Acoustic characteristics of highly distinguishable Cantonese entering and non-entering tones

Puisan Wong and Hoi-Yin Chan
Division of Speech and Hearing Sciences, Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong

(Received 7 June 2017; revised 9 December 2017; accepted 28 December 2017; published online 7 February 2018)

Cantonese has one of the most complex tone systems. Few studies have thoroughly examined or compared the acoustic properties of the full set of Cantonese tones, particularly the entering tones, compromising deeper understanding of Cantonese tone difficulties in various clinical populations. This study (1) describes a theory-driven method for acoustic analysis of tones that successfully normalized the intrinsic pitch of male and female speakers, (2) provides detailed acoustic data on distinctly enunciated Cantonese tones, (3) examines the acoustic similarities and differences between the entering and non-entering tones, and (4) compares the acoustic properties of three easily confused tone pairs. Seventeen male and female native speakers produced 1802 Cantonese tones that were correctly identified by five judges in filtered stimuli. Counter to the established notion that the entering tones are shorter versions of the three level tones, the results revealed that the entering tones have falling contours, suggesting that the entering and non-entering tones should be examined separately in research and clinical settings. The detailed description of the acoustic properties of the nine tones and the acoustic contrasts of the entering and non-entering tones and the three easily confused tone pairs provides references for future Cantonese tone studies with different populations.
 https://doi.org/10.1121/1.5021251

[BVT] Pages: 765–779

I. INTRODUCTION

Tonal languages, which make up a majority of the world's languages (Yip, 2002), use distinctive pitch patterns called lexical tones to contrast word meanings. Cantonese is a major dialect of Chinese (Bauer and Benedict, 1997) with over 72×10^6 speakers (Lewis *et al.*, 2018). Studies have found that even native Cantonese-speaking adults do not always perceive or produce all the Cantonese tones with perfect accuracy (Barry and Blamey, 2004; Ciocca and Lui, 2003; Wong *et al.*, 2017; Wong and Leung, 2018). Various clinical populations such as speakers with neuromotor disorders (e.g., Parkinson's Disease, cerebral palsy, and dysarthria) (Whitehill *et al.*, 2000; Wong *et al.*, 2009), children with profound hearing impairment (Khouw and Ciocca, 2006; Lee *et al.*, 2002a), and children with dyslexia (Cheung *et al.*, 2009; Li and Ho, 2011) also have special difficulties with Cantonese tone processing and production. To date, few studies have systematically examined the acoustic properties of the Cantonese tones, particularly the entering tones. Even fewer studies examined the acoustic properties of tones produced by clinical populations, making it difficult to gain a deeper understanding of the challenges in Cantonese tone acquisition in different typical and atypical populations.

According to traditional Chinese phonology, Cantonese has nine lexical tones (Chao, 1947; Fok Chan, 1974), including six basic tones and three entering tones (Table I). The basic tones occur in open syllables and syllables ending with a nasal consonant (/-m/, /-n/, or /-ŋ/). Based on auditory perception, the six basic tones—Tone 1 (T1) to Tone 6 (T6)—were given the labels of high level (HL), high rising (HR), mid level (ML), low falling (LF), low rising (LR), and low level (LL), respectively. The entering tones (T7, T8, T9) occur in closed syllables ending with voiceless stop consonants (/-p/, /-t/, or /-k/) and were given the labels of high stopped (HS), mid stopped (MS), and low stopped (LS), respectively. The term "entering tone" is a literal translation of the Chinese label "ru-tone" for tones in checked syllables. Thus, Cantonese is often described as having six non-entering tones and three entering tones.

Early studies on Cantonese tones proposed different tone letter systems to describe the levels and shapes of the nine tones based on auditory impression (Chao, 1947). In the systems, typically, the pitch levels at the onset and offset of the tone are notated by a two digit number. Each digit ranges from 1 to 5, with 1 and 5 representing the lowest and highest pitch level of a speaker's regular pitch range, respectively (Chao, 1947, Table I). For example, 35 represents a rising tone starting at the mid pitch range and ending at the highest pitch of the speaker. Most of the systems proposed that the three entering tones [i.e., T7 (HS), T8 (MS), T9 (LS)] have the same pitch heights and shapes as the three non-entering level tones [i.e., T1 (HL), T3 (ML), and T6 (LL)], respectively, but with shorter durations (Hashimoto, 1972; Vance, 1976). The entering tones were, therefore, notated with only the first digit of the tone letters of the corresponding non-entering tones. The tone letter systems (Table I) suggest that the non-entering tones are contrasted by pitch shapes (e.g., level vs rising vs falling tones), pitch heights (e.g., the three level tones and the two rising tones), or both pitch shapes and pitch heights [e.g., (T1) HL vs (T4) LF]. The entering

[a]Electronic mail: pswresearch@gmail.com.

TABLE I. Description of Cantonese tones in pitch height, shape, duration, and tone-letter. Boldface letters indicate inconsistencies in the tone letters among different systems.

				Tone-letter		
Tones	Duration	Height	Shape	Chao (1947)	Hashimoto (1972)	Vance (1976)
T1 (HL)	Long	High	Level	55	55	55
T2 (HR)	Long	High	Rising	35	35	35
T3 (ML)	Long	Mid	Level	33	**44**	33
T4 (LF)	Long	Low	Falling	21	21/**22**	**11**
T5 (LR)	Long	Low	Rising	23	**24**	**13**
T6 (LL)	Long	Low	Level	22	**33**	22
T7 (HS)	Short	High	Level	5	5	5
T8 (MS)	Short	Mid	Level	3/**33**	**4**	3
T9 (LS)	Short	Low	Level	2/**22**	**3**	2

tones are contrasted by pitch heights (high, mid, low), and the non-entering level tones and the entering tones are contrasted by duration.

As presented in Table I, discrepancies are found in the description of the pitch heights and pitch shapes of the tones in different tone-letter systems. Regarding the three level tones, the difference in the pitch level between T1 (HL) and T3 (ML) is larger than the difference in the pitch level between T3 (ML) and T6 (LL) in two systems but is equal in Hashimoto (1972). The pitch level of T3 (ML) and T6 (LL) is lower in two systems but higher in Hashimoto (1972). For the two rising tones, the magnitude of pitch rise is larger in T2 (HR) than in T5 (LR) in Chao (1947) but is equal in the systems of Hashimoto (1972) and Vance (1976). The pitch onset of T5 (LR) is higher in the other two systems but lower in Vance (1976), while the pitch offset of T5 (LR) is lower in two of the systems but higher in Hashimoto (1972). T4 (LF) has a falling contour in Chao (1947) but a level contour in Vance (1976), and either a falling or a level tone in Hashimoto (1972). All the systems notated the entering tones with the same pitch heights and shapes as their non-entering tone counterparts and most considered them having shorter durations than the three level tones, except that Chao (1947) suggested that the durations of T8 (MS) and T9 (LS) may be comparable to those of the non-entering level tones.

Modern Cantonese phonology proposes that Cantonese has six tones and the three entering tones [i.e., T7 (HS), T8 (MS), T9 (LS)] are the allophones/allotones of T1 (HL), T3 (ML), and T6 (LL), respectively. For example, Jyutping, a romanisation system for Cantonese developed by the Linguistic Society of Hong Kong in 1993, uses the same notation for the three entering tones and their non-entering tone counterparts, i.e., T7→T1 (T7 was notated as T1), T8→T3, T9→T6.

Various typical and atypical populations have special difficulties with Cantonese lexical tone processing and production. Three pairs of Cantonese tones [i.e., T2 (HR) – T5 (LR), T3 (ML) – T6 (LL), and T4 (LF) – T6 (LL)] are particularly confusing even for native Cantonese adults. In terms of perception, Lee *et al.* (2015) reported that Cantonese-speaking adults discriminated the tone pairs T2 (HR) vs T5 (LR) and T3 (ML) vs T6 (LL) in monosyllabic words with lower than 90% perceptual accuracy, and T4 (LF) vs T6 (LL) with lower than 95% accuracy. Ciocca and Lui (2003) reported 80% perceptual accuracy and less than 95% perceptual accuracy in adults' discrimination of T2 (HR) vs T5 (LR) and T3 (ML) vs T6 (LL), respectively (see Ciocca and Lui, 2003, Fig. 1). In terms of production, the five adult speakers in Barry and Blamey (2004) produced T2 (HR) and T6 (LL) in monosyllabic words with 88% accuracy and most of the T2 (HR) errors were perceived as T5 (LR) while most T6 (LL) errors were perceived as T3 (ML). Wong *et al.* (2017) reported 78% and 67% production accuracy of adults' monosyllabic T3 (ML) and T6 (LL), respectively, with 16% of the T3 (ML) productions being perceived as T6 (LL) and 20% of T6 (LL) being perceived as T3 (ML), while Wong and Leung (2018) reported 31% of T3 (ML) productions of adults being perceived as T6 (LL), 20% of T6 (LL) being perceived as T3 (ML), 7% of T4 (LF) being perceived as T6 (LL), and about 7%–8% bidirectional confusion of T2 (HR) and T5 (LR).

Some native Cantonese speakers merge the three confusing tone pairs in their perception and production. Two of the eight young speakers in Bauer *et al.* (2003) and six of the 56 speakers in Kei *et al.* (2002) merged T2 (HR) and T5 (LR) and produced them as T2 (HR), T5 (LR), or a hybrid of T2 (HR) and T5 (LR). Mok *et al.* (2013) reported merging of T2 (HR) and T5 (LR), T3 (ML) and T6 (LL), and T4 (LF) and T6 (LL) in 17 young adults. However, not much information was provided on the acoustic characteristics or confusion patterns of their productions.

The three confusion patterns in adults and tone mergers have also been found in young children's perception and production of Cantonese tones. With respect to perception, Lee *et al.* (2015) found that children did not accurately discriminate the six tones in familiar monosyllabic words until after six years of age and they mostly confused T2 (HR) with T5 (LR), T3 (ML) with T6 (LL), and T4 (LF) with T6 (LL). In terms of production, Barry and Blamey (2004) reported that four to six year old children did not produce any of the six tones with adult-like accuracy and exhibited confusions among the three level tones [T1 (HL) vs T3 (ML) and T3 (ML) vs T6 (LL)], between the two rising tones [T2 (HR) vs T5 (LR)], and between the low-falling and low-level tones [T4 (LF) vs T6 (LL)]. Wong *et al.*, (2017), Wong and Leung (2018), and Khouw and Ciocca (2006) found similar error patterns in three-year-old, four- to six-year-old, and 12- to 14-year-old Cantonese-speaking children, respectively.

Tone perception and production difficulties have also been reported in different clinical populations. Children with cochlear implants perform poorer than typical preschool children in discriminating Cantonese tonal contrasts after using cochlear implants for more than two to five years (Lee *et al.*, 2002a) and adolescents with profound hearing loss do not produce the six Cantonese tones accurately (Khouw and Ciocca, 2006). Children with dyslexia have also been found to fall behind age-matched children in Cantonese tone perception and production (Li and Ho, 2011) and Cantonese tone perception accuracy is an indicator for children with and without impaired reading comprehension (Zhang *et al.*, 2014) and with and without dyslexia (Cheung *et al.*, 2009).

Detailed information on the acoustic properties of Cantonese tones can shed light on the tone perception and production difficulties experience by these typical and atypical populations. Yet, no comprehensive acoustic study on the full set of the nine tones is available. Most acoustic studies on Cantonese tones measured the non-entering tones exclusively and reported inconsistent results. Some of the findings supported the description of the tones in the tone letter systems. Whitehill *et al.* (2000) measured the F0 at the beginning, middle, and end of the three level tones (HL, ML, LL) in four sets of monosyllabic words spoken by 18 adult speakers. They supported the observations of Chao (1947) and Vance (1976) that the F0 distances between T1 (HL) and T3 (ML) was significantly larger than between T3 (ML) and T6 (LL), and disagreed with Hashimoto (1972) that the F0 differences between T1 (HL) and T3 (ML) were the same as between T3 (ML) and T6 (LL). With respect to the rising tones, Khouw and Ciocca (2007) found larger F0 rises in T2 (HR) than in T5 (LR), as suggested by Chao (1947), in the tones produced by 10 twelve- to fourteen-year-old teenagers.

Not all acoustic findings supported the description of the tones in the tone letter systems. Based on visual inspection of the tonal contours, Whitehill *et al.* (2000) and Lee *et al.* (2002b) suggested that the three level tones had a slightly falling, rather than a level F0 contour. In contrast to the suggestions in the tone letter systems that all the non-entering tones were of equal durations, Rose (2000) found that T4 (LF) was the shortest and T3 (ML) and T6 (LL) were the longest.

Only two studies measured the acoustic properties of Cantonese entering tones. Rose (2004) plotted the F0 contours of two of the three entering tones T7 (HS) and T9 (LS) and compared them with the contours of the three level tones and the low falling tone collected in Rose (2000). Visual inspection revealed that T7 (HS) had a flat or slightly falling contour while T9 (LS) had a falling contour similar to that of T4 (LF). The author questioned the allotonic relationship between T6 (LL) and T9 (LS), proposed by previous studies.

Bauer and Benedict (1997) was the only acoustic study that measured all the nine tones. Six Cantonese speakers produced each of the nine tones in one to three syllables. By comparing the visual representations of the F0 contours and the mean F0 of the onset, offset and either the peak or the dip of the tones, they agreed with Chao (1947) on the pitch heights of most of the tones. They found that the three level tones (HL, ML, LL) had high, mid and mid-low pitch levels and the three entering tones had the relative pitch levels of 5, 3, 2, similar to those of the non-entering tone counterparts. In terms of duration, they agreed with Chao (1947) that T7 (HS) had the shortest duration. However, they disagreed with Chao (1947) that the durations of T8 (MS) and T9 (LS) were comparable to those of the non-entering tones. They found that all the three entering tones were more than 50% shorter than the non-entering tones. The largest discrepancies were on the contour shapes of the tones. Bauer and Benedict (1997) reported that T1 (HL) had a level or hump shape. T2 (HR) had a falling-rising or rising shape. The contour of T3 (ML) rose to the mid-point then drop. T4 (LF) had a rising-falling or falling contour. T5 (LR) had a dip followed by a rise. T6 (LL) had a rising-falling or a falling-rising contour. The three entering tones all had rising-falling contours. Thus, the findings on the tone shapes did not always match the shapes of the tone contours suggested by the descriptive labels of the tones in the tone letter systems.

Previous acoustic studies provided preliminary data on the acoustic characteristics of Cantonese tones but had some limitations in methodology. None of the studies controlled the impact of segmental structures on F0 and tone durations. Almost all studies compared different tones in words with different vowels and consonant (e.g., Barry and Blamey, 2004; Rose, 2000, 2004). A couple of studies had small sample sizes and a few studies included speakers with different dialects and background. For examples, Bauer and Benedict (1997) examined the nine tones produced by six speakers, with two speakers using a Guangzhou Cantonese dialect, which has a high falling tone contour, rather than a high level contour for T1 (HL) (Rao *et al.*, 1996). Khouw and Ciocca (2007) reported the acoustic characteristics of tones produced by ten 12- to 14-yr-old children, who may not have fully mastered the production of the tones as indicated by the low perceived accuracy among the six tones they produced (range = 58%–84%). Some studies based their findings on a small sample of words [e.g., one to three words for each tone in Bauer and Benedict (1997); four syllables for each tone in Rose (2000, 2004)].

In terms of acoustic analysis, though previous studies have reported inaccurate tone production and tone merging in native Cantonese-speaking adults, none of the acoustic studies performed tone judgment to ensure that only correctly produced tones were included for acoustic analysis. Most studies sampled the pitch contour sparingly and characterized the tones based on the mean F0 and/or the F0 at two to three points in the tonal contours, such as the onset and offset of the tones (Barry and Blamey, 2004); the onset, offset and turning point of the tones (Vance, 1976); the onset, offset and the peak/valley of the tones (Bauer and Benedict, 1997); and the onset, mid-point, and offset of the tones (Whitehill *et al.*, 2000). No study has measured the shapes of the tonal contours (e.g., slopes) and studies that compared the shapes of the tones based their analysis on visual inspection of pitch contours. Some studies compared the tone characteristics on a speaker by speaker basis and did not report group data in consideration of individual and gender differences in voice pitch (e.g., Barry and Blamey, 2004; Rose, 2000, 2004). Few studies performed statistical analysis to confirm whether the observed differences were significant. No study has provided detailed acoustics characteristics of the nine tones, or compared the acoustic characteristics between the three entering and their corresponding level tones, or among the more confusing tones.

To obtain more representative and more comprehensive acoustic data of the nine tones, this study adopted an acoustic analysis method that was based on the results of the theoretical models on lexical tones and has been used in several studies with children and female adults. Research has shown that F0 is the primary and sufficient acoustic cues for lexical tone perception (Fu and Zeng, 2000; Vance, 1976). According to the target approximation model of tonal

contour formation (Xu, 2001, 2004), tones are carried in the vocalic portion of the syllable (Xu and Wang, 2001); the tone contours in the initial part of the syllable are affected by initial consonants and phonetic context. Therefore, contours at the initial portion of the syllable may not be reliable for tone judgment (Xu and Xu, 2003). Several studies found F0 perturbation at the onset of the tones in syllables with initial plosives (Khouw and Ciocca, 2007) or fricatives (see Fig. 2 in Wong and Xu, 2007). When producing tones with a pitch onset much higher or lower than the voice pitch of the speaker [e.g., T1 (HL) or T5 (LR)], a rising/falling contour is found in the initial part of the syllable, showing a transition from the mid pitch range of the speaker to the high/low pitch onset of the target tone (see Fig. 2 in Wong and Xu, 2007). In connected speech, F0 fluctuations are also found in the beginning of the syllable due to carryover coarticulation effect of the preceding tone (Xu and Liu, 2006; Xu, 2001). For examples, when T5 (LR) follows T1 (HL), because T1 (HL) ends at a high pitch level, there is a high falling contour going from the high pitch offset of T1 (HL) to the low pitch onset of T5 (LR). On the other hand, when T5 (LR) follows T4 (LF), there is a low falling contour at the beginning of the second syllable going from the offset of T4 (LF) to the low onset of T5 (LR) (see Xu, 2001, for details). Thus, the tone contours of T5 (LR) in the first half of the second syllable can be very different depending on the preceding tone. Native listeners ignore the tonal transitions and attend to the pitch targets at the end of the syllable for tone discrimination (Xu and Wang, 2001). More evidence was provided by Khouw and Ciocca (2007) who divided the tone contours produced by ten 12- to 14-year-old children into eight sections and asked 30 native adults to label the tones. The mean F0 of the whole tone contour, and the F0 differences between the F0 onset and offset of each of the eight sections were used to predict listeners' perception of the tones using discriminant analysis. The results showed that F0 values in the beginning of the F0 contour deviated from the canonical pattern (i.e., the pitch target) of the tone and did not distinguish the tones. The F0 contours in the later part of the syllable were the most stable, had contours shapes closer to the pitch target of the tones, and could best predict the perception of the tones (Khouw and Ciocca, 2007).

Based on the theoretical models of tone contour formation, this study examined the pitch contours in the whole vocalic portion of the syllable and compared the pitch levels and shapes in the second half of the tones where the pitch targets are located. Tone productions were collected in multiple words with the same segmental structures across the tones from male and female adult speakers. Only tones that were highly distinguishable in filtered speech without lexical support were selected for acoustic analysis. Productions of speakers with different voice pitch and of different genders were normalized using pitch height measures, which have been found to effectively normalize the voice pitch of women and children (Wong, 2012a; Wong *et al.*, 2017; Wong and Leung, 2018). The purpose was to provide detailed acoustic information on the nine Cantonese tones. The specific aims were (1) to measure and characterize the acoustic properties, including the pitch height, shape and duration, of the nine Cantonese tones in distinct productions by male and female native Cantonese-speaking adults in Hong Kong, (2) to compare the acoustic characteristics between the entering tones and their non-entering tone counterparts, and (3) to examine the acoustic similarities and differences in the three easily confused tone pairs in clear productions.

II. METHOD

This study was approved by the Human Research Ethics Committee at the University of Hong Kong. All participants provided written informed consent for their participation.

A. Tone production

1. Participants

Twenty-two (11 male and 11 female, aged 19 to 22 years) phonetically trained Cantonese-speaking young adults majoring in Speech and Hearing Sciences, with no history of speech, language and hearing impairment produced the tones. They were all born in Hong Kong of native Cantonese-speaking parents. Cantonese was their home and strongest language. All participants learnt English as their second language and Mandarin as a third language in school. Reportedly, they all used Cantonese almost exclusively in their daily lives and seldom used Mandarin or another tone language/dialect. They passed a hearing screening at 500, 1000, 2000, and 4000 Hz at 25 dB hearing level bilaterally under headphones using pure-tone audiometry. All participants imitated the nine tones correctly in a phone screening and made no error in a standardized tone perception test [Hong Kong Cantonese Tone Identification Test (CanTIT); Lee *et al.*, 2009].

2. Stimuli

To avoid coarticulation and prosodic effect, 90 monosyllabic words were selected. Among them, thirty-six were combinations of the six non-entering tones with six CV (consonant-vowel) syllables (Table II), while fifty-four of the words were the combinations of the three entering tones with the six CV syllables ending with /-p/, /-t/, or /-k/ (6 syllables $\times$ 3 final consonants $\times$ 3 tones). Very few CV syllables form real words with minimal tone contrasts in all six non-entering tones and very few CVC syllables form real words with minimal contrasts in the three entering tones. Thus, the syllables were selected such that they covered different places and manners of articulation of the initial consonants and formed the maximum number of real monosyllabic words when produced with the nine tones. There were 58 real words and 32 non-sense words. Native Cantonese speakers usually have no problem saying a legitimate syllable in nine tones, regardless of the lexical status of the syllable. A separate set of nine monosyllabic words (6 non-entering tones and 3 entering tones) was used as practice stimuli.

3. Procedures

The participants attended a 1-h session in a sound booth at the University of Hong Kong. They filled out a background questionnaire and received a hearing screening. They

TABLE II. Stimuli for the tone production task. "IPA" refers to International Phonetic Alphabet. "X" indicates that there is no real word for that particular tone category.

		Full Tones							Entering tones				
Sets	IPA	Jyutping	T1(HL)	T2(HR)	T3(ML)	T4(LF)	T5(LR)	T6(LL)	IPA	Jyutping	T7(HS)	T8(MS)	T9(LS)
1	/ji/	ji	衣	椅	意	兒	耳	二	/jip/	jip	X	醃	頁
									/jit/	jit	X	X	熱
									/jɪk/	jik	益	X	疫
2	/tsʰi/	ci	雌	齒	刺	持	似	X	/tsʰip/	cip	X	妾	X
									/tsʰit/	cit	X	切	X
									/tsʰ ɪk/	cik	戚	X	X
3	/tsʰan/	caan	餐	產	燦	殘	X	X	/tsʰap/	caap	X	插	X
									/tsʰat/	caat	X	刷	X
									/tsʰak/	caak	測	策	賊
4	/tsi/	zi	知	紫	志	X	X	字	/tsip/	zip	X	接	X
									/tsit/	zit	澀	節	截
									/tsɪk/	zik	積	X	夕
5	/kim/	gim	兼	檢	劍	X	X	儉	/kip/	gip	X	劫	X
									/kit/	git	X	潔	傑
									/kɪk/	gik	激	X	極
6	/si/	si	詩	史	試	時	市	事	/sip/	sip	X	涉	X
									/sit/	sit	X	屑	舌
									/sɪk/	sit	色	X	食

were then given the list of stimuli (Table II) to practice. After that the nine words for the practice trials were presented on the screen followed by the 90 target words. The participants read the words and their productions were audio-recorded. This procedure was repeated three times to get three productions of each word by each speaker. Finally, CanTIT, a standardized tone perception test, was administered.

B. Tone production judgment

To ensure that only correctly produced tones were selected for acoustic analysis, multiple judges were employed to categorize the speakers' productions.

1. Judges

Five Cantonese-speaking undergraduate students trained to be speech therapists were recruited. They were born and raised in Hong Kong and reported no history of speech, language, and hearing impairment. Cantonese was their home and most proficient language. They all passed a tone screening test on filtered speech with over 90% accuracy prior to performing tone ratings.

2. Stimuli

The recorded productions of the 22 speakers were first normalized for intensity to 60 dB RMS level (number of sound files = 90 words × 3 productions × 22 speakers = 5940). A final year student trained to be a speech therapist chose two productions of each word by each speaker with the best recording quality. Productions of Speaker F01 were excluded due to loud background noise; M04 merged T2 (HR) and T5 (LR) and was excluded. Two productions were excluded due to phonation break. Finally, 3598 productions were chosen (90 words × 2 productions × 20 speakers minus 2 productions) for tone judgment. To minimize lexical bias in the judges' tone ratings following the procedures in Wong (2012a, 2013), all productions were low-pass filtered at 400 Hz using the Hann Band Filter in PRAAT to retain pitch information but eliminate lexical information. Previous studies on Mandarin and Cantonese tones have shown that this cut-off frequency is adequate for the identification of tones produced by adults (e.g., Wong and Strange, 2017; Wong *et al.*, 2017).

3. Procedures

The filtered stimuli were blocked by speakers. Blocks of stimuli and the sound files within each block were presented randomly to the judges for tone rating. Judges listened to the productions one at a time under headphones in a sound-treated room at a comfortable listening level. They could replay the stimuli as many times as needed and categorized the tones in the filtered stimuli by typing the tone number. They re-rated 10% of the stimuli (productions of one female speaker and one male speaker) for determining intra-judge reliability.

C. Acoustic analysis

Sixteen sound files were further excluded due to technical errors during tone judgment, leaving 3582 sound files for acoustic analysis. Following the methods used in Wong (2012a), the production in each sound file was manually segmented into the initial section, the vocalic section, and the final section using a customized PRAAT script (PROSODYPRO version 6.1.4, Xu, 2013). The initial section started at the beginning of the articulation and ended at the zero crossing at the end of the first regular glottal pulse, and, therefore,

TABLE III. Definition of acoustic terms and parameters.

Definition of acoustic terminology used	
Speaker Mean Pitch (St)	Mean pitch across all productions by the same speaker measured in semi-tones (St)
Pitch Height (St)	Pitch level relative to the mean pitch of the speaker, calculated by subtracting the Speaker Mean Pitch from the measured pitch. Positive values indicate pitch levels higher than the speaker's mean pitch; negative values indicate pitch levels lower than the speaker's mean pitch. This procedure normalizes individual and gender differences in vocal pitch level.
Pitch Target	The final 50% of the vocalic section
Definition of the nine acoustic parameters used	
Six pitch height measures	
Initial Pitch Height (St)	Pitch measured at tone onset (i.e., time point 1) minus Speaker Mean Pitch
Final Pitch Height (St)	Pitch at tone offset (i.e., time point 20) minus Speaker Mean Pitch
Mid Pitch Height (St)	Pitch at the midpoint of the tone (i.e., at time point 11) minus Speaker Mean Pitch. This is also the pitch at the onset of the pitch target (i.e., the final 50% of the tone).
Min Pitch Height 50% (St)	Minimum pitch in the 2nd half of the tone (the pitch target) minus Speaker Mean Pitch.
Max Pitch Height 50% (St)	Maximum pitch in the 2nd half of the tone minus Speaker Mean Pitch.
Mean Pitch Height 50% (St)	Mean pitch in the 2nd half of the tone minus Speaker Mean Pitch.
Pitch range 50% (St)	The absolute value of (Max Pitch Height 50% minus Min Pitch Height 50%), an index for degree of pitch fluctuation.
Slope 50% (St/ms)	(Max Pitch Height 50% minus Min Pitch Height 50%) divided by duration between Max Pitch Height 50% and Min Pitch Height 50%, an index of the speed of pitch change.
Pitch duration (ms)	Duration of the vocalic section

included the initial voiceless consonant, the initial irregular vocal cycles with low amplitudes, and the first regular vocal cycle displayed in the waveform. The final section started at the zero crossing at the beginning of the final regular glottal pulse and ended at the end of the articulation. Thus, the final section included the final voiceless consonant, the last regular vocal cycle, and the irregular vocal cycles with low amplitudes. The vocalic section included the whole vocalic segment of the production, except the first and last regular cycles and the irregular glottal pulses with low amplitudes. Each glottal pulse was checked for errors and corrected manually. Creaky productions were included and treated similarly. The vocalic section was segmented into 20 intervals. Because human ears do not respond to frequencies linearly, the Hz scale, which is a linear scale, is not ideal for comparing frequencies. Thus, F0 of the 20 points measured in Hz was converted to semi-tones (St), a logarithmic pitch scale (Russo and Thompson, 2005) in which equal increments are perceived by human ears as roughly equivalent (Attneave and Olson, 1971; Russo and Thompson, 2005), permitting comparisons across frequencies (Grieser and Kuhl, 1988). The data were used to plot time-normalized tone contours for comparisons. Nine acoustic parameters were computed and compared across tones. Table III lists the acoustic terminology used in this study and the definitions of the nine acoustic parameters. To normalize individual differences in the intrinsic pitch level of the speakers and the pitch level differences between male and female speakers, Pitch Heights, which indexed the difference between the measured pitch and the mean pitch of the speaker (Speaker Mean Pitch) were used to represent the pitch level of the productions. For example, a pitch height of −2.58 indicated that the measured pitch was 2.58 semitones lower than the speaker's mean pitch. Because the pitch targets of the tones lie in the second half of the syllable, most of the acoustic parameters were measured from the second half of the tone contours and were marked with "50%."

III. RESULTS

A. Perceived tone production accuracy

1. Interjudge and intrajudge reliability

Tone production accuracy of the speakers was defined by the number of judges who correctly identified the intended tones. Kappa statistics were used to determine interjudge and intrajudge reliability. Based on the conventional interpretation of the kappa values (Landis and Koch, 1977; Posner *et al.*, 1990), Fleiss kappa analysis showed substantial inter-rater reliability among the five judges ($\kappa = 0.76$). Cohen's kappa showed substantial agreement of each pair of the five judges (κ ranged from 0.73 to 0.79), and the five judges also demonstrated perfect intrajudge reliability, $\kappa = 0.93, 0.82, 0.84, 0.91, 0.83$.

2. Results of perceived tone production accuracy

Table IV shows the perceived accuracy and the confusion patterns of the judges' perception of the 3582 productions by the speakers. None of the nine tones were perceived by the judges with 100% accuracy. Judgment accuracy of the tones produced by the speakers ranged from 73% to 96%. Most of the errors with the non-entering tones involved the confusions between the two rising tones [T2 (HR)–T5 (LR)], two of the three level tones [T3 (ML)–T6 (LL)], and the low-level and low-rising tones [T4 (LF)–T6 (LL)]. The error patterns among the entering tones followed the error patterns among their non-entering tone counterparts, with more confusions between the two lower tones, T8 (MS)–T9 (LS). Few entering tone productions were perceived as non-entering tones or vice-versa, except that 3% of T9 (LS) were perceived as the full falling tone, T4 (LF).

Further analysis was performed to gain a better understanding of the judgment errors of the produced tones. Because the data violated the assumptions for parametric

TABLE IV. Judges' categorization of the speakers' tone productions.

	Judges' responses (%)								
Target tones	T1(HL)	T2(HR)	T3(ML)	T4(LF)	T5(LR)	T6(LL)	T7(HS)	T8(MS)	T9(LS)
T1(HL)	94[a]	0	5	0	0	0	1	0	0
T2(HR)	0	85[a]	0	0	15[b]	0	0	0	0
T3(ML)	2	0	73[a]	0	0	23[b]	0	2	0
T4(LF)	0	1	0	86[a]	2	7[b]	0	1	3
T5(LR)	0	6[b]	1	0	93[a]	0	0	0	0
T6(LL)	0	0	7[b]	4	3	85[a]	0	0	1
T7(HS)	1	0	0	0	0	0	96[a]	3	0
T8(MS)	0	0	1	0	0	0	3	78[a]	18[b]
T9(LS)	0	0	0	4	0	0	0	11[b]	85[a]

[a]Perceived accuracy for the tones.
[b]Error patterns that constitute more than 5% in the total trials of the target tones.

statistics, a Mann-Whitney test was performed to determine whether the gender of the speakers affected the perceived tone accuracy of the judges. No significant effect of gender on tone accuracy was found, $U = 37.5$, $p > 0.05$, $r = -0.21$, suggesting that male and female speakers performed similarly and judges rated the tones similarly for female and male speakers. Subsequent analysis on tone production accuracy collapsed the two gender groups.

To determine whether the lexical status of the words had any impact on the judgment accuracy of the produced tones, Wilcoxon Signed Ranks Tests were used to compare the judges' perceptual accuracy of each tone and with all tones combined in real words and non-words. The results showed non-significant results in all the comparisons, all p-values > 0.05. Therefore, real and non-words were combined in subsequent analyses.

A Friedman test was conducted to examine whether the perceived accuracy of the nine tones was comparable. The results revealed significant effect of tone type on tone accuracy, $\chi^2(8) = 86.87$, $p < 0.001$, indicating that the accuracy of the nine tones significantly differed from each other. Pairwise comparisons using Wilcoxon signed-rank tests after adjusting the critical p-value for multiple comparisons showed that the accuracy of T7 (HS) was significantly higher than the accuracy in a majority of the tones (i.e., T4, T2, T9, T6, T8, T3; $ps < 0.01$, $r = -0.81$ to -0.88) while T3 (ML) had significantly lower tone accuracy than a majority of tones (T9, T2, T5, T1, T7; $ps < 0.05$, $r = -0.75$ to -0.88). The judgment accuracy of the entering tones and their non-entering tone counterparts was comparable, T1 (HL) ≈ T7 (HS), T3 (ML) ≈ T8 (MS), T6 (LL) ≈ T9 (LS), $ps > 0.05$.

B. Acoustic characteristics of the nine tones

To characterize the acoustic properties of unambiguous Cantonese tones, only productions in which the tones were correctly identified by all five judges were selected for analysis. Productions of three speakers (F04, F10, and M09) were further excluded from analysis because none of their T2 (HR) or T3 (ML) productions were correctly identified by all five judges. Thus, there were 171, 109, 71, 108, 145, 114, 517, 249, 318 tokens for T1 (HL) to T9 (LS), respectively. Subsequent acoustic and statistical analysis were based on these 1802 100% correctly perceived productions contributed by 8 female speakers and 9 male speakers. The average pitch contours of the tones produced by the male and female speakers are presented in Fig. 1.

Figure 2 presents the pitch contours of the nine tones with relative duration. In each penal, the durations of the tones were plotted in proportion to the duration of T3 (ML), the longest tone. The upper panels show the pitch contours in the whole vocalic section, representing the complete tonal contours. The lower panels show the tone contours in the final 50% of the vocalic section, representing the pitch targets of the tones. Table V presents the means and standard deviations of the acoustic parameters of the nine tones measured in Hz. Creaky phonation was found in the low F0 range in 3% of T2 (HR), 3% of T5 (LR), 6% of T9 (LS), and 69% of T4 (LF) productions.

A two-way mixed analysis of variance using gender as between-subject variable and tone type as within-subject variable was performed on each of the nine acoustic parameters presented in Table III to test whether the tones produced by males and females were acoustically different after using Pitch Height measures to adjust for individual vocal pitch levels, and to examine whether the acoustic parameters differed in male and female speakers in the nine tones. No significant main effect of gender was found for any of the acoustic measures, $ps > 0.05$, except that females had significantly higher initial pitch height (M = 1.58 St) than males (M = 0.80 St), $F(1, 15) = 14.07$, $p = 0.002$, $\eta^2 = 0.484$. The

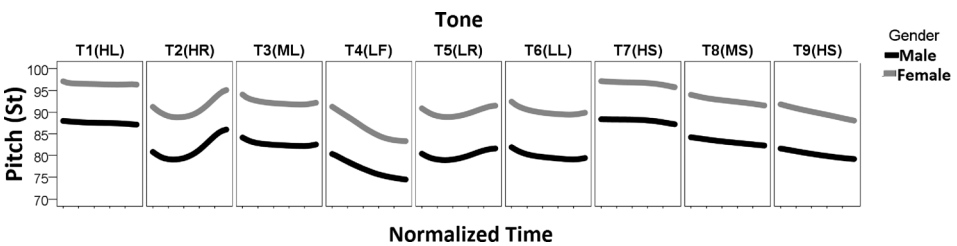

FIG. 1. Mean pitch contours of the tones produced by male and female speakers.

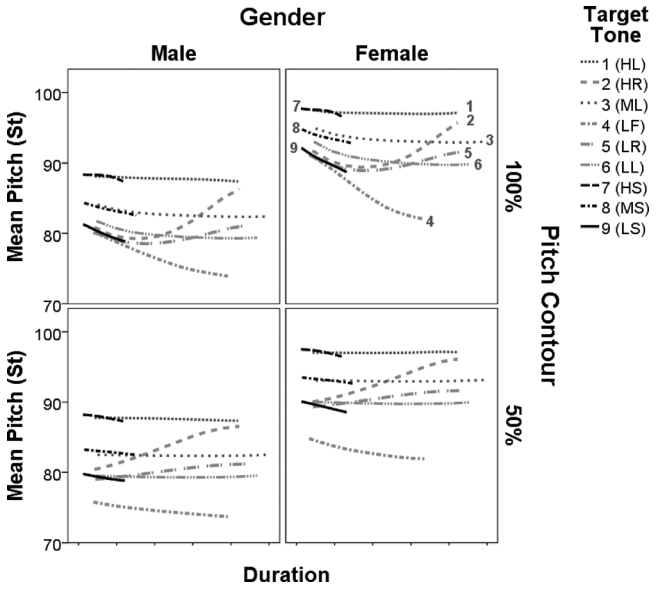

FIG. 2. Pitch contours of the nine tones produced by male and female speakers plotted with proportional duration in the whole vocalic section (100%) (upper panels) and the latter half of the vocalic section (50%) (lower panels). In each panel, the *x* axis represents the duration of T3 (ML), the longest tone. The durations of all other tones were plotted in proportion to the duration of T3 (ML). The contours in the lower panels show the contour of the pitch targets, which are the second half of the pitch contours of the tones.

gender by tone-type interaction effect was also not significant for all acoustic parameters, $ps > 0.05$, suggesting that Pitch Height measurements successfully normalized the intrinsic pitch differences between male and female speakers, and the tones produced by male and female speakers were acoustically similar. Thus, male and female productions were collapsed for subsequent acoustic analyses.

As expected, main effects of tone type were significant for all the nine acoustic parameters, with F values ranged from 109.53 to 348.23, all $ps < 0.01$, $\eta^2 = 0.880$ to 0.959 (Greenhouse-Geisser corrected). Pairwise comparisons with Bonferroni corrections for multiple comparisons were performed to compare the similarities and differences of the acoustic parameters (1) among the nine tones, (2) between the entering and non-entering tone counterparts, and (3) between the tone pairs that have been reported in the literature to be more confusing even for native Cantonese-speaking adults. Tables VI and VII summarize how each acoustic parameter differs in the nine tones. Significant differences are marked by ">" or "<." Non-significant differences are marked by "=." When a tone is not significantly different from two tones that are significantly different from each other, it is put in parenthesis and repeated on both sides of the ">" or "<" sign, e.g., "LL (= LS) > (LS=) HR" means LL is significantly larger than HR, while LS is not significantly different from either HR or LS. In terms of pitch heights, as expected, T1 (HL) and T7 (HS) have the highest mean pitch heights, while T4 (LF) has the lowest mean pitch heights across all tones. T3 (ML) has pitch heights lower than T1 (HL) but higher than T6 (LL). T2 (HR) has higher pitch than T5 (LR). In terms of the pitch range of the pitch targets, contour tones have wider pitch ranges than level tones or entering tones. In terms of pitch shapes, the level tones have very shallow slopes [−0.6 St/

TABLE V. Means and standard deviations of the acoustic parameters of the nine tones by gender. Standard deviations are shown in parentheses.

	Female speakers Acoustic parameters							
Tones	Initial F0 (Hz)	Final F0 (Hz)	Initial F0 50% (Hz)	Min F0 50% (Hz)	Max F0 50% (Hz)	Mean F0 50% (Hz)	Slope 50% (Hz/ms)	Tone duration (ms)
T1(HL)	279 (24)	269 (25)	269 (22)	263 (23)	274 (23)	269 (22)	−11 (141)	421 (69)
T2(HR)	201 (22)	259 (28)	182 (18)	182 (18)	260 (27)	217 (17)	423 (111)	422 (67)
T3(ML)	240 (23)	217 (21)	215 (19)	212 (20)	219 (21)	215 (19)	38 (67)	499 (75)
T4(LF)	194 (26)	116 (30)	135 (23)	112 (24)	139 (25)	123 (24)	−147 (204)	335 (77)
T5(LR)	189 (24)	196 (20)	172 (15)	172 (15)	198 (19)	185 (17)	146 (47)	428 (85)
T6(LL)	212 (22)	179 (18)	179 (15)	173 (16)	182 (16)	177 (15)	37 (115)	454 (90)
T7(HS)	280 (25)	259 (26)	276 (25)	259 (26)	277 (26)	269 (25)	−348 (343)	118 (34)
T8(MS)	238 (20)	209 (18)	220 (17)	209 (18)	220 (17)	215 (18)	−186 (138)	147 (41)
T9(LS)	204 (25)	165 (23)	181 (20)	165 (23)	181 (20)	173 (21)	−305 (185)	131 (47)
	Male speakers Acoustic parameters							
Tones	Initial F0 (Hz)	Final F0 (Hz)	Initial F0 50% (Hz)	MinF0 50% (Hz)	MaxF0 50% (Hz)	MeanF0 50% (Hz)	Slope 50% (Hz/ms)	Tone Duration (ms)
T1(HL)	163 (18)	156 (17)	160 (15)	153 (14)	162 (16)	158 (14)	−10 (135)	396 (82)
T2(HR)	107 (12)	148 (13)	104 (7)	104 (7)	149 (13)	126 (9)	264 (80)	399 (63)
T3(ML)	129 (11)	118 (7)	118 (8)	115 (7)	119 (8)	117 (7)	10 (51)	473 (81)
T4(LF)	103 (11)	71 (11)	79 (9)	70 (10)	80 (9)	75 (10)	−60 (45)	371 (89)
T5(LR)	102 (8)	109 (8)	96 (5)	96 (5)	110 (7)	103 (6)	77 (41)	418 (75)
T6(LL)	113 (13)	99 (9)	99 (10)	96 (8)	101 (10)	98 (9)	11 (49)	444 (103)
T7(HS)	168 (20)	156 (15)	166 (17)	156 (15)	166 (17)	162 (15)	−214 (158)	111 (39)
T8(MS)	132 (12)	118 (8)	123 (9)	118 (8)	124 (9)	121 (8)	−95 (93)	143 (43)
T9(LS)	110 (12)	95 (9)	101 (9)	95 (9)	101 (9)	98 (9)	−117 (94)	115 (49)

ms, 1.45 St/ms, and 3.3 St/ms, for T1 (HL), T3 (ML) and T6 (LL), respectively], suggesting relatively flat pitch contours. The entering tones have significantly larger falling slopes than the level tones, but comparable to the slope of T4 (LF), indicating that the shapes of the entering tones are different from those of the level tones. T2 (HR) has a significantly larger positive slope than T5 (LR). In terms of duration, among the non-entering tones, T3 (ML) is the longest, while T4 (LF) is the shortest. All the other non-entering tones are of comparable duration. The entering tones are shorter than any of the non-entering tones, with T8 (MS) longer than T7 (HS) and T9 (LS) (Tables VI and VII). Table VIII shows the duration differences in different syllable structures in different tones. Future studies will be needed to determine if the observed duration differences are meaningful acoustic cues for identifying the tones in typical native Cantonese tone speakers and other speaker groups.

The acoustic similarities and differences between the non-entering tones and the entering tone counterparts are presented in the upper panels in Table IX, while the acoustic similarities and differences between the tones in the three more confusing tone pairs are presented in the lower panels in Table IX. The cells marked with superscript a indicate the distinctive acoustic characteristics of the tone pairs. The three pairs of full and entering tones all differ in duration and tone shapes, with the entering tones having significantly steeper falling slopes than the level tones. The three confusing tone pairs all differ in pitch heights. In addition, T2 (HR) and T5 (LR) also differ by pitch ranges and pitch slopes, while T4 (LF) and T6 (LL) differ by pitch range, pitch slope and pitch duration (Table IX). The pitch target of T2 (HR) starts at a pitch level lower than that of T3 (ML) but higher than that of T5 (LR), and ends at a pitch level as high as T1 (HL) and T7 (HS) (Table VI). The minimum pitch of T2 (HR) is lower than that of T3 (ML) but higher than that of T5 (LR) and T6 (LL). The pitch target of T5 (LR) starts at a pitch level comparable to that of T6 (LL) and ends at a pitch level higher than that of T6 (LL) but lower than that of T3 (ML). The pitch target of T4 (LF) starts and ends lower than that of T6 (LL) (Tables V and VI).

IV. DISCUSSION

This is the first study that compared the acoustics of lexical tones produced by male and female adults. Given that the mean F0 of male adults, female adults, five-year-olds and seven-year-olds are about 108 Hz (SD = 25 Hz), 205 Hz (SD = 43 Hz), 248 Hz (SD = 50 Hz), and 260 Hz (SD = 48 Hz), respectively (Katz and Assmann, 2001), comparing tones produced by speakers of different genders or different age groups has been challenging. The pitch level of a high level tone produced by a male speaker can be lower than the pitch level of a low level tone produced by a female speaker. This study applied the acoustic method developed by Wong and colleagues (Wong, 2012a; Wong *et al.*, 2017; Wong and Strange, 2017), which has been proven to be effective in normalizing the pitch differences in young children and women. By comparing the pitch values relative to the speaker's average pitch level (i.e., subtracting the mean pitch of the speaker from the measured pitch, the so-called pitch height values), this study successfully normalized the intrinsic pitch differences between male and female adults as evidenced by the lack of statistical difference between male and female speakers in all the acoustic parameters measured in all the tones, except the Initial Pitch Height, which is not the perceptual cue or the perceptual target of tones (Xu, 2001; Khouw and Ciocca, 2007). The pitch shape and duration comparisons of the tones produced by male and female speakers also revealed no significant differences.

To obtain more accurate and detailed acoustic characteristics of correctly produced Cantonese tones, the design of this study differed from that in previous studies in several aspects. First, given that previous studies consistently

TABLE VI. Differences of the acoustic parameters in the nine tones. "=" indicates no statistical difference at 0.05 level. "<" or ">" indicates statistical differences at 0.05 level after Bonferroni corrections for multiple comparisons. When a tone is not significantly different from two tones that are significantly different from each other, it is put in parenthesis and repeated on both side of the ">" or "<" signs, e.g., "T6 LL (= T9 LS) > (T9 LS =) T2 HR" means T6 LL is significantly larger than T2 HR, while T9 LS is not significantly different from either T2 HR or T6 LL. See Table III for the definitions of the acoustic parameters.

Acoustic parameters	Significant differences (Order arranged from largest to smallest values)	Effect sizes of significantly different groups
Initial pitch height	T7 HS = T1 HL > T8 MS = T3 ML > T6 LL (=T9 LS)> (T9 LS=) T2 HR (=T4 LF) > (=T4 LF) =T5 LR	r = 0.74–0.98
Final pitch height	T1 HL = T7 HS = T2 HR > T3 ML = T8 MS > T5 LR > T6 LL > T9 LS > T4 LF	r = 0.72–0.98
Mid pitch height	T7 HS = T1 HL > T8 MS > T3 ML > T2 HR (= T9 LS = T6 LL) > (T9 LS = T6 LL =) T5 LR > T4 LF	r = 0.87–0.99
Min pitch height 50%	T1 HL = T7 HS > T8 MS = T3 ML > T2 HR > T5 LR = T6 LL = T9 LS > T4 LF	r = 0.71–0.99
Max pitch height 50%	T7 HS = T1 HL > T2 HR > T8 MS > T3 ML > T5 LR > T6 LL = T9 LS > T4 LF	r = 0.77–0.99
Mean pitch height 50%	T7 HS = T1 HL > T2 HR = T8 MS = T3 ML > T5 LR > T6 LL = T9 LS > T4 LF	r = 0.82–0.99
Pitch range 50%	T2 HR > T4 LF = T5 LR > T9 LS (= T7 HS = T6 LL) > (T7 HS = T6 LL =) T8 MS = T1 HL = T3 ML	r = 0.70–0.98
Slope 50%	Positive slopes (Rising): T2 HR > T5 LR > T6 LL = T3 ML (=T1 HL)[a]	r = 0.80–0.98
	Negative slopes (Falling):[b] T9 LS (=T7 HS = T4 LF) > (T7 HS = T4 LF=) T8 MS > T1 HL	
Pitch duration	T3 ML (= T6 LL)> (= T6 LL =) T5 LR = T2 HR = T1 HL > T4 LF > T8 MS > T9 LS = T7 HS	r = 0.71–0.99

[a]T1 HL has a negative value in slope 50% (Table III), but is not significantly different from T3 ML, which has a positive slope 50%.
[b]Order is arranged from more negative to less negative slopes, e.g., "T8 MS > T1 HL" means T8MS has a more negative slope and falls more sharply than T1 HL.

TABLE VII. Acoustic characteristics of the nine tones.

Tones	Acoustic parameters	Major characteristics
T1 (HL)	Pitch height measures (see Table III)	Not different from those in T7 (HS), $ps > 0.05$
		Final pitch height not significantly different from T2 (HR), $p > 0.05$
		Higher than the other tones, $ps < 0.001$
	Pitch range 50%	Not significantly different from the other two level tones (ML, LL), or two of the entering tones (HS, MS), $ps > 0.05$
		Smaller than the other tones, $ps < 0.01$
	Slope 50%	Different from the rising and falling tones (HR, LR, LF), $ps < 0.01$
		Not significantly different from the other two level tones (ML, LL), $ps > 0.05$
		Value approaching 0 (mean slope 50% = −0.60 St/ms), indicating level shape
	Pitch duration	Shorter than T3 (ML), longer than T4 (LF) and the entering tones, $ps < 0.01$. Not significantly different from the other tones, $ps > 0.05$
T2 (HR)	Initial pitch height	Lower than T6 (LL), $p < 0.01$
		Higher than T5 (LR), $p < 0.05$
		Not significantly different from T4 (LF), $p > 0.05$
	Final pitch height	Not significantly different from T1 (HL), $p > 0.05$
	Pitch range 50%	Larger than all tones, $ps < 0.001$
	Slope 50%	Significantly more positive (i.e., rises more steeply) than any tones (mean slope 50% = 35.11 St/ms), $ps < 0.001$
	Pitch duration	Shorter than T3 (ML), longer than T4 (LF) and the entering tones, $ps < 0.01$. Not significantly different from the other tones, $ps > 0.05$
T3 (ML)	Pitch height measures	Lower than T1(HL), $ps < 0.001$
		Higher than T6(LL), $ps < 0.001$
	Pitch range 50%	Not significantly different from the other level tones (HL, LL), or two of the entering tones (HS, MS), $ps > 0.05$
		Smaller than the other tones, $ps < 0.001$
	Slope 50%	Different from the rising and falling tones (HR, LR, LF), $ps < 0.001$
		Not significantly different from the other level tones (HL, LL), $p > 0.05$
		The value is close to 0 (mean slope 50 = 1.45 St/ms), indicating level shape
	Pitch duration	Not significantly different from T6 (LL), $p > 0.05$
		Longer than all the other tones, $ps < 0.01$
T4 (LF)	Pitch height measures	Lowest among all tones, $ps < 0.01$, except that initial pitch height was not significantly different from that in T5(LR), $p > 0.05$
	Pitch range 50%	Smaller than T2 (HR), $p < 0.001$.
		Larger than any of the level tones (HL, ML, LL), $ps < 0.01$
		Not significantly different from T5(LR), $p > 0.05$
	Slope 50%	Not significantly different from any of the entering tones (HS, MS, LS), $ps > 0.05$
		More negative (falls more sharply) than any of the non-entering tones (Mean Slope 50 = −17.41 St/ms), $ps < 0.01$
	Pitch duration	Shortest among the non-entering tones, $ps < 0.01$
		Longer than the entering tones, $ps < 0.001$
T5 (LR)	Initial pitch height	Not significantly different from T4(LF), $p > 0.05$
		Lower than all the other tones, $ps < 0.05$
	Final pitch height	Lower than T3 (ML), $p < 0.01$
		Higher than T6 (LL), $p < 0.001$
	Pitch range 50%	Smaller than T2 (HR), $p < 0.001$
		Larger than the non-entering leveltones (HL, ML, LL), $ps < 0.001$
		Not significantly different from T4 (LF), $p > 0.05$
	Slope 50%	More positive than the level tones (HL, ML, LL) and the falling tone, T4 (LF), $ps < 0.01$
		Less positive (rises less sharply) than T2 (HR), (mean slope 50 = 13.36 St/ms), $p < 0.001$
	Pitch duration	Shorter than T3 (ML), longer than T4 (LF) and the entering tones, $ps < 0.01$. Not significantly different from the other tones, $ps > 0.05$
T6(LL)	Pitch height measures	Higher than T4 (LF), $ps < 0.01$
		Lower than all the other non-entering tones, $ps < 0.05$, except Initial Pitch Height
	Pitch range 50%	Not significantly different from the other level tones (HL and ML), or the entering tones (HS, MS, LS), $ps > 0.05$
		Smaller than the other tones, $ps < 0.001$
	Slope 50%	Different from the rising and falling tones (HR, LR, LF), $ps < 0.01$
		Not significantly different from the other two level tones (HL, ML), $ps > 0.05$
		Value is close to 0 (mean slope 50 = 3.30 St/ms), indicating level shape
	Pitch duration	Not significantly different from the non-entering tones, except longer than T4(LF), $ps > 0.05$
		Longer than the entering tones (HS, MS, LS), $ps < 0.001$
T7(HS)	Pitch height measures	Highest among the entering tones, $ps < 0.001$
		Not significantly different from T1(HL), $ps > 0.05$
	Pitch range 50%	Not significantly different from the level tones (HL, ML, LL), and the other entering tones (MS, LS), $ps > 0.05$

TABLE VII. *Continued*

Tones	Acoustic parameters	Major characteristics
		Smaller than the other tones, $ps < 0.01$
	Slope 50%	More negative than T1(HL), (mean slope 50 = −21.20 St/ms), $p < 0.001$
		Not significantly different from T8(MS), T9(LS), or T4 (LF), $p > 0.05$
		Direction different from the other level tones (ML, LL), which have positive slopes, $ps < 0.001$
	Pitch duration	Shorter than any of the non-entering tones, $ps < 0.001$
T8(MS)	Pitch height measures	Lower than T1 (HL) and T7 (HS), $ps < 0.001$
		Higher than T6 (LL) and T9 (LS), $ps < 0.001$
		Not statistically different from T3(ML), except having higher initial pitch height 50% and max pitch height 50%
	Pitch Range 50%	Smallest among the tones, $ps < 0.01$
		Not significantly different from the three level tones (HL, ML, LL), or T7(HS), $ps > 0.05$
	Slope 50%	Less negative (falls less sharply) than T9(LS), (mean slope 50% = −14.55 St/ms), $p < 0.001$
		Not significantly different from T4(LF) or T7(HS), $ps > 0.05$
		More negative than the level tone T1(HL), and direction different from the other two level tones (ML, LL), which have positive slopes, $p < 0.001$
	Pitch duration	Shorter than the non-entering tones, $ps < 0.001$
		Longer than T7(HS) and T9(LS), $ps < 0.01$
T9(LS)	Pitch height measures	Lowest among the entering tones, $ps < 0.001$
		Not statistically different from T6 (LL), except having a lower final pitch height
	Pitch range 50%	Larger than T8(MS), $p < 0.01$
		Not significantly different from T7(HS), $p > 0.05$
	Slope 50%	More negative (falls more sharply) than T8(MS), (mean slope 50% = −24.88 St/ms), $p < 0.001$
		More negative than the level tone T1(HL), and direction different from the other two level tones (ML, LL), which have positive slopes, $p < 0.001$
		Not significantly different from T7 (HS) or T4 (LF), $ps > 0.05$
	Pitch duration	Shorter than the non-entering tones, $ps < 0.001$

reported incorrect Cantonese tone perception and production in native adults (e.g., Barry and Blamey, 2004; Ciocca and Lui, 2003; Wong *et al.*, 2017; Wong and Leung, 2018), to ensure that only correctly produced tones were used for acoustic analysis, this study only included productions in which the tones had been correctly identified by all the five judges in filtered stimuli. Second, to exclude any possible tone mergers, speakers who failed to produce any of the nine tones with 100% perceived accuracy were excluded. Third, to control tone judgment biases caused by the lexical status of the words and word familiarity, this study asked judges to categorize the tones in filtered speech (Wong, 2012b; Wong, 2013). Fourth, in addition to pitch levels and durations, this study also compared the pitch shapes of the tones. Fifth, to gain better understanding of the challenges in perceiving and producing Cantonese tones, this study compared and contrasted the acoustic properties of all the nine tones and the more confusing tones in Cantonese. Sixth, based on previous empirical findings that the initial half of the tones was influenced by factors such as co-articulation (see Fig. 2 in Xu, 2001) and consonantal contexts (see Fig. 2 in Wong and Xu, 2007), and was not as reliable as the second half of the tones for tone identification (Khouw and Ciocca, 2007) and with reference to the proposition of the Target Approximation Model of Tonal Contour Formation that the perceptual targets of tones occur towards the end of the syllable (Xu, 2001, 2004), this study focused the acoustic comparisons on the pitch contours in the second half of the tones.

TABLE VIII. Average duration of the tones in different syllable structures. CV means consonant-vowel syllable structure. CVC means consonant-vowel-consonant syllable structure. Duration is defined by the duration of the whole vocalic potion in the syllable.

Average tone duration (ms)				
	Tones	CV structure	CVC Structure	Collapsing CV and CVC structures
Non-entering tones	T1(HL)	377.98	450.82	400.98
	T2(HR)	388.34	442.80	403.83
	T3(ML)	458.78	518.32	482.26
	T4(LF)	339.25	373.98	350.50
	T5(LR)	396.93	463.60	418.08
	T6(LL)	433.11	482.98	444.92
Entering tones	T7(HS)	—	113.17	—
	T8(MS)	—	143.16	—
	T9(LS)	—	123.16	—

A. Perceived accuracy of adults' tones

Consistent with the findings in previous research (Lee *et al.*, 2015; Wong *et al.*, 2017), this study found that not all tones produced by native Cantonese-speaking adults were correctly identified by the judges. The major error patterns of the non-entering tones involved the three tone pairs that have been reported in previous literature to be particularly confusing [i.e., T2 (HR)-T5 (LR), T3 (ML)-T6 (LL), and T4 (LF)-T6 (LL)] (Table IV). No previous studies reported perceived accuracy or confusion patterns on adults' entering tones. This study found that native Cantonese-speaking adults also made considerable errors with the entering tones, and the confusion patterns of the entering tones followed the

TABLE IX. Acoustic similarities and differences in selected tone pairs. Duration is defined by the whole vocalic potion in the syllable.

Types of comparison	Tone pairs	Acoustic parameters							
		Initial pitch height	Final pitch height	Mid pitch height	Min pitch height 50%	Max pitch height 50%	Mean pitch height 50%	Pitch range 50%	Slope 50%	Tone duration
Non-entering tones vs entering tones	T1(HL) vs T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T1(HL) = T7(HS)	T7(HS) more negative than T1(HL)[a]	T1(HL) > T7(HS)[a]
	T3(ML) vs T8(MS)	T3(ML) = T8(MS)	T3(ML) = T8(MS)	T8(MS) > T3(ML)[a]	T3(ML) = T8(MS)	T8(MS) > T3(ML)[a]	T3(ML) = T8(MS)	T3(ML) = T8(MS)	T3(ML) positive, T8(MS) negative[a]	T3(ML) > T8(MS)[a]
	T6(LL) vs T9(LS)	T6(LL) = T9(LS)	T6(LL) > T9(LS)[a]	T6(LL) = T9(LS)	T6(LL) = T9(LS)	T6(LL) = T9(LS)	T6(LL) = T9(LS)	T6(LL) = T9(LS)	T6(LL) positive, T9(LS) negative[a]	T6(LL) > T9(LS)[a]
Confusing tone pairs	T2(HR) vs T5(LR)	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) > T5(LR)[a]	T2(HR) more positive than T5(LR)[a]	T2(HR) = T5(LR)
	T3(ML) vs T6(LL)	T3(ML) > T6(LL)[a]	T3(ML) > T6(LL)[a]	T3(ML) > T6(LL)[a]	T3(ML) > T6(LL)[a]	T3(ML) > T6(LL)[a]	T3(ML) > T6(LL)[a]	T3(ML) = T6(LL)	T3(ML) = T6(LL)	T3(ML) = T6(LL)
	T4(LF) vs T6(LL)	T6(LL) > T4(LF)[a]	T6(LL) > T4(LF)[a]	T6(LL) > T4(LF)[a]	T6(LL) > T4(LF)[a]	T6(LL) > T4(LF)[a]	T6(LL) > T4(LF)[a]	T4(LF) > T6(LL)[a]	T6(LL) positive, T4(LF) negative[a]	T6(LL) > T4(LF)[a]

[a]Significant difference between tone pairs reported at 0.05 level.

error patterns in the non-entering tone counterparts. There was little confusion between the high and mid entering tones, T7 (HS) and T8 (MS), but substantial confusion between the mid and low entering tones, T8 (MS) and T9 (LS).

Listeners were able to discriminate non-entering tones from entering tones when they were asked to categorize the tones in nine categories. Only 3% of T9 (LS) productions were perceived as T4 (LF). To determine whether the three entering tones could be considered allotones of the three non-entering level tones, we asked five judges to categorize 2084 filtered monosyllabic words (full tones: n = 1383, entering tones: n = 701) produced by 51 native Cantonese speaking females (aged 24 to 41 years) collected in another study into the six non-entering tone categories. The results showed that 93%, 5% and 1% of T7 (HS) were categorized as T1 (HL), T3 (ML), and T6 (LL), respectively; 56% and 42% of T8 (MS) were categorized as T3 (ML) and T6 (LL), respectively; and 89% and 10% of T9 (LS) were categorized as T6 (LL) and T3 (ML), respectively. The findings indicated that when listeners were asked to categorize the entering tones into the six non-entering tone categories, a majority of the entering tones (56%–93%) were categorized as their non-entering tone counterparts. Interestingly, though the entering tones had comparable falling slopes as T4 (LF), only 0% to 2% of the entering tones were perceived as T4 (LF). The reason for listeners not to categorize the entering tones as T4 (LF) but to categorize them as the three level non-entering tones is not clear given the design of the current study. It could be possible that falling contours in brief syllables were not salient and, therefore, difficult to detect, or listeners did not hear the falling contours in entering tones because they were not reliable cues for discriminating the tones and were, therefore, ignored. Another possibility is that the listeners were able to hear the falling contours in entering tones but they classified the tones into tone categories that matched the primary cue for discriminating the tones (i.e., pitch level) or tone categories that best matched the acoustic characteristics of the entering tones (i.e., pitch level and pitch range). Future perceptual studies on naturally produced tones and synthetic pitch contours will be needed to reveal the cues listeners used to categorize entering tones.

B. Acoustic characteristics of the nine Cantonese tones

The acoustic data in this study supported the claim in previous studies that the tone contours were less distinguishable in the first half of the syllable than in the second half of the syllable (e.g., Khouw and Ciocca, 2007; Xu, 2001, 2004). As shown in the upper panels in Fig. 2, the tone contours of T2 (HR), T5 (LR), T6 (LL), and T9 (LS) have similar pitch heights and pitch shapes in the first quarter of the syllable and the tone contours of T2 (HR) and T5 (LR) do not separate until around half way into the syllable. In the following discussions on the acoustic characteristics of the Cantonese tones, other than tone durations, which compare the length of the whole vocalic segment, and initial pitch heights, which compare the pitch at the onset of the vocalic

segment, discussions of all other acoustic properties are focused on the perceptual target of tones in the second half of the syllable.

The primary aim of this study was to examine the acoustic properties of distinctly produced Cantonese tones. Regarding the non-entering tones, as shown in Tables VI and VII, the pitch target of T1 (HL) maintains a pitch level as high as the final pitch of T2 (HR). It is shorter than T3 (ML), longer than T4 (LF) and comparable to the other non-entering tones. T2 (HR) starts at a pitch higher than the pitch onset of T5 (LR), equal to the onset of T4 (LF), and lower than the pitch at the onset of T3 (ML) and the other tones. It ends at a pitch as high as the final pitch of T1 (HL). Its duration is shorter than T3 (ML), longer than T4 (LF), and comparable to all other non-entering tones. T3 (ML) maintains a pitch level significantly higher than the pitch level of T6 (LL) and significantly lower than the pitch in T1 (HL). It is the longest tone though not significantly longer than T6 (LL). T4 (LF) starts at a pitch level similar to the onset of T5 (LR) and T2 (HR). It has a falling slope and usually ends with creaky/glottalized phonation. It is the shortest among the non-entering tones. T5 (LR) starts at a pitch level similar to the pitch at the onset of T4 (LF) and ends at a pitch level higher than that in T6 (LL) but lower than that in T2 (HR) and T3 (ML). It has a rising pitch contour with a pitch range similar to that of T4 (LF), but smaller than that in T2 (HR). It is shorter than T3 (ML), longer than T4 (LF) and comparable to T2 (HR) and all the other non-entering tones. T6 (LL) maintains a pitch level higher than that in T4 (LF) but lower than any of the other non-entering tones throughout the tone. Its duration is longer than that of T4 (LF) and comparable to that of all other non-entering tones.

With respect to the entering tones, the durations of all the three entering tones are about a quarter to one third of the durations of their non-entering tone counterparts in CVN and CV syllables, respectively (Table V), with the duration of T8 (MS) longer than that in T7 (HS) and T9 (LS). The reason for the longer duration of in T8 (MS) than T7 (HS) andT9 (LS) is due to the phonological rule in Cantonese in which "yin" (i.e., upper) entering tone is split into two tones with the higher tone [i.e., T7 (HS)] for short vowels and the lower tone [i.e., T8 (MS)] for long vowels. It could be an additional cue for distinguishing the three entering tones. The pitch targets of T7 (HS), T8 (MS), and T9 (LS) have mean pitch level and pitch range comparable to those in T1 (HL), T3 (ML), and T6 (LL), respectively. All the three entering tones have falling contours and the slopes of the contours are not different from the slope in T4 (LF).

C. Acoustic similarities and differences between the entering tones and the non-entering level tones

Few studies on Cantonese tones examined the entering tones and most of them reported that the entering level tones were not different from the three non-entering tone counterparts, except for shorter durations (e.g., Bauer and Benedict, 1997; Chao, 1947; Hashimoto, 1972). As indicated above, the three entering tones have the same pitch heights and pitch range as the corresponding level tones; however, the entering tones are about three quarters to two thirds shorter than the non-entering tones, and unlike the three level tones which have essentially flat contours, the entering tones have falling contours similar to that of T4 (LF) and significantly different from the slopes of the non-entering level tones. These findings indicate that the entering tones span across the same frequency range as the corresponding level tones, but drop within a much shorter time frame, resulting in a contour with falling slopes. Given the significant acoustic differences in durations and contour shapes between the entering tones and the non-entering tones, the perceptual cues for these two groups of tones are different. Future studies on perception and production difficulties of Cantonese tones in different populations should separate the two sets of tones to gain a clearer understanding of factors contributing to the difficulties.

D. Acoustic similarities and differences between more confusing non-entering tones

The perceptual difficulties between the tones in the easily confused tone pairs, namely, T3 (ML)-T6 (LL), T2 (HR)-T5 (LR), and T4 (LF)-T6 (LL), are likely due to the proximity in the pitch shapes and pitch levels of the tones in the tone pairs. As presented in Fig. 2, among the non-entering tones, the contours of the tones in each confusing tone pair are in much closer vicinity to each other than to other non-entering tones. Also, the difference in the pitch level is larger between T1 (HL) and T3 (ML) than between T3 (ML) and T6 (LL) and, thus, there is less confusion between T1 (HL) and T3 (ML) than T3 (ML) and T6 (LL).

All productions used in the acoustic analysis in this study were highly distinguishable tones, and, therefore, the acoustic differences in the easily confused tone pairs listed in Tables IV and VI can be taken as potential cues that distinguish the tone pairs. The findings suggested that in productions that clearly contrasted T2 (HR) and T5 (LR), T2 (HR) was produced with higher pitch levels throughout the tone contour, and achieved a larger pitch range and a more positive pitch slope than T5 (LR). The offset of T2 (HR) reached a pitch level as high as T1 (HL), while T5 (LR) ended at a pitch level between T3 (ML) and T6 (LL). In unambiguous T3 (ML) and T6 (LL) contrasts, T3 (ML) was produced and maintained at a pitch level higher than the pitch level at the midpoint of T2 (HR) and higher than the pitch levels in T6 (LL) throughout the contour. T6 (LL) was produced with a pitch level similar to the pitch level at the midpoint of T2 (HR) and maintained at pitch levels lower than the pitch of T3 (ML) throughout the tone. In clear T6 (LL) and T4 (LF) contrasts, T6 (LL) was produced with a slightly positive slope with a much smaller pitch range than in T4 (LF). The pitch throughout the contour of T6 (LL) was higher than that in T4 (LF). T4 (LF) was produced with the pitch level at the onset similar to the pitch onset of the two rising tones and had a falling contour with a pitch range larger than that of T6 (LL). These acoustic characteristics could be used to provide feedback when teaching the

production of Cantonese tones and to compare incorrect tone productions in different typical and atypical populations.

E. Summary

In sum, this is the first study that compared the acoustics of the tone produced by male and female adults, examined in greater details the acoustic characteristics of the six non-entering tones and the three entering tones in highly distinctive Cantonese tones, and compared the acoustical differences between non-entering tones and entering tones and between more confusing tone pairs. The acoustic method adopted in this study successfully normalized the intrinsic differences in the vocal pitch between male and female speakers and, therefore, can be used to examine and compare the acoustic similarities and differences in lexical tones and prosodic productions among individuals with different intrinsic pitch. The acoustic findings showed that the entering tones and the non-entering level tones have different contour shapes and durations, and, therefore, should be examined separately in future studies. The detailed acoustic measures presented in Table V can be used for tone modeling and creating synthetic tone stimuli for future research on Cantonese tone perception. The acoustic findings can also be used as references for future studies that examine the acoustic characteristics of tones in different populations, such as tone mergers, young children, and individuals with particular difficulties in tone perception and production such as individuals with hearing impairment, dysarthria, or dyslexia.

Attneave, F., and Olson, R. K. (**1971**). "Pitch as a medium: A new approach to psychophysical scaling," Am. J. Psychol. **84**, 147–166.

Barry, J. G., and Blamey, P. J. (**2004**). "The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese," J. Acoust. Soc. Am. **116**(3), 1739–1748.

Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton de Gruyter, Berlin), pp. 123–144.

Bauer, R. S., Cheung, K. H., and Cheung, P. M. (**2003**). "Variation and merger of the rising tones in Hong Kong Cantonese," Lang. Var. Change **15**(02), 211–225.

Chao, Y. R. (**1947**). *Cantonese Primer* (Cambridge University Press, Cambridge), pp. 1–242.

Cheung, H., Chung, K. K., Wong, S. W., McBride-Chang, C., Penney, T. B., and Ho, C. S. (**2009**). "Perception of tone and aspiration contrasts in Chinese children with dyslexia," J. Child Psychol. Psych. **50**(6), 726–733.

Ciocca, V., and Lui, J. (**2003**). "The development of the perception of Cantonese lexical tones," J. Multiling. Commun. Disorders **1**(2), 141–147.

Fok Chan, Y. Y. (**1974**). *A Perceptual Study of Tones in Cantonese, Occasional Papers and Monographs* (Centre of Asian Studies, University of Hong Kong, Hong Kong), Vol. 18.

Fu, Q. J., and Zeng, F. G. (**2000**). "Identification of temporal envelope cues in Chinese tone recognition," Asia Pacific J. Speech Lang. Hear. **5**(1), 45–57.

Grieser, D. L., and Kuhl, P. K. (**1988**). "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese," Development. Psychol. **24**(1), 14.

Hashimoto, O. Y. (**1972**). *Studies in Yue Dialects 1: Phonology of Cantonese* (Cambridge University Press, Cambridge), pp. 91–93.

Katz, W. F., and Assmann, P. F. (**2001**). "Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing," J. Phon. **29**(1), 23–51.

Kei, J., Smith, V., So, L. K., Lau, C. C., and Capell, K. (**2002**). "Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure," Asia Pacific J. Speech Lang. Hear. **7**, 25–38.

Khouw, E., and Ciocca, V. (**2006**). "Acoustic and perceptual study of Cantonese tones produced by profoundly hearing-impaired adolescents," Ear Hear. **27**(3), 243–255.

Khouw, E., and Ciocca, V. (**2007**). "Perceptual correlates of Cantonese tones," J. Phon. **35**(1), 104–117.

Landis, J. R., and Koch, G. G. (**1977**). "The measurement of observer agreement for categorical data," Biometrics **33**, 159–174.

Lee, K. Y., Chan, K. T., Lam, J. H., van Hasselt, C. A., and Tong, M. C. (**2015**). "Lexical tone perception in native speakers of Cantonese," Int. J. Speech-Lang. Pathol. **17**(1), 53–62.

Lee, K. Y., Van Hasselt, C. A., Chiu, S. N., and Cheung, D. M. (**2002a**). "Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children," Int. J. Pediatr. Otorhinolaryng. **63** (2), 137–147.

Lee, L. Y. S., Chan, T. Y., Ng, K. Y., van Hasselt, C. A., and Tong, C. F. (**2009**). "Validation of the Cantonese Tone Identification Test (CanTIT): Preliminary findings," in *7th Asia Pacific Symposium on Cochlear Implants and Related Sciences (APSCI)*, pp. 100–105.

Lee, T., Lau, W., Wong, Y. W., and Ching, P. C. (**2002b**). "Using tone information in Cantonese continuous speech recognition," ACM Trans. Asian Lang. Inf. Process. (TALIP) **1**(1), 83–102.

Lewis, M. P., Gary F. S., and Charles D. F. (**2018**). *Ethnologue: Languages of the World* (SIL International, Dallas, TX), http://www.ethnologue.com (Last viewed 1/12/2018).

Li, W. S., and Ho, C. S. H. (**2011**). "Lexical tone awareness among Chinese children with developmental dyslexia," J. Child Lang. **38**(4), 793–808.

Mok, P. P., Zuo, D., and Wong, P. W. (**2013**). "Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese," Lang. Var. Change **25**(03), 341–370.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., and Cheney, F. W. (**1990**). "Measuring interrater reliability among multiple raters: An example of methods for nominal data," Stat. Med. **9**(9), 1103–1115.

Rao, B., Ouyang, J., and Zhou, W. (**1996**). *Guangzhouhua Fangyan Cidian* (*Guangzhou Cantonese Dialect Dictionary*) (Commercial Press, Hong Kong), pp. 348–358.

Rose, P. (**2000**). "Hong Kong Cantonese citation tone acoustics: Alinguistic tonetic study," in *Proceedings of the 8th Australian International Conference on Speech Science and Technology*, pp. 198–203.

Rose, P. (**2004**). "The acoustics and probabilistic phonology of short stopped-syllable tones in Hong Kong Cantonese," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology* (Australian Speech Science and Technology Association, Canberra), pp. 445–450.

Russo, F. A., and Thompson, W. F. (**2005**). "An interval size illusion: The influence of timbre on the perceived size of melodic intervals," Atten. Percept. Psychophys. **67**(4), 559–568.

Vance, T. J. (**1976**). "An experimental investigation of tone and intonation in Cantonese," Phonetica **33**, 368–392.

Whitehill, T. L., Ciocca, V., and Chow, D. T. Y. (**2000**). "Acoustic analysis of lexical tone contrasts in dysarthria," J. Med. Speech-Lang. Pathol. **8**(4), 337–344.

Wong, P. (**2012a**). "Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions," J. Phon. **40**(1), 141–151.

Wong, P. (**2012b**). "Monosyllabic Mandarin tone productions by 3-year-olds growing up in Taiwan and in the United States: Interjudge reliability and perceptual results," J. Speech Lang. Hear. Res. **55**(5), 1423–1437.

Wong, P. (**2013**). "Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan," J. Acoust. Soc. Am. **133**(1), 434–443.

Wong, P., Fu, W. M., and Cheung, E. Y. L. (**2017**). "Cantonese-speaking children do not acquire tone perception before tone production—A perceptual and acoustic study of three-year-olds' monosyllabic tones," Front. Psychol. **8**, 1450.

Wong, P., and Leung, C. T. (**2018**). "Suprasegmental features are not acquired early: Perception and production of monosyllabic Cantonese lexical tones in four- to six-year-old children," J. Speech Lang. Hear. Res. (in press).

Wong, P., and Strange, W. (**2017**). "Phonetic complexity affects children's Mandarin tone production accuracy in disyllabic words: A perceptual study," PloS One **12**(8), e0182337.

Wong, P. C., Perrachione, T. K., Gunasekera, G., and Chandrasekaran, B. (**2009**). "Communication disorders in speakers of tone languages: Etiological bases and clinical considerations," Sem. Speech Lang. **30**(3), 162–173.
Wong, Y. W., and Xu, Y. (**2007**). "Consonantal perturbation of f0 contours of Cantonese tones," in *The 16th International Congress of Phonetic Sciences*, Saarbrucken, Germany, August, 2007, pp. 1293–1296.
Xu, C. X., and Xu, Y. (**2003**). "Effects of consonant aspiration on Mandarin tones," J. Int. Phon. Assoc. **33**(2), 165–181.
Xu, Y. (**2001**). "Sources of tonal variations in connected speech," J. Chin. Ling. **17**, 1–31.
Xu, Y. (**2004**). "Understanding tone from the perspective of production and perception," Lang. Ling. **5**(4), 757–797.
Xu, Y. (**2013**). "ProsodyPro—A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France.
Xu, Y., and Liu, F. (**2006**). "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," Italian J. Ling. **18**(1), 125.
Xu, Y., and Wang, Q. E. (**2001**). "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**(4), 319–337.
Yip, M. J. W. (**2002**). *Tone* (Cambridge University Press, Cambridge), pp. 1–16.
Zhang, J., McBride-Chang, C., Wong, A. M. Y., Tardif, T., Shu, H., and Zhang, Y. (**2014**). "Longitudinal correlates of reading comprehension difficulties in Chinese children," Read. Writ. **27**(3), 481–501.