Speaker Variability in the Production of Coarticulated Tones

Ricky KW Chan
Phonetics Laboratory, Department of Theoretical and Applied Linguistics, University of Cambridge
kwrc2@cam.ac.uk

ABSTRACT

Previous studies on lexical tone as a speaker-discriminant feature focused mainly on tone production in isolation or with a fixed tonal context. The present study reports an experiment on the production of Cantonese and Mandarin tones in different tonal contexts and speaking rates. Results show that while speakers show considerable variation in tone production in both languages, speaking rate and tonal context also play a role in the speaker-discriminating power of tone.

Keywords: tone, speaker characteristics, forensic speaker comparison, Cantonese, Mandarin

1. INTRODUCTION

The task of forensic speaker comparison (FSC) often involves comparing speech samples from recordings of a perpetrator and those of a suspect. An important goal in the field of FSC is thus to identify potential speaker-specific variables that can be exploited to discriminate speech samples. While studies of speaker-individuality in the domain of fundamental frequency (f0) abound, few have attempted to investigate dynamic properties in the f0 domain, a potentially rich source of speaker-specific information [3]. Lexical tone is a case in point.

Lexical tone is the use of pitch movement to distinguish one word from another in tone languages. Previous studies have shown that lexical tones can potentially serve as a parameter for discriminating speakers [1, 6]. However, these studies focused on the production of tones either in isolation or with fixed neighbouring tonal contexts. Since the actual realization of tone is not identical in all environments but varies owing to the influence of the neighbouring tones, the speaker-discriminating power of tone cannot be fully understood without taking tonal coarticulation into account.

Xu [7] distinguishes two kinds of tonal context: "compatible" and "conflicting". In a "compatible" context adjacent tones have f0 values identical or similar to the target tones, whereas in a "conflicting" context adjacent tones have f0 values different from the target tone. He demonstrated that a contour tone (rising or falling) in a conflicting context was distorted to the extent that the direction of its contour was sometimes reversed. Given the influence of neighbouring tones on the actual realization of a tone, the primary goal of the paper is to investigate the effect of tonal context on the speaker-discriminating power of tone.

Another potential factor affecting the speaker-discriminating power of a tone lies in the inherent density of tonal contrasts in the tone inventory of the language. Take Hong Kong Cantonese and Beijing Mandarin as examples. Cantonese contrasts six tones: three level tones (T1 high /55/, T3 mid /33/ and T6 low /22/), two rising tones (T2 high /25/ and T5 low /23/) and one falling tone (T4 /21/). The Cantonese "tone space" is so crowded, especially in the lower pitch range, that some tone pairs are confusable (see [4] for details). Coupled with the effort to maintain the contrast among the three level tones and between the two rising tone, Cantonese speakers may have little freedom to stray in their tone production if the tones are to remain perceptually distinguishable. On the other hand, Mandarin has a less crowded 'tone space' and contrasts only four tones: T1 level /55/, T2 rising /35/, T3 fall-rise (/214/ in isolation; /21/ in context/ and T4 falling /51/. The four Mandarin tones have different contours and it is expected that more variability can be tolerated without leading to perceptual confusion. The second goal of the paper is to study the potential link effect of tone inventory density and number of tonal contrast on how speakers are allowed to vary in the production of tones.

In the following, we report an experiment on the production of coarticulated tones in Cantonese and Mandarin.

2. METHOD

2.1. Subjects

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 the University of Hong Kong and had lived in Hong Kong for more than 15 years.

Mandarin: 20 native male speakers of Beijing 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 words whose second syllable carries the target tone (one of the six tones in Cantonese) were used in the experiment. Six words were adopted for each tone in the language (i.e. 6 tones x 6 = 36 Cantonese words and 4 tones x 6 = 24 Mandarin words). Half of the words have a compatible tonal context for the target tone, and the other half have a conflicting tonal context. When there are more than one possible compatible/conflicting contexts for the target tone, the tone with the closest pitch value with the target tone was selected for the compatible condition, and the one with the farthest pitch value from the target tone for the conflicting context. For instance, for the high level tone /55/, a conflicting context can be /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 will change the first fall-rise tone to a rising tone. Tables 1 and 2 show the tone patterns carried by the trisyllabic words.

Table 1 & 2: Tone patterns of the trisyllabic words used. The second syllable carries the target tone (bold) and the first and third syllables form the compatible/conflicting context.

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

Mandarin	
Compatible	Conflicting
/55/-**T1**/**55**/-/55/	/21/-**T1**/**55**/-/21/
/51/-**T2**/**35**/-/51/	/55/-**T2**/**25**/-/21/
/51/-**T3**/**21**/-/35/	/55/-**T3**/**21**/-/55/
/55/-**T4**/**51**/-/21/	/21/-**T4**/**51**/-/55/

Segmental content was also controlled for the target syllables owing to potential effects of segmental structure on f0 [2]. The syllables contain either /si:/, /fu:/, or /a:/ preceded by different consonants. The use of minimal contrast with /a:/ was not possible when only meaningful words were used.

2.3. Procedure

Recordings were made in a quiet room. Before a recording session began, subjects practised the target words once. 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 __" (*CS* condition). Then they read the whole list of words in isolation for four more times (*IS* condition). Items were randomized and presented one by one to avoid list effects. To control for the speaking rate of the speakers, regular beats were played through a virtual metronome at an interval of 2 seconds. Subjects were instructed to produce each word/sentence between two beats. It was expected that the use of the carrier sentence would encourage a higher speaking rate [7].

2.4. Data Extraction

Recordings were analyzed using *Praat* and digitized at a sampling rate of 44.1kHz. For each target syllable, two vertical markers were placed manually at the beginning and the end of periodicity. A *Praat* script was then applied to extract f0 values in all regions delimited by the vertical markers. 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 mostly reflect perturbation by neighbouring consonants. Around 2% of the tokens were so creaky that f0 values could not be extracted and were excluded from the analysis.

3. RESULTS & DISCUSSION

All f0 values were expressed in semitones re 100Hz instead of Hertz as the semitone scale, which is a logarithmic pitch scale, is a more reliable scale for capturing equivalence of intonational span across speakers [5].

3.1. Illustrative example: Mandarin falling tone

Figure 1 shows the mean f0 contours of the Mandarin falling tone for all conditions and contexts across all speakers. The tone has a clear falling shape in the compatible context. However, in the conflicting context the tone has a less steep fall

when produced without a carrier sentence (*IS*), and it even resembles a level tone when produced with a higher speaking rate (in a carrier sentence, *CS*). A closer look at its production by each speaker (Figure 2) reveals considerable variation across speakers; while most speakers' realization of the falling tone resembles the shapes depicted in Figure 1, they differ in not only their f0 level but also their slope. In particular, in the conflicting context some speakers have a relatively level contour, and some even show a small rise at the end of the word.

Figure 1: Mean f0 contours of the Mandarin falling tone for all conditions and contexts. (CS: carrier sentence; IS: isolation)

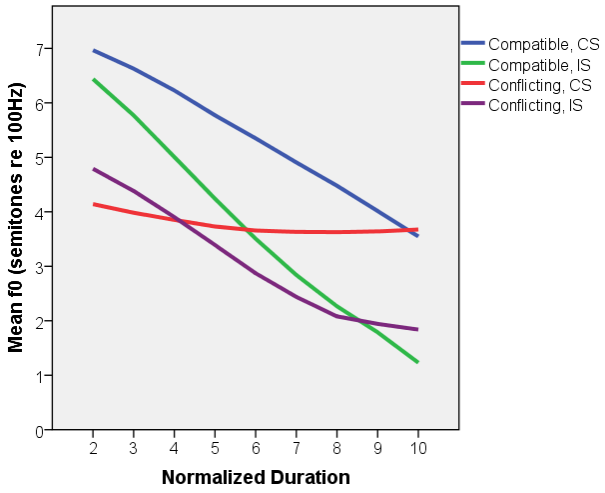
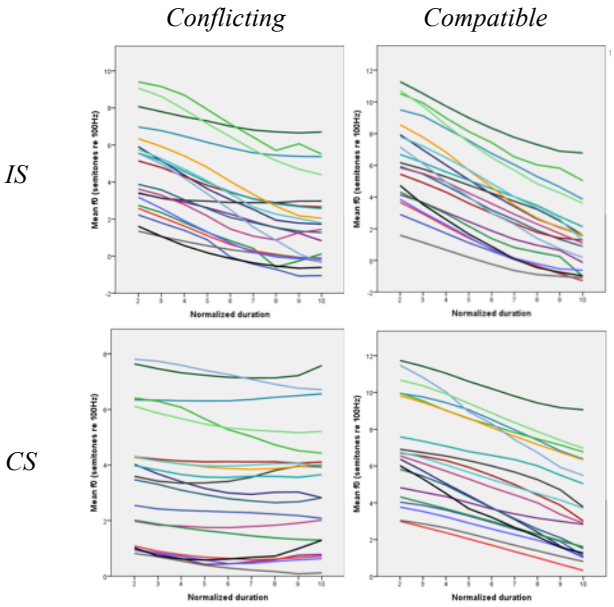


Figure 2: Mean f0 contours of Mandarin falling tone [51] by 20 speakers.



3.2. Discriminant analysis

The raw f0 values reveal between-speaker differences in both absolute frequency and the shape of the tone contours. To determine whether the speakers exhibit idiosyncrasy in the shape of the tone contours, all raw f0 data were normalized in terms of the *z*-score using arithmetic mean and one standard deviation as normalization parameters, i.e.

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

The raw data were normalized separately for each speaker in *CS* and *IS* conditions. The normalized values represent the excursion of tone contours in relation to the speaker's f0 range.

Discriminant analysis (DA) was conducted to evaluate the speaker-discriminating power of the tones by constructing a model for each speaker with a set of known data and attributing "unknown" data to the speaker models. The "leave-one-out" method was adopted: one token in each speaker's data set was regarded as an unknown sample and the remaining tokens were used to build the speaker's model. Every token in the data set was allocated to one of the groups (speaker). The percentage of correctly attributed tokens (the classification rate) was calculated and the best performance was reported as a DA score. Chance performance for the 20 speakers is 5%. DA resembles a closed test situation in which the speaker to be identified is known to be among a group of speakers.

Each tone was quantified based on the 9 measurements used above (i.e. 10%, 20%....90%), resulting in 9 predictors. Univariate and multivariate outliers were removed before DA was conducted; less than 2% of the data was removed. Separate discriminant analyses were run for all the tones in different contexts/conditions, and for both raw and normalized data. The DA scores based on raw data reflect speaker variability in both absolute f0 height and f0 contours, whereas those based on normalized data reflect mainly variability in the shape of the f0 contours. DA results are presented in Table 3 (Cantonese) & 4 (Mandarin) below.

DA scores based on raw frequency values were significantly higher than those on normalized values, $t(46)=10.5$, $p<.0001$ (Cantonese) and $t(30)=7.91$, $p<.0001$ (Mandarin). Although normalisation reduced the DA scores by between a third and a half, discrimination was still generally between three and five times better than chance, suggesting that there is a substantial contribution to the discrimination

potential of tones from between-speaker differences in the shape of contours. Speaking rate has mixed effects on the speaker-discriminating power of tones: overall DA scores for Mandarin tones produced without a carrier sentence were significantly higher than those produced within a carrier sentence, $t(14)=3.56$, $p=.0031$ (raw) and $t(14)=2.35$, $p=.0034$, suggesting that faster speech in general allows less room for speakers to stray in their tone production. However, the difference in DA scores for Cantonese tones produced with or without a carrier sentence only reached marginal significance with raw f0 values, $t(22)=2.26$, $p=.034$ and $t(14)=.352$, $p=.73$, and was not significant after normalization. This shows that the effects of speaking rate on tones as a speaker-discriminant appear to be language-specific.

Table 3: DA results (% correct attribution) for Cantonese tones

Tone	Context	Condition	Raw f0	Normalized f0
T1 [55]	Comp	CS	34.6	17.5
	Comp	IS	32.6	18.8
	Conf	CS	28.0	14.4
	Conf	IS	26.7	13.3
	Average		**30.5**	**16.0**
T2 [25]	Comp	CS	36.6	21.0
	Comp	IS	31.5	17.7
	Conf	CS	31.1	14.9
	Conf	IS	36.2	15.9
	Average		**33.9**	**17.4**
T3 [33]	Comp	CS	28.2	18.5
	Comp	IS	39.6	29.6
	Conf	CS	25.9	13.4
	Conf	IS	28.2	13.0
	Average		**30.5**	**18.6**
T4 [21]	Comp	CS	29.9	20.1
	Comp	IS	40.0	18.7
	Conf	CS	21.2	15.7
	Conf	IS	34.0	24.1
	Average		**31.3**	**19.7**
T5 [23]	Comp	CS	25.5	16.3
	Comp	IS	39.2	15.0
	Conf	CS	22.6	14.9
	Conf	IS	29.4	14.7
	Average		**29.2**	**15.2**
T6 [22]	Comp	CS	35.6	20.5
	Comp	IS	35.1	18.0
	Conf	CS	30.8	20.1
	Conf	IS	30.0	15.2
	Average		**32.9**	**18.5**
Overall mean			**31.4**	**17.6**

With regard to tonal context, DA scores for Cantonese tones were significantly higher in the compatible context than in the conflicting context, $t(22)=2.94$, $p=.0076$ (raw) and $t(22)=2.50$, $p=.02$ (normalized), but no significant effect was found for Mandarin tones; $t(14)=1.03$, $p=.32$ (raw) and $t(14)=0.83$, $p=.42$ (normalized).

When comparing tones with similar canonical pitch contours (i.e. T1 and T2 in Cantonese and Mandarin), the Mandarin ones has better DA scores than the Cantonese counterparts. This provides support to the hypothesis that a less dense tone inventory allows greater speaker variability in realization.

Table 4: DA results (% correct attribution) for Mandarin tones

Tone	Context	Condition	Raw f0	Normalized f0
T1 [55]	Comp	CS	36.7	18.6
	Comp	IS	50.0	24.6
	Conf	CS	32.6	18.4
	Conf	IS	38.8	23.8
	Average		**39.5**	**21.4**
T2 [35]	Comp	CS	30.3	15.8
	Comp	IS	42.6	20.0
	Conf	CS	25.3	18.0
	Conf	IS	46.5	24.1
	Average		**36.2**	**19.5**
T3 [33]	Comp	CS	31.5	21.8
	Comp	IS	35.1	16.6
	Conf	CS	29.1	14.1
	Conf	IS	32.0	20.9
	Average		**31.9**	**18.4**
T4 [21]	Comp	CS	33.6	14.0
	Comp	IS	35.2	15.0
	Conf	CS	22.2	17.1
	Conf	IS	38.6	21.9
	Average		**32.4**	**17.0**
Overall mean			**35.3**	**19.2**

4. CONCLUSION

The present study furthers our understanding of tones as a parameter to distinguish speakers: tonal context and speaking rate may potentially affect the speaker-discriminating power of tone, and should be taken in account in forensic speaker comparison.

5. REFERENCES

[1] Chan, R. (2013). *Individual Differences in the Realization of Cantonese Tones*. M.Phil. Dissertation,

Department of Theoretical and Applied Linguistics, University of Cambridge

[2] Lehiste, I. 1970. *Suprasegmentals* Cambridge, MA: MIT Press.

[3] 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.

[4] Mok, P., Zuo, D., & Wong, P. 2013. Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*, 341-370

[5] Nolan, F. 2003. Intonational equivalence: an experimental evaluation of pitch scales. *Proc. 15th ICPhS* Barcelona.

[6] Thaitechawat, S. & Foulkes, P. 2011. Discrimination of speakers using tone and formant dynamics in Thai. *Proc. 17th ICPhS,* Hong Kong.

[7] Xu, Y. 1994. Production and perception of coarticulated tones. *J. Acoust. Soc. Am*. 95(4), 2240-53.