Modeling lexical tones for speaker discrimination

Ricky K.W. Chan[1], Bruce Xiao Wang[2]

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

[2]Department of English and Communication, Hong Kong Polytechnic University

rickykwc@hku.hk, brucex.wang@polyu.edu.hk

Acknowledgments

The authors would like to acknowledge the financial support by the International Association for Forensic Phonetics and Acoustics (IAFPA) research grants.

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

Modeling lexical tones for speaker discrimination

Abstract: Fundamental frequency (F0) has been widely studied and used in the context of speaker discrimination and forensic voice comparison casework, but most previous studies focused on long-term F0 statistics. Lexical tone, the linguistically structured and dynamic aspects of F0, has received much less research attention. A main methodological issue lies on how tonal F0 should be parameterized for the best speaker discrimination performance. This paper compares the speaker-discriminatory performance of three approaches to lexical tone modelling: discrete cosine transform (DCT), polynomial curve-fitting, and quantitative target approximation (qTA). Results show that using parameters based on DCT and polynomials led to similarly promising performance, whereas those based on qTA generally yielded relatively poor performance. Implications modelling surface tonal F0 and the underlying articulatory processes for speaker discrimination are discussed.

Keywords: speaker discrimination, fundamental frequency, lexical tone, Mandarin, Cantonese

1. Introduction

Speech carries a certain degree of indexical information about a speaker such as gender and regional background, and we can often identify familiar speakers even when we cannot see them (e.g. on the phone) (Nolan, 1999). Whilst it is theoretically interesting to determine the extent to which a person's voice is idiosyncratic, there are specific circumstances where it is important to determine the identity of a speaker solely based on speech. For example, a witness/victim of a crime may have heard but not seen the offender, and be asked to identify the offender from a voice lineup; or an anonymous speech sample related to a crime (e.g. a bomb threat) may have to be compared with the voice of a suspect (Nolan, 1999). With the widespread availability of speech recordings, law enforcement and courts are increasingly relying on specialists who can analyze/provide expert opinions on speech samples during court proceedings or as part of the investigation process. The majority of such work concerns forensic voice comparison (FVC) (French & Stevens, 2013).

FVC typically involves the comparison of two or more speech samples in forensic contexts (e.g., hoax emergency calls, ransom demand, conversation with accomplice), with the goal of assisting the trier-of-fact in determining if the speech samples are from the same person or different people (French & Stevens, 2013). One of the main objectives in FVC research is to identify useful parameters for distinguishing speakers. Fundamental frequency (F0) has been widely researched and used in FVC casework due to its ease of extraction and robustness in channel transmission and poor recording conditions (e.g., Hudson et al., 2007; Kinoshita et al., 2009). Commonly used F0-based parameters are all long-term static parameters such as mean, standard deviation, and median (Gold & French, 2011). However, many forensic experts noted that long-term fundamental frequency (LTF0) parameters are mostly of limited help in forensic casework and used mainly for elimination rather than speaker identification (Gold & French, 2011), primarily due to its notable within-speaker variability as a result of factors such as health, emotional state, Lombard effect, and intoxication (Braun, 1995). Empirical tests using the likelihood ratio (LR) approach also show poor evidential strength of LTF0 parameters (e.g., Kinoshita, 2005; Kinoshita et al., 2009; Rose, 2017; Rose & Zhang, 2018). Nonetheless, few studies have explored the linguistically structured and dynamic aspects of F0, which may carry rich speaker-specific information (Chan, in press; McDougall, 2004).

Lexical tone is a case in point. Lexical tones are often defined as syllable-based pitch patterns for contrasting lexical or grammatical meaning. For instance, in Cantonese, the syllable [ji] means "clothes" when it carries a high-level tone, but "doubt" when it carries a low falling tone. Yip (2002) estimated that as much as 60%–70% of the world's languages are tone languages, and a number of tone languages have millions of native speakers (e.g., Mandarin—921 m, Cantonese—85 m, Vietnamese—76 m; Eberhard et al., 2021). Yet, most existing research on FVC involves non-tonal languages such as English or a few European languages such as German. FVC research on tone languages not only will be relevant to a large number of people in the world but also is necessary for developing a comprehensive theory of speaker idiosyncrasy in speech production.

Although lexical tones function at the lexical level like consonants and vowels, tones are typically regarded as a suprasegmental feature because the primary acoustic correlate of tone is F0, determined mainly by the rate of vibration of the vocal folds (Bauer & Benedict, 1997). However, F0 variation in tone languages cue not only lexical tones but also intonation that may convey discourse, attitudinal and affective information alike, and indexical information such as the speaker's age, gender, regional background, health, and psychological state (Braun, 1995). These kinds of information are transmitted virtually simultaneously. Given the multiplicity of information carried by F0 in tone languages, the extent to which speaker-specific information is encoded in F0 is an empirical question. It is not until the past decade that research has focused on speaker discriminatory power and evidential value of lexical tones (see Chan, in press for a review). For example, based on discriminant analysis of 20 Cantonese male speakers, Chan (2016) found that rising tones generally perform better than other tones at discriminating speakers across speech rates and voice levels. Also, between-speaker differences in tone realization manifest in terms of F0 height and the shape of the tone contours. Chan (2016) found that after normalizing for F0 height, in general around 70% of the discriminatory power of lexical tones was preserved. This shows that the dynamic changes of tonal f0 make a substantial contribution to speaker discrimination. Similarly, using both Cantonese and Mandarin data, Chan (2020) found that tones in their citation forms generally yield better speaker discriminatory performance than tones undergoing coarticulation, and the inclusion of duration as an additional predictor leads to significantly better performance. However, a few studies used the LR framework to assess the evidential strength of tonal F0-based parameters (e.g., Pingjai, 2019; Rose, 2017; Rose & Wang, 2016). For instance, based on spontaneous speech data from Standard Thai young male speakers, Pingjai demonstrated that tonal F0 from the low tone in Thai outperformed LTF0 parameters but the falling tone performed worse than LTF0 parameters.

A key methodological issue lies in how lexical tones should be parameterized to capture maximal speaker-specific information. Existing studies either used a series of instantaneous measurements of the tonal F0 contours directly (e.g., dividing a tone-bearing unit into equidistant intervals and taking F0 measurements accordingly), or fitted polynomial curves (typically quadratic or cubic) with these measurements based on the following general equation:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots a_1 x + a_0 = 0 \ (1)$$

where n refers to the degree (e.g., $n = 2$ for quadratic and $n = 3$ for cubic). $a_n \ldots a_1$ are the coefficients that describe the vowel formant or tonal f0 trajectories, a_0 is the intercept with the y-axis. These coefficients are used for subsequent statistical analysis. It should be noted that there are many other ways to model tonal F0 production that have not been thoroughly explored for speaker characterization. This study is an extension of Chan (2020), who investigated speaker discriminatory power of lexical tones parameterized using quadratic or cubic polynomials. We aim to compare the effectiveness of modeling surface F0 contours (i.e., instantaneous F0 measurements of lexical tones) directly versus modeling underlying articulatory mechanisms of F0 production for speaker discrimination, using discrete cosine transform (DCT) and quantitative target approximation (qTA) as test cases, respectively.

DCT involves estimating time-varying data points using cosine basis function(s) and has been widely used in signal processing (see Rao & Yip, 1990 for details). For tone modeling, DCT involves representing F0 contours with coefficients that are in proportion to the mean, linear slope, and curvature of F0 (Yu et al., 2022). qTA, however, is quantitative implementation of the parallel encoding and target approximation (PENTA) model (Xu, 2005). PENTA assumes that prosody carries multiple levels of communicative functions in parallel and surface F0 is the result of realizing the underlying syllable-based pitch target, which is the ideal F0 contour associated with each syllable and can be static ([high], [low], or [mid]) or dynamic ([rise] or [fall]) (see Figure 1 below).

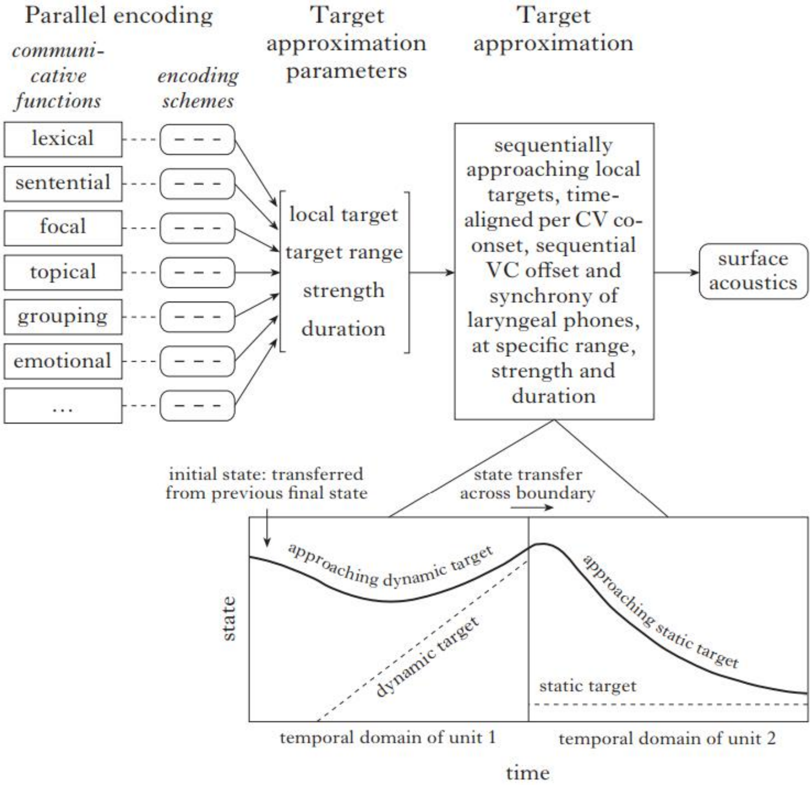

Parallel encoding
communi-cative functions
encoding schemes
Target approximation parameters
Target approximation
lexical
sentential
focal
topical
grouping
emotional
...
- - -
local target
target range
strength
duration
sequentially approaching local targets, time-aligned per CV co-onset, sequential VC offset and synchrony of laryngeal phones, at specific range, strength and duration
surface acoustics
initial state: transferred from previous final state
state transfer across boundary
approaching dynamic target
dynamic target
approaching static target
static target
state
temporal domain of unit 1
temporal domain of unit 2
time

Figure 1: A schematic representation of the PENTA model. The upper panel illustrates how multiple levels of parallel communicative functions are encoded in surface F0 through a number of encoding schemes and target approximation parameters. The lower panel illustrates how each syllable is assumed to have an underlying F0 target (dotted lines), and that the surface F0 contour (solid line) is the result of asymptotic approximation of the target in full synchrony with the syllable (Xu et al., 2022).

In qTA, a pitch target is estimated based on a third-order critically damped linear system as illustrated in the formula below (Prom-on et al., 2009):

$$F0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \quad (2)$$

where *F0(t)* refers to the surface F0 of a syllable as a function of time *t*, and the first term *mt* + *b* the underlying pitch target with *m* representing its slope and *b* representing its height. The second term represents the natural response of the system, with *λ* representing the strength of

F0 movement and three transient coefficients ((c_1, c_2, and c_3) calculated based on the initial F0 level, velocity and acceleration of the current syllable involved (see Xu et al., 2022 for details on the conceptual framework of PENTA and the principles behind the development of qTA). The parameters m, b, and λ can be extracted from speech data automatically based on analysis-by-synthesis machine learning algorithms in PENTA (Prom-on et al., 2009). DCT and qTA have been used to parameterize lexical tones for a range of research questions but not in the context of speaker characterization and discrimination. As an extension of Chan (2020), this short report involves an exploratory study that compares the speaker discriminatory power of the tonal parameters generated from polynomial curve fitting, DCT, and qTA. This is assessed based on the speaker classification results from discriminant analysis as discussed below.

2. Methods

2.1 Databases

We used the same speech corpora reported in Chan (2020), which involve 20 male Hong Kong Cantonese speakers and 20 male Beijing Mandarin speakers aged 19–25 years. The speech data consist of all the Cantonese and Mandarin phoneme tones[1] produced under normal or fast speech rate, and in a compatible or a conflicting adjacent tonal context (i.e., adjacent tones have F0 values either similar to or different from the target tone (Chan, 2020; Xu, 1994)). Normal speech rate and a compatible tonal context generally facilitate the realization of tones in their citation forms, whereas fast speech rate and a conflicting tonal context induce tonal coarticulation. In total, 5,760 tokens of Cantonese tones and 3,840 tokens of Mandarin tones

[1] Cantonese contrasts six phonemic tones: high level[55], mid level[33], low level[22], high rising[25], low rising[23], and low falling[21] (Bauer & Benedict, 1997), and Mandarin distinguishes four phonemic tones: high level[55], rising[35], dipping[214], and falling[51]. See Chan (2020) for details.

were analyzed. Details of the recording procedure and reading materials for eliciting the target tones can be found in Chan (2020).

2.2 Data extraction and parameterisation

Syllables that carry the target tones were manually segmented in *Praat* (Boersma & Weenink, 2024), and two different segmentation approaches were adopted. Before conducting DCT, the vocalic portion of the target syllable was segmented and F0 was estimated using the STRAIGHT package in VoiceSauce (Shue et al., 2011). Eleven measurements were taken for each token; the first and the last measurements were excluded due to potential perturbation effects by neighboring consonants. The remaining nine measurement points then served as the input for zeroth to second and zeroth to third DCT tone modeling, resulting in three or four coefficients for subsequent speaker discrimination. In qTA modeling, for each tone, the entire target syllable was segmented as syllable is argued to be the best target interval for modeling its underlying pitch target (Xu et al., 2022). Three coefficients related to the underlying pitch target—slope m, height b, and strength λ—were extracted using the qTAtrainer[2]. The coefficients generated from these two approaches were used as the predictors for subsequent (linear) discrimination analysis (DA). Discriminant analysis is a multivariate statistical test that determines if a given set of predictors may be used in conjunction to predict group membership (Tabachnick & Fidell, 2014). In the forensic phonetics literature, DA has been used to evaluate the speaker-specificity of a given feature and its potential usefulness in forensic casework (e.g., Chan, 2020; Eriksson & Sullivan, 2008; McDougall, 2004, 2006). In the current context, each speaker is treated as a “group” and parameters generated from tone modeling with DCT/qTA were used as predictors to predict speaker identity (see Tabachnick & Fidell, 2014 for details

[2] Details can be found on http://www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/.

on the mathematical procedure). The results were then compared with those reported in Chan (2020) based on polynomial curve fitting.

DA can be divided into two parts: (1) constructing discriminant functions, and (2) classification. Taking into account both between- and within-speaker variations, DA first uses Wilks' lambda to assess the overall relationship between predictors and groups (speakers); if the relationship is significant, it is concluded that the groups can be distinguished on the basis of the combinations of predictors. Discriminant functions that best separate different speakers based on linear combinations of predictors are then constructed. The classification part evaluates the extent to which group membership can be predicted with the data provided. A classification equation is developed for each group, and every case (i.e., each tone token) in the dataset is allocated to one of the groups (speakers) based on the classification equations. For each case, the data on each predictor are inserted into the classification equation to compute a classification score, and the case is assigned to the group that gives the highest classification score. The percentage of correctly attributed cases (or a classification rate) is calculated and is reported as a DA score. Classification was cross-validated with the "leave-one-out" method, which involved leaving each case out in turn when the classification equations were calculated (Tabachnick & Fidell, 2014). This allows testing of the generalizability of the classification equations to new data as that case is not used in the formulation of any classification equation. With 20 speakers for each language, the chance performance was 5%.

3. Results

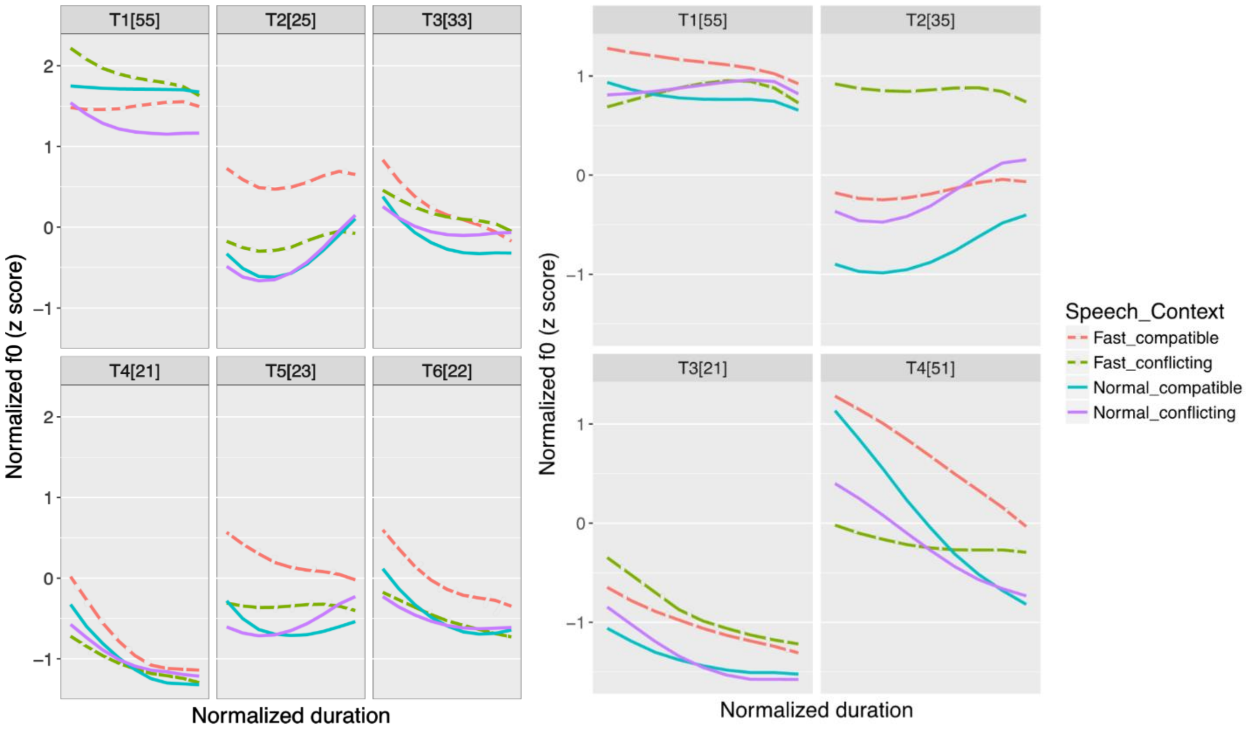
3.1 Descriptive statistics

To highlight the between-speaker differences in f0 contours visually, all raw f0 data were *z* normalized using the following formula (Chan, 2016; Rose, 1987):

$$\mathrm{f0_{norm}} = (\mathrm{f0_i} - \mathrm{f0_{mean}})/s \quad (3)$$

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 2 shows the six Cantonese tones and the four Mandarin tones, respectively, under different speech rates and tonal contexts: fast + compatible, fast + conflicting, normal + compatible, and normal + conflicting, which have observable effects on the shape of the tone contours. For Cantonese tones, even tones of the same types display different patterns. For the level tones, T1[55] shows a modest fall in a conflicting context but rises gradually to the peak in a fast + compatible context. T3[33] and T6[22] generally exhibit f0 declination, especially under a fast speech rate. 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[23], however, 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 2 (left panel) and 2 (right panel): 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 (see Appendices A and B for the production of tones by individual speakers).

As for Mandarin tones, 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[35] 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".

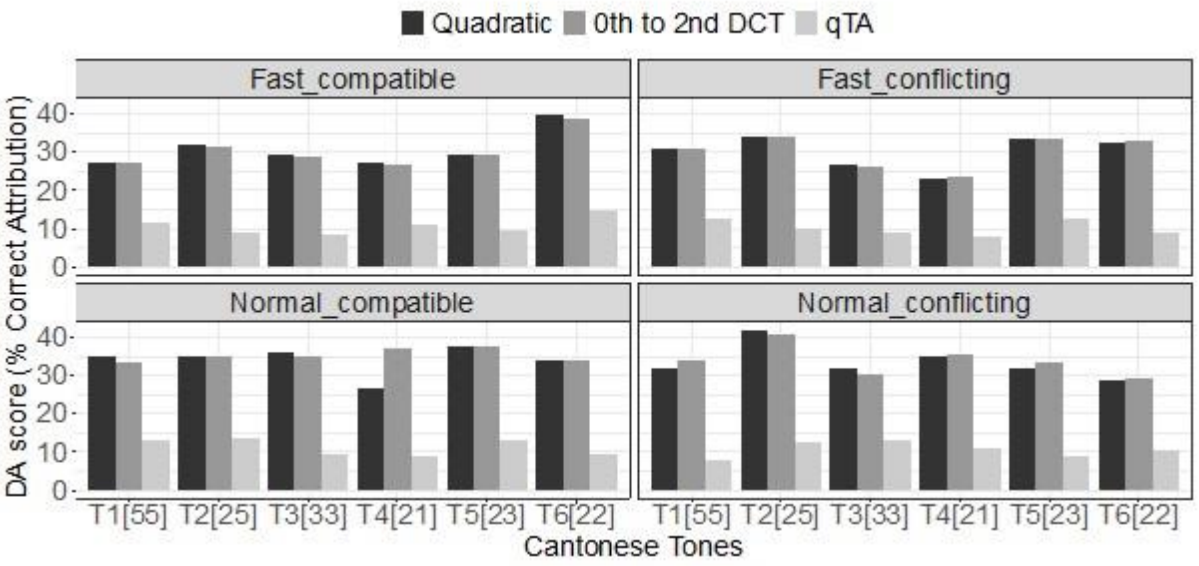
One might imagine that because the realizations of some of these coarticulated Cantonese and Mandarin tones deviate considerably from their citation forms, they might not be perceived as their intended targets. In real-life interactions, semantic information may be used by listeners to unravel the tones of the target words. However, Xu (1994) found that even when Mandarin tones deviated drastically from their canonical forms due to coarticulation and when semantic information was removed, native Mandarin listeners were able to identify coarticulated tones with a high accuracy with the help of the adjacent tonal context. When heavily coarticulated tones were presented without the adjacent tonal context, tone identification dropped below the chance level. However, even with the presence of adjacent tonal context, tone identification accuracy was higher for tones in compatible contexts than for those in conflicting contexts, suggesting that listeners do not always fully compensate for tonal variation due to coarticulation. In this study, we followed the procedure outlined in Xu (1994) when eliciting naturally produced Cantonese and Mandarin tones under different speech rates and tonal contexts. We would expect that if a tone identification task was conducted for the tones reported in this study, the results would be largely similar to those of Xu (1994).

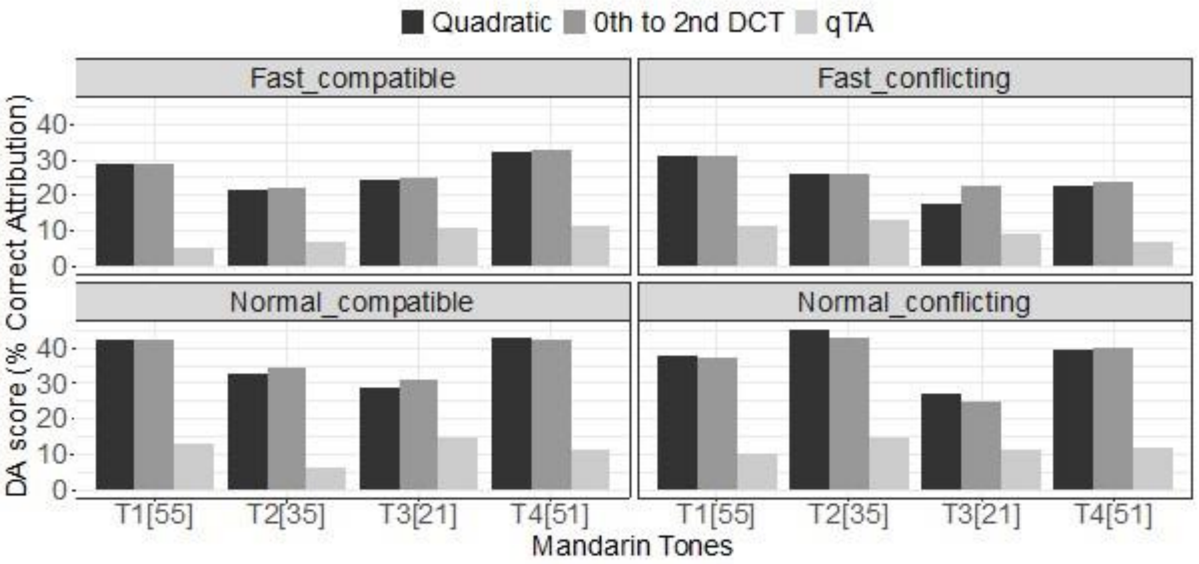
3.2 Discriminant analysis

Figures 3 and 4 show the DA scores (% correct classification) of Cantonese and Mandarin tones based on parameters generated from 0th to 2nd DCT, quadratic polynomial and the qTA model (qTA hereafter), each of which involved three parameters as predictors for DA (chance level = 5% for 20 speakers). In general, parameters based on DCT appeared to yield significantly-higher-than-chance DA scores, ranging from 23.6% to 40.8% (mean = 32.5%, SD = 4.2%, $t(23) = 32$, $p < .001$, $d = 6.5$) for Cantonese tones and 22.0% to 43.1% (mean = 31.8%, SD=7.5%, $t(15) = 14.3$, $p < .001$, $d = 3.6$) for Mandarin tones. Modelling with quadratic polynomial yielded similar results: 23.2% to 42.0% (mean = 32.2%, SD = 4.4%, $t(23) = 30.1$, $p < .001$, $d = 6.1$) for Cantonese tones and 17.7% to 45.4% (mean = 31.3%, SD = 8.3%, $t(15) = 12.7$, $p < .001$, $d = 3.2$) for Mandarin tones. Parameters based on qTA modelling led to relatively poor but still significantly-above-chance performance: 7.9% to 15.0% (mean =10.7%, SD = 2.1%, $t(23) = 13.5$, $p < .001$, $d = 2.7$) for Cantonese tones and 5.0% to 14.6% (mean = 10.5%, SD = 2.9%, $t(15) = 7.7$, $p < .001$, $d = 1.9$) for Mandarin tones. Separate one-way ANOVAs revealed significant overall differences in the DA scores of the three modelling approaches, $F(2, 69) = 270.6$, $p < 0.001$ for Cantonese tones and $F(2, 45) = 53.5$, $p < 0.001$ for Mandarin tones. For both Cantonese tones and Mandarin tones, *post-hoc* Tukey's HSD tests for multiple comparisons (Table 1) revealed no significance difference in DA scores based on quadradic polynomial and 0th to 2nd DCT, but significant differences between quadradic polynomial and qTA and between 0th to 2nd DCT and qTA.

	Cantonese		Mandarin	
Comparison	Estimated difference	Adjusted *p*	Estimated difference	Adjusted *p*
0th to 2nd DCT - qTA	21.81	0.00	21.24	0.00
Quadratic polynomial - qTA	21.54	0.00	20.80	0.00
0th to 2nd DCT – Quadratic polynomial	0.27	0.96	0.43	0.98

Table 1: Post-hoc Tukey pairwise multiple comparisons for the overall DA scores based on the three modelling approaches.





Figures 3 and 4: DA scores (% correct attribution) of Cantonese tones (upper panel) and Mandarin (lower panel) tones under different speech rates (normal vs. fast speech) and tonal contexts (compatible vs. conflicting), based on parameters generated from 0th to 2nd DCT, quadratic polynomial and qTA (chance level = 5%).

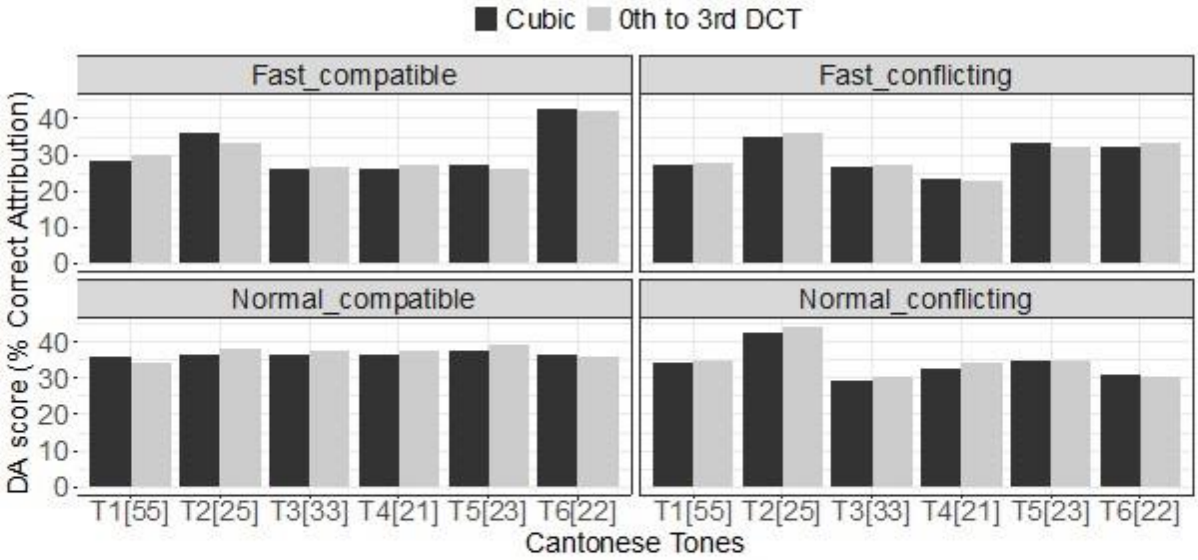
Figures 5 and 6 show the DA scores of Cantonese and Mandarin tones based on parameters generated from 0th to 3rd DCT and cubic polynomial, each of which involved four parameters as predictors for DA (qTA modelling was not involved in the comparison here as it only yielded three parameters). Overall, DA scores based on zeroth to third DCT coefficients are significantly above chance, ranging from 22.7% to 44.5% (mean = 33.3%, SD = 5.3%, $t(23) = 25.9$, $p < .001$, $d = 5.3$) for Cantonese tones and 23.6% to 43.5% (mean = 33.2%, SD = 7.3%, $t(15) = 15.4$, $p < .001$, $d = 3.9$) for Mandarin tones. Cubic polynomials yielded similar results: 23.2% to 42.9% (mean = 32.9%, SD = 5.2%, $t(23) = 26.1$, $p < .001$, $d = 5.3$) for Cantonese tones and 18.1% to 48.7% (mean = 33.5%, SD = 8.8%, $t(15) = 12.9$, $p < .001$, $d = 3.2$) for Mandarin tones. The DA scores based on zeroth to third DCT are not significantly different from those based on cubic polynomial, $t(23) = -0.22$, $p = 0.828$ for Cantonese tones and $t(15) = 0.099$, $p = 0.922$ for Mandarin tones.

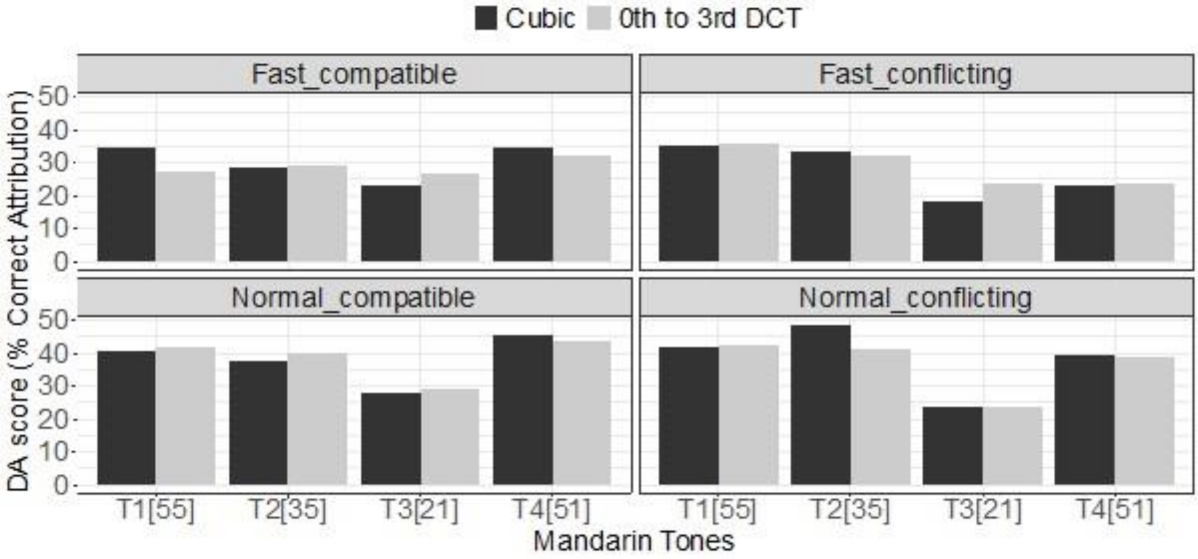
Comparing with the results above, independent sample t-tests showed no significant difference in overall DA scores based on 0th to 3rd DCT vs. 0th to 2nd DCT, $t(46) = -0.559$, $p = 0.580$ for Cantonese tones and $t(30) = -0.559$, $p = 0.581$ for Mandarin tones. Similarly, there is no significant difference in overall DA scores based quadratic vs. cubic polynomials, $t(46) = 0.512$, $p = 0.611$ for Cantonese tones and $t(30) = 0.721$, $p = 0.477$ for Mandarin tones. This suggests that having one extra predictor in DCT or polynomial curve-fitting does not improve speaker-discriminatory performance.

Separate one-way ANOVAs revealed that in general speech rates and tonal contexts (i.e. normal_compatible, normal_conflicting, fast_compatible, and fast_conflicting) do not have any significant effect on the DA scores of Cantonese tones, $F(3, 116) = 1.61$ $p = 0.192$ but have significant effect on the DA scores of Mandarin tones, $F(3, 76) = 4.12$, $p = 0.0092$. However, a *post-hoc* Tukey's HSD Test for multiple comparisons found no significant difference for any of the pairwise comparisons as shown in Table 2.

Comparison	**Estimated difference**	**Adjusted *p***
normal_compatible - normal_conflicting	0.24	1.00
normal_compatible - fast_compatible	8.00	0.094
normal_compatible - fast_conflicting	9.03	0.046
normal_conflicting - fast_compatible	7.76	0.11
normal_conflicting - fast_conflicting	8.79	0.054
fast_compatible - fast_conflicting	1.03	0.99

Table 2: Post-hoc Tukey pairwise multiple comparisons for the DA scores based on different speech rates and tonal contexts.





Figures 5 and 6: DA scores of Cantonese tones (upper panel) and Mandarin (lower panel) tones under different speech rates and tonal contexts, based on parameters generated from 0th to 3rd DCT and quadratic polynomial (chance level = 5%).

4. Discussion

In this study, we compared the speaker discriminatory performance of three approaches with lexical tone parameterization: polynomial curve fitting, DCT, and the extraction of underlying pitch target parameters using qTAtrainer. Results show that, broadly speaking, speech rates and tonal contexts do not have any significant effect on the correct classification rate (DA scores) on Cantonese or Mandarin tones. DA scores based on all three modeling approaches are significantly higher than the chance level, suggesting that some degree of speaker idiosyncratic information can be captured by these approaches. Using the same number of predictors, polynomial curve fitting and DCT did not yield significantly different results, revealing that both approaches capture similar amount of speaker-specific information based on surface tonal F0 contours. This is not surprising, as the fitting points of both DCT and polynomial curves, generated with a cosine function and a sine function, respectively, are related to the mean, slope, and curvature of tonal F0 contours. Their mean discrimination rates are about 33%, which might appear not ideal for a closed set of 20 potential speakers. However, it should be noted that only one speech feature (i.e., lexical tone) was analyzed in this study, and one can analyze a combination of speech features for maximal speaker discriminatory performance, especially in forensic situations (French & Stevens, 2013). Future research may explore how lexical tones may be combined with other features for better speaker discrimination.

However, parameterizing tonal F0 using qTA generally led to poor speaker discriminatory performance, with some of the DA scores close to the chance level, suggesting that this modeling approach captures limited information for separating speakers. A possible reason is that the underlying pitch targets which encode communicative functions in a tone language may be shared among speakers in the same speech community, leaving little room for encoding speaker-specific information. Speaker idiosyncrasy may be manifested in the implementation of these underlying pitch targets, which may be observable in the surface realization of tonal F0.

Overall, our findings point to the conclusion that modeling surface tonal F0 leads to better speaker discriminatory performance than approximating underlying articulatory mechanisms of F0 production. However, it should be noted that there are other computational models that parameterize lexical tones based on their surface F0 contours (e.g., the tilt model (Taylor, 2000), the superposition of functional contours (Bailly & Holm, 2005), the linear alignment model (van Santen & Möbius, 2000), the quadratic spline model (Hirst & Espesser, 1993), and the tone transformation model (Ni et al., 2006)), and other models that simulate the process of F0 production (e.g., the soft-template model (Kochanski & Shih, 2003) and the command response model (Fujisaki, 1983; Fujisaki et al., 2005)). Future research should test the relative effectiveness of different tonal parameterization approaches for maximal speaker discrimination based on lexical tones.

In this study, DA was used to evaluate the speaker discriminatory power of tonal F0-based parameters. While DA is a useful statistical method for this purpose, it should be noted that DA is not a proper method for evaluating the evidential strength of speech features in forensic contexts. In order for the results to be directly relevant to forensic contexts, the LR framework, which evaluates both the similarity and the typicality of evidence, should be used to assess the evidential strength of tonal F0 data (see Morrison et al., 2021 for a discussion). Also, although a total of 5,760 tokens of Cantonese tones and 3,840 tokens of Mandarin tones were analyzed, only 20 male speakers were involved per language. Future studies should ideally involve a larger number of speakers (at least 60–90 speakers for LR-based studies; e.g., Hughes, 2017; Kinoshita & Ishihara, 2014), and involve both male and female speakers. Forensically relevant speech styles (e.g., police interview) and recording conditions (e.g., telephone recordings) should also be involved (see Morrison et al., 2012 for a discussion). This study should be treated as a controlled study on testing various modeling approaches under different speech rates and tonal contexts, and our findings lay the groundwork for large-scale studies using forensically relevant databases on tone languages in the future.

References

Bailly, G., & Holm, B. (2005). SFC: A trainable prosodic model. *Speech Communication*, *46*(3–4), 348–364. https://doi.org/10.1016/j.specom.2005.04.008

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese Phonology*. Berlin: Mouton De Gruyter.

Braun, A. (1995). Fundamental frequency: how speaker-specific is it?. *Beiträge zur Phonetik und Linguistik, 64,* 9-23.

Boersma, P. & Weenink, D. (2024). Praat: doing phonetics by computer [Computer program] (Version 6.4.10). http://www.praat.org.

Chan, R. (2016). Speaker variability in the realization of lexical tones. *International Journal of Speech, Language and the Law, 23*(2), 195-214.

Chan, R. (2020). Speaker discrimination: citation tones vs. coarticulated tones. *Speech Communication*, *117,* 38-50.

Chan, R. (in press). Tone languages. In F. Nolan, K. McDougall & T. Hudson (Eds), *Oxford Handbook of Forensic Phonetics*. Oxford University Press.

Eberhard, M., Simons, G. & Fennig, C. (2021). *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

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.

French, P., Stevens, L., Jones, M., & Knight, R. A. (2013). Forensic speech science. *Bloomsbury Companion to Phonetics*, 183-197.

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.

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. https://doi.org/10.1016/j.specom.2005.06.009

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, *18*(2). https://doi.org/10.1558/ijsll.v18i2.293

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.

Hudson, T., de Jong, G., McDougall, K. & Nolan, F. (2007). f0 statistics for 100 young male speakers of standard Southern British English. In Trouvain, J. & Barry, W. J. (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrucken, Germany, pp. 1809–1812.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough?. *Speech Communication*, *94*, 15-29.

Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term F0. *International Journal of Speech, Language and the Law*, *12*(2), 235–254. https://doi.org/10.1558/sll.2005.12.2.235

Kinoshita, Y., & Ishihara, S. (2014). Background population: how does it affect LR based forensic voice comparison?. *The International Journal of Speech, Language and the Law*, *21*(2), 191-224.

Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech Language and the Law, 16*(1). https://doi.org/10.1558/ijsll.v16i1.91

Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3–4), 311–352. https://doi.org/10.1016/s0167-6393(02)00047-x

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law, 11*(1), 103–130. https://doi.org/10.1558/sll.2004.11.1.103

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech Language and the Law*, *13*(1), 89–126. https://doi.org/10.1558/ijsll.v13i1.89

Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*(2), 155-167.

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., ... & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 299-309.

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. https://doi.org/10.1121/1.2165071

Nolan, F. (1999). Speaker identification and forensic phonetics. In Hardcastle, W. J. & Laver, J. (Eds.), *Handbook of phonetic sciences*. Oxford: Blackwell.

Pingjai, S. (2019). *A Likelihood-Ratio Based Forensic Voice Comparison in Standard Thai*. PhD Thesis. Australian National University.

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. https://doi.org/10.1121/1.3037222

Rao, K.R., & Yip, P.C. (1990). *Discrete Cosine Transform - Algorithms, Advantages, Applications*. Elsevier Science.

Rose, P (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communication, 6*, 343–351.

Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: research and reality. *Computer Speech & Language*, *45*, 475–502. https://doi.org/10.1016/j.csl.2017.03.003

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.

Rose, P., & Zhang, C. (2018). Conversational style mismatch: its effect on the evidential strength of long-term F0 in forensic voice comparison. *Proceedings of ASSTA*, 157-160.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

Tabachnick, B., & Fidell, L. (2014). *Using Multivariate Statistics* (6th ed.). Boston: Allyn and Bacon.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *The Journal of the Acoustical Society of America*, *107*(3), 1697–1714. https://doi.org/10.1121/1.428453

van Santen, J. P., & Möbius, B. (2000). A quantitative model of F0 generation and alignment. In *Intonation* (pp. 269-288). Springer, Dordrecht.

Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, *95*(4), 2240–2253. https://doi.org/10.1121/1.408684

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3-4), 220-251.

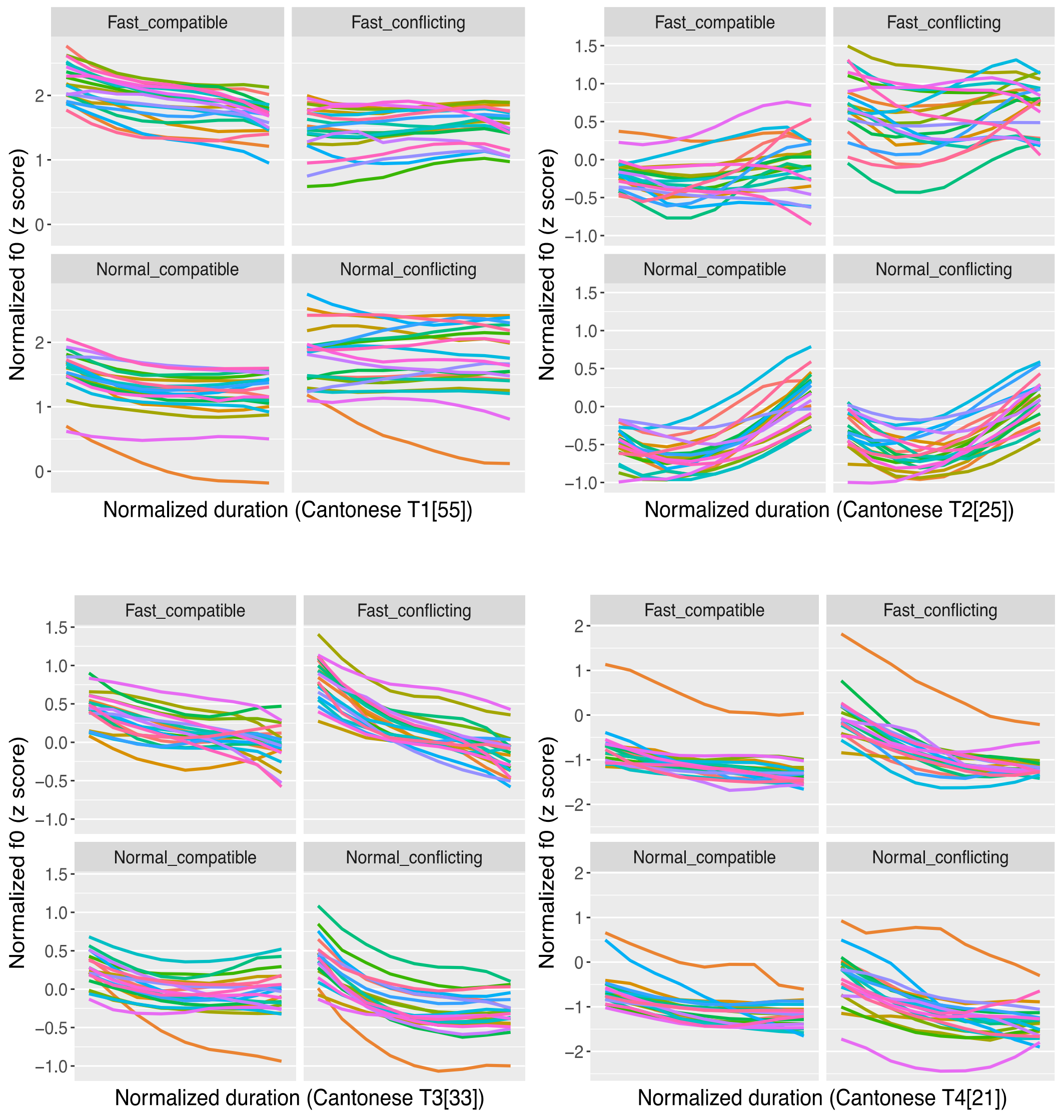
Xu, Y. (2022). The PENTA model: Concepts, use and implications. In *Prosodic Theory and Practice*. S. Shattuck-Hufnagel and J. Barnes (eds.). Cambridge: The MIT Press. pp. 377-407

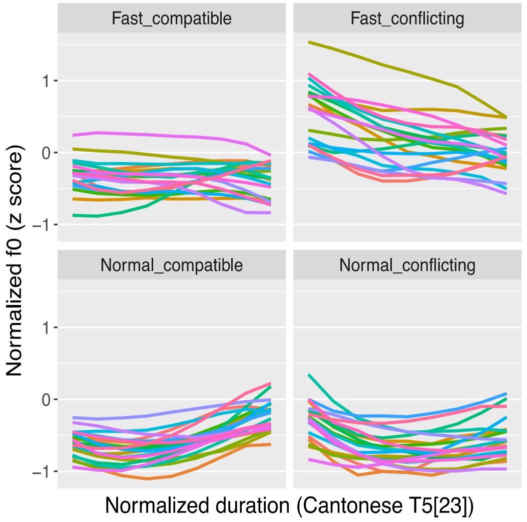
Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

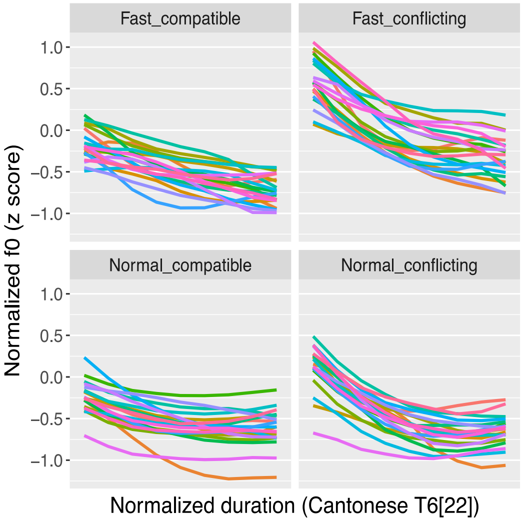
Yu, A. C., Lee, C. W., Lan, C., & Mok, P. P. (2022). A new system of Cantonese tones? Tone perception and production in Hong Kong South Asian Cantonese. *Language and speech*, *65*(3), 625-649.

Appendix A

Mean f0 contours of the six Cantonese tones in different speech rates and tonal contexts by 20 speakers.

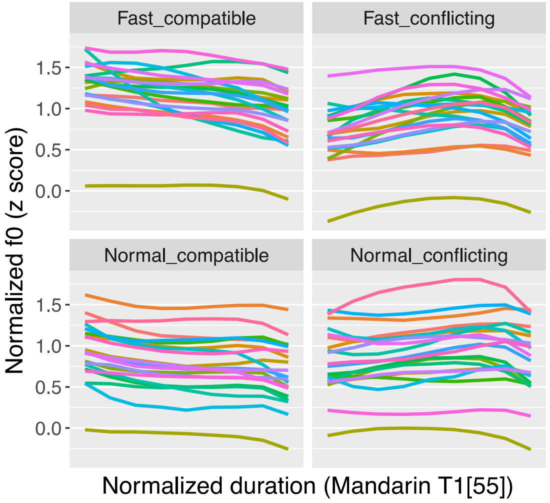


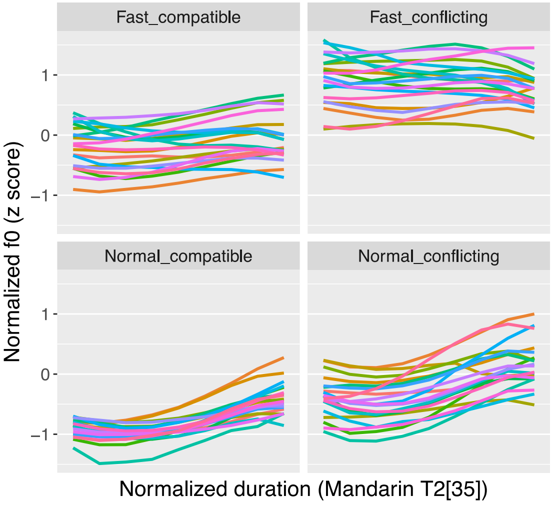


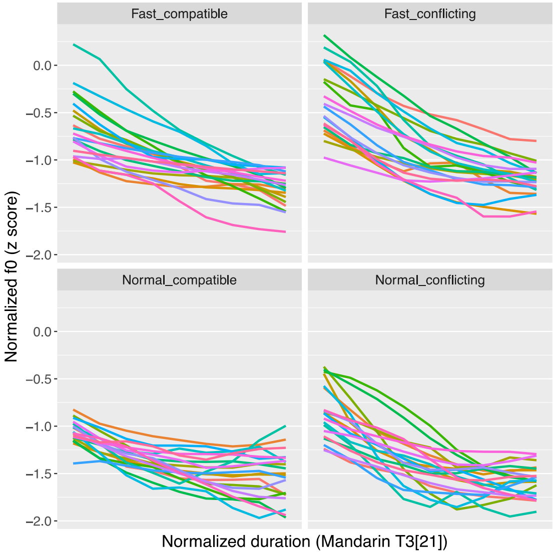


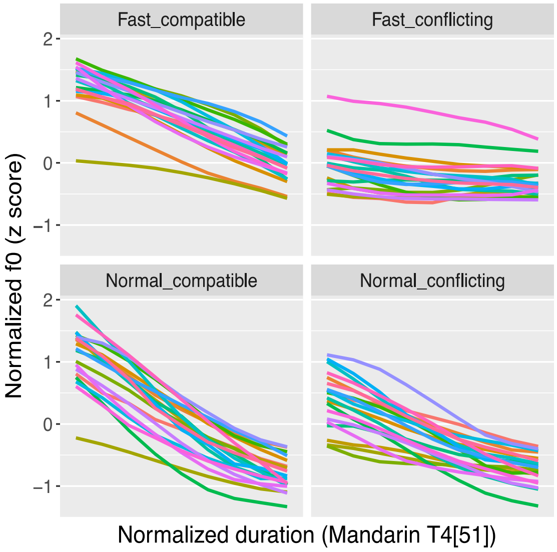
Appendix B

Mean f0 contours of the four Mandarin tones in different speech rates and tonal contexts by 20 speakers.






Fast_compatible
Fast_conflicting
Normal_compatible
Normal_conflicting
0.0
−0.5
−1.0
−1.5
−2.0
Normalized f0 (z score)
Normalized duration (Mandarin T3[21])


Fast_compatible
Fast_conflicting
Normal_compatible
Normal_conflicting
2
1
0
−1
Normalized f0 (z score)
Normalized duration (Mandarin T4[51])