The complex interplay of perceived pitch and formant frequencies in lexical tone perception in Cantonese

Qian Min Feng, Amy Wu, Jon Nissenbaum

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 other than 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, including whether a model like Yip's two-feature system [4] is viable (table 1).

Table 1: 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 a set of 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 paradigm as a list of citation forms. For the second condition, the words were embedded in carrier sentences and randomized. 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 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: if the two features encode distinct pitch-adjusting articulations (e.g. laryngeal raising/lowering vs. vocal fold lengthening/shortening), the two "mid"-tones would each simultaneously encode one f0-raising and one f0-lowering articulation.

Fig. 1a representative f0 contours for one speaker

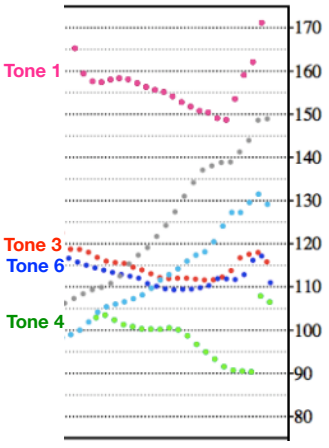
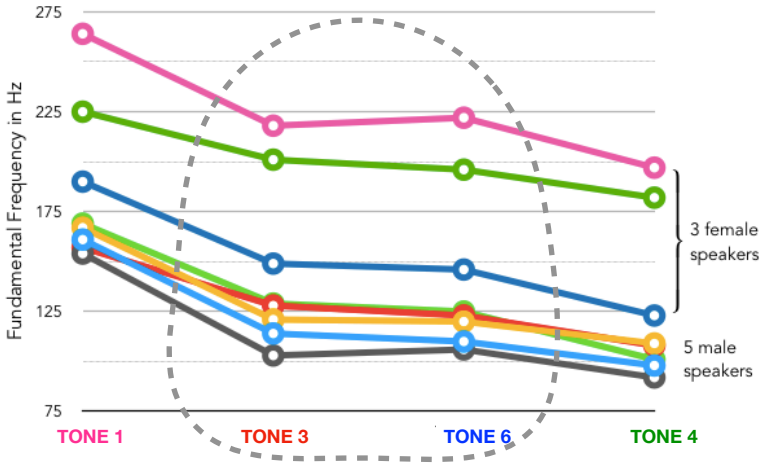


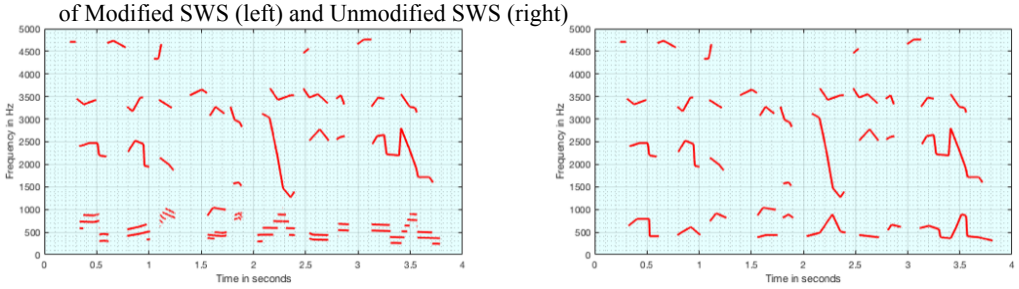
Fig. 1b Cantonese Production Study: Average f0 for the four level tones (by speaker)



Our second study was a perception study, designed to isolate f0 from all other cues for tone using modified Sinewave Speech (SWS). SWS represents formants as 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 f0, creating a simultaneous cue for harmonic direction and F1 direction. We synthesized a set of

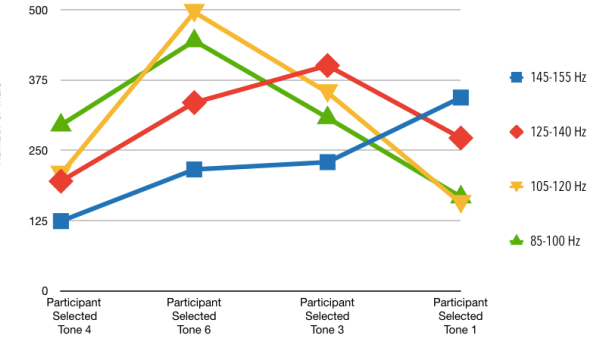
SWS replicas of Cantonese syllables, replacing F1 with this complex tone, to induce a range of implied f0 levels. Our stimuli thus had a minimal cue for f0 but lacked other potential cues for tone. We varied the implied f0 of six target syllables from 85-155Hz in 5Hz steps, and embedded them in one of four carrier sentences (also modified SWS replicas of sentences [fig 2] 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.

Fig. 2: Example of stimuli for perception study: Target word *JAU* embedded in carrier sentence. Pseudo-spectrograms of Modified SWS (left) and Unmodified SWS (right)



If f0 alone is a sufficient cue for tone identification in the crowded tone space of Cantonese, listeners should perceptually divide the f0 range into four regions with distinct peaks. Our results at least partially support this prediction (fig. 3); even in the absence of any cue besides f0, listeners are more likely to identify the four level tones according to perceived pitch. However, the results of this perception study appear to contradict the results of the production study, which indicates that *speakers* tend to divide the f0 space into three categories rather than four (other than in citation form) — the two mid-tones appear merged in production but not perception.

Fig 3: Results of perception study: Number of trials on which listeners identified each level tone based on induced f0



We conclude by reporting pilot results from a third study, similar to the second study except that in addition to manipulating the implied f0, we also raise/lower the first and third formants. This modification was designed to test an articulatory 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 lower formant frequencies as well. Preliminary results confirm the prediction that (implied) f0s in the range of the two mid-tones can be disambiguated to tone 3 or tone 6 (high/low register) by shifting the frequencies of F1 and F3. The combined results of all three studies thus argue against mid-tone merger, and support an articulatory model of tone/register features that underlie a complex interplay of f0 and formant frequencies.

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. *Linguistic Society of America* 93rd annual meeting.