Speaker discrimination: Citation tones vs. coarticulated tones

Ricky K. W. Chan

Speech, Language and Cognition Laboratory, School of English, University of Hong Kong

rickykwc@hku.hk

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2019.06.006.

Correspondence concerning this article should be addressed to Ricky Chan, Speech, Language and Cognition Laboratory, School of English, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: rickykwc@hku.hk

Abstract

The task of forensic voice comparison (FVC) often involves the comparison of a voice in an offender recording with that in a suspect recording, with the aim to assist the investigating authority or the court in determining the identity of the speaker. One of the main goals in FVC research is to identify speech variables that are useful for differentiating speakers. While French and Stevens (2013) stated that connected speech processes (CSPs) vary across speakers and thus CSPs may be included in the 'toolbox' for forensic voice comparison casework, little empirical research has been done to test how effective various CSPs are in speaker discrimination. This paper reports an exploratory study comparing the speaker-discriminatory power of lexical tones in their citation forms and coarticulated tones. 20 Cantonese and 20 Mandarin speakers were instructed to produce tones under different speech rates and tonal contexts. Results based on discriminant analysis show that the combination of normal speech rate and compatible tonal context appears to have yielded the best speaker discrimination. On the other hand, the combination of fast speech and a conflicting tonal context, which in principle led to the greatest tonal coarticulatory effects, yielded the worst speaker discrimination. The addition of duration on top of tonal f0 significantly improved the classification rates in both languages. Furthermore, for the same tone categories, the Mandarin ones generally discriminate speakers better than the Cantonese counterparts, suggesting that tone inventory density affects the speaker-discriminatory power of tones. Implications of the findings for forensic speaker comparison are discussed.

Keywords: Speaker discrimination, coarticulation, tone, Cantonese, Mandarin

1. Introduction

Forensic voice comparison (FVC) typically concerns the comparison of a voice in an offender recording with that in a suspect recording, with the aim to assist the investigating authority or the court in determining the identity of the speaker. The task of FVC by phoneticians often includes auditory and/or acoustic analysis (Gold and French, 2011), which involves decomposing the speech signal into separate variables for analysis and comparison (French and Stevens, 2013). However, for any variable (e.g. vowel formant, speaking rate), different speakers' range will inevitably overlap. In principle, the more variables included, the more likely individual speakers can be discriminated. Therefore, although individual research papers on FVC often focused on only one or a few variables in the speech signal, the maximal speaker-discriminatory power should lie in a combination of variables. An ultimate research goal in the field of FVC is to identify a combination of speaker-specific variables that will enable forensic phoneticians to best discriminate voice samples.

French and Stevens (2013) noted that connected speech processes (CSPs) such as assimilation and elision vary across speakers and thus CSPs may be included in the 'toolbox' for forensic voice comparison casework, but so far there has been very little empirical research on the speaker-specificity of different CSPs and how such specificity may inform decisions in the forensic analysis of voice recordings. This paper reports an exploratory study comparing the speaker-discriminatory power of citation tones and coarticulated tones in Cantonese and Mandarin, and discusses how the findings may be of potential relevance to the task of forensic voice comparison.

1.1 What is lexical tone?

Lexical tones are often defined as distinctive pitch patterns used for contrasting word meaning in a tone language (Bauer and Benedict, 1997). 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 fundamental frequency (f0), which is mainly determined by the rate of vibration of the vocal folds (Bauer and Benedict, 1997). Other acoustic correlates of tone include amplitude envelope (fluctuation in the overall amplitude of a sound) (Fu et al., 1998, Zhou and Martin, 2012), voice quality (e.g. creaky voice for low tones in Mandarin (Belotel-Grenié and Grenié, 2004) and Cantonese (Yu and Lam, 2014)), and duration (Howie, 1976, Yu, 2010).

The tone systems of Hong Kong Cantonese and Standard Mandarin as spoken in Beijing, China, which are the focus of this paper, are illustrated below.

1.2 Cantonese and Mandarin Tone Systems

Hong Kong Cantonese has a relatively complex tone system in which tones are distinguished by both pitch height and pitch direction. Cantonese contrasts six lexical tones: three level tones (high, middle and low), two rising tones (high and low) and a low falling tone (Bauer and Benedict, 1997). Table 1.1 illustrates how the syllable /ji:/ exploits the six tones for lexical contrast. On top of these six tones which occurs in open syllables or nasal-final syllables, the three level tones T1[55], T3[33], and T6[22] also have three allotones— T7[5], T8[3], and T9[2] (often called ‘entering tones’)—which are shorter in duration and occur only in syllables ending with an unreleased stop /p/, /t/, or /k/.

By contrast, Mandarin has a relatively simple tone system in which tones are mainly distinguished by pitch direction. Mandarin contrasts four phonemic tones[1] on full syllables: T1[55] high tone, T2[25] rising tone, T3[214] dipping/low tone, and T4[51] falling tone (Li and Thompson, 1989, Norman, 1988). Table 1.2 illustrates how the syllable /da/ exploits the four tones for lexical contrast. However, T3 is often realised as [21] (i.e. without the rise) except on a monosyllable or when the syllable is emphasised (Duanmu, 2007). The paper will focus only on the phonemic tones in both languages.

Tone	**Example in Cantonese**	**Gloss**	**Phonemic Transcription[2]**
T1 High level	衣	clothing	/ji: 55/
T2 High rising	椅	chair	/ji: 25/
T3 Mid level	意	idea	/ji: 33/
T4 Low falling	疑	suspicious	/ji: 21/
T5 Low rising	耳	ear	/ji: 23/
T6 Low level	二	two	/ji: 22/

Table 1.1: Illustration of the six Cantonese tones.

Tone	**Example in Mandarin**	**English Translation**	**Phonemic Transcription**
T1 High Level	答	to answer	/da 55/
T2 Rising	達	to arrive	/da 25/
T3 Dipping/Low	打	to hit	/da 214/
T4 Falling	大	large	/da 51/

Table 1.2: Illustration of the four Mandarin tones.

[1] Apart from these phonemic tones, Mandarin also has a 'neutral tone' (T0) which can be found on weak syllables and never appears in the initial position of a word. T0 has a mid-pitch target and its acoustic realisation is affected by the preceding tone (Chen & Xu, 2006). The paper will focus on the speaker-discriminating power of the four phonemic tones only.

[2] Tones in Chinese languages are often transcribed using Chao letters (Chao, 1930), in which 1 represents the lowest pitch and 5 the highest in the speaker's normal pitch range. In most cases each tone is numerically represented by two to three digits: the first digit indicates its starting pitch, the final one its

In terms of tonal fundamental frequency (f0), Cantonese has a relatively crowded tonal space: while T1[55] is well separated from other tones, the other five tones cluster in the mid and low tonal f0 range and they can be confusable even for native speakers (Mok et al., 2013). On the other hand, Mandarin has a relatively less dense tone system in which the four tones are well separated from one another (Duanmu, 2007). See Figures 1 and 2 of Francis et al. (2008) for illustrations.

1.3 What is tonal coarticulation?

Like vowels and consonants, the f0 realisation of tones shows gradient variation under the influence of the neighbouring tones in connected speech. This phenomenon is called tonal coarticulation. While segmental coarticulation is essentially assimilatory (e.g. the English phoneme /k/ is fronted (i.e. [k̟]) when followed by a front vowel, and is articulated further back (i.e. [k̠] when followed by a back vowel) (Kühnert & Nolan, 1999), both assimilatory and dissimilatory tonal contextual effects have been reported, and the term 'tonal coarticulation' has been used in the literature to refer to general contextual effects and include both types of contextual effects. Dissimilatory effects have been observed in both progressive and regressive tone coarticulation in Cantonese and Mandarin. An example of progressive dissimilation concerns post-low bouncing, which involves raising f0 after is a low pitch target (Gu & Lee, 2009; Prom-on et al., 2012). An example of regressive dissimilation concerns pre-low raising, where the f0 of a high tone is raised when preceding a low pitch target (Gu & Lee, 2007; Xu, 1999). These two dissimilatory effects have been attributed to the activities of both the intrinsic laryngeal muscles (mainly the cricothyroids) and extrinsic laryngeal muscles (mainly the sternohyoids and the thyrohyoids) (Gu & Lee, 2007; Prom-on et al., 2012). In addition, whilst coarticulation has long been assumed to be biomechanical in nature (i.e. due to

final pitch, and the middle one its mid pitch which is optional. For instance, [33] represents a midlevel tone and [214] a low dipping tone. Whilst Chao letters will be used throughout this paper to represent tonal pitch patterns, it should be noted that Chao letters were designed in perceptual terms with reference to a speaker's tonal pitch range. One should not expect the letters to translate readily into f0 differences or phonological features.

physiological constraints on articulators when executing a motor plan and reaching production targets), recent studies have shown that at least some degree of planning is involved in both segmental coarticulation (Solé, 2007; Whalen, 1990) and tonal coarticulation (Franich, 2015). Thus, both the mind and the vocal folds (and related muscle activities) may contribute to any observed speaker-related variation in tone realisation.

Conceptually, tonal coarticulation should be distinguished from a closely related process—*tone sandhi* (see Chen (2000) for an overview). Both processes involve contextual effects on tonal variation; tone coarticulation concerns gradient, non-neutralising effects which vary across speech styles and speaking rates, whereas tone sandhi concerns categorical, neutralising effects which are stable across speech styles and speaking rates (Zhang & Liu, 2011). A classic example of tone sandhi is the T3 sandhi in Mandarin: a T3 changes to a T2 when followed by another T3 in Mandarin.

1.4 Tonal coarticulation and forensic voice comparison

The task of FVC often involves auditory and/or acoustic phonetic analysis, which exploits the componentiality of the speech signal. Specifically, the speech signal is often conceptualised as consisting of different components (e.g. vowel, consonants, intonation) for separate analysis and comparison (French & Stevens, 2013). However, it should be noted that the maximal speaker-discriminatory power should lie in a combination of variables in the speech signal, even though individual research in FVC often focused on only one or a few variables. Most existing FVC research has focused on English and some other European languages, and the usefulness of tonal variables for FVC remains under-researched. The study of tonal variables for FVC will be beneficial to forensic casework in places where tone languages are used. An important feature of lexical tone is that tones are defined not in absolute terms by the language, but in relative terms mainly with reference to the realisation of other tones and the speaker's pitch range (Bauer & Benedict, 1997). A few early studies provided a description on the between- and within-speaker variability of tones of different languages in their citation forms (e.g. Gandour et al., 1991; Rose, 1996). Studies on perceptual normalisation of lexical tones (e.g. Morre and Jongman, 1997; Leather, 1983; Wong and Diehl, 2003) provide indirect evidence for

speaker-specific realisation of lexical tones and that listeners may use tonal f0 information as a cue to speaker identity. Recently, there has been a few experimental studies investigating the speaker-discriminatory power of lexical tones mainly in their citation forms (Chan, 2016; Li & Rose, 2012; Thaitechawat & Foulkes, 2011; Wang & Rose, 2012) and the potential implications for forensic voice comparison, but so far no study has focused on tones undergoing coarticulation. The study of speaker discriminatory power of coarticulated tones are of particular relevance to forensic casework in that forensic recordings mostly consist of spontaneous speech where lexical tones are likely to coarticulate with one another. At a broader level, while French and Stevens (2013) noted that connected speech processes (CSPs) such as assimilation and elision vary across speakers and thus CPSs may be included in the 'toolbox' for FVC casework, they made no prediction as to whether features undergoing CSPs are better speaker-discriminants than features in their citation forms. Also, there is little empirical research on the speaker-specificity of different CSPs and how such specificity may be relevant to the forensic analysis of voice recordings, let alone comparing the speaker-specificity of speech variables undergoing CSPs with those in their canonical forms.

1.5 Research question

The primary goal of this paper is to compare the speaker-discriminatory power of citation tones and coarticulated tone, and draw potential implications for forensic voice comparison. Xu (1997) noted that when tones are produced in isolation (i.e in their citation forms), their f0 contours appear well-defined and relatively stable. When tones are produced in context, their f0 contours may exhibit a range of possible realisations depending on adjacent tones (see Xu, 2001 for a detailed discussion). It is thus hypothesised that, tones undergoing coarticulation should have greater speaker-discriminatory power since speakers have more freedom to vary in their articulatory 'pathways' and achieve their production targets when producing coarticulated tones than citation tones.

2. Materials and Method

2.1 Participants

Cantonese: 20 native male speakers of Hong Kong Cantonese (aged from 19 to 25, mean = 22.7) were recruited. All of them were undergraduates at University of Cambridge or University of Hong Kong, and had lived in Hong Kong for more than 15 years.

Mandarin: 20 native male speakers of Standard Mandarin (aged from 19 to 25, mean = 21.9) were recruited. All of them were undergraduates at the Communication University of China or Beijing Normal University, and had lived in Beijing for more than 15 years.

2.2 Materials

Trisyllabic meaningful words whose second syllable carries the target tone were used in the experiment. The neighbouring tones served as the tonal context (i.e. x**X**x). Six words were used for each target tone in each language (i.e. 6 tones x 6 = 36 Cantonese words and 4 tones x 6 = 24 Mandarin words; see Table 2.1 and 2.2 for the words used). To control for segmental content, the syllables carrying the target tones contained either /a:/ preceded by different consonants (the use of minimal contrast with /a:/ was not possible when only meaningful words were used), /si:/ or /fu:/. As only real words were used, it is acknowledged that the segmental content of the neighbouring syllables could not be controlled for.

To elicit citation tones and coarticulated tones, the tonal contexts (i.e. the neighbouring tones) either formed a 'compatible context' (i.e. adjacent tones have f0 values identical or similar to the target tone) or a 'conflicting context' (i.e. adjacent tones have f0 values different from the target tone) (Xu, 1994). For example, a compatible context for the Cantonese high level tone /55/ could be /**55**/-/55/-/**55**/, and a conflicting context could be /**21**/-/55/-/**21**/. Half of the words had a compatible tonal context, and the other half had a conflicting tonal context. When there were more than one possible compatible/conflicting context for the target tone, the tones with the closest pitch value with the target tone was selected for the compatible condition, and the ones with the farthest pitch value from the target tone for the conflicting context. For instance, for the high level tone /55/ in Cantonese, a conflicting context could have been /33/_/33/ or /22/_/22/, but only /22/_/22/ was used. One exception was that juxtaposition of two fall-

rise tones in Mandarin was avoided owing to the tone sandhi which would change the first fall-rise tone into a rising tone. The same compatible/conflicting tonal contexts were used across both languages whenever possible to facilitate comparison (e.g. the conflicting contexts for T1[55] and T2[25] in both languages). Table 2.3 summarises the tone patterns carried by the trisyllabic words in Cantonese and Mandarin.

2.3 Recordings

Recordings were made in a sound-treated room with a Zoom H6 portable recorder at a sampling rate of 44.1kHz/16 bits. To minimize potential lexical and frequency effects on tone production, participants were asked to practise the target words until they felt ready for the actual recordings. The speakers first recorded the whole list of words four times, with each word embedded in a carrier sentence *佢未聽過 xxx 呢個詞語* (Cantonese)/ *他沒聽過 xxx 這個詞語* (Mandarin) 'He/She has never heard of the word __'. Then they read the whole list of words in isolation four more times. The target words were presented in a random order for each repetition. To control for the speaking rate of the speakers, regular beats were played through a virtual metronome at an interval of 2 seconds and participants were instructed to produce each word/sentence between two beats. In other words, for each trial, participants had to produce, within two seconds, 11 syllables when the target word is embedded in a carrier sentence (fast speech), and 3 syllables when the target word was in isolation (normal speech). The combination of normal speech and a compatible tonal context encouraged the realisation of tones in their citation forms, whereas the combination of fast speech and a conflicting tonal context were supposed to encourage the highest degree of tonal coarticulation.

Items were randomized and presented one by one on a computer to avoid list effects. In total, there were 5760 tokens for Cantonese (36 words x 2 conditions x 4 repetitions x 20 speakers) and 3840 tokens for Mandarin (24 words x 2 conditions x 4 repetitions x 20 speakers).

Cantonese Stimuli						
	Compatible			Conflicting		
Tone	**/i/**	**/u/**	**/a/**	**/i/**	**/u/**	**/a/**
T1 [55]	金絲貓 Fighting spider /kɐm55 si55 mau55/	歡呼聲 Cheering sound /fun55 fu55 sɪŋ55/	炆花膠 Stewed fish maw /mɐn55 fa55 kau55/	男詩人 Male poet /nam21 si55 jɐn21/	皮膚癌 Skin cancer /pʰei21 fu55 ŋam21/	荷花池 Lotus pond /hɔ21 fa55 t͡sʰi21/
T2 [25]	麵豉湯 Miso soup /min22 si25 tʰɔŋ55/	爛苦瓜 Rotten bitter gourd /lan22 fu25 kʷa55/	玩耍區 Playing area /wun22 sa25 kɵɥ55/	芝士迷 Cheese lover / t͡si55 si25 mɐi22/	甘苦茶 Bitter tea /kɐm55 fu25 t͡sʰa21/	花灑頭 Showerhead /fa55 sa25 tʰɐu21/
T3 [33]	怪嗜好 Weird hobbies /kʷai33 si33 how33/	配賦稅 Apportioned tax /pʰui33 fu33 sɵɥ33/	素炸醬 Vegetarian soy bean sauce /sow33 t͡sa33 t͡sœŋ33 /	天使光 Angel light /tʰin55 si33 kʷɔŋ55/	天賦高 Highly gifted /tin55 fu33 kow55/	轟炸機 Bomber /kʷɐŋ55 t͡sa33 kei55/
T4 [21]	長時期 Long period /t͡sʰœŋ21 si21 kʰei21/	黃芙蓉 Yellow hibiscus /wɔŋ21 fu21 joŋ21 /	矇查查 Blurred /moŋ21 t͡sʰa21 t͡sʰa21/	超時空 Hyperspace /t͡sʰiw55 si21 hoŋ55/	陰符經 Yinfu Jing (Book) /jɐm55 fu21 kɪŋ55/	蔘茶包 Ginseng tea bag /sɐm55 t͡sʰa21 paw55/
T5 [23]	舊市鎮 Old town /kɐw22 si23 t͡sɐn33/	孕婦照 Photo of a pregnant woman /jɐn22 fu23 t͡siw33/	大馬戲 Circus /tɔi22 ma23 hei33/	都市人 City dweller /tou55 si23 jɐn21 /	天婦羅 Tempura /tʰin55 fu23 lɔ21/	班馬牌 Zebra (brand) /pan55 ma23 pʰai21/
T6 [22]	豆豉飯 Soy bean rice /tɐw22 si22 fan22/	豆腐飯 Bean curd rice /tɐw22 fu22 fan22/	仲夏夜 Mid-summer night /t͡soŋ22 ha22 jɛ22/	公事包 Briefcase /koŋ55 si22 pau55/	煎腐衣 Fried bean curd sheet /t͡sin55 fu22 ji55/	天下間 Among the world /t͡sin55 ha22 ka55/

Table 2.1: Cantonese words (with gloss and phonemic transcriptions) used in experiment.

Mandarin Stimuli						
	Compatible			Conflicting		
Tone	**/i/**	**/u/**	**/a/**	**/i/**	**/u/**	**/a/**
T1 **[55]**	黑西裝 Black suit / xei55 ɕi55 tʂuɑŋ55/	乾膚機 Skin dryer /gan55 fu55 tɕi55/	批發商 Wholesaler /pʰi55 fa55 ʂɑŋ55/	小吸管 Little straw /ɕiɑu21 ɕi55 kuan21/	緊膚水 Firming lotion /tɕin21 fu55 ʂuei21/	手發抖 Shivering hands /ʂəu21 fa55 təu21/
T2 **[25]**	不習慣 Not accustomed to /pu51 ɕi25 kuan51/	不服氣 Recalcitrant /pu51 fu25 tɕʰi51/	木筏戰 Raft war /mu51 fa25 tʂan51/	出席表 Attendance list /tʂʰu55 ɕi25 piɑu21/	開服表 Server list /kʰai55 fu25 piɑu21/	交罰款 Paying fine /tɕiɑu55 fa25 kʰuan21/
T3 **[21]**	不洗頭 Not wash one’s hair /pu51 ɕi21 tʰəu25/	抗腐蝕 Anti-corrosion /kʰɑŋ51 fu21 ʂi25/	立法員 Legislator /li51 fa21 ɥyɛn25/	乾洗機 Dry-cleaning machine /kan55 ɕi21 tɕi55/	香腐絲 Bean curd slices /ɕiɑŋ55 fu21 si55/	書法班 Calligraphy class /ʂu55 fa21 pan55/
T4 **[51]**	新系統 New system /ɕin55 ɕi51 tʰuŋ21/	親父母 Parents /tɕʰin55 fu51 mu21 /	曲髮捲 Hair-curling comb /tɕʰy55 fa51 tɕuan21/	腦細胞 Brain cell /nau21 ɕi51 pau55/	老富翁 Rich old man /lau21 fu51 wuŋ55/	理髮廳 Barber shop /li21 fa51 tʰiŋ55/

Table 2.2: Mandarin words (with gloss and phonemic transcriptions) used in experiment.

Cantonese		Mandarin	
Compatible	Conflicting	Compatible	Conflicting
/55/ - **T1[55]** - /55/	/21/ - **T1[55]** - /21/	/55/ - **T1[55]** - /55/	/21/ - **T1[55]** - /21/
/22/ - **T2[25]** - /55/	/55/ - **T2[25]** - /21/	/51/ - **T2[25]** - /51/	/55/ - **T2[25]** - /21/
/33/ - **T3[33]** - /33/	/55/ - **T3[33]** - /55/	/51/ - **T3[21]** - /25/	/55/ - **T3[21]** - /55/
/21/ - **T4[21]** - /21/	/55/ - **T4[21]** - /55/	/55/ - **T4[51]** - /21/	/21/ - **T4[51]** - /55/
/22/ - **T5[23]** - /33/	/55/ - **T5[23]** - /21/		
/22/ - **T6[22]** - /22/	/55/ - **T6[22]** - /55/		

Table 2.3: Tone patterns of the trisyllabic words used. The second syllable (bold) carries the target tone and the first and third syllables form either a compatible or conflicting context.

2.4 Data Extraction

Data were segmented (as illustrated in Figure 2.1) and annotated in *Praat* (Boersma and Weenink, 2014). f0(Hz) values were estimated using the STRAIGHT software package (Kawahara et al. 1998) in VoiceSauce (Shue et al., 2011). f0 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 low tones) were so creaky that f0 values could not be extracted and were excluded from the analysis.

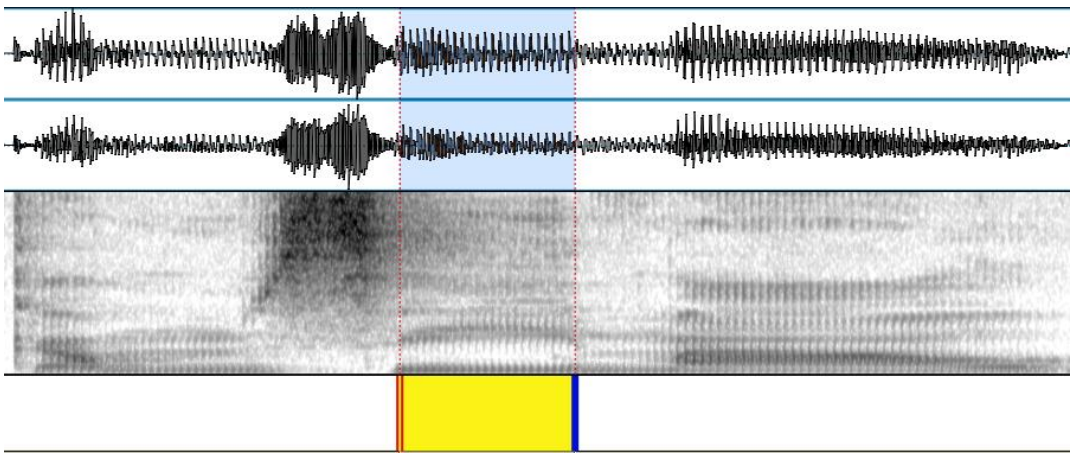


Figure 2.1: illustration of the segmentation of the target vowel in Praat. For each trisyllabic word, two vertical markers were inserted manually from the beginning to the end of the vocalic portion in the second syllable.

3. Results and discussion

3.1 Descriptive statistics

Figures 3.1 and 3.2 show the tonal durational data of the Cantonese and Mandarin speakers respectively, under different speech rates and tonal contexts. Tones appear to be generally longer in fast speech than in normal speech, but tonal context does not seem to have a considerable effect on tone duration. Duration of individual tones alone does not seem to be highly speaker-specific as considerable overlap among speakers can be observed for all the tones.

Figures 3.3 and 3.4 show the boxplots of the Cantonese and Mandarin speakers' f0 data based on their production of all the target tones. While most Cantonese speakers' f0 median values fall in the range of 100-125 Hz, they appear to have considerably different tonal f0 range and distribution. Mandarin speakers, on the other hand, appear to exhibit greater between-speaker differences in the f0 median, range and distribution. While a few speakers' f0 statistics appear to be idiosyncratic (e.g. CC, HF, and KM for Cantonese and CX and LM for Mandarin), f0 statistics does not appear to be highly speaker-specific in general as considerable overlapping in tonal f0 can be observed.

Figures 3.5 and 3.6, which show the distributions of their f0 data, echo the above observations: both Cantonese and Mandarin speakers exhibited considerable differences in terms of overall tonal f0 distribution and range. Still, in general most Cantonese speakers show a right-skewed distribution and some of them also have another smaller peak in the high f0 region. This can be attributed to the fact that most of the Cantonese tones occupy the low-mid, mid and high pitch regions (2, 3 and 5 in terms of Chao letters; see Table 1.1). Similarly, most Mandarin speakers display a bimodal distribution as the four Mandarin tones mostly occupy the low-mid and the high pitch regions (see Table 1.2).

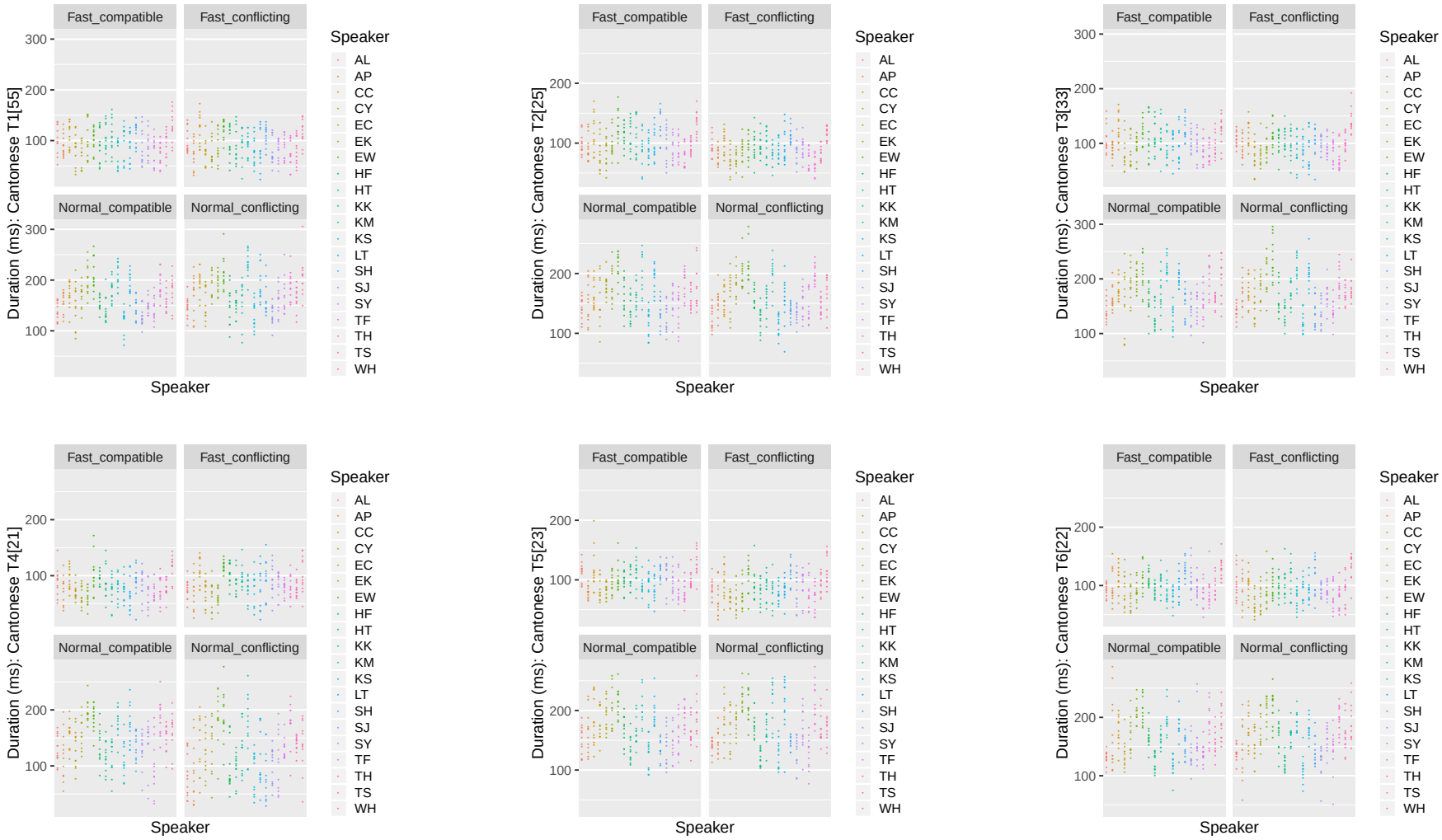


Figure 3.1: duration (ms) of the six Cantonese tones by 20 speakers under different speech rates and tonal contexts.

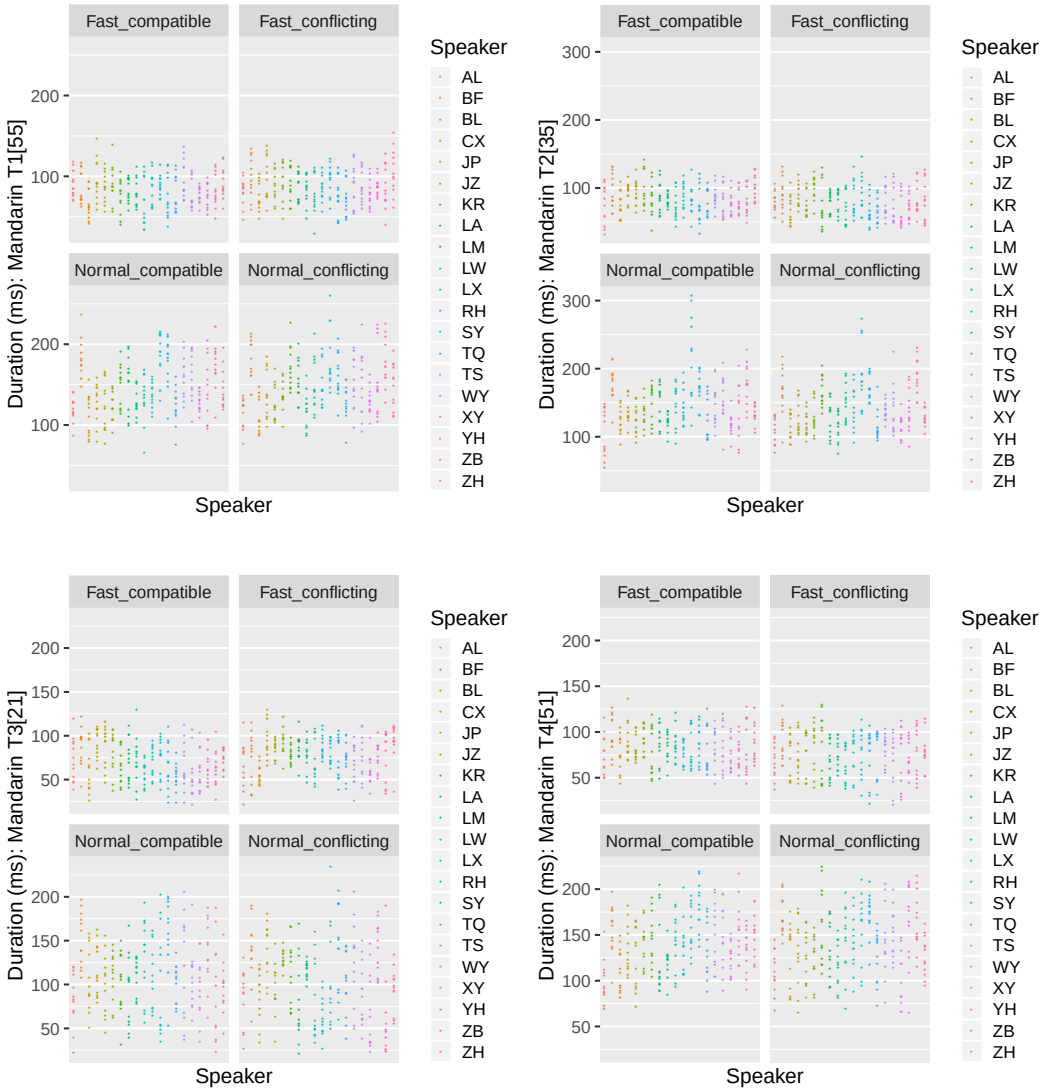


Figure 3.2: duration (ms) of the four Mandarin tones by 20 speakers under different speech rates and tonal contexts.

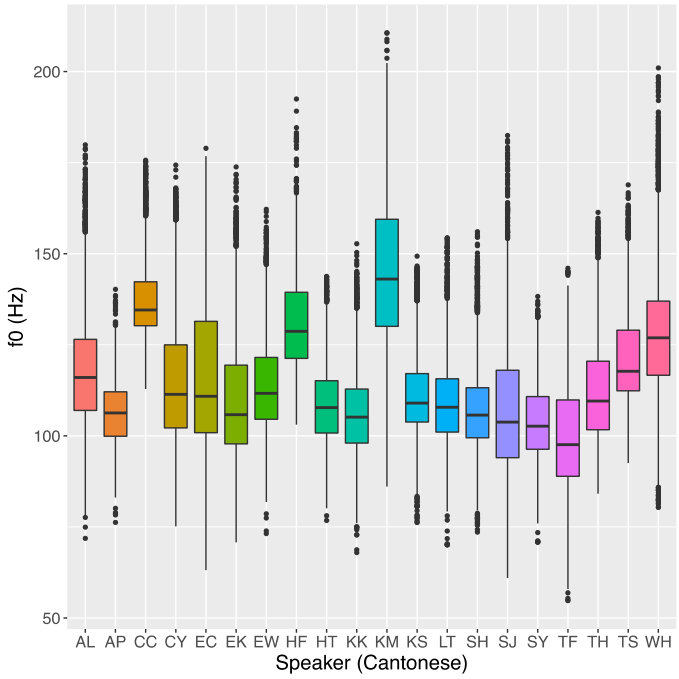


Figure 3.3: Box and whisker plots of the 20 Cantonese speakers' f0 data.

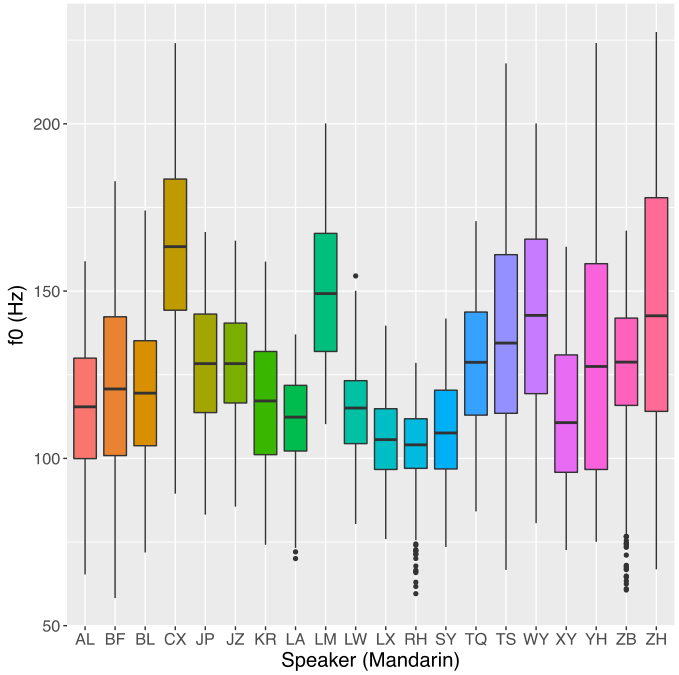


Figure 3.4: Box and whisker plots of the 20 Mandarin speakers' f0 data.

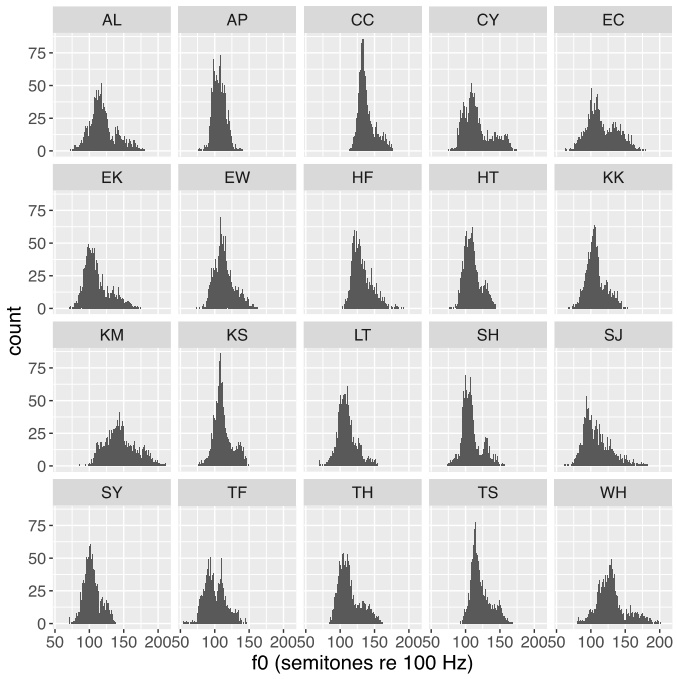


Figure 3.5: Distributions of the 20 Cantonese speakers' f0 data.

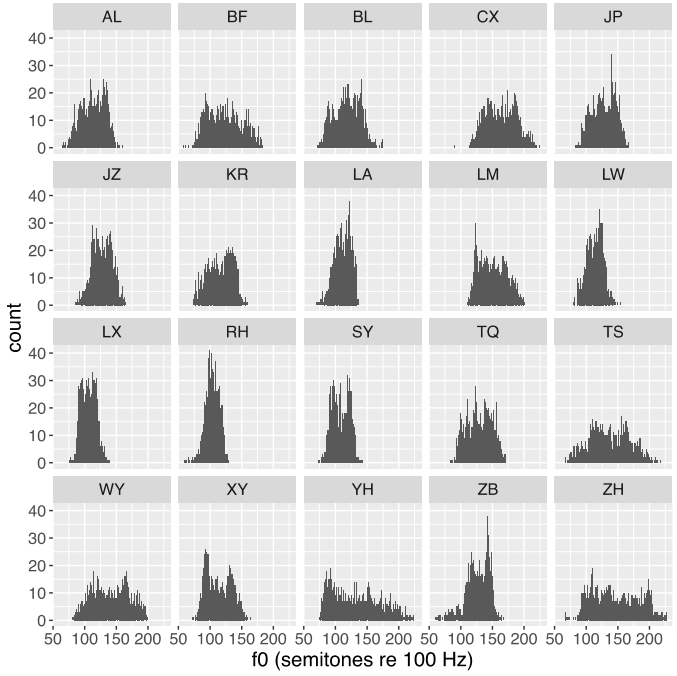


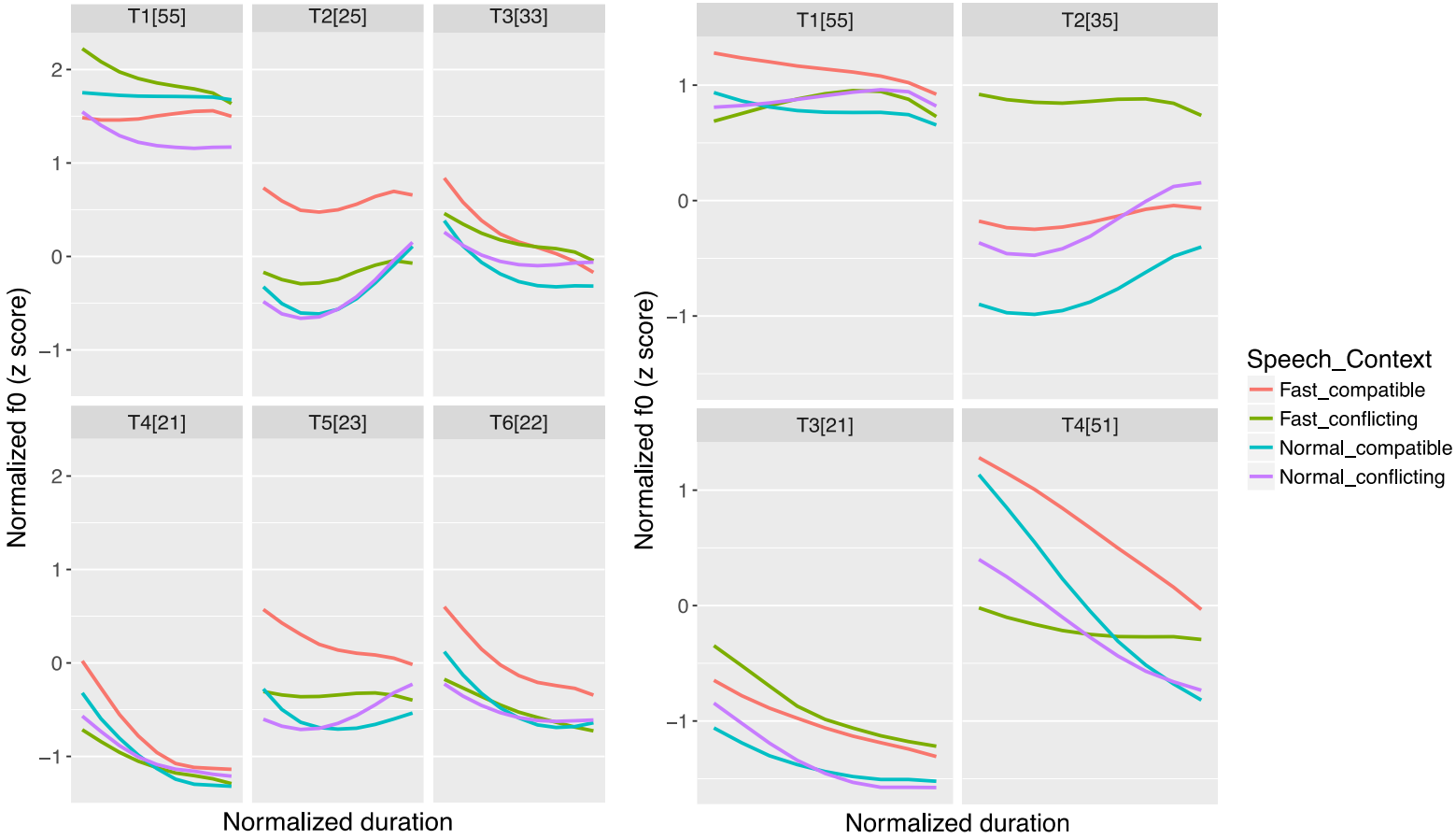
Figure 3.6: Distributions of the 20 Mandarin speakers' f0 data.

It is by no means surprising that speakers differ in their f0 statistics (mean, range, standard deviation, etc.); what is theoretically more interesting lies in whether speakers exhibit idiosyncratic differences in the trajectory of their tone contours. To highlight the between-speaker differences in f0 contours visually, all raw f0 data were normalised on a *z*-score scale (Chan, 2016; Rose, 1987), which involve expressing an observed f0 value in a standard score based on the following formula:

$$f0_{norm} = (f0_i - f0_{mean})/s$$

where $f0_{mean}$ 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 in each language.

Figure 3.7 shows the realisations of the six Cantonese tones under different conditions and tonal contexts, which have observable effects on the shape of the tone contours. Even tones of the same types display different patterns. For the level tones, T1[55] shows a modest fall in a compatible context but rises gradually to the peak in a conflicting context. T3[33] and T6[22] exhibit a small f0 declination in a compatible context but a greater f0 drop in a conflicting context. Speech rate does not appear to affect the general shape of level tones. For the rising tones, T2[25] resembles its canonical citation forms (i.e. shows a small dip and then rises to the peak) in normal speech, but shows a considerably smaller rise and a subtle fall at the end in fast speech. Tonal context has little effects on the overall shape of T2[25]. T5[25], on the other hand, varies drastically: it changes from a canonical low rising tone to a level tone and even a falling tone in the order of normal speech + compatible context → normal speech + conflicting context → fast speech + compatible context → fast speech + conflicting context, revealing the influence of both speaking rate and tonal context. T4[21] displays a consistent falling pattern in the first half of the tone, with a steeper fall in a conflicting context.



Figures 3.7 (left) and 3.8 (right): Mean f0 contours of the six Cantonese tones (left) and the four Mandarin tones (right) by 20 speakers under different speech rates and tonal contexts.

Figure 3.8 shows the Mandarin tones whose trajectories are also affected by tonal context and speaking rate. Overall, T1[55], T2[35] and T3[21] in Mandarin exhibit comparable patterns to Cantonese T1[55], T2[25] and T4[21] respectively. Similar to Cantonese T1[55], the Mandarin T1[55] shows a small declination in a compatible context but rises gradually to the peak in a conflicting context. T3[21] in Mandarin exhibits similar patterns to T4[21] in Cantonese and has a consistent falling pattern in the first half of the tone, with a steeper fall in a conflicting context. T2[25] in Mandarin, just like the Cantonese T2[25], resembles its citation forms in normal speech. However, unlike the Cantonese counterpart, it becomes more like a level tone and even shows a subtle fall at the end in fast speech, especially in the conflicting context. A possible explanation lies in the need to maintain perceptual contrast: Cantonese speakers have to maintain a rising pattern for the T2[25] lest it should be perceived as a mid-level tone or a low rising tone; by contrast, the Mandarin T2[25] is less likely to be confused as another tone in the language even when it becomes more like a level tone due to contextual tonal effect. T4[51] shows a sharp fall in most cases, but becomes more like a level tone in fast speech and a conflicting context. The observations in T2[25] and T4[51] of Mandarin chime with Xu's (1994) findings that the contour of a rising/falling tone can be drastically 'distorted'.

The deviations of the tones from their citation forms observed in Figures 3.5 and 3.6 can be largely attributed to the tonal coarticulatory effects. Progressive assimilation plays an important role in shaping the first half of some of the tones. For instance, the high level tones in both languages are preceded by a low falling tone in a conflicting context, and it takes time for them to rise steadily to the peak (in the second half of the vowel) (except for normal speech + conflicting context in Cantonese), showing progressive assimilatory effects from the preceding low falling tone. Similarly, while the rising tones in both Cantonese (T2[25] and T5[23]) and Mandarin (T2[35]) no longer resemble a rising tone in fast speech, these tones have consistently higher f0 when in a conflicting context, in which the rising tone is preceded by a high level tone T1[55].

Figure 3.9 show the mean f0 contours of each of the Cantonese tones produced in speech rates and tonal contexts by the 20 speakers. Speakers did not respond uniformly to the change in speaking rate and tonal context, and considerable variation across speakers can be observed. For T1[55], most speakers have a steady falling contour and a few speakers show a small rising contour in a compatible context. In a conflicting context,

while a few speakers maintain similar patterns as in a compatible context, most speakers exhibit a small rise-fall contour since it takes time for them to reach the f0 peak. For T2[25], in normal speech most speakers display similar dip-rise contours in a compatible context, but a few speakers already show a much smaller rise when the tonal context becomes conflicting. Speakers exhibit diverse realisation of the rising tone in fast speech and a compatible context: rising, level, and even falling. Most speakers no longer have a clear dip-rise patterns in fast speech and a conflicting context. For T5[25], speakers have similar tone contours in a compatible context—rising in normal speech and level in fast speech—with a few exceptions. But their realisation of the low rising tone varies considerably in a conflicting context, with fewer speakers showing a rising contour in fast speech than in normal speech. For T3[33], T4[21] and T6[22], speakers show similar patterns across conditions and contexts and they differ mainly in the magnitude of fall.

Figures 3.10 shows the mean f0 contours of each of the Mandarin tones produced in different speech rates and contexts by different speakers. Similar to the Cantonese speakers, Mandarin speakers vary in their tone productions in response to different speaking rates and tonal contexts. For T1[55], Mandarin speakers show similar patterns to Cantonese speakers: in a conflicting context most speakers have a small rise-fall tone contour whose peak is in the second half of the tone and a few speakers show a small fall; in a compatible context mostly speakers have a roughly level contour. As for T2[25], while almost all speakers exhibit a dip-rise contour in normal speech, in fast speech more speakers exhibit a level or even small falling contour, increasing from a compatible context to a conflicting context. For T3[21], speakers display consistent patterns across conditions and contexts and they differ mainly in the magnitude of fall in the first half of the tone, but differ considerably in the second half (with a fall, level or even rise). As for T4[51], speakers have similar falling contours in a compatible context, and interestingly in fast speech and a conflicting context (roughly level or even a small rise) as well. On the other hand, their realisation of the tone differs considerably in a conflicting context and normal speech, with speakers doing a steep fall, a small fall, a roughly level and a small fall-rise contour.

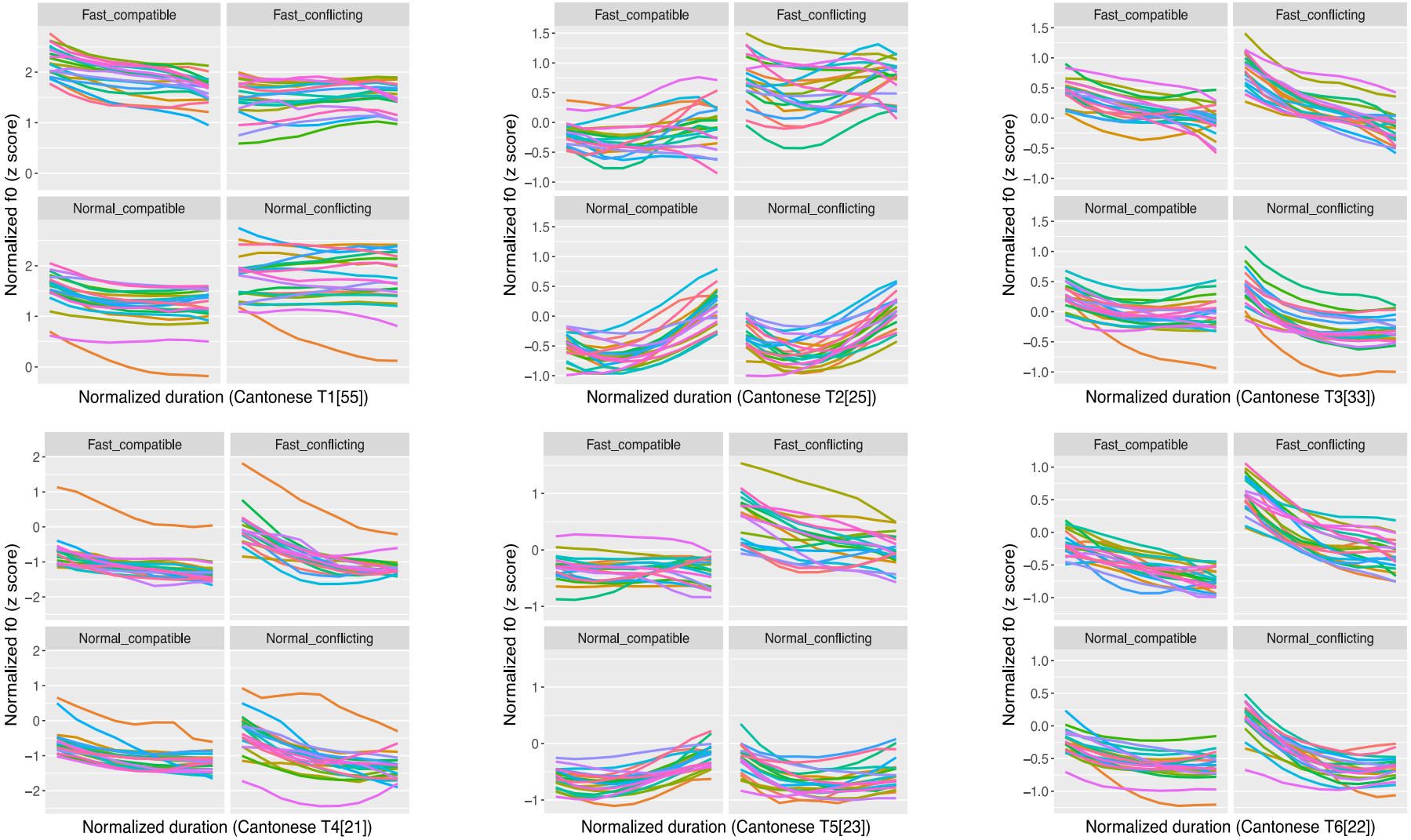


Figure 3.9: mean f0 contours of the six Cantonese tones in different speech rates and tonal contexts by 20 speakers.

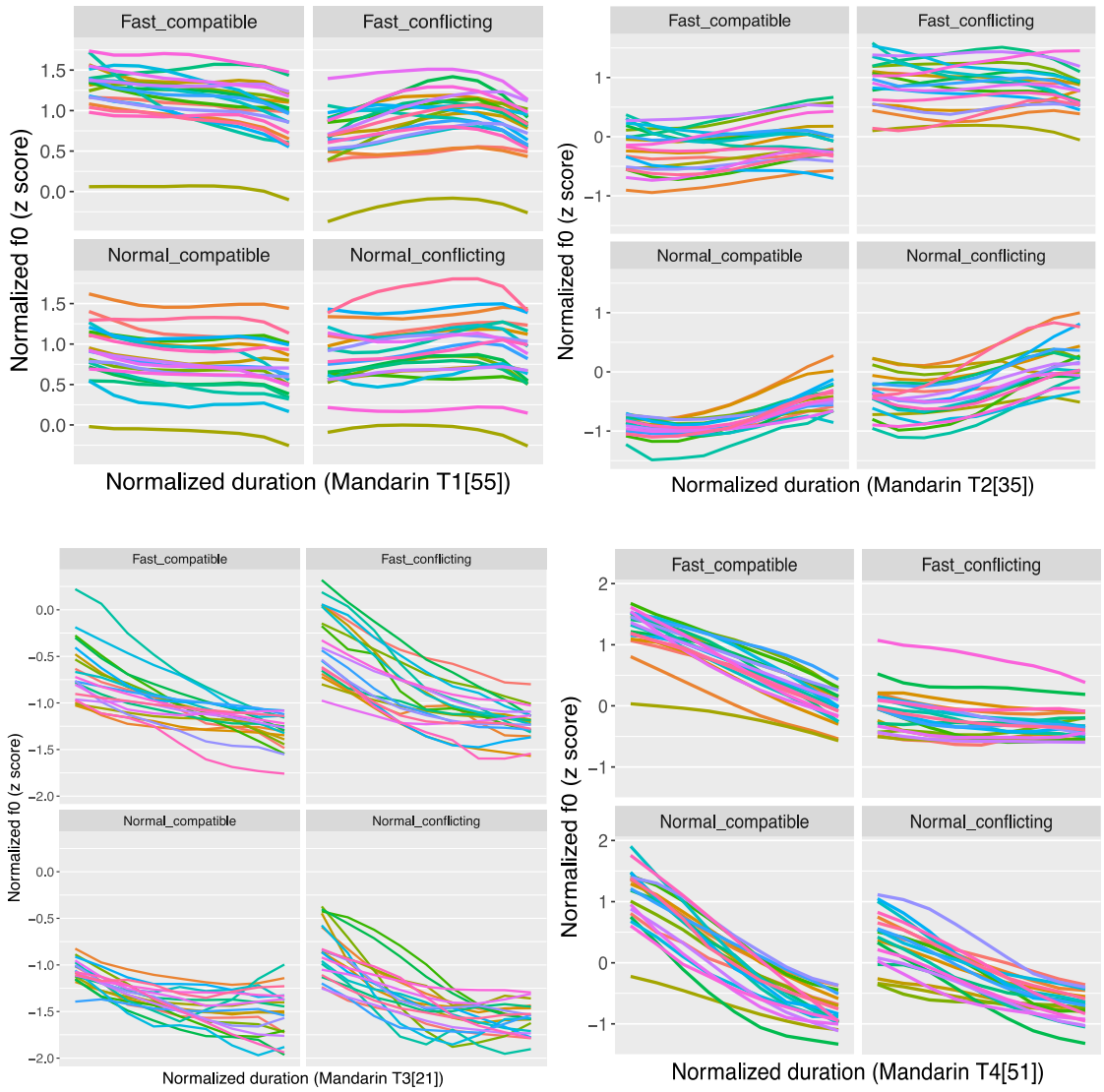


Figure 3.10: mean f0 contours of the four Mandarin tones in different speech rates and tonal contexts by 20 speakers.

There are cases in which a tone is so distorted that its tonal direction is changed (e.g. Cantonese T5[23] and Mandarin T4[51] in fast speech and a conflicting context), but their contours do not resemble any other tones in the language. One might imagine that these cases can be regarded as examples of tone sandhi. Nonetheless, in spite of the fact that the magnitude of the contextual tonal effects was so high than the changes involved are far from gradient, such changes can only be observed in a conflicting tonal context and fast speech. In addition, as discussed below, speakers differ considerably in response to the change in tonal context and speaking rate, suggesting that the observed

changes may not have been phonologised in the language yet. Therefore, these cases should not be treated as cases of tone sandhi.

Discriminant analysis (DA) was used to explore the speaker-discriminatory power of citation tones and coarticulated tones in both languages. As a multivariate statistical technique, DA determines if a given set of predictors can be combined to predict group membership (Tabachnick & Fidell, 2007), and were often used as a statistical tool to evaluate the speaker-specificity of a given feature and its potential usefulness in forensic casework (e.g. Eriksson & Sullivan, 2008; McDougall, 2004, 2006). In the present study, raw (instead of normalised) tonal f0 data of each tone were used as predictors and each speaker as a group. However, as the 9 raw measurement points of each tone were likely to correlate with one another, especially for the adjacent ones, it may be undesirable to use the 9 raw measurement points of each tone directly to serve as predictors. Therefore, the raw tonal f0 contours were modelled with quadratic and cubic polynomials, and the resultant polynomial parameters (3 parameters for quadratic polynomials and 4 parameters for cubic polynomials) were taken as predictors for DA. Tone duration was added as an additional predictor in a separate set of analysis to determine the effect of tonal duration on the speaker-discriminatory power of lexical tones.

DA first constructed discriminant functions that could best separate different groups (speakers) based on the predictors. The discriminant functions were then used to assign the predictors of each tone token to one of the speakers and the accuracy of the classification was computed (classification rate). Classification was cross-validated with the 'leave-one-out' method, which involved leaving each case out in turn when the classification equations are calculated (Tabachnick & Fidell, 2007). This allows testing of the generalizability of the classification involved to new data. With 20 speakers in the data set, the chance performance was 5%.

Separate DAs were run for each tone in different tonal contexts and speech rates. 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.

Figures 3.11 and 3.12 show the DA scores (% correct attribution) of the Cantonese and Mandarin tones under different contexts and speech rates based on quadratic and

cubic polynomials and tone duration. Based on tonal f0 alone, the classification rates range from 23.2 to 42.9 for Cantonese tones (M=32.6; SD=4.81) and 17. 7 to 48.7 for Mandarin tones (M=32.4, SD=8.47), which are much higher than chance level (5% for 20 speakers in each language). This indicates that although a single lexical tone is not sufficient to discriminate 20 speakers from one another, lexical tones carry some degree of speaker-specific information and appear to be good speaker-discriminants. With tone duration as an additional predictor, the classification rates range from 23.7 to 51.7 for Cantonese tones (M=37.8, SD=6.20) and 21.3 to 51.9 for Mandarin tones (M=36.8, SD=8.59). Independent samples t-tests show that the inclusion of tone duration as an additional predictor significantly improves the classification rate in both languages, $t(94)=4.65$, $p<0.0001$, $d=0.95$ for Cantonese tones and $t(62)=2.05$, $p=0.044$, $d=0.52$ for Mandarin tones.

As for the effects of speech rate and tonal context, there does not seem to be any single combination of speech rate and tonal context that yielded the best speaker-discriminatory power for all tones; the effects of speech rate and tonal context appear to be tone-specific. Table 3.1 shows the average DA scores of Cantonese and Mandarin tones under different speech rates and tonal contexts. On average there appears to be a trend in terms of in the order of discriminability with both quadratic and cubic fitting and duration: normal speech + compatible context > normal speech + conflicting context > fast speech + compatible context > fast speech + conflicting context. In other words, the combination of normal speaking rate and a compatible tonal context, which are supposed to have facilitated the realisation of tones, appears to yield the best speaker discrimination in both languages. On the other hand, the combination of fast speech and a conflicting tonal context, which are two major factors contributing to tonal coarticulation, appear to yield the worst DA scores in both languages. This may seem surprising, as one might imagine that speakers will have more freedom to vary in their articulatory 'pathways' and achieve their production targets when producing coarticulated tones than citation tones. Still, the speaker-discriminatory power of tones does not necessary increase with a greater degree of coarticulation. This is partly because speakers may be forced to realise their tone trajectories in a similar way under extreme coarticulation (as shown in Figures 3.9 and 3.10, fast speech + conflicting context). Also, it is speculated that there might also be

greater within-speaker variation for coarticulated tones, potentially due to the difficulty in maintaining consistent articulatory pathways and targets.

	Normal + compatible	Normal + conflicting	Fast + compatible	Fast + conflicting
Cantonese				
Cubic	36.6	34.3	31.1	29.7
Cubic + duration	45.9	40.5	35.0	33.0
Quadratic	34.1	33.6	30.9	30.2
Quadratic + duration	42.6	39.2	34.5	32.3
Mandarin				
Cubic	37.9	38.5	30.3	27.4
Cubic + duration	45.1	41.3	32.4	31.5
Quadratic	36.7	37.5	26.8	24.4
Quadratic + duration	43.6	40.4	30.6	28.8

Table 3.1: average DA scores (% correct attribution) of Cantonese and Mandarin tones under different speech rates (normal vs. fast speech) and tonal contexts (compatible vs. conflicting)

Previous studies have shown that better results were achieved by fitting cubic rather than quadratic polynomial curves to formant trajectories when duration is normalised (McDougall, 2006; Morrison, 2008; Morrison & Kinoshita, 2008). As for the DA results based on curve-fitting of lexical tones with different polynomials, while cubic polynomial appears to have achieved roughly similar or slightly better speaker discrimination than quadratic polynomial for duration-normalised tones, independent-samples t-tests reveal no significant effect of different polynomials on the classification rate, $t(94)=0.86$, $p=0.39$ for Cantonese tones and $t(62)=0.88$, $p=0.38$ for Mandarin tones.

Finally, Cantonese and Mandarin share two schematically similar tone categories: high level and high rising[3]. A comparison of average DA scores between these tone categories in both languages based on tonal f0 data, as shown in Figure 3.13, reveals that for the high level and the high rising tones, the Mandarin ones outperformed the Cantonese counterparts. This is attributable to the difference in tone inventory density:

[3] The Mandarin low/dipping tone T3[214] has been reported to be realised as [21] in connected speech which in schematically similar to the low falling tone T4[21] in Cantonese. However, as shown in Figures 3.7 and 3.8, the Mandarin low tone is realised by a number of speakers with a final rise when produced in normal speech and in a compatible tonal context, where no final rise can be observed for the Cantonese low falling tone. Given such realisational difference, these two tones are not compared here.

Cantonese has a more crowded tone inventory with three level tones and two rising tones, whereas Mandarin has a less dense tone inventory with only one level tone and one rising tone. As such, there is potentially more freedom for speakers to stray in the realisation of the Mandarin level and rising tones while maintaining perceptual contrast. On the other hand, there is potentially less room for the Cantonese high level tone and high rising tone to vary without being perceptually confusable with the mid level tone and low rising tone respectively.

These findings are also in line with the output constraints hypothesis (which was originally proposed by Manuel (1990) to explain cross-linguistic differences in the degree of vowel-to-vowel coarticulation): the number of tonal contrasts in the language poses constraints on the degree of tone coarticulation and thus the speaker-specificity in the production of coarticulated tones. Still, it should be noted that f0 is the primary acoustic cue for tonal contrast in Cantonese and Mandarin and other acoustic correlates only play a secondary role. In languages where more than one articulatory or acoustic dimension is involved for tonal contrast, speakers will have more than one dimension to maintain phonological contrast. For example, Brunelle (2009) studied how the magnitude of tone coarticulation is limited by phonological contrast in northern and southern Vietnamese. Northern Vietnamese uses both pitch and voice quality for tonal contrasts but southern Vietnamese relies exclusively on pitch. Brunelle (2009) found that northern Vietnamese displays greater tonal pitch variation across speakers than southern Vietnamese, suggesting that the number of dimensions for phonological contrasts should be taken into account when testing the output constraints hypothesis. Further research may test the output constraint hypothesis for tone coarticulation with languages that employ more than one primary acoustic parameter for tonal contrast.

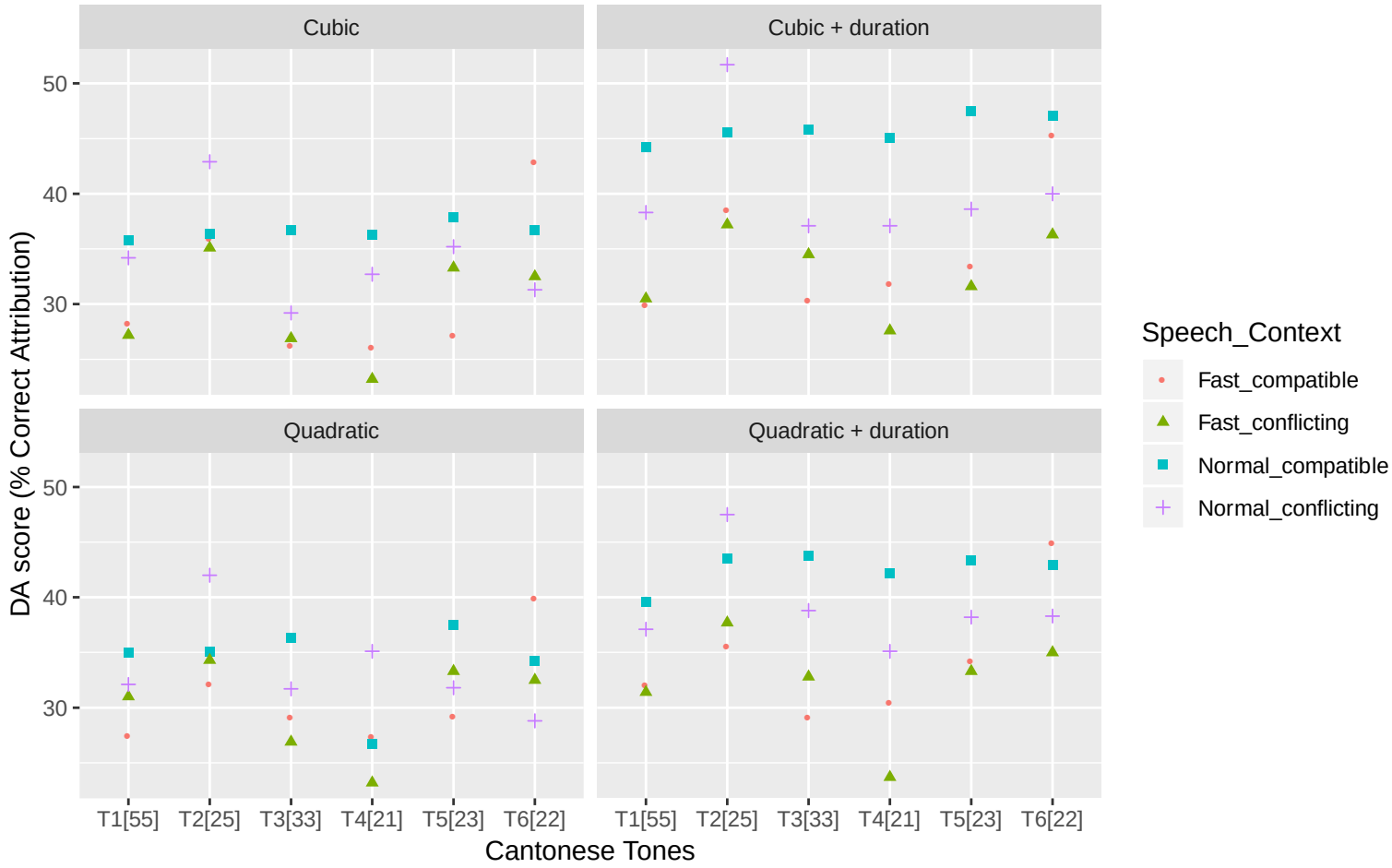


Figure 3.11: DA scores of the Cantonese tones under different speech rates and tonal contexts, based on polynomial parameters and tone duration.

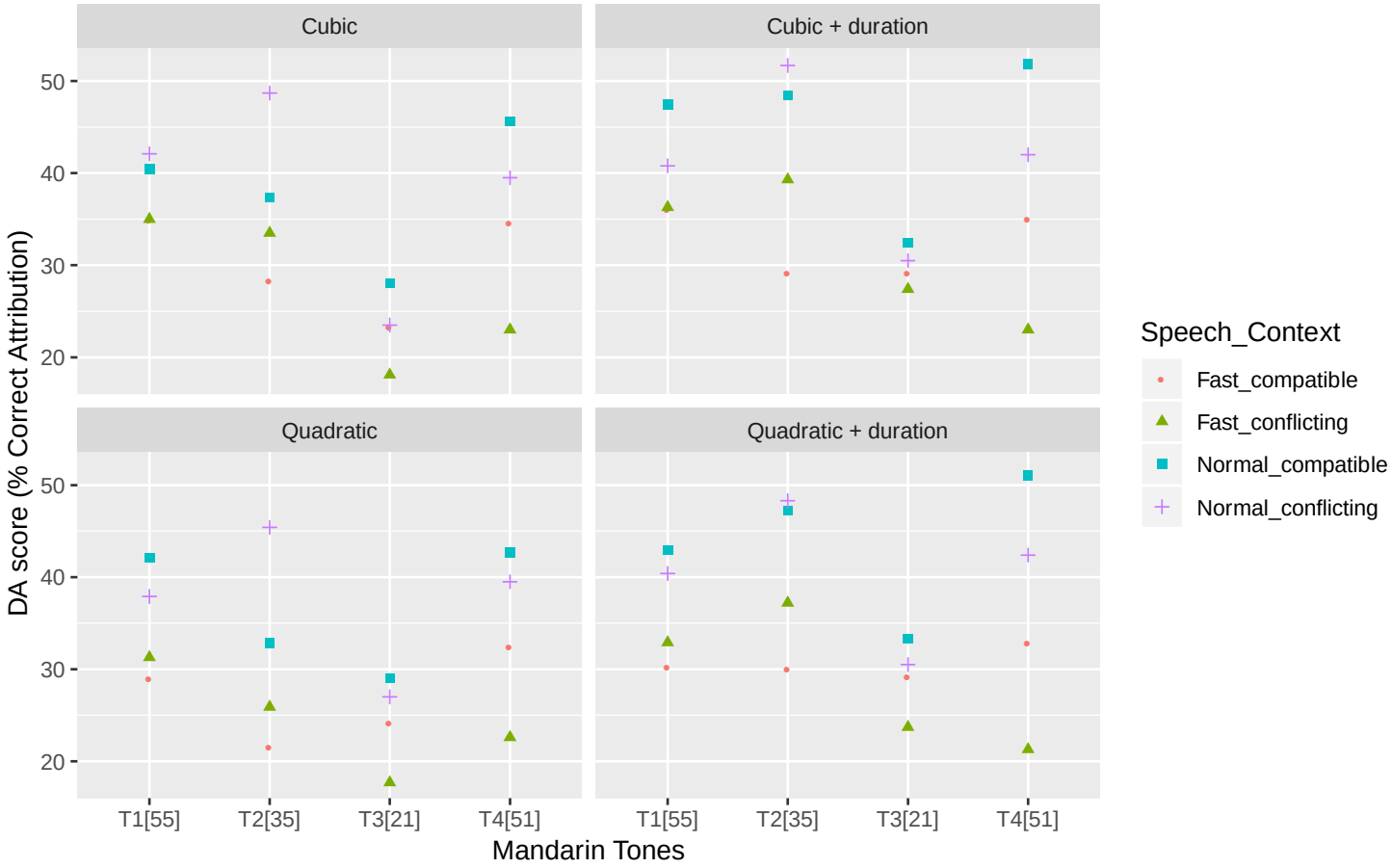


Figure 3.12: DA scores of the Mandarin tones under different speech rates and tonal contexts, based on polynomial parameters and tone duration.

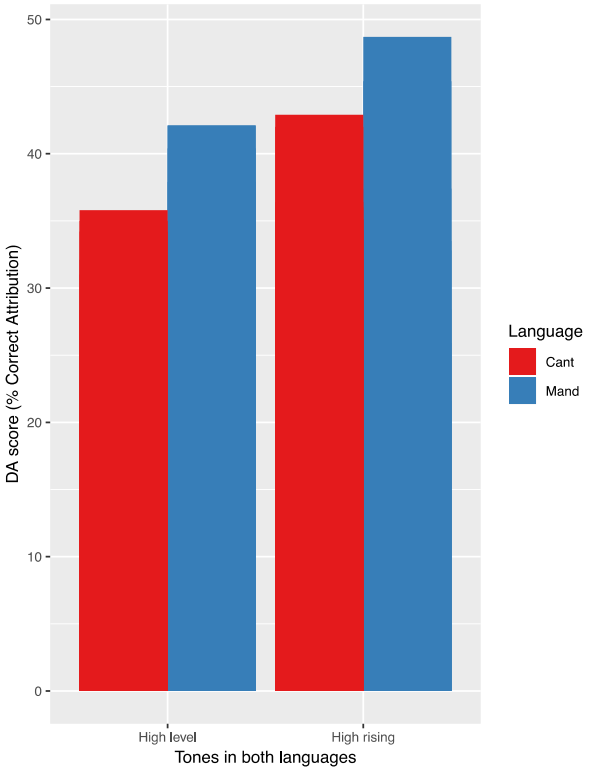


Figure 3.13: comparison of the average DA scores of the high level tone and the high rising tones in Cantonese and Mandarin (based on tonal f0 data).

4. Conclusions

The present study sought to compare the speaker-discriminatory power of lexical tones in their citation forms with those undergoing coarticulation. Two major factors which contribute to tonal coarticulation—speech rate and tonal context—were manipulated to elicit citation tones and coarticulated tones from Cantonese and Mandarin speakers. Contrary to our hypothesis, results show that the combination of normal speech rate and compatible tonal context seems to have yielded the best speaker discrimination. On the other hand, while a faster speech rate or a conflicting tonal context triggered a greater amount of tonal coarticulation, the combination of both fast speech and a conflicting tonal context yielded the worst speaker discrimination. More coarticulation does not necessarily lead to greater speaker-specificity in tone realisation. Moreover, the inclusion of duration as an additional parameter significantly improved the classification rates in both languages. This suggests that both duration and f0 contour should be taken into account in the analysis of tones for speaker discrimination. Finally, the high level tone

and high rising tone in Mandarin outperformed the Cantonese counterparts, potentially attributable to the difference in tone inventory density in the two languages.

The present study modelled tone contours with polynomials, which have been frequently used for fitting dynamic features in the speech signal in the forensic phonetics literature (e.g. McDougall, 2006; Hughes & Foulkes, 2015; Morrison, 2008; Rose, 2017; Rose & Wang, 2016). However, it should be noted that for (tonal) f0 contours, there are a range of computational models available for f0 modelling (e.g. the superposition of functional contours (SFC) model (Bailly and Holm, 2005), the tone transformation model (Ni et al., 2006), the tilt model (Taylor, 2000), the linear alignment model (van Santen and Möbius, 2000), and the quadratic spline model (Hirst & Espesser, 1993)), some of which have functions that represent underlying articulatory mechanisms of tone production (e.g. the quantitative target approximation (qTA) model (Prom-on, Xu & Thipakorn, 2008), the command response (CR) model (Fujisaki, 1983; Fujisaki et al., 2005), and the soft-template model (Kochanski & Shih, 2003)). Future research may explore these models and identify the one(s) that may best serve the purpose of speaker discrimination/voice comparison.

The present study assessed the speaker-discriminatory power of lexical tones using discriminant analysis (DA). Although DA is a useful statistical tool for evaluating the speaker-specificity of a (set of) feature(s) within a group of known speakers, it is not a proper method for assessing the strength of speech evidence in forensic casework. In order for the results to have direct relevance to forensic casework, ideally the likelihood-ratio (LR) framework should be used to provide a gradient assessment of the strength of evidence for the tonal f0 data (see Gold, 2014; Morrison, 2009; Rose and Morrison, 2009 for a detailed discussion). However, while it is possible to carry out LR analysis with the data set in the present study (20 speakers for each language), the sample size may be too small and the results may be misleading. This is supported by Kinoshita & Ishihara (2014) who found that LR analyses of f0 data with a small data set (i.e. $N<30$) may be unreliable. Therefore, the DA results presented above are exploratory in nature; further large-scale studies should compare the strength of evidence between citation tones and coarticulated tones based on the LR framework.

5. References

1. Bailly, G., & Holm, B. (2005). SFC: a trainable prosodic model. *Speech communication*, *46*(3-4), 348-364.
2. Bauer, R., & Benedict, P. (1997). *Modern Cantonese Phonology*. Berlin: Mouton de Gruyter.
3. Belotel-Grenié, A., & Grenié, M. (2004). The creaky voice phonation and the organisation of Chinese discourse. *Proceedings of Tonal Aspects of Language*, 5-8.
4. Boersma, P., Weenink, D., 2014. Praat: Doing Phonetics with Computers. Retrieved from www.praat.org.
5. Brunelle, M. (2009). Northern and southern Vietnamese tone coarticulation: A comparative case study. *Journal of the Southeast Asian Linguistics Society*, *1*, 49-62.
6. Chan, R. (2016). Speaker variability in the realization of lexical tones. *International Journal of Speech, Language and the Law, 23*(2), 195-214.
7. Chen, M. Y. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press.
8. Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford, USA: Oxford University Press.
9. Eriksson, E, J., & Sullivan, K. P. H. (2008). An investigation of the effectiveness of a Swedish glide + vowel segment for speaker discrimination. *International Journal of Speech, Language and the Law*, *15*(1), 51-66.
10. Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268-294.
11. Franich, K. (2015). The effect of cognitive load on tonal coarticulation. *Proceedings of International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow.
12. French, P., & Stevens, L. (2013). Forensic speech science. *The Bloomsbury companion to phonetics*, 183-197.
13. Fu, Q., Zeng, F., Shannon, R., & Soli, S. (1998). Importance of tone envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America*, *104*(1), 505-510.

14. Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech* (pp. 39-55). Springer, New York, NY.
15. Fujisaki, H., Wang, C., Ohno, S., & Gu, W. (2005). Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model. *Speech communication*, *47*(1-2), 59-70.
16. Gandour, J., Potisuk, S., Ponglorpisit, S., & Dechongkit, S. (1991). Inter-and intraspeaker variability in fundamental frequency of Thai tones. *Speech Communication*, *10*(4), 355-372.
17. Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD dissertation. University of York.
18. Gold, E. & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law*, *18*(2), 293-307.
19. Gu, W., & Lee, T. (2007). Effects of tonal context and focus on Cantonese F0. *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrcken, Germany.
20. Gu, W., & Lee, T. (2009). Effects of tone and emphatic focus on F0 contours of Cantonese speech—A comparison with standard Chinese. *Chinese Journal of Phonetics*, *2*, 133-147.
21. Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, *15*, 75–85.
22. Howie, J. (1976). *Acoustical Studies of Mandarin Vowels and Tones.* Cambridge: CUP.
23. Hughes, V., & Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, *66*, 218-230.
24. Kawahara, H., de Cheveign, A., & Patterson, R. D. . (1998). An instantaneous-frequency-based pitch extraction method for high quality speech transformation: Revised TEMPO in the STRAIGHT-suite. *Proceedings of ICSLP'98*, Sydney, Australia.

25. Kinoshita, Y., & Ishihara, S. (2014). Background population: how does it affect LR-based forensic voice comparison?. *International Journal of Speech, Language & the Law*, *21*(2).
26. Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech communication*, *39*(3-4), 311-352.
27. Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, Data and Techniques in Speech Production* (pp. 7-30). Cambridge: CUP.
28. Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*.
29. Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.
30. Li, J. J., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with F-pattern and tonal F0 from the Cantonese /eu/ diphthong. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Macquarie University. Sydney, Australia.
31. Manuel, S. Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, *88*(3), 1286-1298.
32. McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law, 11*, 103-130.
33. 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.
34. Mok, P. P., Zuo, D., & Wong, P. W. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language variation and change*, *25*(03), 341-370.
35. Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*(3), 1864-1877.
36. Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/," *International Journal of Speech, Language and the Law, 15*, 247–264.

37. Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298-308.
38. Morrison, G. S., & Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. *Proceedings of Interspeech 2008 Incorporating SST 2008*, International Speech Communication Association, 1501–1504.
39. Ni, J., Kawai, H., & Hirose, K. (2006). Constrained tone transformation technique for separation and combination of Mandarin tone and intonation. *The Journal of the Acoustical Society of America*, *119*(3), 1764-1782.
40. Norman, J. (1988). *Chinese*. Cambridge: CUP.
41. Prom-on, S., Liu, F., & Xu, Y. (2012). Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America*, *132*(1), 421-432.
42. Prom-On, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, *125*(1), 405-424.
43. Rose, P (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communi*cation, *6*, pp. 343–351.
44. Rose, P. (1996). Between- and within-Speaker variation in the fundamental frequency of Cantonese citation tones. In Davis, P. & Fletcher, N. (eds.). *Vocal Fold Physiology: Controlling Complexity and Chaos*. Singular Press: 307-324.
45. Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: research and reality. *Computer Speech & Language*, *45*, 475-502.
46. Rose, P., & Morrison, G. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law*, *16*(1), 139.
47. Rose, P., & Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Odyssey 2016*, 326-333.
48. Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

49. Solé, M. J. (2007). Controlled and mechanical properties in speech: a review of the literature. In M. J. Solé, P. Beddor & M. Ohala (Eds.), *Experimental Approaches to Phonology*. Oxford: OUP.
50. van Santen, J. P., & Möbius, B. (2000). A quantitative model of F_0 generation and alignment. In *Intonation* (pp. 269-288). Springer, Dordrecht.
51. Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics*. Boston: Allyn and Bacon.
52. Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the acoustical society of America*, *107*(3), 1697-1714.
53. Thaitechawat, S., & Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in Thai. *Proceedings of International Congress of Phonetic Science*s, Hong Kong.
54. Wang, C. Y., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with Cantonese /i/ F-pattern and tonal F0. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, Macquarie University. Sydney, Australia.
55. Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3- 35.
56. Wong, P. C., & 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.
57. Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of Acoustic Society of America*, *95*, 2240-2253.
58. Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61-83.
59. Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics*, monograph series #17. 1-3.
60. Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*(1), 55-105.
61. Yip, M. (2002). *Tone*. Cambridge: CUP.
62. Yu, A. (2010). Tonal effects on perceived vowel duration. *Laboratory phonology*, *10*, 151-168.

63. Yu, K., & Lam, H. W. (2014). The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America*, *136*(3), 1320-1333.
64. Zhang, J., & Liu, J. (2011). Tone sandhi and tonal coarticulation in Tianjin Chinese. *Phonetica*, *68*(3), 161-191.
65. Zhou, Y., & Martin, B. (2012). The role of amplitude envelope in Cantonese lexical tone perception: Implications for cochlear implants. *Proceedings of Speech Prosody*, Shanghai, China.