JSLHR

Research Article

Development of Phonetic Contrasts in Cantonese Tone Acquisition

Peggy Pik Ki Mok,[a] Vivian Guo Li,[a] and Holly Sze Ho Fung[a]

Purpose: Previous studies showed both early and late acquisition of Cantonese tones based on transcription data using different criteria, but very little acoustic data were reported. Our study examined Cantonese tone acquisition using both transcription and acoustic data, illustrating the early and protracted aspects of Cantonese tone acquisition.

Method: One hundred fifty-nine Cantonese-speaking children aged between 2;1 and 6;0 (years;months) and 10 reference speakers participated in a tone production experiment based on picture naming. Natural production materials with 30 monosyllabic words were transcribed by two native judges. Acoustic measurements included overall tonal dispersion and specific contrasts between similar tone pairs: ratios of average fundamental frequency height for the level tones (T1, T3, T6), magnitude of rise and inflection point for the rising tones (T2, T5), magnitude of fall, H1*–H2*, and harmonic-to-noise ratio for the low tones (T4, T6). Auditory assessment of creakiness for T4 was also included.

Results: Children in the eldest group (aged 5;7–6;0) were still not completely adultlike in production accuracy, although two thirds of them had production accuracy over 90%. Children in all age groups had production accuracy significantly higher than chance level, and they could produce the major acoustic contrasts between specific tone pairs similarly as reference speakers. Fine phonetic detail of the inflection point and creakiness was more challenging for children.

Conclusion: Our findings illustrated the multifaceted aspects (both early and late) of Cantonese tone acquisition and called for a wider perspective on how to define successful phonological acquisition.

Supplemental Material: https://doi.org/10.23641/asha.11594853

Cantonese is a tone language spoken widely in southern parts of China, including Guangzhou, Hong Kong, and Macau, and many overseas Chinese communities. The Cantonese tone system is very complex with six lexical tones contrasting in both pitch height and pitch contour (Gandour, 1981). Recently, there are a few new studies demonstrating that Cantonese tone acquisition by children is a protracted process (Mok et al., 2019; P. S. Wong et al., 2017; P. S. Wong & Leung, 2018), contrary to findings in earlier studies that tones are all acquired by age 2;0 (years;months; So & Dodd, 1995; Tse, 1978). Most of these studies were based on auditory analysis of children's production. Only some general acoustic data of twenty 3-year-old children were reported so far. Our study fills the research gap by reporting acoustic data from 159 children aged 2;1–6;0 (divided into eight 6-month age bands for finer analysis), focusing on how important phonetic contrasts among several similar tone pairs developed with age. Our data provide a comprehensive picture of the process of Cantonese tone acquisition and call for reconsideration of how best to define successful acquisition.

Early Versus Late Acquisition of Cantonese Tones

Figure 1 shows the six lexical tones in Cantonese produced by reference speakers: T1 [55], a high-level tone; T2 [25], a high-rising tone; T3 [33], a midlevel tone; T4 [21], a low-falling tone; T5 [23], a low-rising tone; and T6 [22], a low-level tone. They contrast in both pitch height and pitch contour (Gandour, 1981). The six tones appear in open syllables or syllables with nasal codas [–m, –n, –ŋ]. There are three allotones that are traditionally called the *entering tones* in Chinese phonology. They only appear in syllables with unreleased stop codas [–p, –t, –k]: T7 [5], high-stopped; T8 [3], midstopped; and T9 [2], low-stopped. They are much shorter in duration. Although they have a low-falling contour in actual realization (Rose, 2004; P. S. Wong & Chan, 2018), they are not contrastive with any

[a]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin

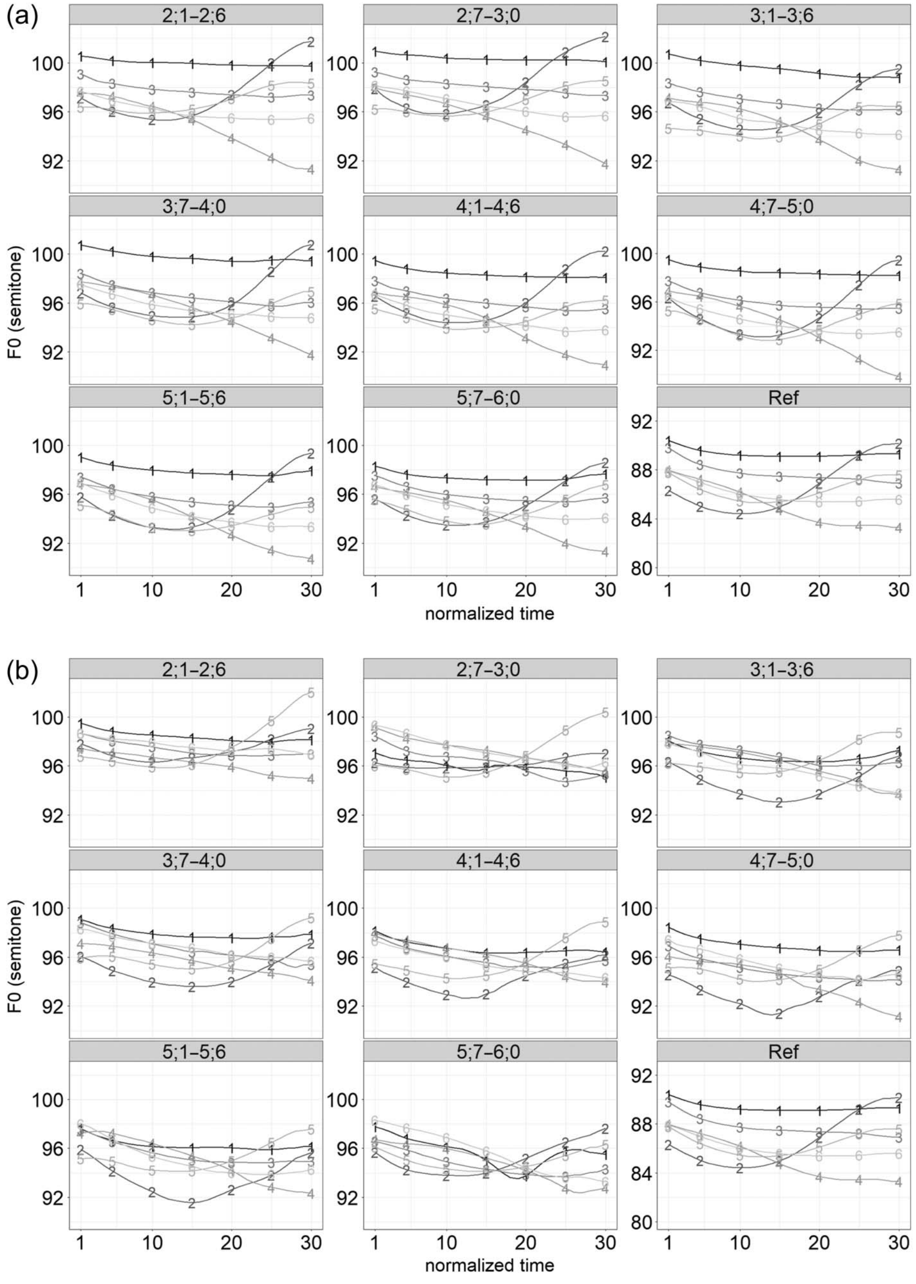
Correspondence to Peggy Pik Ki Mok: peggymok@cuhk.edu.hk

Editor-in-Chief: Bharath Chandrasekaran
Editor: Chao-Yang Lee

Received January 31, 2019
Revision received June 4, 2019
Accepted October 13, 2019
https://doi.org/10.1044/2019_JSLHR-19-00152

Disclosure: The authors have declared that no competing interests existed at the time of publication.

Figure 1. Fundamental frequency (F0) contours of the six Cantonese tones produced by children and reference speakers for (a) correct tokens and (b) incorrect tokens. Data from reference speakers are the same in both panels.



short level tones, and it is unclear if listeners could perceive the contour given the very short duration. Most phonological analyses of Cantonese considered them allotones of the three corresponding unstopped level tones T1, T3, and T6, respectively (Bauer & Benedict, 1997; Chao, 1947).

Despite the complexity of the Cantonese tone system, some earlier studies have collectively shown that Cantonese monolingual children could produce all the six tones accurately very early, by age 2;0, using longitudinal conversational speech data of just a few children (So & Dodd, 1995; Tse, 1978) or cross-sectional picture-naming data of 268 children (So & Dodd, 1995). Also using cross-sectional picture-naming data of many more children (1,726), To et al. (2013) confirmed that Cantonese tones were acquired by age 2;6, because their youngest age group of 2;6 already had ceiling production accuracy.

In addition to Cantonese, early acquisition of lexical tone has been reported for other tone languages as well. Studies using simple transcription data of natural production showed an early acquisition of Mandarin tones by the age of 2 years (Li & Thompson, 1977; Zhu, 2002; Zhu & Dodd, 2000). Thai tones were also reported to be acquired by the age of 2 years already (Clumeck, 1980). These findings suggest that tone production is generally acquired early by around the age of 2 years, regardless of the complexity of the tone systems (four tones in Mandarin vs. five tones in Thai vs. six tones in Cantonese). Early production accuracy is mirrored by some perception studies showing that even infants could distinguish simple tone contrasts (e.g., Mattock & Burnham, 2006; Mattock et al., 2008; Singh & Fu, 2016). Nonetheless, the ability of infants learning tone languages to maintain sensitivity to acoustic differences between simple stimuli (e.g., contrasting just two tones in a repetitive experiment) in their first year of life, as demonstrated by these studies, is not the same as the ability to distinguish all possible tonal contrasts in authentic situations related to meaning later in life. Mok et al. (2019) demonstrated that the averaged perception accuracy of all possible Cantonese tone pairs for children aged 2;1–2;6 was only about 60%, although it was still significantly higher than chance level (50%).

The above studies (So & Dodd, 1995; To et al., 2013; Tse, 1978) showing early acquisition of all the six tones in Cantonese were based on transcription data of children's natural production by one native transcriber (with only a small portion of the data cross-checked by another transcriber). Wong and colleagues (P. S. Wong & Leung, 2018; P. S. Wong et al., 2017) used a novel method to do transcription. They low-pass filtered children's production at 500 Hz to remove segmental information so the judges could rely on the fundamental frequency (F0) information only to judge the accuracy of children's tone production based on cross-sectional picture-naming tasks. They also had multiple judges (five). Production accuracy was defined as the percentage of judges who correctly identified the target tones. With such stringent judgment criteria, it is understandable why they found that children had not fully acquired all Cantonese tones even by age 6;0; that is, not all tones produced by children had adultlike accuracy. Similarly, late acquisition of Mandarin tone was also reported by Wong using filtered materials for transcription (P. S. Wong, 2012, 2013). Their results are a stark contrast to earlier studies.

The findings of early acquisition versus late acquisition of Cantonese tones discussed above are not directly comparable due to important methodological differences among the studies, even though they were all based on auditory analysis, for example, with or without low-pass filter and the number of judges. Earlier studies used lenient criteria (no filter, one judge), while Wong used very stringent criteria (with filter, multiple judges). Mok et al. (2019) revisited the whole issue by having four judgment conditions incorporating the two extremes: unfiltered one judge (like earlier studies), unfiltered two judges, filtered one judge, and filtered two judges (like studies by Wong). They had 111 Cantonese-speaking children aged between 2;0 and 6;0 divided into eight 6-month age bands and 10 reference speakers. As expected, production accuracies varied with judgment criteria: having very high accuracy (> 90%) by age 3;0 in the most lenient unfiltered-one-judge condition, while even the reference speakers were only 74% accurate in the most stringent filtered-two-judges condition, let alone the children. The other two judgment conditions gave comparable patterns showing increasing accuracy with age for children and very high accuracy for reference speakers (> 90%). As listening to filtered materials is more about auditory perception than normal speech perception and that perceptual differences were found for these two types of materials with the same contours (Mok & Zuo, 2012), they argued that the unfiltered-two-judges condition should be a more realistic criterion. Based on this set of data, children's production accuracy by age 6;0 (~80%) was still not on a par with that of the reference speakers (94%). Thus, Mok et al. (2019) also confirm that Cantonese tone acquisition is a protracted process.

Among the above studies on Cantonese tone acquisition, only P. S. Wong et al. (2017) reported some acoustic data produced by 20 children with a mean age of 3;7 (range: 3;1–3;11). Their Figure 1 shows the six tone contours in adults' correct production, children's correct production, and children's incorrect production, with similar patterns demonstrated by the first two types of production. In addition, they calculated six acoustic parameters for each token: mean pitch height, initial pitch height, final pitch height, minimum pitch height, maximum pitch height, and the slope of the second half of the tone contour. Statistical analyses were conducted to compare these parameters among the three types of production for each tone to determine whether children could produce them in an adultlike manner (i.e., whether children's values are significantly different from those of adults). Their Table 5 summarizes the comparisons, but it is very hard to decipher. Moreover, the same six acoustic parameters were used for all tones regardless of whether they were level tones or contour tones, and the parameters were not contrasted among similar tone pairs. Their discussion of the data suggests that their main concern was whether the acoustic characteristics of the correct and incorrect tone productions justified the judges' perceptual judgments of the tones. Thus, although they provided detailed acoustic data, their data only give a general picture of the acoustic patterns of child Cantonese tone production at one age, but not how specific tonal contrasts develop acoustically during the process of tone acquisition.

This Study

Many studies on segments have demonstrated that the time course of development in child's speech production is much more protracted when production accuracy is assessed by acoustic analysis than by transcription alone (Edwards & Beckman, 2008; Edwards et al., 2015; Munson et al., 2011). Transcription studies using more rigorous methods discussed above have already revealed a similar protracted process in Cantonese tone acquisition. Moreover, Wong and colleagues (P. S. Wong, 2012; P. S. Wong et al., 2017) demonstrated that, even in 3-year-old children (both Mandarin- and Cantonese-speaking) whose tone production was judged to be accurate, the acoustic patterns of their tones were still different from those of adults. Nevertheless, it is probably unrealistic to expect children's production to be the same as those of adults, but if both were considered to be correctly produced by the native judges, it will be interesting to compare the differences in fine phonetic detail between the two groups of speakers, especially for important phonetic contrasts among similar tone pairs, and to examine how these differences develop with age. No such data are reported in the literature so far. Our study aims to fill this important gap.

The six lexical tones in Cantonese can be divided into three main groups based on their acoustic contrasts (see Figure 1): the three level tones (T1, T3, and T6), the two rising tones (T2 and T5), and the low tone pair (T4 and T6). The three level tones T1 [55], T3 [33], and T6 [22] are differentiated mainly by pitch height, and the acoustic distance between T1 and T3 is larger than that between T3 and T6 (as little as about 20 Hz in adult female speech). The three level tones often have a slight falling contour in actual realization. The two rising tones are mainly differentiated in the second half of the tone: T2 [25] having a steeper rising slope than T5 [23]. In addition, there is a dip in F0 contour in the first half of both rising tones, with the minimum of the dip in T2 [25] often appearing slightly earlier than that in T5 [23]. This is likely a covert contrast as defined by Edwards and Beckman (2008): statistically reliable acoustic differences that are not perceptible to naïve listeners. They suggested that covert contrast is of interest because it provides a finer-grained window into children's phonetic development. Although studies on covert contrasts often focus on children's production of covert contrasts (e.g., Edwards & Beckman, 2008; Scobbie et al., 2000), covert contrasts do appear in adult speech, for example, incomplete neutralization and near mergers. So far, no study on Cantonese tone has examined the F0 dip of the two rising tones in detail. We will demonstrate the covert perceptual aspect of this dip. For the low tone pair T4 [21] and T6 [22], the main difference is also in the second half of the tone: T4 having a more consistent and obvious falling contour than T6. Additionally, T4 falls so low that it is often realized as creaky voice, a phenomenon with a physiological basis that is also found in the third tone in Mandarin [214] because of the low pitch target (Kuang, 2017). The presence of creak increases adult listeners' perception of T4 over T6 significantly (Yu & Lam, 2014), which demonstrates that it is a perceptually salient feature.

It is important to examine the acoustic differences between the above similar tone pairs because the complex Cantonese tone system is undergoing changes in recent years in that some similar tone pairs began to merge. Mok et al. (2013) reported that some young Cantonese speakers in Hong Kong may not clearly distinguish the two rising tones (T2 vs. T5), the two level tones (T3 vs. T6), and the low tone pair (T4 vs. T6) in their production and perception. The merging is in an incipient stage, as these speakers still had six tone categories. Acoustic similarity between these tone pairs is one of the reasons for the merge. The high-level tone T1 is not involved in tone merging presumably because of its better separation from the other tones acoustically (see Figure 1). In addition, both children and adults find these merging pairs much more difficult to distinguish than other tone pairs perceptually (Lee et al., 2015; Mok et al., 2019), so the merging phenomenon is likely related to tone acquisition as well. Therefore, the focus of this study is on the development of the acoustic differences between these specific contrasts.

This study is an expansion of Mok et al. (2019) with more children (111 vs. 159). Production data between aged 2;1–6;0 were divided into eight 6-month age bands for finer analysis of the development of acoustic differences. Comparisons with reference speakers are included. Given the ongoing tone merging mentioned above, there will be variability in tone production accuracy even among adult speakers. In order to simplify the comparisons, we only used reference speakers who did not merge any tones so their production accuracy should be very high (> 90%), comparable to the reference speakers in previous studies (So & Dodd, 1995; To et al., 2013; P. S. Wong et al., 2017).

Nonetheless, unlike some previous studies, we do not expect values in children's correct production to be the same as those of the reference speakers. Rather, we were interested to see if similar contrasts were maintained in children's correct production regardless of their actual values. If so, it can be concluded that, although overall children were not as accurate as adults in their tone production (as expected, because they were still acquiring the tones), they were aware of the important phonetic features of the Cantonese tone system when they do produce the tones correctly. This can give us a wider perspective on how to define successful phonological acquisition and a window into the development of their higher level phonological knowledge, which is a protracted process (Munson et al., 2011). Studies have shown that the acquisition of segmental phonology can be influenced by literacy development (Burnham, 2003). Children in Hong Kong start to learn to read and write rather early from 4 years of age onwards. Nonetheless, Cantonese tones are not presented in the Chinese writing system, which is logographic in nature. Thus, we can be confident that the findings reported below were not complicated by literacy factors.

Method

Participants

One hundred fifty-nine children aged 2;1–6;0 from five local kindergartens-cum-nurseries and 10 adolescents aged 15;9–16;7 from local secondary schools were recruited. The adolescent speakers were screened for their accuracy of tone production: Their production of syllables /fɐn/, /jɐn/, /ji/, and /si/ in six tones (all attested real words in Cantonese) was independently checked by two phonetically trained native speakers of Cantonese who do not merge any tones.

Ten adolescent speakers who were confirmed to clearly distinguish the six Cantonese tones were chosen as reference speakers (screened from 22 adolescents recorded). The reference speakers were all born in Hong Kong and spoke Cantonese as their native language. Previous studies demonstrated that adultlike perceptual patterns were found at the age of 10 years (Ciocca & Lui, 2003; Lee et al., 2015) and that our adolescent speakers were screened for their tone production accuracy and they could safely serve as reference speakers in our project. Adolescent speakers were used as reference in a previous study on Cantonese tone as well (Khouw & Ciocca, 2007). In total, recordings from 169 speakers were analyzed. Table 1 presents a breakdown of the participants by age. All of the participants were native speakers of Hong Kong Cantonese. None reported any speech, hearing, or learning impairment. Ethics approval from the Survey and Behavioural Research Ethics Committee of the Chinese University of Hong Kong was obtained for the project.

Materials

Productions of Cantonese monosyllables were elicited through a picture-naming task. The tones produced by both children and reference speakers were tones in isolation, comparable to other studies on Cantonese tone acquisition (P. S. Wong et al., 2017; P. S. Wong & Leung, 2018). The stimuli were the same as those reported in Mok et al. (2019). There were 30 colored pictures, and each picture was designed to elicit one monosyllabic word (6 tones × 5 words). Nineteen of the words were adopted from the Hong Kong Cantonese Articulation Test (Cheung et al., 2006), and the other 11 words were supplemented to create a balanced stimulus set. All words were familiar concepts to children.

Procedure

With parental consent, experimenters, who were phonetically trained native speakers of Cantonese and were able to distinguish the six lexical tones clearly in both production and perception, recorded the participants one at a time in quiet rooms in their kindergartens-cum-nurseries or schools. We followed the question instructions in the Hong Kong Cantonese Articulation Test to elicit production (questions such as *ni1 go3 hai6 me1 aa3?* Or "What is this?"). Similar questions were used for the Supplemental Materials. The children produced the target words in isolation. They were prompted to say the target word again for the second recording. Both recordings were used for analysis. In case of failure to produce a target word despite hints given, they would be asked to repeat after the experimenters. This happened occasionally for the two youngest age groups (2;1–3;0).

Table 1. Number of speakers and exclusion rate in each age group.

Age group (years;months)	No. of speakers	No. of excluded tokens	% of exclusion
2;1–2;6	19	454	40.0
2;7–3;0	20	247	20.8
3;1–3;6	20	347	28.8
3;7–4;0	20	430	35.9
4;1–4;6	20	293	24.5
4;7–5;0	20	255	22.0
5;1–5;6	20	201	16.8
5;7–6;0	20	98	8.7
Reference	10	49	8.2
Total	169	2374	23.7

Data Processing

Each recording was auditorily analyzed by two phonetically trained native Cantonese speakers who transcribed the tones they heard (as any of the six tones or "uncategorized" for tokens that did not fit any of the six). Since there are three level tones in Cantonese, in order to minimize misperception due to failure in speaker normalization (P. C. M. Wong & Diehl, 2003), recordings for transcription were blocked by speaker and a short recording of the exchange between the child and the experimenters was presented before each speaker block to familiarize the transcribers with the speaker's lexical pitch range.

Transcription was performed for both unfiltered and filtered materials: A low-pass filter at 500 Hz for children's recordings and 400 Hz for reference speakers' recordings was applied to remove segmental and lexical information, following Wong and colleagues (P. S. Wong et al., 2017; P. S. Wong & Leung, 2018). Two judges transcribed all recordings with high interrater reliability (83.2% agreement, Fleiss' κ = .798). Each recording was then labeled as correct or incorrect under four judgment conditions: unfiltered one judge, unfiltered two judges, filtered one judge, and filtered two judges. As mentioned in the introduction, Mok et al. (2019) demonstrated that the unfiltered-two-judges condition is the most realistic and appropriate one; only accuracy data in this condition will be reported here. Acoustic measurements of F0 were done through the automatic F0 tracking performed using ProsodyPro (Xu, 2013) and Praat (Boersma, 2001) on 30 equidistant points along the tone contour. Creaky tokens were excluded from acoustic analysis (except in the comparison of Tone 4 vs. Tone 6; see details below). Analyses were based on correct tokens, except for F0 range dispersion, which also included the incorrect tokens because the exclusion of incorrect tokens may skew the range of dispersion undesirably, especially for younger children with lower accuracy rates. The right panels of Table 1 detail the number and percentage of excluded tokens in each age group (6 tones × 5 words × 2 repetitions = 60 tokens per speaker in total).

Results

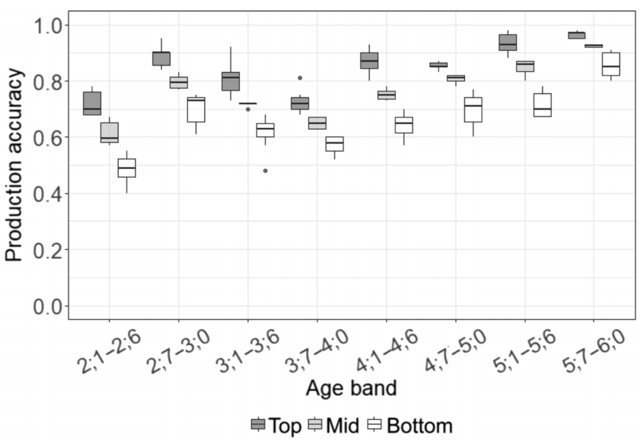
Figure 1 shows the F0 contours of the six Cantonese tones produced by children and reference speakers. The upper panels show the F0 contours based on correct tokens, while the lower panels show those of incorrect tokens (data from reference speakers are the same in both panels). It is obvious that the correct contours of the youngest children already resemble those of reference speakers very much. Incorrect tokens generally had a narrower F0 range. This shows that including incorrect tokens in the calculation of F0 range dispersion would not exaggerate the overall pattern. Another noticeable pattern of the wrong tokens is that incorrect T2 and T5 had reverse patterns with T5 higher than T2, possibly because children had confused these rising tokens as T2 and T5 are merging.

Based on the transcription results, children were ranked by their mean accuracy across tokens. Data from children in each age band were divided into three parts, representing the top, mid, and bottom performers. This gives a comprehensive yet layered picture of children's tonal development (see Figure 2). There is a clear trend of rising accuracy as children grow older in all three performance groups, although the trend is not simply linear. The average accuracy of the reference speakers is 94%. Only the top and mid performers in the eldest group (aged 5;7–6;0) and the top performers in the second eldest group (aged 5;1–5;6) can reach an accuracy level of over 90%. One-sample *t* tests confirmed that the accuracy of top-, mid-, and bottom-ranking children of each age group was all significantly higher than chance level (1/6; see Supplemental Material S1). Thus, it can be concluded that children in even the youngest age group were using the tones meaningfully in their Cantonese production, although their overall accuracy rates fall short of that of reference speakers as they were still acquiring the tones. Only the eldest group (aged 5;7–6;0) can possibly be considered close to adultlike in production accuracy, with two thirds of the children having an adultlike accuracy of over 90%.

Overall Tonal Dispersion

Before comparing specific contrasts between similar tone pairs, the development of overall F0 range dispersion was examined to see if younger children would have a more expanded "tone space" than older children and reference speakers. It is possible that younger children would have more exaggerated tone production to maximize contrasts as finer contrasts are harder to produce. This hypothesis is supported by the findings in Thai tone development by Burnham et al. (2006) and Xu Rattanasone et al. (2013). By comparing acoustic data from infants between 18 and 33 months old, they showed that the Thai tone space peaked at 27 months old. They argued that such a peak might be due to the vocabulary spurt around this age and the consequent crowding of the lexical space, resulting in more exaggerated productions of tones, especially in newly learned words.

Figure 2. Children production accuracy divided into three performance groups across age groups.



F0 range expansion in our study was evaluated in two ways. The first one is tone space dispersion, which represents the "mean Euclidean distance of individual tones from the center of the speakers' F0 space (acoustic tone space)," according to Zhao and Jurafsky (2009). Following their procedure, first, a central F0 at each measurement point was obtained by averaging over all the tokens by the same speaker at that measurement point. Thus, for a measurement point k of a speaker, its central F0, $\mathrm{CF0}_k$, is the F0 of token i at k ($\mathrm{F0}_k^i$) averaged over the number of tokens j.

$$\mathrm{CF0}_k = \frac{1}{j}\sum_{i=1}^{j} \mathrm{F0}_{k}^{i}. \tag{1}$$

Then, for any measurement point, a tonal distance (TDCF0) to central F0 in semitone can be obtained for each token:

$$\mathrm{TDCF0} = 12\left|\log_2 \frac{\mathrm{F0}_k^i}{\mathrm{CF0}_k}\right|. \tag{2}$$

The overall tonal dispersion (TD) is the mean of tonal distance across all tokens and all measurement points:

$$TD = \frac{1}{30j}\sum_{k=1}^{30}\sum_{i=1}^{j} \mathrm{TDCF0}_{k}^{i}. \tag{3}$$

Figure 3a shows the distribution of TD values of children of different age groups and the reference speakers. A slight general decrease in mean TD and variance with age can be observed from age 2;7 onwards. A one-way analysis of variance (ANOVA) indicated that there was a significant difference in the TD range (in semitone) between age groups, $F(8, 160) = 2.772$, $p = .007$. Pairwise comparisons using the Tukey's honestly significant difference (HSD) test revealed that the TD for children aged 5;7–6;0 was smaller than that of children aged 2;7–3;0 ($p = .003$), and it was marginally so between the 2;7–3;0 group and reference speakers ($p = .062$). See Supplemental Material S2 for details of the pairwise comparisons.

In addition to the overall F0 range dispersion, a phonological pitch range measure was adopted by calculating the proportion of the last parts (Point 27) in Tone 1 [55] over Tone 4 [21], representing the phonologically highest

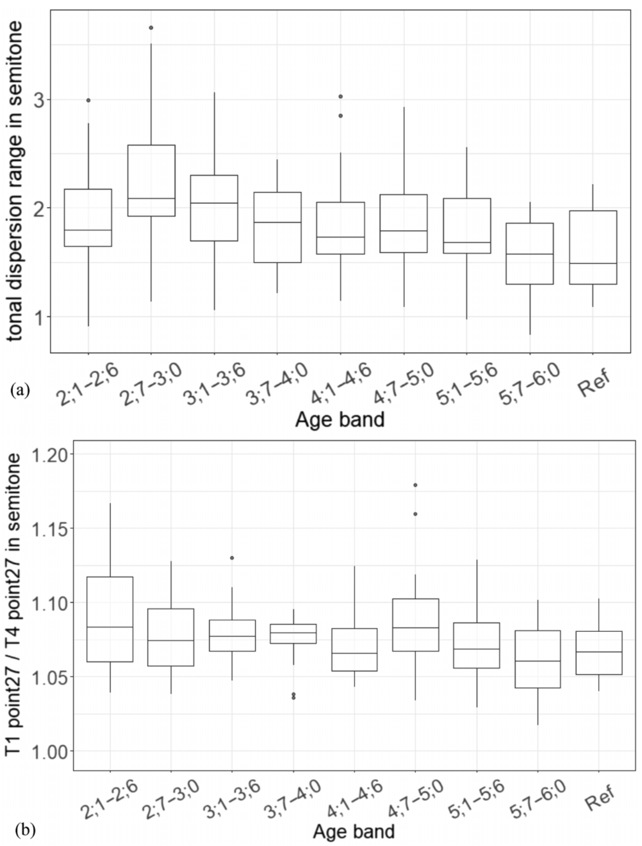


Figure 3. Overall (a) tonal dispersion and (b) phonological pitch range across age groups.

and lowest pitch values in the tone inventory, respectively, following Mok et al. (2013). Results of children's phonological pitch range are shown in Figure 3b.

There was a near-significant difference between age groups with respect to their phonological pitch range at the 27th point on the tonal contours as indicated by one-way ANOVA, $F(8, 160) = 1.989$, $p = .051$. Pairwise comparisons using the Tukey's HSD test revealed that only the phonological range for children aged 5;7–6;0 was smaller than that of children aged 2;1–2;6 ($p = .033$), but no significant difference was found between children and reference speakers (see Supplemental Material S2). Thus, incorporating the results from TD and phonological pitch range, it seems that, as a whole, only the two youngest age groups could be considered having more exaggerated tone production, similar to the Thai findings mentioned above.

In what follows, important acoustic features of the similar tone pairs will be examined.

Level Tone Contrasts (T1, T3, and T6)

There is a larger F0 difference between T1 [55] and T3 [33] than that between T3 [33] and T6 [22] (see reference speakers' data in Figure 1). As the three level tones often have a slightly falling contour in actual realization, relative mean F0 height is used for comparison: the ratio of Tone 1's mean F0 over Tone 3's mean F0 and the ratio of Tone 3's mean F0 over Tone 6's mean F0. It can be seen in Figure 4 that all age groups produced a larger ratio of T1/T3 than that of T3/T6, while the scale of difference varied across age groups. A two-way ANOVA showed a significant main effect of age group, $F(8, 318) = 2.900$, $p = .004$, and that the T1/T3 ratio was significantly larger than the T3/T6 ratio, $F(1, 318) = 68.821$, $p < .001$. However, the interaction between age groups and the ratios was not significant, indicating that all children and reference speakers had a similar larger distinction between the T1/T3 ratio than that between the T3/T6 ratio. Tukey's HSD post hoc tests showed that the age group main effect was due to the significantly lower ratios in children aged 5;7–6;0 than those in children aged 3;1–3;6 ($p = .018$), 3;7–4;0 ($p = .001$), and 4;7–5;0 ($p = .005$), but the ratio patterns were the same. It is unclear what contributed to the lower ratios of the 5;7–6;0 group, but Figure 3 shows that the TD of the 5;7–6;0 group was among the lowest. Figure 1a also shows that their pitch range was narrower than those of other groups, especially at the tone onset. Nonetheless, the three level tones were still clearly distinctive in this age group regardless of the actual ratios.

The Rising Tone Pair (T2 and T5)

The two rising tones differ mainly in the magnitude of rise, with T2 [25] having a much steeper rising slope in the second half of the tone than T5 [23] (see reference speakers' data in Figure 1). The magnitude of rise was quantified by calculating the F0 difference between the 27th measurement point (offset of the tone) and the 10th measurement point (average position of the minimum F0 between the two tones; see below). A larger difference represents a steeper rising slope, as the two rising tones do not differ in duration (P. S. Wong & Chan, 2018).

Figure 4. Ratios between mean fundamental frequency (F0) height of T1/T3 and T3/T6 across age groups.

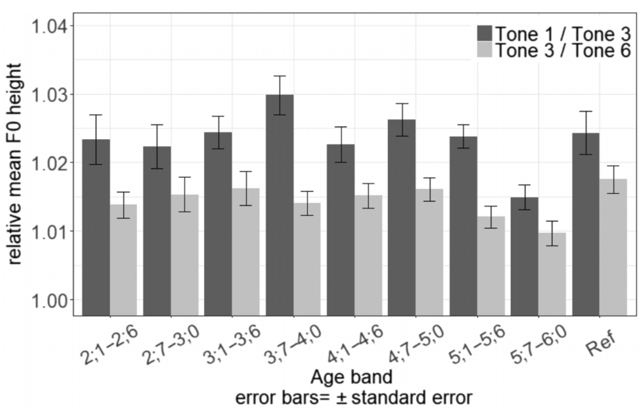
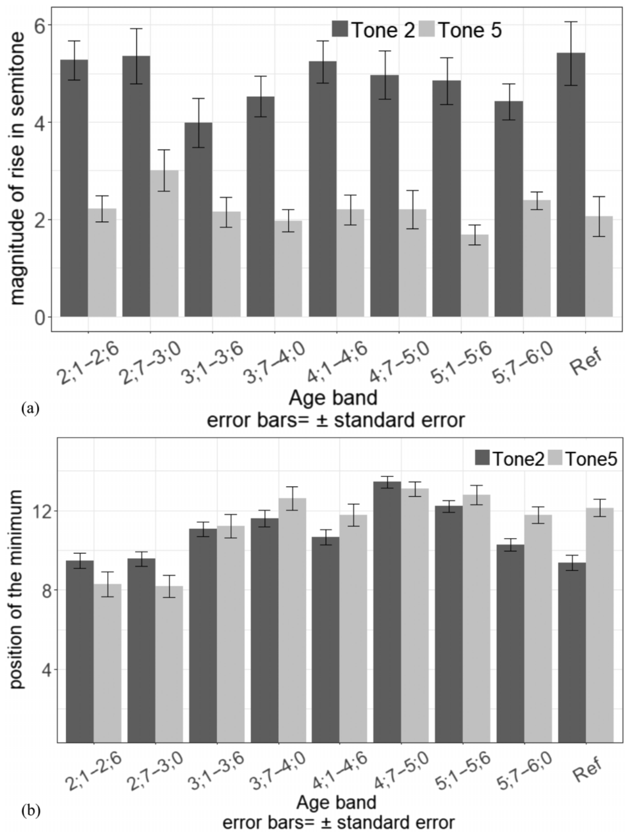


Figure 5a illustrates that all speakers had a much steeper slope for T2 than T5.

A two-way ANOVA showed that the magnitude of rise in Tone 2 was significantly larger than that of Tone 5, $F(1, 319) = 190.760$, $p < .001$. Two-tailed paired t tests confirm that the slope differences were significant in all age groups ($p < .001$ for all; see Supplemental Material S3). The effect of age group and the interaction between age groups and the tones were not significant. This suggests that all age groups had a similar distinction between the rise magnitude of Tone 2 and that of Tone 5, although the actual magnitude difference between the two rising tones varied across age groups.

In addition to the magnitude of rise, there is a dip in the F0 contour in the first half of the tone before the rising contour in both T2 and T5 (see Figure 1). Another difference between the two rising tones is the inflection point: The minimum F0 value appears slightly earlier in T2 than in T5, which we believe to be a covert contrast—reliable acoustic difference not perceivable by naïve speakers (Edwards & Beckman, 2008). Positions of the inflection point in the rising tones were identified by locating the minimum F0 measurement point in the first two thirds of the tone (corresponding to the first 21 measurement points). Figure 5b shows where the position falls for Tones 2 and 5 in children and reference speakers' speech. A two-way ANOVA showed a significant effect of age group, $F(8, 2467) = 22.503$, $p < .001$, and that the inflection position in Tone 2 was significantly earlier than that in Tone 5, $F(1, 2467) = 5.212$, $p = .023$. There was a significant interaction as well, $F(8, 2467) = 4.182$, $p < .001$. Post hoc pairwise comparisons with Tukey correction (see Supplemental Material S3) indicated that the difference between the inflection position in Tones 2 and 5 was significant for children aged 2;7–3;0 ($p = .021$) and 5;7–6;0 ($p = .005$) and reference speakers ($p < .001$).

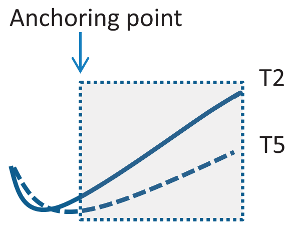
Figure 5. The (a) magnitude of rise and (b) inflection point of T2 versus T5 across age groups.



In addition to having statistically reliable acoustic differences, covert contrasts should not be perceptible to naïve listeners as well. As no previous study has examined the T2/T5 inflection point, a simple perception experiment was conducted to confirm its nonperceivability. Four minimal pairs contrasting T2/T5 in CANTIT (Lee, 2012) were used: /k^{h}ei/: 棋 "chess" vs. 企 "stand"; /lei/: 梨 "pear" vs. 你 "you"; /ŋɔ/: 鵝 "goose" vs. 我 "I/me"; and /mou/: 帽 "hat" vs. 冇 "have not." The T2 tokens of these four pairs all underwent tone change (pinjam); that is, the base tones were changed into T2 (Alan, 2007; M. Wong, 1982), a very common phenomenon in Cantonese conversational speech. One male and one female adult native Cantonese speakers produced these tokens naturally. F0 were tracked at 30 equidistant points on the rime. For each minimal pair, the time point where the minimum F0 value in the T5 token appeared was used as the anchoring point (see Figure 6). Both the T2 and T5 tokens of the minimal pair were truncated after this anchoring point (i.e., the shaded part in Figure 6 was deleted), resulting in two first-half tokens of equal duration. The F0 of both first-half tokens starts to fall from the tone onset, with one having an earlier F0 minimum plus any following rise (T2) and one having a later F0 minimum with no rise (T5), capturing the covert contrast of the different positions of the inflection point in the first part of the tone. Results from a two-tailed paired-samples t test indicated that T2 tokens ($M = 5.25$, $SD = 1.91$) had a significantly earlier inflection position than T5 tokens ($M = 8.75$, $SD = 2.60$), $t(7) = -4.78$, $p = .002$.

A simple identification task using Chinese characters was conducted with 20 adult native Cantonese listeners who did not merge the two rising tones (14 women; $M_{\text{age}} = 24.7$ years). When listeners heard a token, they had to click

Figure 6. Schematic illustration of stimulus manipulation for the perception experiment of the T2/T5 covert contrast.



on the corresponding Chinese character. Three repetitions of the materials were included in semirandomized order such that the T2 and T5 tokens of the same syllable would not appear consecutively. The presentation order of Chinese character was counterbalanced. Thus, in total, there were 96 trials for each listener (4 minimal pairs × 2 tones × 2 speakers × 3 repetitions × 2 presentation orders). Mean identification accuracy was 0.41 (SD = 0.11), significantly lower than chance level (0.5), $t(19) = -3.46$, $p = .003$. The results confirm that the difference in inflection point is indeed a covert contrast between T2 and T5.

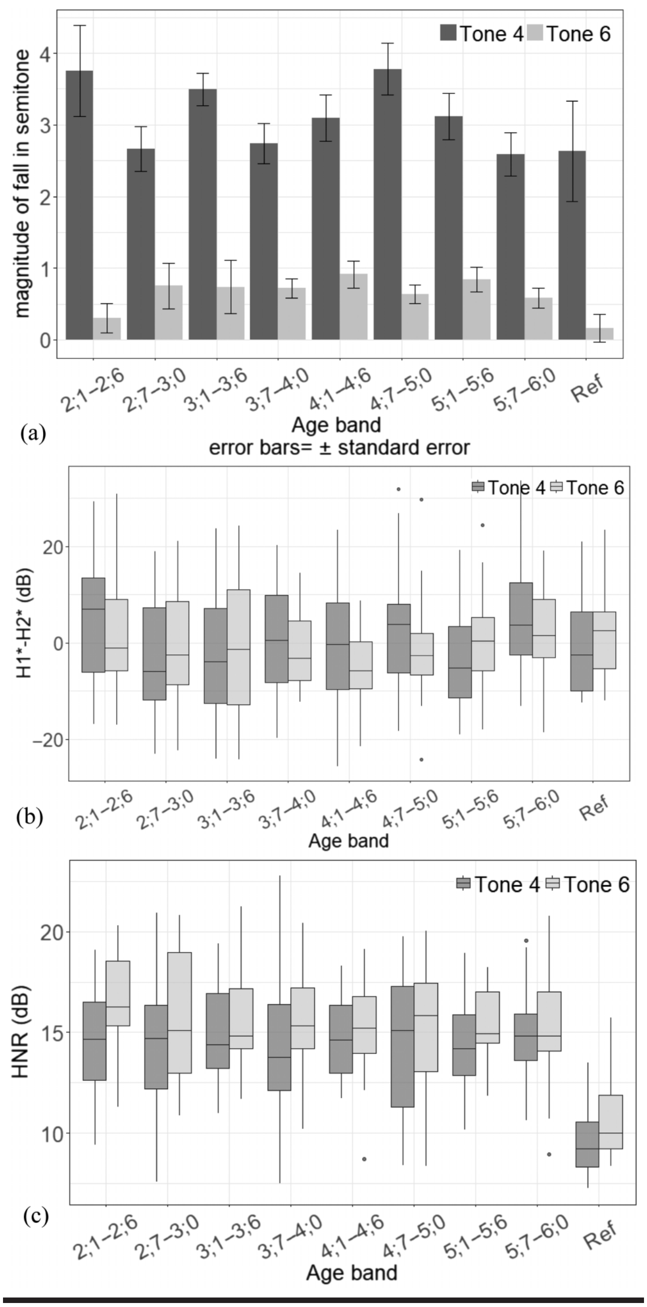
Although all the children's tokens included in the analyses were considered correct by the transcribers, their inflection point results deviate from those of reference speakers. Two-tailed paired *t* tests (see Supplemental Material S3) show that the differences were significant in children aged 2;7–3;0 ($p = .021$) and 5;7–6;0 ($p = .005$). Notice, however, the direction of the difference in younger children (aged 2;7–3;0) was the reverse of the reference pattern. Older children aged 5;1–6;0 had the same pattern of reference speakers, but the difference was not significant. Only the eldest group (aged 5;7–6;0) had the same significant pattern as reference speakers. This suggests that, unlike other more obvious phonetic contrasts discussed above, children take time to acquire this covert contrast between the two rising tones.

The Low Tone Pair (T4 and T6)

The two low tones T4 [21] and T6 [22] differ mainly in the second half of the tone (see Figure 1), with T4 falling to the lowest F0 level in the speaker's pitch range, often resulting in creakiness. Creaky phonation often has irregular vocal fold vibration, rendering automatic F0 tracking unreliable. Manual pulse-fixing was performed using ProsodyPro (Xu, 2013) for some T4 and T6 tokens. It is because the vocal pulse markings generated by the autocorrelation algorithm in Praat were not error free, especially for irregular voicing typical of creak. Vocal pulses that were missed by the automatic F0 tracking were manually rectified, but still, some tokens were too irregular to fix. F0 data for T4 were only obtained from tokens with regular or fixed pulses. The difference between T4 and T6 was quantified by calculating the F0 difference between the 27th measurement point and the 15th measurement point, that is, the magnitude of fall from midpoint toward the end. T4 (a low falling tone) should have a larger difference than T6 (a low-level tone, which often has a slightly falling contour in actual realization; see Figure 7a). A two-way ANOVA showed that the magnitude of fall was larger in Tone 4 than in Tone 6, $F(1, 317) = 278.987$, $p < .001$. The effect of age group or its interaction with tone was not significant. All speakers had a significantly larger F0 difference between the two measurement points for T4 than T6 ($p \leq .007$; see Supplemental Material S4).

As creakiness is a salient concomitant feature of T4, both acoustic and auditory analyses were conducted to examine this feature. Acoustically, H1*–H2* (corrected for formants, following Iseli et al., 2007) as a common measurement of creakiness in phonation was taken at the final one-tenth portion of Tone 4 and Tone 6 tokens by using ProsodyPro (Xu, 2013). A lower H1*–H2* value indicates more glottal constriction. Creaky voice generally has low values of H1*–H2*, because the glottis is usually constricted (Keating et al., 2015). Again, only tokens with regular or fixed pulses were used. Outliers (beyond 1.5 times the length of the box in Figure 7b) were trimmed

Figure 7. The (a) magnitude of fall, (b) H1*–H2*, and (c) harmonic-to-noise ratios of T4 versus T6 across age groups.



from the data. It can be observed from Figure 7b that there did not seem to be any consistent difference between the two tones across all groups, although there were more negative values for T4 for reference speakers. A two-way ANOVA showed that none of age group, tone, or their interaction was significant. Paired *t* tests also show no difference for all age groups (see Supplemental Material S4).

In addition to H1*–H2*, harmonic-to-noise ratios (HNRs) of the two low tones were also measured (see Figure 7c) using ProsodyPro (Xu, 2013), because spectral tilt measure (e.g., H1*–H2*) needs to be interpreted with respect to noise measure such as HNR (Garellek, 2019). Irregular F0 can be measured as spectral noise. When F0 is irregular, the signal's noise will increase. Low HNR values indicate less strong periodic excitation relative to glottal noise (Keating et al., 2015; Garellek, 2019). A two-way ANOVA showed a significant effect of age group, $F(8, 314) = 6.662$, $p < .001$, and that the HNR of Tone 4 was significantly lower than that of Tone 6, $F(1, 314) = 10.228$, $p = .002$. There was no significant interaction. Tukey's HSD post hoc tests (see Supplemental Material S5) indicated that all children groups showed higher HNR values than the reference speakers for both tones ($p < .001$).

Prototypical creaky voice has a lower spectral tilt and a lower HNR (Garellek, 2019), as demonstrated by T4 versus T6 by the reference speakers (see Figures 7b and 7c). The acoustic measures above suggest that T4 produced by all speakers was generally creakier than T6.

One probable reason why there was no difference in H1*–H2* is that very creaky tokens were already excluded from acoustic analysis because of irregular pulses. Moreover, there are different kinds of creaky voice. Some can sound creaky but have regular pulses, for example, vocal fry (Keating et al., 2015). Thus, auditory analysis was conducted to complement acoustic measurements. Auditory perceptual judgment was collected from one phonetically trained researcher who did not know Cantonese to avoid any lexical bias. She listened to all the Tone 4 tokens and judged whether or not each one sounded creaky. Among all the correct tokens, there appeared a general trend for children to produce a higher percentage of creaky tokens as they grow older, although reversion was also found (see Table 2). Nearly half of the correct tokens from the reference speakers were auditorily creaky, while the percentage of creaky tokens was much smaller for children.

Pearson chi-square test indicates that age group and the number of perceived creaky tokens are dependent ($\chi^2 = 73.415$, $df = 8$, $p < .001$). Post hoc pairwise comparisons using Fisher's exact test of independence showed that all children groups produced significantly less creaky tokens than the reference speakers (mostly $p < .001$; see Supplemental Material S6). Among the children, children aged 4;1–4;6 had significantly more creaky tokens than other children ($p < .05$), except those aged 4;7–5;0. Children aged 4;7–5;0 also showed significantly more creaky tokens than younger children aged 2;1–2;6 ($p = .03$) and 3;1–3;6 ($p = .009$).

Table 2. Number and percentage of creaky tokens in the correct production of T4 by each age group.

Age group (years; months)	No. (%) of modal tokens	No. (%) of creaky tokens
2;1–2;6	111 (90)	13 (10)
2;7–3;0	132 (85)	23 (15)
3;1–3;6	151 (90)	17 (10)
3;7–4;0	131 (86)	22 (14)
4;1–4;6	122 (70)	52 (30)
4;7–5;0	138 (78)	39 (22)
5;1–5;6	154 (85)	28 (15)
5;7–6;0	143 (82)	31 (18)
Reference	47 (55)	39 (45)

Discussion

The current study examined the development of phonetic contrasts in Cantonese tone acquisition by comparing important acoustic features of several similar tone pairs produced by children aged 2;1–6;0 and reference speakers. The data show that only the eldest group (aged 5;7–6;0) could be considered close to adultlike in production accuracy based on transcription by two native judges, concurring the results in P. S. Wong and Leung (2018) and Mok et al. (2019) that Cantonese tone acquisition finishes late. Nonetheless, children, even the youngest age group (2;1–2;6), already had accuracy significantly higher than chance level and were generally producing major acoustic contrasts between tone pairs similarly as reference speakers did (although the actual values between tone pairs still varied across age groups). Children differed from reference speakers in three ways. First, the tone space of the youngest children (before aged 3;0) was more dispersed; that is, they produced more exaggerated tones. Second, only the eldest children (aged 5;7–6;0) could produce the covert contrast of the inflection points of T2 versus T5. Third, children produced much fewer T4 tokens with the concomitant feature of auditory creakiness than reference speakers did.

Our tone acquisition data concur well with Munson et al. (2011) and other acoustic studies on the acquisition of segments that phonological development takes place over an extensive period (if fine phonetic detail and covert contrast are considered), not simply in the first few years of life. They stated that acquisition involves two aspects: the acquisition of productions that are sufficiently adultlike to be perceived and transcribed as accurate and the development of adultlike speech motor control as reflected in acoustic and kinematic measures (Munson et al., 2011, p. 297). The time course of development is much more protracted when these measures were used than what is revealed by transcription data alone. While the protracted nature of the acquisition process is better understood with finer measures, a critical question remains: When should a child be considered having acquired a sound/phoneme? Is being completely adultlike in all aspects (including fine motor control) a precursor of successful acquisition? Can a more lenient and realistic definition be acceptable? They did not provide a clear answer to these crucial questions.

Our data give us a wider perspective on how to define successful acquisition by providing layered production accuracy data and acoustic data. Previous studies on Cantonese tone acquisition (and indeed most previous studies on phonological acquisition) either defined successful acquisition as having near-perfect accuracy (90% accurate in So & Dodd, 1995, or correct in all environments in To et al., 2013) or compared the production accuracy between children and adults who essentially had ceiling accuracy (Mok et al., 2019; P. S. Wong et al., 2017; P. S. Wong & Leung, 2018). However, the criterion of (near) perfect production accuracy is quite demanding for children. Given the ongoing tone merging phenomenon, even some adult speakers may not have ceiling accuracy (Mok et al., 2013). The average accuracy data (see Table 1) suggest that only children in the eldest group (aged 5;7–6;0) could be considered having acquired all the tones. Nevertheless, the layered data (see Figure 2) demonstrate that only the top and mid performers in that age group had accuracy over 90%. With one third of the children failing this criterion, it is unclear if successful acquisition can still be assumed for that age group, but saying that children have not acquired the tones by age 6;0 appears to be counterintuitive.

Another stage of sound acquisition was proposed in the literature—stabilization. A sound was considered stable when the child produced the sound correctly on at least two of three opportunities. When 90% of the children in an age group achieved an accuracy rating of at least 66.7% (i.e., 2/3) for a sound, the sound would be considered to have been stabilized by that age group (So & Dodd, 1995, p. 17). The criterion of 90% of children achieving an accuracy rate of 66.7% seems to be arbitrary (although not unreasonable). Using this arbitrary criterion, as a group, tones appeared to be stabilized by ages 2;7–3;0 using the average data (see Table 1), but the layered data illustrate that even stabilization can vary across age groups (see Figure 2). Thus, simply using production accuracy to define successful acquisition seems to be, although being a useful measure, inadequate because of large individual variations both within and across age groups.

It is probably unrealistic to expect young children to have the same accuracy level as adults, but not having perfect accuracy does not mean that children were not using tones meaningfully. So and Dodd (1995) also mentioned that phoneme development is a continuum ranging from the initial stage of being able to articulate a sound in isolation to the final stage of being able to articulate a sound both phonetically and phonologically accurately (p. 17). As such, the youngest age group (2;1–2;6) was already using the six lexical tones phonologically meaningfully in a sense that their production accuracy was significantly higher than chance even for the worst performers. This demonstrates that they had the six tones in their phonology. In addition, their correct tokens had major acoustic contrasts very similar to those of reference speakers already (see Figure 1), indicating that they knew what was important phonetically for these tones and they could produce these contrasts accordingly. Our study is the first to demonstrate such phonetic sensitivity of young children, which should not be overlooked. Considering both lines of reasoning, we could argue that Cantonese tones were basically acquired by ages 2;1–2;6. What is left for the following few years is mainly about maturation (getting adultlike accuracy) and fine tuning (e.g., getting the fine phonetic detail and covert contrast right). Cantonese tone acquisition does start rather early. Instead of insisting on having (near) perfect accuracy, as long as children could produce the tones meaningfully (e.g., better than chance level), there are good reasons to believe that they have acquired the tones phonologically already.

One aspect of child motor development has very similar developmental patterns to tone acquisition, which can give us insights into how best to understand and define successful acquisition: learning to walk. A baby typically starts to crawl around 7 months of age and starts to stand up and cruise (walking while holding onto sturdy objects such as furniture) at around 9 months of age. These ways of locomotion are not defined as walking yet. Around 12 months of age, babies start to take independent steps and become toddlers. Toddlers' walking is still unstable and wobbly and is different from adult walking in many ways. It takes years to mature and walk exactly like an adult in fine motor control. If adult-likeness in all aspects is the criterion to judge whether a child has "acquired" walking, then the age of acquisition could be around 7 years old the earliest (Hung et al., 2013; Wu et al., 2015). However, such conclusion is unrealistic. Most people would happily accept that a child is walking when he or she can take independent steps, albeit walking clumsily.

In a similar vein, it is not unreasonable to say that children had acquired the Cantonese tones by ages 2;1–2;6 because they could produce tones contrastively both phonological and phonetically ("the acquisition of productions that are sufficiently adultlike to be perceived and transcribed as accurate" stated by Munson et al., 2011), while it takes at least another few years for them to fine-tune the acoustic patterns and achieve adultlike production accuracy ("the development of adultlike speech motor control assessed by acoustic and kinematic measures" stated by Munson et al., 2011). In fact, similar developmental patterns apply to many motor skills learned during childhood.

Tone perception data of 78 out of the 159 children in the current study were also reported in Mok et al. (2019). Their perception data of 111 children show that children at ages 2;1–2;6 could already distinguish the six tones significantly better than chance level. There is a clear pattern of perception accuracy increasing with age. Except T1, perception accuracy of the eldest group (aged 5;7–6;0) was still worse than that of the reference group. These perceptual patterns concur well with the above argument based on production data that even the youngest age group was using Cantonese tones phonologically, while it takes another few years for their tone perception to mature. Adultlike perceptual patterns in Cantonese tones could only be found at the age of 10 years the earliest (Ciocca & Lui, 2003; Lee et al., 2015). The protracted development of tone perception is also echoed by Chen et al. (2017) on categorical perception. Thus, Cantonese tone acquisition is both early (being able to use tones meaningfully in both production and perception) and protracted (full acquisition including maturation). For parents and laymen, early acquisition of tone would match their understanding and experience well. Even for clinical settings where trained auditory judgment is the main form of analysis, the early definition would seem reasonable. Insisting on late acquisition based on fine acoustic patterns and perfect production accuracy would be unnecessary and impractical. As for speech researchers with access to various acoustic and kinematic techniques, more efforts should be made to uncover covert aspects of the protracted process of tone acquisition using fine measures. Early and protracted acquisition can be mutually compatible depending on situations and use.

The perception data in Mok et al. (2019) also illustrate that there is a watershed in tone perception development. Before age 4;0, there was a steadier increase in perception accuracy, while the perception development had slowed down after age 4;0 with reduced improvement. It will be interesting to see if any similar watershed can be found in the production data. For the accuracy data (see Figure 2), more fluctuation was found before age 4;0. There was a consistent and steady increase in accuracy from ages 4;0 to 6;0. As for the acoustic data, no obvious difference can be observed for the major contrasts between similar tone pairs. Nevertheless, there seems to be such a watershed in the degree of perceived creakiness in T4 (see Table 2), with more creaky tokens produced by children from age 4;0 onwards. It could be possible that younger children (before age 4;0) develop perception better and faster than production because of the perceptual importance of tones and the difficulty in finer motor control for laryngeal configurations. Higher TD produced by the two youngest age groups supports this possibility. When children could distinguish tones with relatively higher accuracy perceptually, they possibly could incorporate the concomitant feature better in their own production as well. The current study is the first to report on creakiness in T4 in child Cantonese. Further study can investigate whether incorporating the creakiness feature can increase T4 perception for children at different ages as well, like that reported for adult listeners (Yu & Lam, 2014). In any case, the process of tone acquisition is nonlinear. In addition to “vocabulary spurts” (Ganger & Brent, 2004), there could be “phonological spurts” as well. Admittedly, more refined longitudinal studies on production and perception developments of both tones and segments, and how the two types of development are linked, can help investigate whether there are indeed “phonological spurts,” as the watershed in production is not unequivocal.

Our data were based on tone produced in isolation. Tone undergoes much contextual tonal coarticulation in connected speech (Gandour et al., 1994; Xu, 1997). So far, not much work has been done on tonal coarticulation in adult Cantonese (Flynn, 2003). Our findings demonstrate that young children can contrast tones in isolation meaningfully. It will be interesting to compare their tonal coarticulation patterns to those of adults in both compatible and conflicting contexts. It is possible that discrepant coarticulation patterns can be found, especially in conflicting contexts, based on our findings on the T2/T5 covert contrast. Further study can explore the acquisition of subtle tonal coarticulation by children.

In conclusion, this study demonstrated that tone acquisition is multifaceted. Major acoustic contrasts among similar Cantonese tone pairs were acquired early by age 2;1–2;6, while covert contrast and fine phonetic detail were not fully acquired even by age 6;0. Maturation in tone acquisition goes well beyond age 6;0, whereas tones were used phonologically meaningfully even at age 2;1. It is necessary to define successful acquisition using a wider perspective with various criteria and applications, not just relying on adult-likeness in all aspects. More studies are warranted to further investigate both the early and protracted processes of tone acquisition using different tone languages.

Acknowledgments

This study was supported by the Hong Kong SAR Government Research Grant Council General Research Fund 2015/16 Project Reference 14602715 and the funding support by the department of the authors. We would like to thank Mandy Cheung, Crystal Lee, and Mercy Wong for their help in data collection and analysis. We are grateful for the five kindergartens participating in our study: ELCHK Shatin Lutheran Kindergarten, Heng On Baptist Nursery School, Homantin Yang Memorial Methodist Pre-School, Lutheran Philip Hse Oi Lun Nursery School, and NTW&JWA Cheung Fat Estate Nursery School. We especially thank the parents and children for their participation. We also thank Carol To for her assistance in earlier stages of the project.

References

Alan, C. L. (2007). Understanding near mergers: The case of morphological tone in Cantonese. *Phonology, 24*(1), 187–214.

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing: An Interdisciplinary Journal, 16*(6), 573–609.

Burnham, D., Tsukada, K., Jones, C., Rungrojsuwan, S., Krachaikiat, N., & Luksaneeyanawin, S. (2006). The development of lexical tone production in Thai children, 18 months to 6 years: Relationships with language milestones? In H. C. Yehia, D. Demolin & R. Laboissière (Eds.), *Proceedings of the 7th International Seminar on Speech Production*, (pp. 107). Cefala.

Chao, Y. R. (1947). *Cantonese primer*. Greenwood Press.

Chen, F., Peng, G., Yan, N., & Wang, L. (2017). The development of categorical perception of Mandarin tones in four- to seven-year-old children. *Journal of Child Language, 44*(6), 1413–1434.

Cheung, P. S. P., Ng, A., & To, C. K. S. (2006). *Hong Kong Cantonese Articulation Test (HKCAT)*. Language Information Sciences Research Centre, City University of Hong Kong.

Ciocca, V., & Lui, J. (2003). The development of lexical tone perception in Cantonese. *Journal of Multilingual Communication Disorders, 1*(2), 141–147.

Clumeck, H. (1980). The acquisition of tone. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Production* (pp. 257–275). Academic Press.

Edwards, J., & Beckman, M. E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics & Phonetics, 22*(12), 937–956.

Edwards, J., Beckman, M. E., & Munson, B. (2015). Cross-language differences in acquisition. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 530–554). Wiley.

Flynn, C.-Y.-C. (2003). *Intonation in Cantonese*. Lincom.

Gandour, J. (1981). Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics, 9*(1), 20–36.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics, 22*(4), 477–492.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology, 40*(4), 621–632.

Garellek, M. (2019). The phonetics of voice. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics* (Chap. 4). Routledge.

Hung, Y. C., Meredith, G. S., & Gill, S. V. (2013). Influence of dual task constraints during walking for children. *Gait & Posture, 38*(3), 450–454.

Iseli, M., Shue, Y.-L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America, 121*(4), 2283–2295.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)* (pp. 0821.1–0821.5). The University of Glasgow.

Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics, 35*(1), 104–117.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America, 142*(3), 1693–1706.

Lee, K. Y. S. (2012). *The Cantonese Tone Identification Test (CANTIT)*. Department of Otorhinolaryngology, Head & Neck Surgery, The Chinese University of Hong Kong.

Lee, K. Y. S., Chan, K. T. Y., Lam, J. H. S., Van Hasselt, C. A., & Tong, M. C. F. (2015). Lexical tone perception in native speakers of Cantonese. *International Journal of Speech-Language Pathology, 17*(1), 53–62.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language, 4*(2), 185–199.

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*(3), 241–265.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition, 106*(3), 1367–1381.

Mok, P., & Zuo, D. (2012). The separation between music and speech: Evidence from the perception of Cantonese tones. *The Journal of the Acoustical Society of America, 132*(4), 2711–2720.

Mok, P., Zuo, D., & Wong, P. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*(3), 341–370.

Mok, P. P. K., Fung, H. S. H., & Li, G. V. (2019). Assessing the link between perception and production in Cantonese tone acquisition. *Journal of Speech, Langauge, and Hearing Research, 62*(5), 1243–1257.

Munson, B., Edwards, J., & Beckman, M. E. (2011). Phonological representations in language acquisition: Climbing the ladder of abstraction. In A. C. Cohn, C. Fougeron, & M. K. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (pp. 288–209). Oxford University Press.

Rose, P. (2004). The acoustics and probabilistic phonology of short-stopped syllable tones in Hong Kong Cantonese. In S. Cassidy, F. Cox, R. Mannell & S. Palethorpe (Eds.), *Proceedings of the 10th Australian International Conference on Speech Science & Technology* (pp. 445–450). Australian Speech Science and Technology Association.

Scobbie, J. E., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Language acquisition and the lexicon*. Cambridge University Press.

Singh, L., & Fu, C. S. L. (2016). A new view of language development: The acquisition of lexical tone. *Child Development, 87*(3), 834–854.

So, L. K. H., & Dodd, B. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language, 22*(3), 473–495.

To, C. K. S., Cheung, P. S. P., & McLeod, S. (2013). A population study of children's acquisition of Hong Kong Cantonese consonants, vowels and tones. *Journal of Speech, Language, and Hearing Research, 56*(1), 103–122.

Tse, J. K. P. (1978). Tone acquisition in Cantonese: A longitudinal case study. *Journal of Child Language, 5*(2), 191–204.

Wong, M. (1982). *Tone change in Cantonese* (Unpublished PhD thesis). University of Illinois at Urbana-Champaign.

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intra-talker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46*(2), 413–421.

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

Wong, P. S. (2013). Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan. *The Journal of the Acoustical Society of America, 133*(1), 434–443.

Wong, P. S., & Chan, H. Y. (2018). Acoustic characteristics of highly distinguishable Cantonese tones. *The Journal of the Acoustical Society of America, 143*(2), 765–779.

Wong, P. S., Fu, W. M., & 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. *Frontiers in Psychology, 8,* 1450.

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

Wu, J. H., Ajisafe, T., & Beerse, M. (2015). Children display adult-like kinetic patterns in the time domain, but not in the frequency domain, while walking with ankle load. *Journal of Applied Biomechanics, 31*(5), 292–308.

Xu Rattanasone, N., Attina, V., Kasisopa, B., & Burnham, D. (2013). How to compare tones. In H. Winskel & P. Padakannaya (Eds.), *South and Southeast Asian psycholinguistics* (pp. 233–246). Cambridge University Press.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61–83.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10).

Yu, K., & Lam, H. W. (2014). The role of creaky voice in Cantonese tone perception. *The Journal of the Acoustical Society of America, 136*(3), 1320–1333.

Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics, 37*(2), 231–247.

Zhu, H. (2002). *Phonological development in specific contexts: Studies of Chinese-speaking children*. Multilingual Matters.

Zhu, H., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language, 27*(1), 3–24.