The role of creaky voice in Cantonese tonal perception

Kristine M. Yu and Hiu Wai Lam

Department of Linguistics, University of California, Los Angeles
krisyu@ucla.edu and hiuwai208@ucla.edu

ABSTRACT

Cantonese has anecdotally been claimed to have a low tone (Tone 4, 21/11) inconsistently realized with creaky voice. This paper shows that Cantonese listeners are sensitive to creaky voice in tonal perception: the presence of creak increases tonal identification accuracy for Tone 4 and biases listeners toward Tone 4 responses. Moreover, listeners are sensitive to details of creak production, suggesting that a more detailed representation of creak than a binary flag for the presence of creak is needed in tonal representation and automatic tonal recognition.

Keywords: Cantonese, tone, phonation, creak, voice quality

1. INTRODUCTION

In some tonal languages, one or more tone categories may consistently be realized with a particular non-modal phonation type, e.g. in (H)mong [1]. For these languages, perceptual studies suggest that listeners use phonation cues in tonal perception, e.g. [1]. In other tonal languages, one or more tone categories may *inconsistently* be realized with non-modal phonation type. A well-known example is Mandarin: Tone 3 (T3), the lowest tone in the inventory, is sometimes creaky [3, 6]. F0-based features are assumed to be sufficient for discrimination of T3 from the other tones of Mandarin: phonation cues are considered to be redundant or concomitant. Another language of this type is Cantonese, which has six tones: high level (T1, 55), high rising (T2, 35/25), mid level (T3, 33), low falling (T4, 21/11), low rising (T5, 23/13), and low level (Tone 6, 22) [11]. (Some descriptions also distinguish shorter entering tones in syllables with unreleased stop codas.)

Cantonese has anecdotally been reported to have an (inconsistently) creaky T4 [11, 13]. This paper addresses the role of voice quality cues in tonal perception in Cantonese, a language with redundant phonation cues. Since the presence of creaky voice in Cantonese may be both frequent and systematically conditioned on tonal category, a natural question is: *are listeners sensitive to creaky voice in tonal perception in tonal languages with redundant phonation cues, such as Cantonese*? [8] used a 2-way forced choice tonal identification task for T3 and T4 in Mandarin and tested listeners on a continuum of timepoints for a minimum (turning point) in the f0 contour. They found that added creak, resynthesized by introducing vowel-medial pitch halving, had very little or no effect on tonal identification. However, [3] performed a tonal identification gating task in Mandarin with both creaky and non-creaky Tone 3 natural stimuli and found that the recognition point for the listeners came sooner for creaky Tone 3s.

In this paper, we report on two experiments in Cantonese tonal perception to build on these results. The first experiment was an initial test for sensitivity of listeners to creak in Cantonese tonal perception. It was a 6-alternative forced choice tonal identification task of Cantonese monosyllables extracted from a corpus of multispeaker connected speech. Drawing on the natural variation in the corpus, we chose half of the Tone 4 (T4) stimuli to be creaky. We hypothesized that identification accuracy for the creaky T4s would be higher than that of the non-creaky T4s.

The second experiment elaborated on our knowledge of *how* listeners use creaky voice in tonal perception. It addressed two questions: (i) *Do listeners subsume creak under f0 as signalling a low absolute f0?* and (ii) *Are listeners sensitive to details about the creak production such as the duration of creak and the spacing of the glottal pulses*? The experiment was a 2-alternative forced choice tonal identification task between Tone 4 (21/11) and Tone 6 (22), the tone most confusable with Tone 4 [7]. Level f0 Tone 6 productions from a male and female speaker were resynthesized and cross-spliced with different natural creaky productions of Tone 4 and presented to the listener. Additionally, the tone to be identified was preceded by a syllable, whose f0 was shifted up and down to see how listeners integrate creak with contextual f0 information.

The results of both experiments bear on a range of issues for tonal languages with redundant phonation cues: (1) how tones are represented in the acoustic space, i.e. if there is reason to define tones in an elaborated space with voice quality parameters, and (2) if automatic tonal recognition might benefit by parameterizing creak in the speech signal.

2. EXPERIMENT 1

2.1. Methods

2.1.1. Materials

The stimuli were 596 tokens of sentence-medial /lau/ syllables drawn from a Cantonese tonal production corpus consisting of sentences [lei$^{25/35}$ jiu^{33} lau lau jak$^{˧33}$ tʃoeŋ^{33}/kap$^{˧33}$/sou^{33}] 'you want lau-lau to eat sauce/pigeon/vegetarian' with the target bitone /lau-lau/ over all 36 possible combinations of tones T1 to T6. For all tones, *lau* is a real word, although lau^{33} is uncommon. There were 72 examples for each tone, drawn from a total of 4 males and 4 females.

Half of the T4 tokens were chosen to be creaky and the other half non-creaky by listening and manual inspection of the waveform and spectrogram in Praat [4]. A token was defined to be creaky if it had the auditory percept of creak, as determined by the authors, and if there were: alternating cycles of amplitude and/or frequency or irregular glottal pulses in the waveform, missing values or discontinuities in the f0 track determined by Praat's autocorrelation algorithm with default settings, and/or strong subharmonics or regions lacking harmonic structure in the narrowband spectrogram. All tokens were resynthesized using PSOLA in Praat to have equal average amplitude, and the duration of each token was equalized to 313ms, the grand mean of token durations.

2.1.2. Participants

The participants were 16 native Cantonese speakers (11M, 20.6±1.6 years, 5F, 21.2±0.8 years), all of whom spoke Cantonese on a daily basis. All were born in Hong Kong or Macau.

2.1.3. Procedure

Participants were tested in a soundbooth at UCLA in Los Angeles, California. The experiment was run in Matlab using Psychophysics Toolbox extensions [5]. Stimuli were played from an Echo Indigo IO sound card on a laptop over studio monitor headphones at a standardized, comfortable volume and the responses and reaction times of the participants were recorded. The interstimulus interval was 3s. Participants were asked to identify each stimulus by a keyboard press of one of six keys labeled with characters corresponding to the minimal tone set over *lau*. The order of the stimuli presentation as well as which key was labeled with which word was randomized across participants.

2.2. Results

Statistical analysis was performed in R [12]. The T4 stimuli subset was analyzed using mixed effects regression [2]. Correctness in responses for T4 stimuli was analyzed with mixed effects logistic regression. Forward model selection was used to test the partial effect of CREAK (present, absent) on correctness of T4 response while controlling for other stimuli variables: SYLLABLE (S1, S2 in the bitone), SPEAKER SEX (male, female), and SPEAKER (8 levels). Each of these fixed effects was coded using numeric indicator variables and mean-centered [9], and SPEAKER was included as a non-interacted fixed effect. Random intercepts for the listener were included, providing individual by-listener adjustments for T4 response correctness. Successive nested models in the forward model selection were compared using likelihood ratio tests, with chi-squared tests to test for significance.

Overall, identification accuracy for T4 was high (70.51%) compared to that of other tones, and the tone most confusable with T4 was T6 (13.41%). (Only T1 identification accuracy was higher, 85.94%.) Identification accuracy for T4 was 82.03% (SE=2.27%) for creaky T4s, but only 58.98% (SE=3.57%) for noncreaky T4s.

Model comparison showed that the best model for the probability of correct identification of T4 included the following factors: CREAK, SYLLABLE, SPEAKER SEX, and the interaction SYLLABLE:SPEAKER SEX ($\chi^2(1) = 17.02$, p = 2.0E-4 for the likelihood ratio comparison between a model with all main effects and a main effects model without creak; $\chi^2(1) = 13.20$, p = 2.8E-4 for the additional inclusion of the interaction). Overall, the presence of creak significantly increased the probability of correct identification of T4.

2.3. Discussion

Experiment 1 demonstrated that the presence of creak significantly improved identification accuracy of Cantonese T4. To our knowledge, this is the first experimental result suggesting that native speakers of a tone language with phonation as a redundant cue for tone show improved tonal identification accuracy due to phonation cues. However, the stimuli were drawn from naturally produced speech. Not only were the creaky T4s heterogeneous in duration and production mechanism of creak, but also, absolute f0 or f0 movement in noncreaky regions of the naturalistic stimuli was not controlled. Thus we were unable to obtain direct information about listener sensitivity to details of creak, and we could not factor out the presence of creak from other possible cues, such as low and/or falling f0 preceding the creaky region, or even the precept of low f0 in creak.

3. EXPERIMENT 2

In Experiment 2, our goals were to disentangle a low f0 percept from creaky voice as a cue for discriminating Cantonese T4 and T6 and to examine whether variation in the realization of creaky voice influenced tone perception. Using [14]'s result that Cantonese listeners use the f0 of preceding context to judge relative pitch for identifying tones, we manipulated f0 perception by resynthesizing a 8-step f0 continuum in semitone increments for the syllable preceding the target syllable to be identified. The target syllable was resynthesized by cross-splicing in stimuli with different creak qualities. To preclude the availability of absolute low f0 as a cue, we chose stimuli from a high-pitched male and female and resynthesized the f0 of the target syllable to be ambiguous between that of T4/6 for the speakers in the corpus. Because the vocal fry mechanism is contingent on low f0 [10], we selected instances of laryngealization due to period doubling, in which f0 perception is ill-defined due to a bitonal percept.

3.1. Methods

3.1.1. Materials

Disyllables /jiu lau/ were selected from one male and one female with a high pitch range from the corpus described in Experiment 1. For each speaker, the utterance with the lowest level contour instance of T6 was selected from the corpus, and three additional utterances were selected to exhibit a range of variation in creaky realizations of T4. The T4 utterances were chosen to be period doubled based on spectrographic evidence of strong subharmonics and a bitonal auditory percept: one had a wider pulse width ("wide"), another a narrower width ("narrow"), and one had a strong f0 percept ("pitched"). The creaky T4 [au]s were cross-spliced with the T6 /jiu lau/ utterances, with three splice points chosen for each creaky /au/ token: "light", "medium", and "heavy", where "light" included the minimal amount of speech material to induce a creaky percept in /au/. Durations of /jiu/, /l/ and /au/ were equalized to their averages between all utterances. Three repetitions of each of the two modal T6 stimuli were included, for a total of 192 stimuli. Monosyllables--only the target syllables--were also presented and included 9 repetitions of the modal stimuli, 36 stimuli in total.

3.1.2. Participants

The participants were 20 native Cantonese speakers born and raised in Hong Kong (10M, 20.3±1.9 years, 10F, 21.8±1.7 years).

3.1.3. Procedure

Participants were tested in a soundbooth at City University of Hong Kong. The procedure was the same as in Experiment 1, except that the choices were limited to two choices corresponding to T4 and T6. Each participant heard each stimulus set twice in separate blocks.

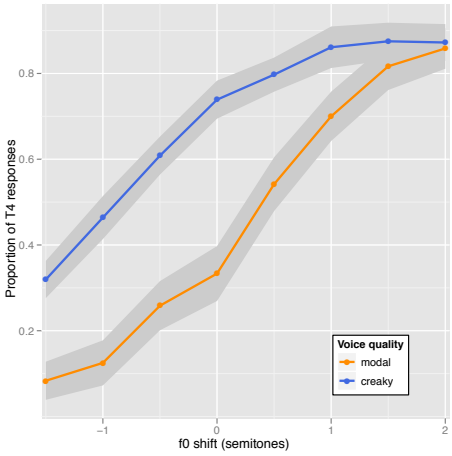
3.2. Results

The partial effects on response choice (T4 or T6) of: CREAK (present, absent), CREAK TYPE (none, pitched, wide, narrow), CREAK PROPORTION (none, light, medium, heavy), and F0 SHIFT (treated as an interval-scale variable, following [14]) were analyzed using mixed effects regression. CREAK TYPE and PROPORTION were also crossed to create CREAK QUALITY, with ten levels.

Analysis of the monosyllable results indicated that the presence of creak biases for T4 responses in the absence of immediate contextual f0 information. Likelihood ratio tests showed that the best model for the probability of a T4 response included the fixed effects: SPEAKER SEX, CREAK, and SPEAKER SEX:CREAK ($\chi^2(1) = 22.10$, p = 2.6E-6 for the likelihood ratio comparison between a model with all main effects and a main effects

model without CREAK; $\chi^2(1) = 9.50$, p = 2.1E-3 for inclusion of the interaction). Analysis of the modal stimuli subset of bisyllable data indicated that contextual f0 information biased responses: the probability of a T4 response increases as f0 increases on the preceding syllable relative to the target syllable. Moreover, likelihood ratio tests supported the inclusion of both F0 SHIFT and CREAK and their interaction in modeling T4 responses for the full male and female disyllabic stimuli sets (Fig. 1). For both creaky and modal stimuli, there was a higher probability of T4 response with higher preceding f0, but for creaky stimuli, the slope of the T4 response curve as a function of F0 SHIFT was less steep than for modal stimuli. Finally, within the creaky stimuli subset of the bisyllable data, likelihood ratio tests supported the inclusion of CREAK QUALITY above F0 SHIFT for both the male and female stimuli, showing that listeners were sensitive to details of creak production.

Figure 1: Overall T4 ID conditioned on f0 shift of preceding syllable and presence of creak for female stimuli; aggregated results. Ribbons show ±1SE.



3.3. Discussion

Experiment 2 showed that even when the stimuli were carefully controlled to tease apart creaky voice as a cue from f0-based cues, Cantonese listeners were still sensitive to creaky voice in tonal identification. In a forced choice task between T4 and T6, they were biased towards T4 responses in the absence of contextual f0 information, and in the presence of contextual f0 information from the preceding syllable, this bias was maintained above the effect of contextual f0 information. Finally, the proportion of creak and glottal pulse width also affected listener response choices, showing that listeners are sensitive to details of creak production.

4. CONCLUSIONS

The results of both experiments bear on a range of issues for tonal languages with redundant phonation cues. Since they show that Cantonese listeners use creaky voice in tonal perception, there is reason to define tones in an elaborated perceptual space with voice quality parameters, even for tonal languages with redundant phonation cues. The results also suggest that automatic tonal recognition might benefit by parameterizing phonation in the speech signal, and that a detailed parameterization of phonation, rather than a binary flag for the presence of creak, is necessary.

5. REFERENCES

[1] Bates, D., Maechler, M. 2010. lme4: linear mixed-effects models using S4 classes, http://lme4.r-forge.r-project.org.

[2] Andruski, J.E. (2006). Tone clarity in mixed pitch/phonation-type tones. *J. Phon.* 34, 388-404.

[3] Belotel-Grenié, A., Grenié, M. 1997. Types de phonation et tons en chinois standard. *Cahiers de Linguistique – Asie Orientale* 26, 249-279.

[4] Boersma, P., Weenink, D. 2008. Praat: doing phonetics by computer (Version 5.0.13).

[5] Brainard, D.H. 1997. The psychophysics toolbox. *Spatial vision*. 10, 443-446.

[6] Davison, D.S. 1991. An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics* 78, 50-57.

[7] Fok, C. 1974. *A perceptual study of tones in Cantonese*. Hong Kong: University of Hong Kong, Centre of Asian Studies.

[8] Gårding, E., Kratochvil, P., Svantesson, J.-O. Tone 4 and tone 3 discrimination in modern Standard Chinese. *Lg. & Sp.* 29, 281-293.

[9] Gelman, A., Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

[10] Gerratt, B.R., Kreiman, J. 2001. Toward a taxonomy of nonmodal phonation. *J. Phon.* 29, 365-381.

[11] Matthews, S., Yip, V. 1994. *Cantonese: a comprehensive grammar.* New York: Routledge.

[12] R Development Core Team. 2010. R: a language and environment for statistical computing, http:/www.R-project.org.

[13] Vance, T.J. 1977. Tonal distinctions in Cantonese. *Phonetica* 34, 93-107.

[14] Wong, P.C.M., Diehl, R.L. 2003. Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *JSLHR* 46, 413-421.

ACKNOWLEDGMENTS

Thank you to E. Zee for use of the soundbooth at CUHK and support from NSF grant BCS-0720304 and a NSF graduate fellowship to the first author.