Perceived pitch and formant frequencies in the perception of lexical tone in Cantonese

Qian Min Feng[1], Amy Wu[2], Jon Nissenbaum[1]
[1] *Brooklyn College, CUNY,* [2] *Brandeis University*

Cantonese has six lexical tones (four level, two rising) that distinguish otherwise identical syllables. Four level tones should create a crowded fundamental frequency (f0) space requiring more fine-grained distinctions than simple systems that use just two categories, raising the question how listeners are able to identify the intended tone level. It is known that acoustic cues besides f0 enter into tone perception (eg. voice quality, spectral tilt [1–3]). Less understood is whether f0 in the absence of other cues could reliably support distinctions among the four level tones of Cantonese, and how f0 interacts with other factors to produce tone perception. These questions are relevant for longstanding debates about the phonological representation of tone, and specifically whether a model like Yip's two-feature system [4] is viable.

Yip's (2002) Two-feature model:

	Upper Register		**Lower Register**	
High Tone	Tone 1 (highest)	Tone 2 (rise)	Tone 6 (mid-low)	Tone 5 (rise)
Low Tone	Tone 3 (mid-high)		Tone 4 (lowest)	

We conducted two experiments to investigate whether f0 on its own is sufficient for perception of lexical tone in Cantonese. The first was a production study: we recorded eight native speakers of Cantonese (five male, three female) reading word paradigms where all six tones were present (for six distinct syllables). In one condition, subjects read each of the paradigms as a list of words in citation form. For the second condition, the words were embedded in carrier sentences and presented in randomized order. The results were striking: in citation form, the level tones were distributed roughly evenly throughout each speaker's f0 range. However, in the carrier sentences, speakers were quite consistent in dividing the f0 space into *three* categories rather than four: the two *mid*-tones (tones 3 and 6) were produced at essentially the same f0. Fig 1a shows a representative set of contours for one male speaker. The same pattern held across subjects (Fig 1b). These results are consistent with Yip's model. Tones 3 and 6 just instantiate opposite pairings of values for Register and Tone, and are not predicted specifically to be separated in f0 space.

Fig. 1a representative f0 contours

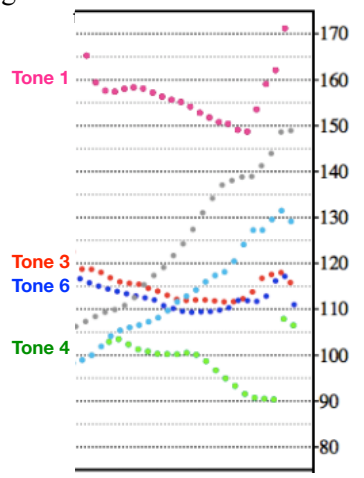
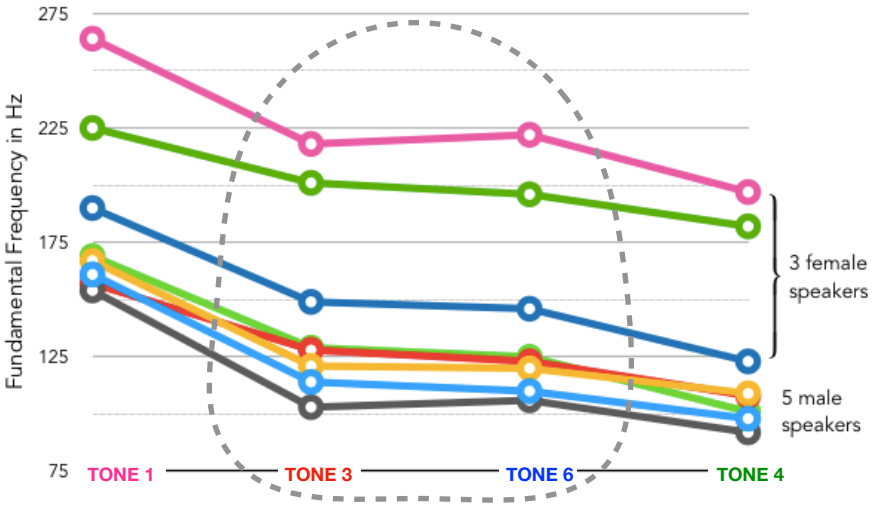


Fig. 1b Cantonese Production Study: Average f0 for the four level tones (by speaker)

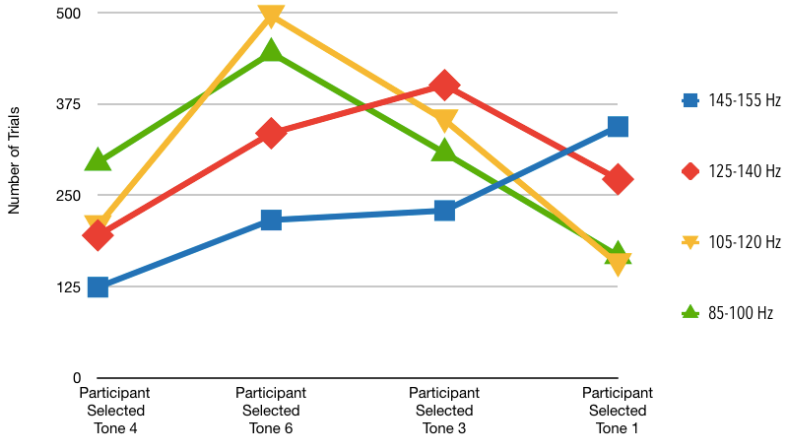


Our second study was a perception study, designed to isolate f0 from all other conceivable cues for tone, using modified Sinewave Speech (SWS). SWS replaces vocal tract formants with sinusoids. However, formant trajectories by themselves omit information about f0, making standard SWS unsuitable for studying tone perception [5–7]. To create our stimuli, we replaced the lowest sinusoid of the SWS replica (representing F1) with a complex tone constructed using a time-varying bandpass whose center frequency tracks F1, wide enough at any timepoint for two harmonics of an experimentally controllable f0 contour [8]. The resulting two-component tone implies a missing fundamental, allowing simultaneous impression of harmonic direction and F1 direction. We synthesized a set of SWS replicas of Cantonese syllables, replacing F1 with this complex tone, to induce a range of implied f0 levels. Our stimuli thus had a minimal cue for f0 but lacked other potential cues for tone.

Our tone perception experiment used a 15-step continuum, varying the implied f0 from 85-155Hz in 5Hz increments. Six target syllables, each synthesized to cue f0 at all 15 steps, were presented randomly in one of four carrier sentences. Carrier sentences were likewise modified SWS replicas of sentences where words preceding/following the target syllable contained high, mid, or low tones. With each auditory presentation, the carrier sentence was displayed on a screen, with the target syllable left blank. Listeners were asked to identify which word they heard at the target syllable, choosing from four words with distinct level tones displayed in random order below the carrier.

If f0 alone is a sufficient cue for tone identification even in the crowded tone space of Cantonese, listeners would be predicted to perceptually divide the f0 range into four regions with distinct peaks, with Tone 1 occupying the highest f0s, Tone 3 the next highest, Tone 6 the mid-low region, and Tone 4 the lowest. Our results at least partially support this prediction. Listeners identified the target word as having Tone 1 on 929 out of 4544 trials. More Tone 1 choices were made when the induced f0 on the target word was in the highest frequency bin (145-155Hz) than for the three lower bins. Similarly, each of the other level tones was chosen more often for induced f0s in the expected bin than for other bins. Thus, even in the absence of any cue except f0, listeners are more likely to identify the four level tones according to perceived pitch.

Fig 2: Tone identification study: Number of trials on which listeners identified each level tone based on induced f0



While the results of our perception study were promising, two rather glaring puzzles emerged. The first concerns the surprisingly large number of trials on which, for instance, listeners identified a word as having Tone 1 (high) despite the induced f0 being in a lower bin; the same holds for the other tones. This may be related to another question, namely the uneven distribution of choices (for instance, listeners identified many more words as having Tone 6, mid-low level, than any of the other tone levels). Perceived pitch clearly influences tone identification, then, but might not be a sufficient condition.

The second, and more pressing (to our thinking) puzzle concerns the joint results of our production and our perception studies. The production study indicates that *speakers* tend to divide the f0 space into three categories, suggesting that f0 may not even be a *necessary* condition for perception at least of the two mid-tones. We conclude the talk with some speculation about how to reconcile these two results and briefly describe a follow up perception study that we are conducting, testing a physiological model for Yip's two-feature system whereby Register distinctions are achieved by laryngeal lowering, which is known to have a lowering influence on f0, and should have an effect on formant freqencies as well.

References: **[1]** Khouw, E, & V Ciocca. 2007. Perceptual correlates of Cantonese tones. *J Phon* 35: 104:117. **[2]** Whalen, D, & Y Xu. 1992. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica* 49: 25–47. **[3]** Yu, KM, & HW Lam. 2014. The role of creaky voice in Cantonese tonal perception. *J Acoustical Soc of Amer* 136.3: 1320–1333. **[4]** Yip, M. 2002. *Tone*. Cambridge University Press. **[5]** Feng, YM, L Xu, N Zhou, G Yang, & SK Yin. 2012. Sine-wave speech recognition in a tonal language. *J Acoustical Soc of America* 131(2), EL133. **[6]** Han, Y, & F Chen. 2017. Relative contributions of formants to the intelligibility of sine-wave sentences in Mandarin Chinese. *J Acoustical Soc of America* 141.6 EL: 495–499. **[7]** Remez, RE., & PE. Rubin. 1984. On the perception of intonation from sinusoidal sentences. *Attention, Perception & Psychophysics*, 35(5), 429-440. **[8]** Nissenbaum, J. 2019. Modifying sinewave speech with a minimal cue for pitch: a new tool for perception studies. *Proc Linguistic Society of America* 93rd annual meeting.