on Acoustic Perceptual Mapping in Cantonese and Mandarin Tone Systems

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

J Speech Lang Hear Res 2012;55;579-595; originally published online Dec 29, 2011;
DOI: 10.1044/1092-4388(2011/11-0025)