

# **Speaker Variability in the Realization of Lexical Tones**

**Ricky Chan**

**Department of Linguistics and English Language, Lancaster University**

**k.w.chan2@lancaster.ac.uk**

## **Abstract**

While previous studies on the speaker-discriminatory power of static f<sub>0</sub> parameters abound, few have focused on the dynamic and linguistically-structured aspects of f<sub>0</sub>. Lexical tone offers a case in point for this endeavour. This paper reports an exploratory study on the speaker-discriminatory power of individual lexical tones and of the height relationship of level tone pairs in Cantonese, and the effects of voice level and linguistic condition on their realization. Twenty native Cantonese speakers produced systematically controlled words either in isolation or in a carrier sentence under two voice levels (normal and loud). Results show that f<sub>0</sub> height and f<sub>0</sub> dynamics are separate dimensions of a tone and are affected voice level and linguistic condition in different ways. Moreover, discriminant analyses reveal that the contours of individual tones and the height differences of level tone pairs are useful parameters for characterizing speakers.

## **Keywords**

Speaker characteristics, Lexical Tone, Forensic Speaker Comparison, Cantonese

## 1.0 Introduction

The realization of the same phonological unit exhibits considerable variation across speakers. Such between-speaker variation has been exploited in the task of forensic speaker comparison (FSC), which typically involves the comparison of the speech sample from a perpetrator and that from a suspect (French and Stevens, 2013; Nolan, 1983). A major goal in FSC research is to identify a set of parameters that can potentially discriminate speakers. One of the most widely-used acoustical parameters for FSC is fundamental frequency ( $f_0$ ) (Rose, 2002). Most previous studies in the  $f_0$  domain focused on static parameters such as average  $f_0$ , range, standard deviation or  $f_0$  alignment (e.g. Boss, 1996; Braun, 1995; Jessen, Köster and Gfroerer, 2005; Künzel, 2000; Nolan, 2002). A recent survey on international practices in FSC also reveals that while all respondents measured  $f_0$  in FSC, most of them used only static  $f_0$  parameters in their analysis (Gold and French, 2011). Relatively few studies have investigated the dynamic and linguistically-structured aspects of  $f_0$ , which are a potentially rich source of speaker-specific information (see McDougall, 2004, 2006 for dynamic measures – of formants in her case). Lexical tone is a case in point.

Lexical tone mainly involves the use of dynamic pitch patterns to contrast word meanings. Around 60-70% of the world languages are tone languages, which are mostly found in Africa, East and South-East Asia and the Pacific, and the Americas (Yip, 2002). The primary acoustic correlate of lexical tone is  $f_0$ , which is mainly determined by the rate of vibration of the vocal folds (Bauer and Benedict, 1997).

Previous studies on the use of tonal  $f_0$  information for characterising speakers have got mixed results. Thaitechawat and Foulkes (2011) studied the speaker-specificity of lexical tones and formant dynamics in standard Thai. Five male speakers of standard Thai were instructed to read aloud a word list that contained a balanced number of the five tones in Thai. Discriminant analyses showed that tonal  $f_0$  data alone yielded 72-88% correct attribution, with the rising tone producing the best discriminatory power. Still, the generalizability of their results is limited in that only a small sample size and production of isolated words were involved. Wang and Rose (2012) studied the speaker discriminatory power of the low level tone in Cantonese carried by the vowel /i:/ in the word “二” (*two*). Speech samples of 26 male Cantonese speakers were obtained from a database which contains two non-contemporaneous recordings of

responses to questions about the Hong Kong Mass Transit Railway. They found that the log-likelihood ratio cost ( $C_{llr}$ ) for the low level tone were 0.68 with an equal error rate at 19%, suggesting that the tone is of potential use in FSC casework. On the other hand, using similar methods, Li and Rose (2012) focused on the high rising tone [25] in Cantonese carried by the diphthong /ɔy/ elicited from 15 young male speakers. They found that  $C_{llr}$  for the tone was 0.86 (close to 1) with the equal error rate (EER) at 40%, indicating that the rising tone is of limited use in identifying speakers. These conflicting results point to the need of a more comprehensive study on the role of speaker discriminating potential of lexical tones. The present study investigates the speaker-discriminatory power of lexical tones in Hong Kong Cantonese.

Hong Kong Cantonese contrasts six lexical tones: three level tones (high, middle and low), two rising tones (high and low) and a falling tone (Bauer and Benedict, 1997). Table 1 illustrates how the syllable /ji:/ exploits the six tones for lexical contrast. With such a rich tone inventory, Cantonese offers excellent scope for comparing the speaker discriminatory powers of different types of tone. Moreover, the two rising tones have recently been reported to be merging, especially among young speakers (Bauer, K. H. Cheung and P. M. Cheung, 2003; Fung and C. Wong, 2011; Mok, Zuo and P. W. Wong, 2013). Three patterns of merging have been identified: 1) merging T2[25] to T5[23]; 2) merging T5[23] to T2[25]; and 3) realizing a new intermediate tone between the canonical forms of the two rising tones. Studies have shown that diachronically-changing sounds are more likely to display between-speaker variation than the relatively stable counterparts (DeJong, McDougall and Nolan, 2007; Moosmüller, 1997), as some speaker may have more conservative or more novel realisations. It is thus hypothesised that the merging tones will possess higher speaker-discriminatory powers than the non-merging ones.

*Table 1: Illustration of the six Cantonese tones. The numbers in the phonemic transcriptions represent the pitch level of the tone with reference to a speaker's tonal pitch range (1=lowest; 5= highest) (Chao, 1947).*

	<b>Tone</b>	<b>Example in Cantonese</b>	<b>English Translation</b>	<b>Phonemic Transcription</b>
1	High level	衣	clothing	/ji: 55/
2	High rising	椅	chair	/ji: 25/

3	Mid level	意	idea	/ji: 33/
4	Low falling	疑	suspicious	/ji: 21/
5	Low rising	耳	ear	/ji: 23/
6	Low level	二	two	/ji: 22/

An important feature of lexical tone is that tones are defined not in absolute terms by the language but in relative terms with reference to the speaker's pitch range (Bauer and Benedict, 1997). In Cantonese, for example, the identity of the three level tones is determined by taking into account the speaker's pitch range and adjacent tonal context (Wong and Diehl, 2003). Li (2006) postulates that while the absolute  $f_0$  level of different tokens of a tone may vary greatly in an utterance, the relative height between two adjacent tones produced by the same speaker should be largely consistent locally (i.e. between neighbouring tones) for maintaining communication accuracy. While this predicts restricted within-speaker variation in the relative height of two adjacent tones, the degree of between-speaker variation remains unclear. Wong and Diehl (2003) provide indirect evidence for speaker-specific realization of the relative height of Cantonese level tones. In one of their experiments, native Cantonese-speaking listeners were asked to identify isolated Cantonese level tones produced by 7 different speakers. The presentation of the level tones was either grouped by speaker or mixed across speakers. They found that identification accuracy was significantly higher when items were blocked by speakers than when items were mixed across speakers, suggesting that there were considerable between-speaker differences and/or small within-speaker differences in the realisations of the level tones for the listeners to exploit in the tone identification task. The second goal of the present study is to explore speaker-specificity in the relative realization of tones. As a start, the present study focuses on the relative height of two adjacent level tones.

In addition, to determine the potential value of a parameter for FSC, it is necessary to assess how the parameter may be affected by changes in speaking conditions, as in forensic casework there is often a mismatch in speaking styles between the known and unknown speech samples. The present study focuses on the effects of different speaking rates (normal vs. fast) and voice levels (normal vs. loud) on the speaker discriminatory powers of tonal parameters. These two factors are particularly relevant

to lexical tone in that, in acoustic terms, change in voice levels often lead to differences in tonal f0 mean and ranges, and change in speaking rates may result in differences in tone duration and differences in tonal dynamics (e.g. the timing of the turning point in dynamic tones, Sereno, Lee and Jongman, 2015).

In sum, this paper reports an exploratory study on the speaker-discriminatory power of 1) the six tones in Cantonese; and 2) the relative height relationship of two consecutive level tones in different speaking rates and voice levels, in a bid to identify potentially useful tonal parameters for FSC casework.

## 2. Method

### 2.1 Participants

20 native male speakers of Hong Kong Cantonese (aged from 19 to 26, mean = 22.4) were recruited for the experiment. All of them were born and brought up in Hong Kong, and have resided in Hong Kong for more than 15 years.

### 2.2 Materials

*Realisation of the six Cantonese tones.* 6 disyllabic meaningful words were adopted to elicit the production of the 6 Cantonese tones (see (a) in Table 2). The first syllable carries T3 [33] which occupies the middle tonal space and serves as a constant tonal context, and the second syllable carries the target tone.

*Realisation of two level tones in sequence.* To study the relativity of tone realisation, we focused on the relative height relationship of the three level tones. 9 tone pairs were concatenated from the 3 level tones in Cantonese: high-high (HH), high-mid (HM), high-low (HL), mid-high (MH), mid-mid (MM), mid-low (ML), low-high (LH), low-mid (LM) and low-low (LL). Nine disyllabic words were used to elicit the production of the above 9 tone pairs (see (b) in Table 2). Three of the 15 disyllabic words (in bold) overlapped, and thus a total of 12 disyllabic words were used in the present study.

*Table 2: Disyllabic words used in the experiment and their phonemic transcriptions. H denotes the high level tone; M the mid level tone; and L the low level tone.*

Tones	Disyllabic word	Gloss
<b>(a) 6 tones</b>		

<b>T3-T1 (M-H)</b>	<b>至知</b> /t̃si: t̃si:/	<b>to realize</b>
T3-T2	廁紙 /t̃sʰi: t̃si:/	tissue paper
<b>T3-T3 (M-M)</b>	<b>次次</b> /t̃sʰi: t̃sʰi:/	<b>every time</b>
T3-T4	致詞 /t̃si: t̃sʰi:/	to deliver a speech
T3-T5	嗜柿 /si: t̃sʰi:/	to love persimmon
<b>T3-T6 (M-L)</b>	<b>試事</b> /si: si:/	<b>exam</b>
<b>(b) 9 level tone pairs</b>		
T1-T1 (H-H)	痴痴 /t̃sʰi: t̃sʰi:/	to stick
T1-T3 (H-M)	之至 /t̃si: t̃si:/	very much
T1-T6 (H-L)	私事 /si: si:/	private matter
<b>T3-T1 (M-H)</b>	<b>至知</b> /t̃si: t̃si:/	<b>to realize</b>
<b>T3-T3 (M-M)</b>	<b>次次</b> /t̃sʰi: t̃sʰi:/	<b>every time</b>
<b>T3-T6 (M-L)</b>	<b>試事</b> /si: si:/	<b>exam</b>
T6-T1 (L-H)	自知 /t̃si: t̃si:/	self-consciousness
T6-T3 (L-M)	自置 /t̃si: t̃si:/	privately-owned
T6-T6 (L-L)	事事 /si: si:/	everything

All the words share the same nucleus (the vowel /i:/ with no coda) and similar onsets: either a voiceless fricative or a voiceless affricate. This served to control for potential differences in f0 perturbation effects and vowel intrinsic f0 effects (Lehiste, 1970).

### 2.3. Procedure

Recordings took place inside the sound-treated booth in the Phonetics Laboratory in the Department of Linguistics, University of Cambridge. Subjects were recorded through a Sennheiser MKH 40P48 condenser microphone set about 15 cm away from the subject's mouth, sampling at 44.1kHz/16 bits. All materials were presented on a computer screen in a random order.

To explore how the acoustic realization of the lexical tones varies across different speaking rates and voice levels, subjects were instructed to produce, in both normal and loud voice, the 12 disyllabic words in 1) isolation (*IS condition*); and 2) in a carrier sentence (*CS condition*):

彼得未聽過 XX 呢個詞語。  
xxxxMXXHxxx  
(Peter has never heard of the word “XX”.)

where XX stands for the target disyllabic word, M a syllable with a mid level tone, H a syllable with a high level tone, and x other syllables in the sentence. It was expected that the use of a carrier sentence would encourage a higher speaking rate.

Production data were obtained at two different voice levels: normal voice and loud voice. To elicit a loud voice from the speakers, the experimenter sat far away from the subject and the computer screen was moved further away as well. These created a sense of “distance” for the subject. The subject was then told to imagine speaking to a person far away from him and was instructed to “speak up”. A dummy microphone was set far away from the subject, while the position of the microphone used for recording remained unchanged.

To minimize potential lexical effects, participants were given enough time to practise and familiarize themselves with the disyllabic words before the actual recordings. In the actual recording, participants first read aloud the target words in the carrier sentence, in normal voice and then loud voice. They then read aloud the words in isolation, in normal voice and then loud voice. Each target word was produced 10 times in each condition and voice level, resulting in a total of 480 tokens from each speaker (12 words x 2 conditions x 2 voice levels x 10 times).

#### *2.4. Data Extraction*

This study focuses on f<sub>0</sub> since it is the primary acoustic correlate of Cantonese tones (Vance, 1976; Khouw & Ciocca 2007). Data were analysed using *Praat* (Boersma and Weenink, 2014). For each target word, two vertical markers were inserted manually from the beginning to the end of periodicity (from the start of F1 to the end of F2) of the /i:/ vowel (which carries the lexical tone) in the spectrogram. A *Praat* script was then applied to extract f<sub>0</sub> values with the autocorrelation method in all regions delimited by the vertical markers. As all tokens have different durations, the f<sub>0</sub> contours were equalised by dividing the delimited regions into 10 equal intervals. f<sub>0</sub> values were extracted at each 10% step of each delimited region (i.e. 0%, 10%, 20%, 30%...90%, 100%), giving 11 values in total. Values at onset (0%) and offset (100%) have been excluded in the analysis as these values are unreliable and mostly reflect perturbation by neighbouring consonants. Around 2% of the tokens (mostly

T4[21] and a few T6[22]) were so creaky that f0 values could not be extracted and were excluded from the analysis.

### 3.0 Results and Discussion

#### 3.1 Realization of the six tones

##### 3.1.1 Descriptive data

Figure 3.1 shows the average duration of the six tones in IS and CS conditions.

Overall speaking, duration of the six tones in descending order is

T2>T6>T1>T5>T3>T4.

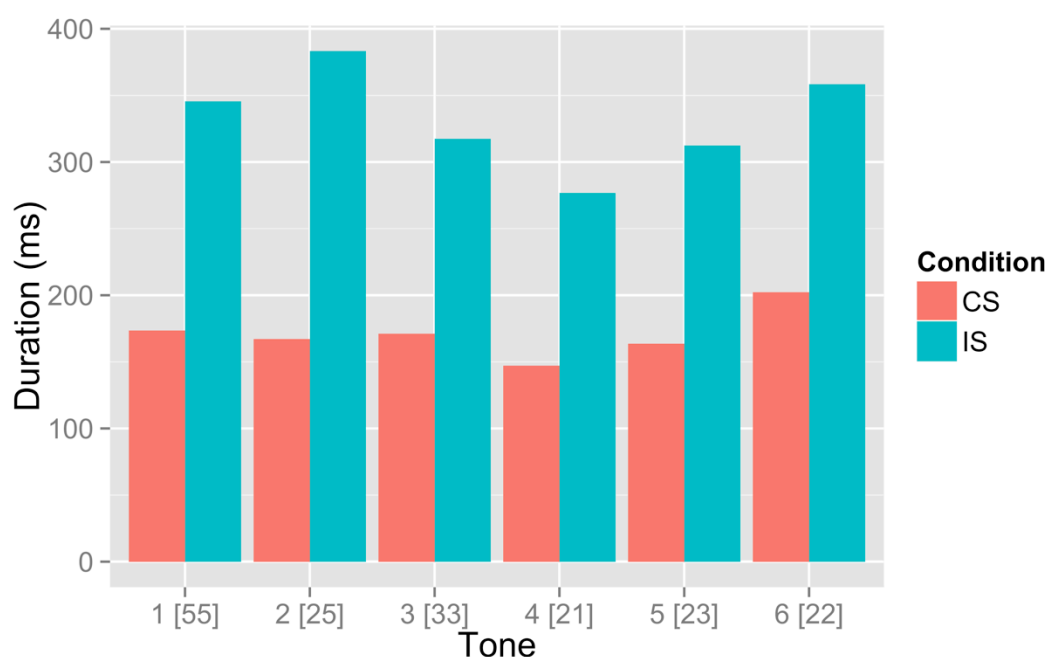


Figure 3.1.: Duration of the six tones produced in isolated words (IS condition) and in a carrier sentence (CS condition).

Figure 3.2 shows the boxplot and figure 3.3 shows the distributions of the 20 speakers' f0 data based on their realization of the six tones in all voice levels and linguistic conditions. Impressionistically, speakers show considerable variation in their f0 range and the distribution of the f0 values. For instance, speakers HC and JW had a relative small tonal f0 range whereas speakers KT and PL had a relatively large tonal f0 range.



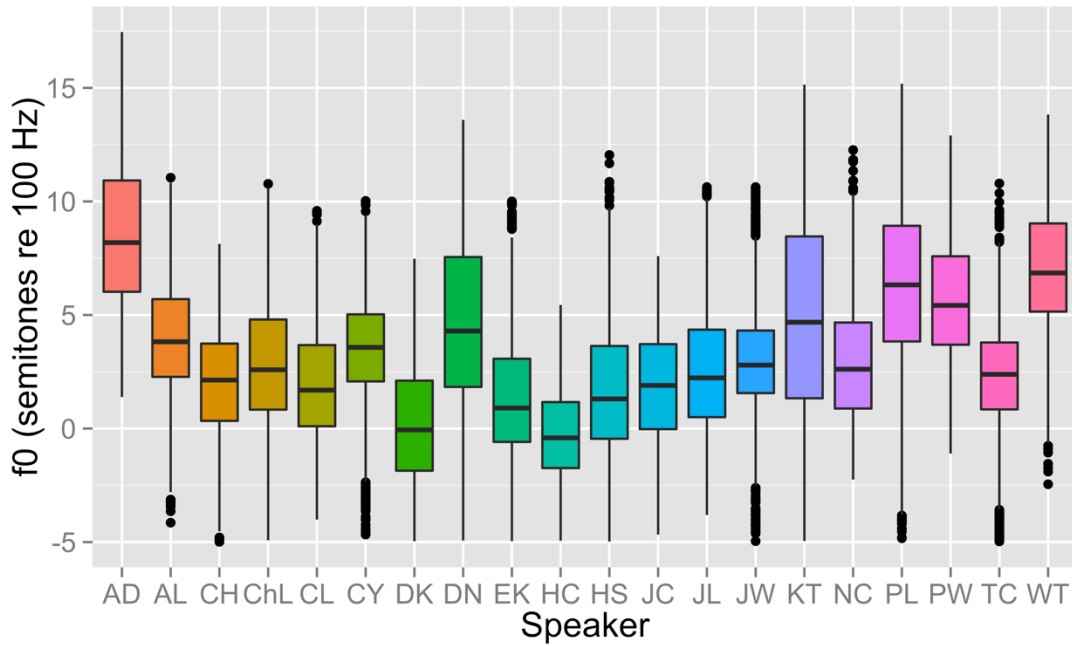


Figure 3.2.: Box and whisker plots of the 20 speakers'  $f_0$  data based on their realization of the six tones (x-axis: Speaker; y-axis:  $f_0$  (semitones re 100Hz)). The bottom and top of the boxes represent the first and third quartiles respectively, and the band inside the box the median. The ends of the whiskers represent the minimum and maximum.

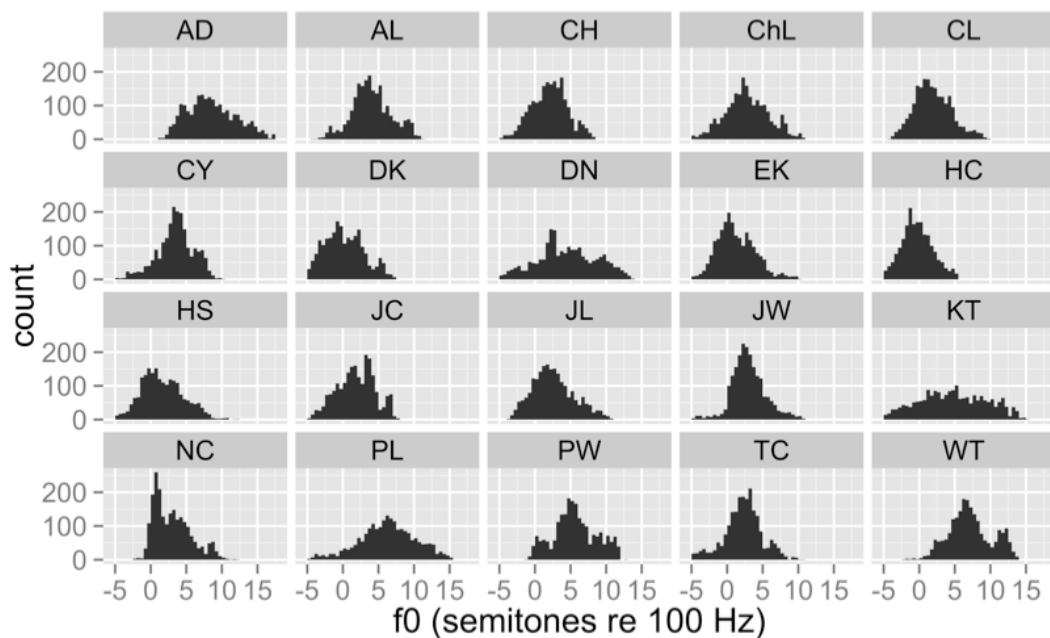


Figure 3.3: Distribution of the 20 speakers'  $f_0$  data (in semitones re 100 Hz) based on their realization of the six tones.

Linear mixed-effects models (LMMs) were used to determine the effect of *Condition* on tone duration and the effect of *Voice level* on speakers’ mean  $f_0$  across all measurement points of all the tones, with the R package *lme4* (Bates and Maechler, 2012) in R (R Core Team, 2012). *Condition*, *Tone* and *Voice Level* are treated as fixed factors, and *Speaker* and *Token* as random factors with by-*Condition* random slopes. Table 3.1 summarize the levels of each the factor. Effects were tested by likelihood ratio tests of a full model against a reduced model that excluded the effect to be tested (i.e. *Condition/Voice Level*), using the R code “`anova(full_model, reduced_model)`”. A p-value was obtained for each model comparison using standard likelihood ratio tests. Results showed that in general tone duration is shorter in the CS condition than in the IS condition by  $122.04 \pm 9.04\text{ms}$ ,  $\chi^2(1) = 44.4$ ,  $p \ll 0.001$ . This is attributable to the faster speaking rate in CS condition. On the other hand,  $f_0$  of the speakers at loud voice is higher than that at normal voice by  $2.17 \pm 0.258\text{st}$ ,  $\chi^2(1) = 30.6$ ,  $p \ll 0.001$ . These suggest that our procedure of eliciting loud speech did lead to a significantly higher  $f_0$  in general.

*Table 3.1: Summary of different levels for each factor*

<b>Factor</b>	<b>No. of Levels</b>	<b>Details</b>
Condition	2	CS and IS
Voice Level	2	Normal and Loud Voices
Tone	6	The six tones
Speaker	20	20 speakers
Token	10	10 repetitions

By presenting the tone contours on a frequency scale (e.g. Hz or semitones), between-speaker differences in both absolute frequency and the shape of the tone contours will be revealed. To determine whether the speakers exhibit idiosyncratic differences in the dynamic changes of their tone contours, all raw  $f_0$  data were normalised on a  $z$ -score scale (Rose, 1987), which involves expressing an observed  $f_0$  value in a standard score based on the following formula:

$$f_{0\text{norm}} = (f_{0i} - f_{0\text{mean}})/s$$

where  $f_{0\text{mean}}$  stands for the mean of all sampled data for a given speaker and  $s$  one standard deviation from the mean. The  $z$ -score then represents the degree of dispersion by the number of standard deviations from the mean. Data were normalised separately for each speaker and for normal and loud voices.

Figure 3.4 shows the average  $f_0$  contours of the six Cantonese tones across all conditions and voice levels by each speaker based on the normalised data. While across speakers the  $f_0$  contours of all tones generally show some degree of resemblance, the density of their tone system seems to differ. For example, the low falling tone [21] produced by speakers such as ChL, JC and TC are well separated from their realisations of the other tones, but the same tone by speakers such as HC, JW and NC is close to the other tones. The speakers also exhibit different patterns of merging of the two rising tones. While some speakers (e.g. CY, HC, NC) seem to distinguish the two rising tones, others show different patterns of merging, with speakers such as AD, JW, and TC completely merging the two tones.

Figure 3.5 and 3.6 shows the mean  $f_0$  contours of the six tones across speakers. Figure 3.5 adopts the same y-axis limits with the aim of accurately capturing the general  $f_0$  height and shape of the six tones. The production of the six tones is generally consistent with the canonical forms reported in the literature. Figure 3.6 provides zoom-in views, each with a scale that best captures the individual differences among the 20 speakers.

For the high level tone, most speakers had a stable rise to the peak with a terminal fall, but they differ in the timing and magnitude of  $f_0$  declination and the terminal fall. The two other level tones display similar patterns: both resemble a falling tone owing to  $f_0$  declination, and speakers differ in the the gradient of their drop in pitch and the onset of a levelling off. The two rising tones both depict a dip-rise patterns in general, but speakers vary in terms of the timing and degree of the dip, if present, and magnitude of rise in the second half of the tone. For the low falling tone, while some have demonstrated a straight and constant lowering of  $f_0$ , others have terminated the fall half-way and level off.

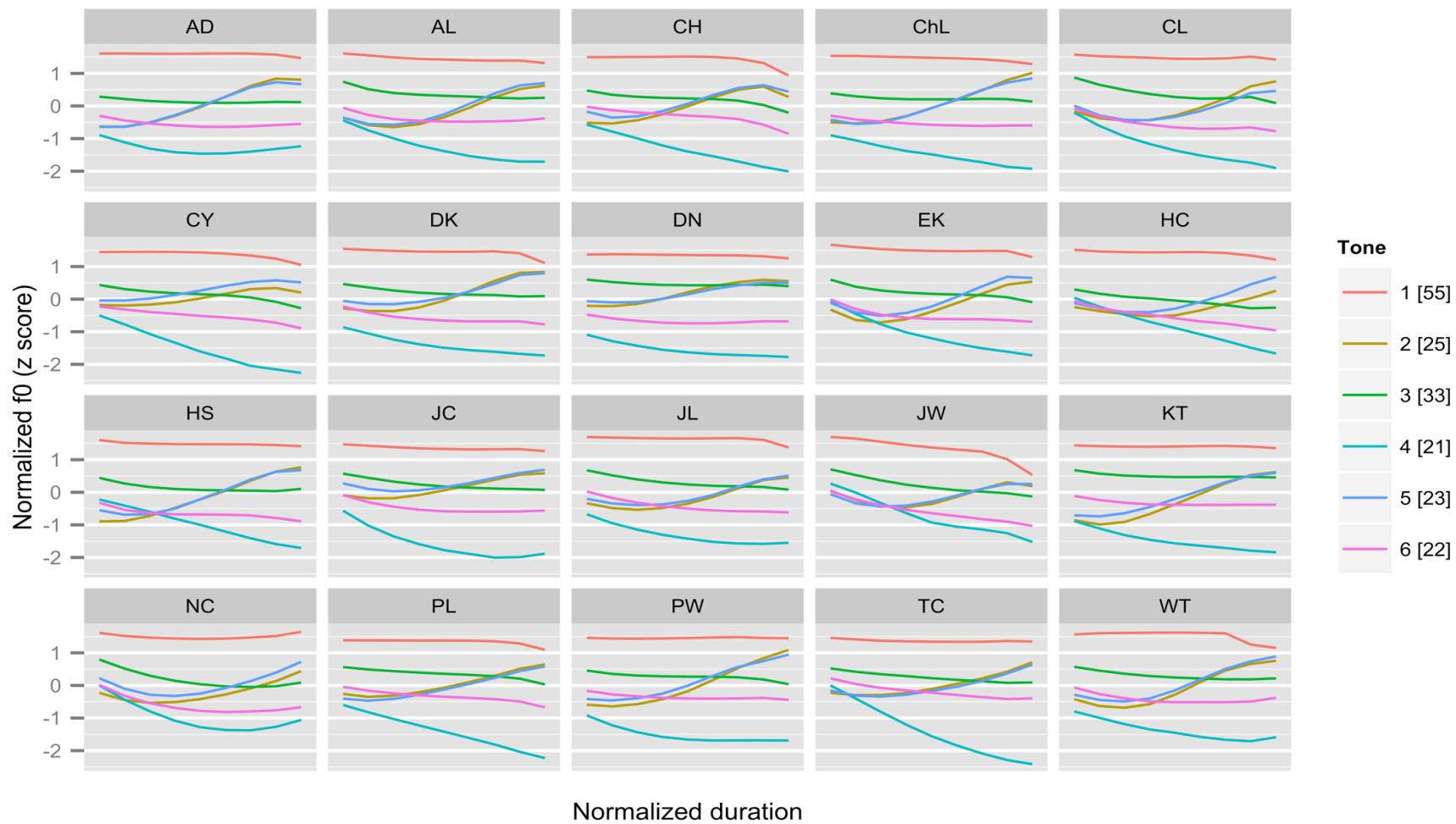


Figure 3.4: Average  $f_0$  contours of the six Cantonese tones by 20 speakers.

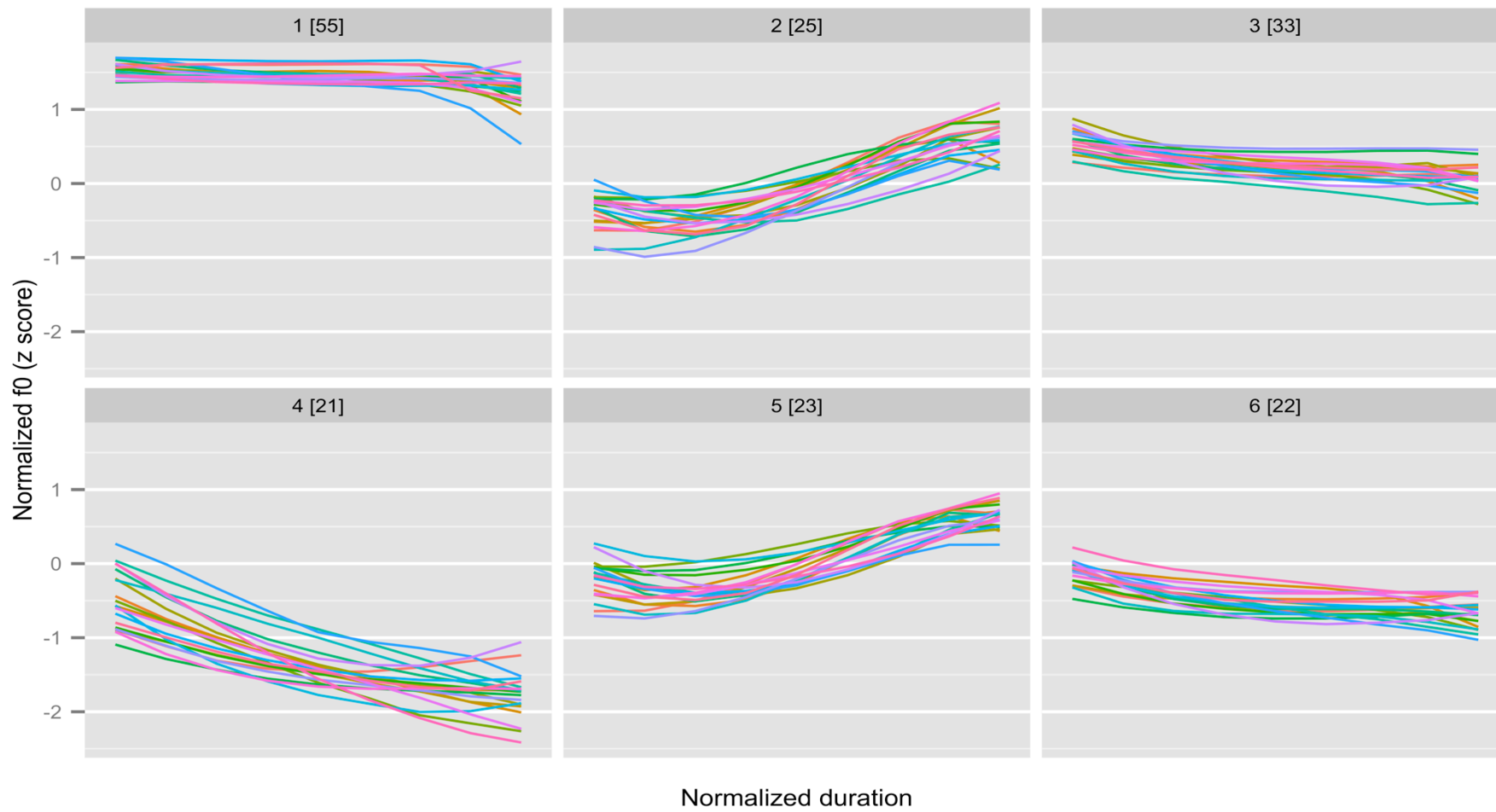


Figure 3.5: mean  $f_0$  contours of the six tones by 20 speakers (fixed y-axis limits).

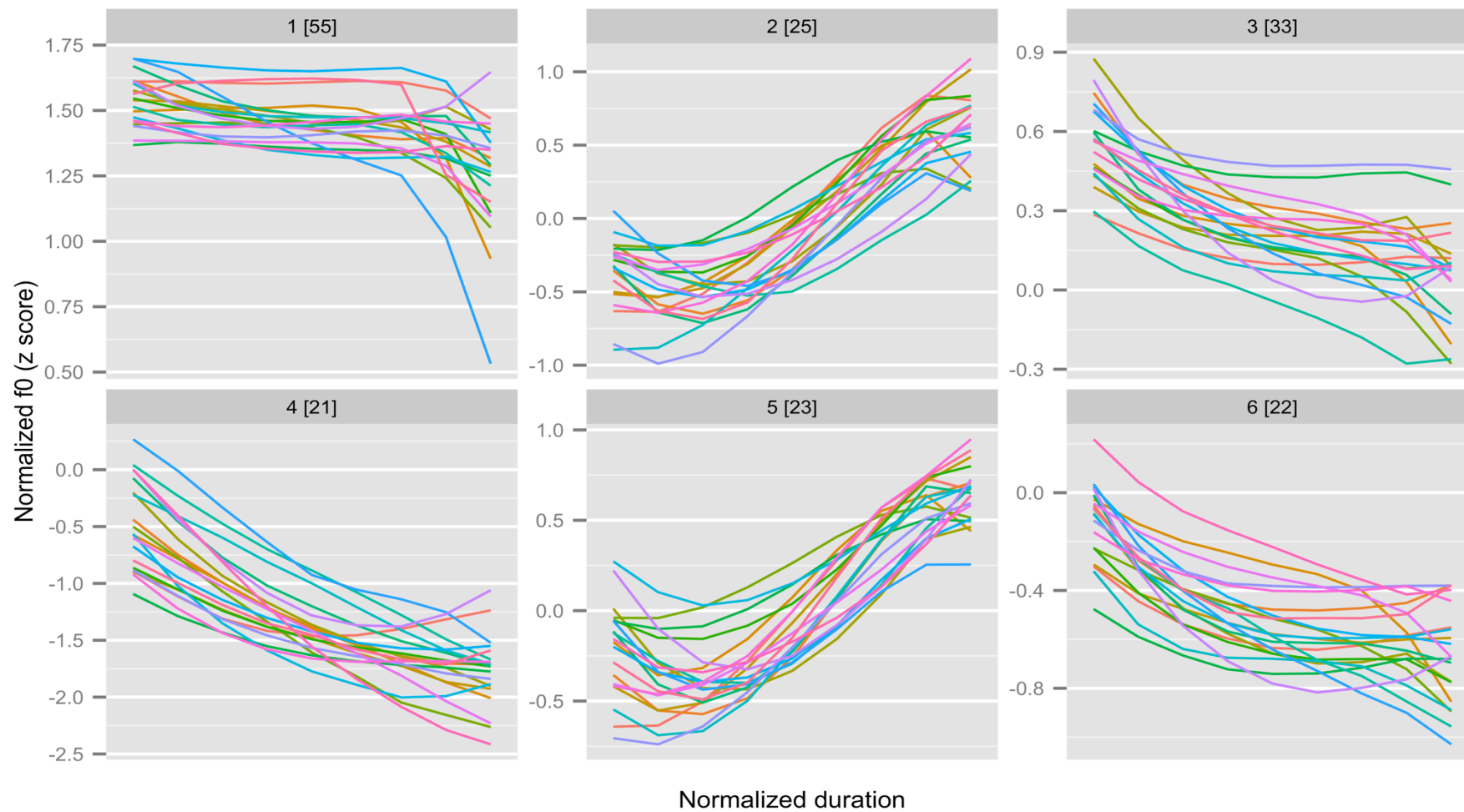


Figure 3.6: mean f0 contours of the six tones by 20 speakers (different y-axis limits).

LMMs were used to determine the effect of *Condition*, *Voice Level*, and their interaction with *Interval* (the 9 measurement points which represents the dynamic change) on  $f_0$ . *Condition*, *Voice Level*, *Tone*, and *Interval* were treated as fixed factors, and *Token* and *Speaker* as random factors with random slopes on the factor under investigation. Effects for each individual factor were tested by likelihood ratio tests of a full model against a reduced model that excluded the effect to be tested. For instance, to test the effect of *Condition*, the R code for the full model is “full model <- lmer( $f_0 \sim$ Condition +Voice Level + Tone + Interval + (Condition|Speaker) + (Condition|Token))”; the reduced model was coded as “reducedmodel <- lmer( $f_0 \sim$ Voice Level + Tone + Interval + (Condition|Speaker) + (Condition|Token))”. Effects for interaction were tested by comparing models with and without the interaction (e.g. R code: *Condition \* Interval* vs. *Condition + Interval*), with random slopes on the factor under investigation (i.e. *Condition/Voice Level*) for each random factor. Whilst the factor *Condition* or *Voice Level* alone shows how different levels in each factor account for baseline differences in  $f_0$ , their interaction with *Interval* reveals their effects on the dynamic changes of  $f_0$  contours. Table 3.2 summarizes the main results for the raw  $f_0$  data and normalized data.

Table 3.2: Summary of the statistics of the mixed model comparisons for *Condition*, *Voice Level*, and their interaction with *Interval*.

Factor	Result	
	Raw	Normalized
Condition	$\chi^2(1) = 2.73, p = 0.0983$	$\chi^2(1) = 2.50, p = 0.114$
Condition * Interval	$\chi^2(8) = 349.46, p \ll 0.001$	$\chi^2(8) = 482.89, p \ll 0.001$
Voice Level	$\chi^2(1) = 49.60, p \ll 0.001$	$\chi^2(1) = 1.24, p = 0.2651$
Voice Level * Interval	$\chi^2(8) = 8.06, p = 0.4275$	$\chi^2(8) = 8.122, p = 0.4216$

Analyses based on raw  $f_0$  data reveal that although *Condition* does not have a significant effect on  $f_0$ , its interaction with *Interval* does, suggesting that different speaking rates affect  $f_0$  dynamics (e.g. more tone compression at a faster speech rate) but not general  $f_0$  height. On the contrary, while *Voice level* has a significant effect on  $f_0$ , its interaction with *Interval* does not, suggesting that different voice levels affect general  $f_0$  height but not  $f_0$  dynamics. Similar results were obtained for normalized data except that *Voice level* no longer has a significant effect on general  $f_0$  height, indicating that the degree of excursion of tones with reference to a speaker’s  $f_0$  range is largely

independent of the speaker's voice level. Overall, these results show that speaking rate and voice level affect f0 height and f0 dynamics of a tone in a different way, and the two dimensions of tones should be considered separately in characterising speakers.

### *3.1.2 Speaker discrimination*

To determine whether the observed speaker variations are of potential value for forensic speaker comparison, the speaker-discriminating powers of the six Cantonese tones were evaluated using discriminant analysis (DA). DA is a multivariate statistical technique that determines if a given set of predictors can be combined to predict group membership (Tabachnick & Fidell, 2007). DA requires that the number of tokens must be greater than the number of predictors. For the present study, each speaker was treated as a group with 10 tokens for each tone in each condition and voice level, and the 9 f0 measurement points of each tone as predictors. Taking into account both between- and within-speaker variations, DA constructs discriminant functions that could best separate different speakers based on the predictors, and the discriminant functions were used in classification. The "leave-one-out" cross-validation method was adopted: one token in each speaker's data set was regarded as an unknown sample and the remaining tokens were used to build the speaker's model. Every token in the data set was allocated to one of the group (speaker). The percentage of correctly attributed tokens (or a classification rate) is calculated and the best performance is reported as a DA score. With 20 speakers in the data set, the chance performance is 5%.

Separate DAs were run for each tone in different voice levels and conditions, and for both raw f0 (semitones) data and z-score normalized data. DA scores based on raw f0 data reflect classification results based on both absolute f0 height and the dynamic changes of the f0 contours, whereas those based on normalized data reflect mainly the latter. As DA is sensitive to outliers, the data were scanned for univariate ( $z > 3.29, p < .001$ ) and multivariate outliers ( $\chi^2 \geq \chi^2_{crit}, p < .001$ ) for each speaker (Tabachnick and Fidell, 2007). These outliers were removed from the analysis. The results are presented in Table 3.3.

*Table 3.3: DA scores (% correct attribution) of the 6 tones, with the chance level at 5% (N: normal voice; L: loud voice; IS: words produced in isolation; CS: words produced in a carrier sentence).*



<b>Tone</b>	<b>Voice</b>	<b>Condition</b>	<b>Raw</b>	<b>Normalized</b>	<b>Overall</b>
T1 [55] High Level	N	CS	47.0	33.3	<b>41.3</b>
		IS	50.0	41.3	
	L	CS	46.0	27.5	
		IS	48.5	37.0	
	<b>Mean</b>		<b>47.9</b>	<b>34.8</b>	
T2 [25] High Rising	N	CS	57.5	40.5	<b>51.5</b>
		IS	62.0	44.3	
	L	CS	51.5	31.0	
		IS	72.0	53.3	
	<b>Mean</b>		<b>60.8</b>	<b>42.3</b>	
T3 [33] Mid Level	N	CS	47.5	34.0	<b>40.6</b>
		IS	47.0	36.0	
	L	CS	45.0	21.8	
		IS	59.0	34.8	
	<b>Mean</b>		<b>49.6</b>	<b>31.7</b>	
T4 [21] Low Falling	N	CS	34.0	31.0	<b>35.5</b>
		IS	45.5	30.5	
	L	CS	39.5	26.6	
		IS	43.5	33.0	
	<b>Mean</b>		<b>40.6</b>	<b>30.3</b>	
T5 [23] Low Rising	N	CS	43.5	34.5	<b>43.1</b>
		IS	45.5	36.5	
	L	CS	50.5	27.9	
		IS	62.0	44.0	
	<b>Mean</b>		<b>50.4</b>	<b>35.7</b>	
T6 [22] Low Level	N	CS	44.5	34.8	<b>39.0</b>
		IS	44.5	34.8	
	L	CS	44.0	26.6	
		IS	49.0	33.5	
	<b>Mean</b>		<b>45.5</b>	<b>32.4</b>	
<b>Overall Mean</b>			<b>49.1</b>	<b>34.5</b>	

DA scores based on raw f0 data and normalized data are both much higher than chance (5%), and such high DA scores are generally preserved across different conditions and voice levels. This demonstrates lexical tones are potentially useful for separating speakers. DA scores based on raw frequency values were significantly higher than those on normalized values,  $t(46) = 7.24, p \ll .001, d = 2.09$ , suggesting that z-normalization has significantly reduced speaker-specificity in absolute frequency. Still, in general around 70% (34.5/49.1) of the discriminatory power was preserved after normalization;

this shows that the dynamic changes of tonal f0 make a substantial contribution to the discrimination.

Of the six tones in Cantonese, the two rising tones yielded the highest DA scores (51.5 for T2 and 43.1 for T5). This is consistent with previous findings that the rising tone in Thai performed best in differentiating speakers. The dynamic nature of the rising contours may have afforded more between-speaker differences in, for example, magnitude of rise and speed and rate f0 change. Furthermore, the two rising tones have been reported to be merging (Bauer et al, 2003; Mok et al, 2013), and our data reveal that the 20 speakers exhibit different degree of merging of these two tones. This is in line with the idea that diachronically dynamic features may be more valuable in separating speakers (DeJong et al., 2007; Moosmüller, 1997). On the other hand, the three level tones (T1 [55], T3 [33] and T6 [22]) scored slightly lower than the rising tones and have similar DA scores. This may be attributable to the fact that when producing a level tone, speakers have to maintain a relatively steady f0 contour and the main source of between-speaker difference is the degree of f0 declination. T4 [21] ranks last among all tones, indicating that it is produced with relatively high consistency across speakers as the low end of a person’s speaking tessitura is more bound by physiological limits on vocal cord vibration.

### 3.2 Relative Height of Level Tone Pairs

#### 3.2.1 Descriptive data

The f0 height of each tone is defined as the mean value of the 9 measurement points (based on raw f0 values). The relative height relationship for a tone pair is therefore defined as  $Ta_{mean} - Tb_{mean}$ , where  $Ta$  and  $Tb$  denote the first and second tones correspondingly. Since f0 data were expressed on a semitones scale, the same f0 height difference corresponds to a perceptually equivalent difference in pitch. The mean height differences of the nine tone pairs across all tokens are summarized in Table 3.4.

*Table 3.4: Mean height differences of the 9 level tone pairs (H: high[55]; M: mid[33]; L: low[22]; scale: semitones re100Hz). Positive values denote that the first tone is higher than the second tone, and vice versa for negative values.*

<b>Tone Pair</b>	<b><math>Ta_{mean} - Tb_{mean}</math></b>
HH	0.39
HM	4.54

HL	6.05
MH	-2.41
MM	0.54
ML	2.54
LH	-4.38
LM	-1.08
LL	0.65

According to Chao (1947), the high level tone is approximately three semitones higher than the mid level tone, and the mid level tone is approximately two semitones higher than the low level tone. The present data differ from the results reported in Chao (1947), and this can be explained by general f0 declination. When the two identical tones (i.e. LL, MM and HH) are produced in a row, the second tone in general has lower f0 values. Besides, while Li (2006) asserted that the f0 frequency spacing of two consecutive tones should not be affected by their order (e.g. the frequency spacing between HL should be approximately equal to that between LH), the present data do not support the claim. The height difference is always larger than what Chao reported when a tone pair starts with a tone with higher f0 (e.g. HM > 3 semitones; ML > 2 semitones), and vice versa when a tone pair starts with a tone with lower f0 (e.g. MH < 3 semitones; LM < 2 semitones). This is not surprising, given the well-known phenomenon of f0 declination (Cohen, Collier and Hart, 1982) and the fact that upward pitch change tend to take longer than a downward pitch change for a given pitch interval (Ohala & Ewan, 1973). Thus, the order of tones should be taken into account when investigating the height relationship between two tones.

LMMs were used to determine the effect of *Condition* and *Voice Level* on height differences of the 9 tone pairs. *Condition*, *Voice Level*, and *Tone Pair* were treated as fixed factors, and *Token* and *Speaker* as random factors with random slopes on the factor under investigation. Effects for each individual factor were evaluated by likelihood ratio tests of a full model against a reduced model that excluded the effect to be tested. Results show that the f0 height difference in the IS condition is higher than that in the CS condition by  $0.677 \pm 0.072$  semitones in general,  $\chi^2(1) = 33.32$ ,  $p << 0.001$ , but *Voice Level* does not have a significant effect,  $\chi^2(1) = 1.40$ ,  $p = 0.236$ . This shows that although “speaking up” may raise f0 mean and range, the height differences of level tone pairs remain largely constant.

### 3.2.2 Speaker discrimination

DA was used to assess the speaker-specificity of the height differences of level tone pairs. The f0 difference within each tone pair was used as the sole predictor with “speaker” as the dependent variable. DA was run separately for each condition and voice level; the results are presented in Table 3.5.

*Table 3.5.: DA scores (% correct attribution) of the height difference of the 9 tone pairs.*

Tone Pair	Condition and Voice Level				Mean
	CS, N	IS, N	CS, L	IS, L	
HH	12.0	12.0	18.0	15.5	<b>14.4</b>
HM	13.5	14.0	17.0	11.5	<b>14.0</b>
HL	16.0	18.0	16.5	17.0	<b>16.9</b>
MH	15.0	14.0	17.5	13.5	<b>15.0</b>
MM	12.0	10.0	10.0	7.5	<b>9.9</b>
ML	17.5	18.0	14.5	15.0	<b>16.3</b>
LH	23.0	19.0	24.0	13.0	<b>19.8</b>
LM	12.0	11.0	10.5	14.5	<b>12.0</b>
LL	8.5	8.0	11.5	9.0	<b>9.3</b>

In general DA scores of all level tone pairs appear to be low (from 9.3 to 19.8); this may be explained by the limits for the frequency spacing between two adjacent level tones, and exceeding the limits may lead to misidentification of the tone (e.g. HM may be perceived as HL if the height difference in HM is too large). However, it should be noted that these DA scores are based on only 1 predictor and are higher than chance (5%), indicating between-speaker variation is considerably greater than within-speaker variation. Variations for pairs of same level tones (i.e. HH, MM, LL) are attributable to speaker variability in f0 declination, potentially due to individual physiological differences and intonational preferences. On the other hand, tone pairs which involve a change of tone (e.g. ML, HM) in general show greater speaker-discriminatory powers than pairs of the same level tone. A change in tone may have provided more freedom for speakers to realize the frequency spacing of a tone pair. Noticeably, tone pairs HL and LH display the greatest between-speaker differences among all the tone pairs. This may be related to the fact that these tone pairs involve the biggest change in f0 and may have allowed for more space for variations.

## 4.0 Conclusions

This paper set out to explore the potential value of tonal parameters—individual tone contours and height differences of level tone pairs—for characterising speakers, and

how different voice levels and linguistic conditions may affect their realization. Results show speaker-specific realization of both individual tone contours and height differences of level tone pairs; such specificity is preserved across different voice levels and linguistic conditions. We conclude that lexical tones offer useful parameters for discriminating speakers and may potentially be useful for FSC casework. Also, speaking rate and voice level affect the f<sub>0</sub> height and f<sub>0</sub> contour of a tone in different ways, thus the two dimensions of a tone may be treated as separate parameters for characterising speakers. Future research should explore how the two dimensions of a tone may be affected by various within-speaker factors such as health and emotional states (Braun, 1995).

While speakers exhibit significant variation in various aspects of tone realization such as f<sub>0</sub> slope (both rise and fall), timing of f<sub>0</sub> turning points, and density of the whole tone system, the general shapes of all the tones appear to be consistent across speakers. This suggests that the observed speaker variability may mainly be attributed to articulatory factors. Since f<sub>0</sub> production involves the coordination of vocal folds and various muscles, cartilages, tissues and bones, individual differences in the properties of these articulatory components such as their size, mass, stiffness, compressibility and stretchability will all contribute to between-speaker differences in f<sub>0</sub> production (Xu, 2001). Such individual differences also to some extent give rise to between-speaker differences in speed of pitch change, speed of pitch direction shift, and preferred tonal pitch range (Xu, 2001), and future research should explore these individual differences in detail.

Despite the promising results, further research is required to evaluate the evidential value of the reported tonal parameters for two reasons. First, while the present study used DA which is a useful statistical tool for evaluating the speaker-specificity of a (set of) feature(s) within a group of known speakers, DA resembles a closed-set identification test (i.e. assuming the offender is among a list of reference speakers) which is not common in forensic casework. Furthermore, in FSC the job of the forensic scientists is to assist the trier of fact with their decision-making by taking into account both the prosecution hypothesis (the probability of the evidence assuming that the suspect is the person who produced the incriminating speech sample) and the defence hypothesis (the probability of the evidence assuming that the offender sample coming

from another speaker in the relevant population) (Aitken and Taroni, 2004). One way to achieve this is to use the likelihood ratio, which provides a gradient measure of the strength of evidence under a Bayesian framework (e.g. see Rose and Morrison, 2009 for a detailed discussion). Still, assessing typicality for the defence hypothesis in the likelihood ratio approach requires a large amount of reference data. By demonstrating the speaker-specificity of tonal parameters, the present study serves as a foundation for future study on tonal parameters with large-scale forensically-oriented datasets.

Second, while the goal of the present study was to examine the speaker-specificity of tonal parameters, the experiment was not designed to match real-life forensic conditions. The present study used systematically constructed read speech to test the effects of speaking rate and voice level while keeping other confounds (e.g. segmental  $f_0$  effects and intonation patterns) under control. Also, the data were of studio quality and collected in a single session. However, forensic casework mostly involves non-contemporaneous spontaneous speech samples and the quality analysis is often affected by adverse factors such as noise, compressed file formats, short speech samples, and reverberation. Further research should examine whether speaker-discriminatory powers observed in the present study may also be found under more forensically realistic conditions.

## References

- Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, UK: Wiley. <http://dx.doi.org/10.1002/0470011238>
- Bates, D.M. and Maechler, M. (2009). *lme4: Linear Mixed-Effects Models Using S4 Classes*, R Package Version 0.999375-32
- Bauer, R. and Benedict, P. (1997). *Modern Cantonese Phonology*. Berlin: Mouton de Gruyter. <http://dx.doi.org/10.1515/9783110823707>
- Bauer, R. S., Cheung, K. H. and Cheung, P. M. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change* 15(2): 211--225. <http://dx.doi.org/10.1017/S0954394503152039>
- Boersma, P. and Weenink, D. (2014). *Praat: Doing Phonetics with Computers*. <[www.praat.org](http://www.praat.org)>
- Boss, D. (1996). The problem of F0 and real-life speaker identification: a case study. *International Journal of Speech, Language and the Law* 3(1): 155--169. <http://dx.doi.org/10.1558/ijssl.v3i1.155>
- Braun, A. (1995). Fundamental frequency – how speaker-specific is it? In A. Braun and O. Köster (eds.) *Studies in Forensics Phonetics*. Beiträge zur Phonetik und Linguistik: 64. Trier: Wissenschaftlicher Verlag.
- Chao, Y. R. (1947). *Cantonese Primer*. Cambridge: Cambridge University Press. <http://dx.doi.org/10.4159/harvard.9780674732438>
- Cohen, A., Collier R., and 't Hart J. (1982). Declination: construct or intrinsic feature of speech pitch? *Phonetica* 39: 254-273. <http://dx.doi.org/10.1159/000261666>
- DeJong, G., McDougall, K. and Nolan, F. (2007). Sound change and speaker identity: an acoustic study. In C. Müller and S. Schötz (eds.) *Speaker Classification*. Springer.
- French, P. and Stevens, L. (2013). Forensic speech science. In M. Jones & R.-A. Knight (eds.) *The Bloomsbury Companion to Phonetics*. London: Bloomsbury.
- Fung, R. and Wong, C. (2011). The acoustic analysis of the new rising tone in Hong Kong Cantonese. In *Proceedings of the 17th International Congress of Phonetic Sciences*.

- Gold, E. and French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law* 18(2): 293--307. <http://dx.doi.org/10.1558/ijsl.v18i2.293>
- Jessen, M., Köster, O. and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law* 12(2): 174--213. <http://dx.doi.org/10.1558/sll.2005.12.2.174>
- Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics*, 35(1), 104-117. <http://dx.doi.org/10.1016/j.wocn.2005.10.003>
- Künzel, H. (2000). Effects of voice disguise on speaking fundamental frequency. *International Journal of Speech Language and the Law* 7(2): 150--179. <http://dx.doi.org/10.1558/sll.2000.7.2.149>
- Lehiste, J. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Li, J. J. and Rose, P. (2012). Likelihood ratio-based forensic voice comparison with F-pattern and tonal F0 from the Cantonese /eu/ diphthong. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)*.
- Li, Y. (2006). Tone ratios combined with F0 register in Cantonese as speaker-dependent characteristic. In *Proceedings of Speech Prosody 2006*.
- McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law* 11: 103--130.
- McDougall, K. (2006). Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13(1): 89--126. <http://dx.doi.org/10.1558/sll.2004.11.1.103>
- Mok, P., Zuo, D. and 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. <http://dx.doi.org/10.1017/S0954394513000161>
- Moosmüller, S. (1997). Phonological variation in speaker identification. *International Journal of Speech, Language and the Law Linguistics* 4(1): 29--47. <http://dx.doi.org/10.1558/ijsl.v4i1.29>



- Nolan, F. (2002). Intonation in speaker identification: an experiment on pitch alignment features. *International Journal of Speech, Language and the Law* 9(1): 1-21. <http://dx.doi.org/10.1558/sll.2002.9.1.1>
- Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: CUP. [http://dx.doi.org/10.1016/0167-6393\(87\)90039-2](http://dx.doi.org/10.1016/0167-6393(87)90039-2)
- Nolan, F., McDougall, K., DeJong, G. and Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1): 31--57. <http://dx.doi.org/10.1558/ijssl.v16i1.31>
- Ohala, J., & Ewan, W. (1973). Speed of pitch change. *Journal of the Acoustical Society of America*, 53(1): 345--345. <http://dx.doi.org/10.1121/1.1982441>
- Osanai, T., Tanimoto, M., Kido, H. and Suzuki, T. (1995). Text-dependent speaker verification using isolated word utterances based on dynamic programming [In Japanese]. *National Research Institute for Police Science Report* 48: 15--19.
- Pang, J. L. and Rose, P. (2012). Likelihood ratio-based forensic voice comparison with the Cantonese diphthong /ei/ F-pattern. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*.
- Protopapas, A. and Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* 101(4): 2267--2277. <http://dx.doi.org/10.1121/1.418247>
- R Core Team. (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Version 3.0.0. <<http://www.R-project.org>>.
- Rose, P (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communication* 6: 343--351.
- Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis. <http://dx.doi.org/10.1201/9780203166369>
- Rose, P. and Morrison, G. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law* 16: 139--163. <http://dx.doi.org/10.1558/ijssl.v16i1.139>
- Sereno, J., Lee, H. and Jongman, A. (2015). Effects of speaking rate and context on the production of Mandarin tone. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

- Tabachnick, B. and Fidell, L. (2007). *Using Multivariate Statistics*. Boston: Allyn and Bacon.
- Vance, T. J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetica* 33: 368—392. <http://dx.doi.org/10.1159/000259793>
- Wang, C. Y. and Rose, P. (2012). Likelihood ratio-based forensic voice comparison with Cantonese /i/ F-Pattern and tonal F0. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology (SST 2012)*.
- Wong, P. C. and Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research* 46(2): 413--421. [http://dx.doi.org/10.1044/1092-4388\(2003/034\)](http://dx.doi.org/10.1044/1092-4388(2003/034))
- Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics* Monograph series #17: 1--31.
- Yip, M. (2002). *Tone*. Cambridge: CUP. <http://dx.doi.org/10.1017/CBO9781139164559>