The role of creaky voice in Cantonese tonal perception[a)]

Kristine M. Yu[b)]
Department of Linguistics, University of Massachusetts Amherst, Amherst, Massachusetts 01003

Hiu Wai Lam
Department of Linguistics, University of California Los Angeles, Los Angeles, California 90095

(Received 2 January 2014; revised 10 June 2014; accepted 24 June 2014)

There are few studies on the role of phonation cues in the perception of lexical tones in tonal languages where pitch is the primary dimension of contrast. This study shows that listeners are sensitive to creaky phonation in native tonal perception in Cantonese, a language in which the low falling tone, Tone 4, has anecdotally been reported to be sometimes creaky. First, in a multi-speaker corpus of lab speech, it is documented that creak occurs systematically more often on Tone 4 than other tones. Second, for stimuli drawn from this corpus, listeners identified Tone 4 with 20% higher accuracy when it was realized with creak than when it was not. Third, in a two-alternative forced choice task of identifying stimuli as Tone 4 or Tone 6 (the low level tone) isolating creak from any concomitant pitch cues, listeners had a higher proportion of Tone 4 responses for creaky stimuli. Finally, listeners had more Tone 4 responses for creaky stimuli with longer durations of nonmodal phonation. These results underscore that differences in voice quality contribute to human perception of tone alongside f0. Automatic tonal recognition and clinical applications for tone would benefit from attention to voice quality beyond f0 and pitch.
 [http://dx.doi.org/10.1121/1.4887462]

PACS number(s): 43.71.Es, 43.70.Gr, 43.66.Ki [BRM] Pages: 1320–1333

I. INTRODUCTION

There are two means of distinguishing words by their laryngeal source articulations: pitch and phonation. Traditionally, tone languages refer to languages where pitch is lexically contrastive, while register languages, at least in opposition to tone languages, refer to languages where phonation is lexically contrastive: "phonation type is to a register language what tones are to a tone language" (DiCanio, 2009, p. 162).[1] However, there is growing evidence of languages that make use of both pitch and phonation. Some languages make use of pitch and phonation independently, such that all tones can be produced with the same distinct phonations. One example of such a language is Jalapa Mazatec (Silverman *et al*., 1995), where all phonation types (breathy, modal, creaky) may co-occur with all tonal categories (high, mid, low). [Additionally, dimensions of contrast for tone and phonation type in this language also overlap and co-vary in systematic patterns of coarticulation (Garellek and Keating, 2011); see also Itunso Trique (DiCanio, 2012).]

In contrast, other languages combine particular phonations with particular tones. Such languages differ in whether pitch or phonation is more important perceptually as a dimension of contrast, suggesting that there is a "fuzzy boundary" between register and tone languages (Abramson and Luangthongkum, 2009). There are cases where pitch has been shown to be perceptually primary, although there are consistent differences in phonation, e.g., Suai (Abramson *et al*., 2004) and Cham (Brunelle, 2012). There are also cases where phonation has been shown to be perceptually primary, although there are consistent differences in pitch, e.g., Hmong (Garellek *et al*., 2013), Northern Vietnamese (Brunelle, 2009), Sgaw Karen (Brunelle and Finkeldey, 2011).

For the majority of tone languages, pitch has long been assumed to be the primary—if not the only—dimension of tonal contrast. This assumption is reflected in the pervasive lack of attention to phonation in automatic tonal recognition (see Yu, 2010) and in clinical applications such as speech therapy for linguistic tone and improving tone perception for patients with cochlear implants, e.g., Barry and Blamey (2004); Lee *et al*. (2002). For tone languages where pitch is thought to be the primary dimension of tonal contrast, it is not clear that phonation cues may be useful in tone perception, or that listeners use such cues, if they are available. This is in contrast to register languages, where it is clear that pitch cues are available and used by listeners, e.g., Abramson *et al*. (2004); Brunelle (2012). Evidence that listeners use phonation cues in tone languages where pitch cues are thought to be the primary dimension of contrast would contribute further support to the idea of a continuum between the use of phonation and pitch cues in register and tonal contrasts in the world's languages, as suggested in the typological literature (Abramson and Luangthongkum, 2009; Zhu, 2012). Such evidence would also be important to draw attention to phonation in research on automatic tonal recognition and in clinical applications involving tone. This study provides such evidence.

The most well-studied case where phonation may be a secondary cue to pitch cues in tonal perception is Mandarin: Tone 3 (T3, 214, ˨˩˧),[2] the lowest tone in the inventory, is

[a)]Portions of this work were presented at Interspeech 2010, the Spring 2010 meeting of the Acoustical Society of America, and the 17th International Congress of Phonetic Sciences in 2011.
[b)]Author to whom correspondence should be addressed. Electronic mail: krisyu@linguist.umass.edu

sometimes creaky (Davison, 1991; Kong, 2001, *inter alia*). Zhu (2012) reports that creak occurs on the lowest tone in "hundreds of local varieties of Chinese and other tonal languages in China." But even for Mandarin, there are only two small corpus studies documenting that speakers frequently and systematically use creak (Belotel-Grenié and Grenié, 1997, 2004); similar studies for other tonal languages where phonation may play a secondary role in tonal contrast do not exist. Moreover, tonal perception experiments in Mandarin have largely abstracted away from creak, e.g., Whalen and Xu (1992, pp. 27–29); Wang *et al.* (2010).

The tonal perception experiments in Mandarin on creak that exist are the only previous experiments on the role of phonation cues in well-studied tone languages where pitch is thought to be the primary dimension of contrast. These have provided only weak evidence that listeners use creak in tonal perception. Gårding *et al.* (1986) used a two-alternative forced choice tonal identification task for T3 and T4 (51, ˥˩) in Mandarin and tested listeners on a continuum of time points for a minimum (turning point) in the f0 contour. They compared the identification curve for stimuli without creak, and those resynthesized to be creaky by introducing pitch halving in the middle of the vowel. They were unable to show an effect of pitch halving on identification. Belotel-Grenié and Grenié (1997) performed a tonal identification gating task in Mandarin with both creaky and non-creaky T3 natural stimuli. They found that in non-creaky T3s, the percentage of correct tonal identification was 58% after 60% of the syllable was played. In contrast, for creaky T3s, identification accuracy was 93.4% after 60% of the syllable. Thus, the presence of creak appears to have slightly sped recognition of T3. However, identification accuracy for both creaky and non-creaky T3s after the entire syllable was played was 100%. No results show that creak can improve tonal identification accuracy of T3.

The above review shows the difficulty of demonstrating that creak plays a substantial role in tonal perception in Mandarin. This may be due to the fact that it has a small tonal inventory of four tones that have very distinctive pitch contours (high level, rise, low, fall): Mandarin listeners were already at ceiling in Belotel-Grenié and Grenié (1997) in tone perception using only pitch cues. We assumed that phonation differences are more likely to be used when a language has more tones, which thus are less reliably distinguished by pitch alone. Therefore, for this study, we chose Cantonese as an exemplar of a tone language where pitch is the primary dimension of tonal contrast. It has six tones, many with similar pitch contours: 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 (T6, 22, ˨) (Matthews and Yip, 1994).[3] (Note that Matthews and Yip list two different contours for some tones to reflect the variability in tonal transcription labels in the literature.) Studies have shown that pitch is perceptually important in Cantonese tonal perception (Fok, 1974; Wong and Diehl, 2003; Khouw and Ciocca, 2007). However, Cantonese has also anecdotally been reported to have an (inconsistently) creaky T4 (Vance, 1977, p. 105; Matthews and Yip, 1994, p. 22), although the incidence of creak in T4 has not been documented in the literature.

Since reports of creak in Cantonese T4 have been anecdotal, a first goal of this study was *to ascertain whether creak occurs frequently and systematically in T4 in Cantonese*. A second goal of this study was *to investigate to what extent Cantonese listeners use creak in native tonal perception*. Even if phonation information in the speech signal may be useful in tonal perception, it is not at all obvious that listeners use this information. There are only a handful of studies on the role of phonation when phonation is correlated with pitch. Among these few studies, evidence has been split—sometimes even within the same study—between whether or not listeners use phonation cues or only pitch cues. Belotel-Grenié and Grenié (1997); Brunelle (2009); Brunelle and Finkeldey (2011); Brunelle (2012); Garellek *et al.* (2013); Kuang (2013) found evidence that listeners do use phonation cues. However, Brunelle (2009); Brunelle and Finkeldey (2011); Brunelle (2012); Garellek *et al.* (2013) found evidence that listeners do not. It is thus important to add to the small body of case studies on the use of phonation cues in tonal perception so that we can begin to understand when listeners attend to phonation cues and when they do not.

In this study, we: (1) documented the tendency for creak in productions of T4 in Cantonese in a multi-speaker corpus of lab speech and (2) performed two perception experiments to demonstrate that Cantonese listeners use creak in native tonal perception. The first perception experiment tested whether natural occurrences of creak in Cantonese T4 improved listeners' accuracy in identifying T4 in a six-alternative forced choice task involving all six tones. The use of naturalistic stimuli in the first perception experiment left open the possibility that improvement in tonal identification in creaky T4s may have been due to concomitant pitch cues in those creaky stimuli, such as falling pitch movement.

The second perception experiment thus tested if the presence of creak in T4 biased listeners toward T4 responses when pitch cues were controlled. Listeners were asked to identify cross-spliced creaky and non-creaky stimuli as either the low fall T4 or the low level T6, the tone most confusable with T4 (Fok, 1974; Ma *et al.*, 2005; Khouw and Ciocca, 2007). Creaky stimuli were selected and resynthesized to eliminate the role of low and falling pitch cues as much as possible in monosyllabic stimuli. In disyllabic stimuli, f0 on a syllable preceding the target syllable was manipulated in an eight-step continuum to systematically control listeners' percept of pitch in the target syllable. Wong and Diehl (2003) previously showed that raising f0 on a preceding syllable with a fixed tone biased Cantonese listeners toward perceiving a lower level tone. Thus, we tested if listeners were biased toward T4 when creak was present even if lowered f0 on the preceding syllable biased them toward a T6 percept. Finally, because we manipulated variability in the quality of creak in a controlled way in the second experiment, we were also able to test if variability in the properties of the creaky region, such as the duration of the creaky region, affected listeners' biases toward a T4 response.

In the rest of this paper, we report on the documentation of creak in Sec. II and the two perception experiments in Secs. III and IV, and conclude with a general discussion (Sec. V).

II. EXPERIMENT 1: DOCUMENTATION OF CREAK IN CANTONESE T4

Previously, two small corpus studies have documented the occurrence of creak in Mandarin tonal production. Belotel-Grenié and Grenié (1997) documented that creak appeared in almost 80% of T3s in 204 syllables of lab speech from seven speakers (4M, 3F), compared to in 32% of T4s (51) and 52% of tones overall; Belotel-Grenié and Grenié (2004) examined a TV newscast of 298 syllables uttered by a single female and found creak in 26.5% of T3s, compared to 3.3% in T4s and 9.42% of tones overall. Together, these studies suggest that although the frequency of occurrence of creak may be highly variable across speakers and speech styles, the tendency for creak to be produced in Mandarin T3 is systematically higher than in other tones. The Cantonese corpus described in this section documents a systematically higher tendency for creak to be produced for Cantonese T4 than other tones. Sec. II A describes how the corpus was collected and analyzed for the occurrence of creak, and Sec. II B presents results of the corpus analysis.

A. Materials and methods

1. Materials

A corpus of Cantonese read speech was collected including all licit bitones in the language, following Xu (1997). The materials were designed to elicit contextual tonal variability for a larger study. It consisted of sentences [lei$^{25/35}$ jiu^{33} lau lau jak$^{\neg 3}$ tʃoeŋ^{33}/kap$^{\neg 3}$/sou^{33}] "*you want Lau-Lau to eat* {*sauce*/*pigeon*/*vegetarian*}" with the target bitone /lau-lau/ over all 36 licit combinations of tones T1 to T6.[4] Five fluent repetitions of each sentence were elicited from each speaker, for 180 tokens per tone and 1440 tokens in total for each speaker. The syllable /lau/ is a minimal sextuplet in terms of tone, although *lau*^{33} is uncommon. The words in the sentence were chosen to be sonorant to minimize segmental perturbation of f0. The words flanking the target bitone were chosen not only to be sonorant, but also to be mid level tones. This was done so that the f0 neighboring the target tones would have a neutral effect on inducing/not inducing the occurrence of creak due to low/high f0. The lexical meanings of the orthographic characters used for Tones T1–T6, respectively, were: "angry," "twist," "instigate," "stay," "willow," and "leak."

2. Participants

Eight native speakers (4M, 4F, 24.5 ± 4.7 yr) from Hong Kong and Macau were recruited from the student population at the University of California, Los Angeles; all spoke Cantonese on a daily basis. They were recorded in a sound-attenuated booth at the university using a Shure SM10A head-mounted microphone, whose signal ran through an XAudioBox pre-amplifier and A-D device. The recording was done using PCquirerX at a sampling rate of 22 050 Hz. Speakers were asked to think of the target bitones as proper names, since the bitone combinations formed nonce words. To control for speech rate, the speakers were asked to listen to a metronome beat during the production experiments with a beat of 20 bpm following Xu (1997), which resulted in speech rates around 3.2 syllables/s, comparable to speech rates in Belotel-Grenie and Grenie (1994). During the recording, a native speaker monitored utterances to check that tones were produced accurately and that the repetitions were fluent. Two native Cantonese speakers also checked recorded utterances to verify that the collected repetitions were fluent and that the tones were produced accurately.

3. Data analysis

Target syllables in individual tokens were determined to be creaky by listening and visual inspection of the waveform and spectrogram in Praat (Boersma and Weenink, 2010). A token was defined to be creaky if it had the auditory percept of creaky voice, as determined by the authors and if: (1) there were alternating cycles of amplitude and/or frequency or irregular glottal pulses in the waveform or wide-band spectrogram, (2) missing values or discontinuities in the f0 track determined by Praat's autocorrelation algorithm with default settings,[5] or (3) the appearance of strong subharmonics or lack of harmonic structure in the narrow-band spectrogram. Generally these three indicators occurred simultaneously. Examples of alternating cycles of frequency can be seen in the waveform in Fig. 1, top right, and in the wide-band spectrogram in Fig. 2, left. A clear example of irregular glottal pulses is the waveform in Fig. 1, bottom left. An example of strong subharmonics in the narrow-band spectrogram is Fig. 2, right. Examples explicitly illustrating the criteria for creak are available at http://www.krisyu.org/blog/posts/2014/06/supp-material-cantonese-creak-perception/#exp1.

Statistical analysis was performed in R (R Development Core Team, 2010). The probability of the occurrence of creak in the stimulus set was analyzed using mixed effects logistic regression implemented by the lme4 package of Bates and Maechler (2010).

For modeling the probability of the occurrence of creak, the logistic model included the following fixed effects: *tonal category* (Tone 4, not Tone 4), *prosodic position* in target bitone (second syllable, not second syllable), *speaker sex* (male, not male), and their interactions. Each of the fixed effects was a categorical variable and was mean-centered. To avoid anticonservativity, the random effects structure was chosen to be the maximal random effects structure justified by the experimental design that led to convergence (Barr *et al.*, 2013); this procedure resulted in the inclusion of random slopes for tonal category by subject as well as random intercepts by subject. Backward elimination was used to test for the inclusion of factors in the model. Model summaries presented throughout are from the final model after backward elimination. Listed *p*-values for fixed-effects coefficients are from Wald

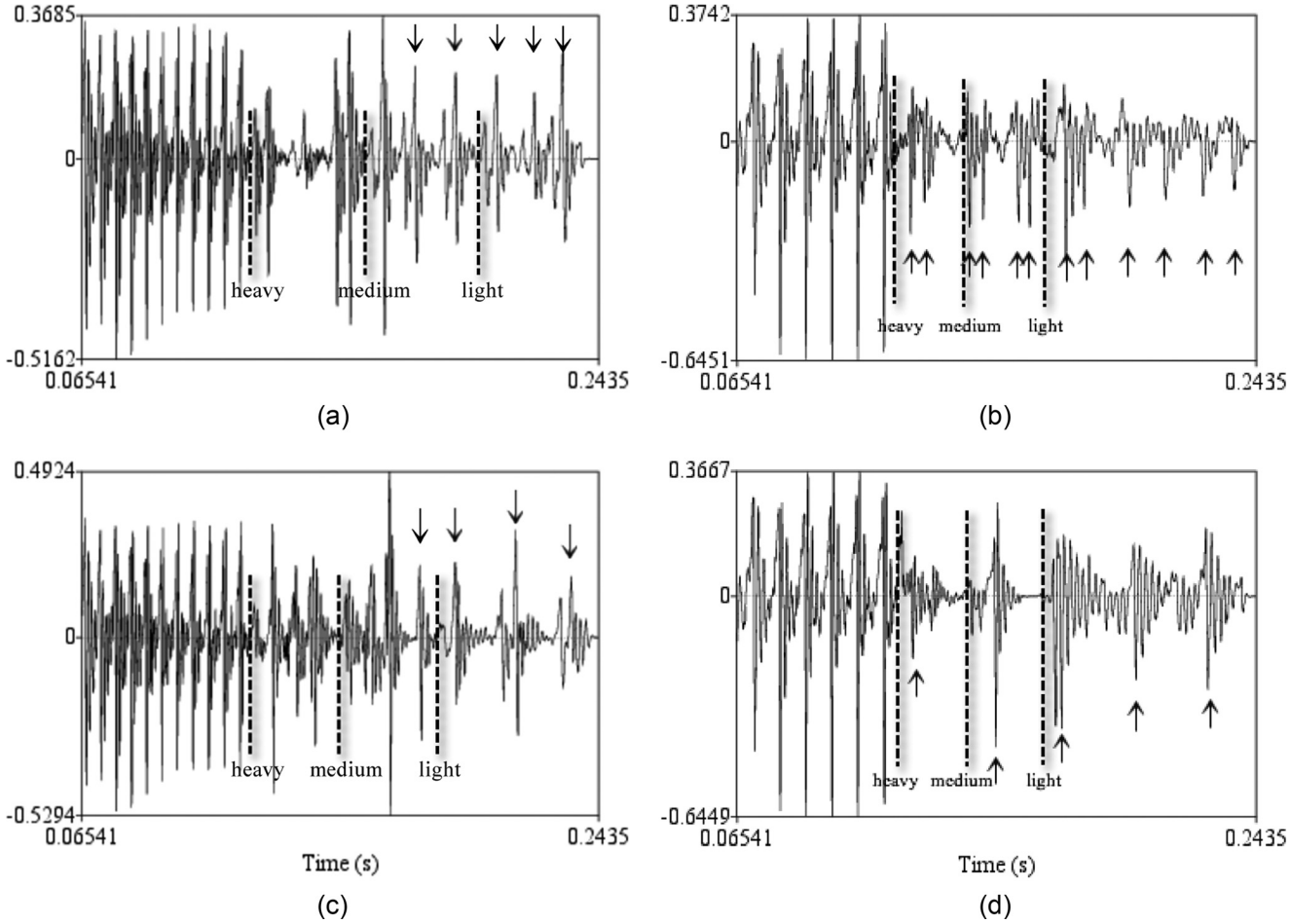

FIG. 1. Waveforms of /au/, from "narrow" pulse width (top) and "wide" pulse width (bottom) disyllabic stimuli for Experiment 3 with heavy creak proportion for the female (left) and male (right). Pulses are indicated with arrows. Splice points for "light," "medium," and "heavy" creak proportion are indicated with dashed lines; the creaky material included for each cross-spliced stimulus was the material to the right of the dashed line, and the material to the left of the dashed line came from the resynthesized T6.

z-statistics. Significance was determined at an alpha level of 0.05.

B. Results and discussion

Creak occurred on 24.2% (SD 17.6%) of the T4s in the corpus, compared to 4.7% overall in the corpus; the tone with the next highest incidence of creak was T2, with 2.6%. Thus, creak occurred an order of magnitude more often in T4 than in other tones, and our results confirm anecdotal reports noticing a tendency for T4 to be produced with creak.

The large standard deviation across speakers (17.6%) in the occurrence of creak on T4 reflects that the use of creak was highly variable across speakers. The high magnitude of variability in the occurrence of creak is consistent with the high magnitude of variability in the occurrence of creak in Mandarin T3 between the two Mandarin corpora discussed above (Belotel-Grenié and Grenié, 1997, 2004). Another source of variability in the incidence of T4 creak was whether the syllable was the first or second in the /lau-lau/ bitone. While creak occurred on 35.9% (SD 24.3%) of T4s in the second /lau/, it occurred in only 12.5% (SD 11.9%) of T4s in the first /lau/. A breakdown of percentage of creaky T4s by speaker and by syllable is given in Table I.

As shown in Table II, the mixed effects logistic model showed that the probability of the presence of creak on a tone was greater for T4 relative to the other tones ($p_{T4} = 1.6e^{-7}$) and that there was also an interaction between *tonal category* and *prosodic position* ($p_{T4\times S2} = 8.4e^{-6}$). Subset models unpacking interactions showed that: (1) the higher probability of creak on T4 generalized across both syllables ($p_{T4} < 2e^{-16}$ for both syllable 1 and syllable 2 data subsets), and (2) moreover, the probability of creak was higher in syllable 2 within T4s, but not within the other five tones.

The increased occurrence of creak on the second /lau/ may have been a reflex of marking the right edge of a prosodic constituent, e.g., final lowering. Indeed, the mean f0 measured over the vowel was significantly lower in the second syllable than the first, dropping 3.1 Hz on average (SD 3.2 Hz), paired t-tests by speaker: $t(7) = 2.72$, $p = 0.03$. [f0 was extracted using the STRAIGHT algorithm using VoiceSauce (Shue *et al.*, 2011) with mean f0 calculated from means over nine evenly divided intervals over the vowel. Intervals containing untrackable f0 regions were dropped from mean calculations.] To check for durational correlates of a prosodic phrase edge after the second /lau/,

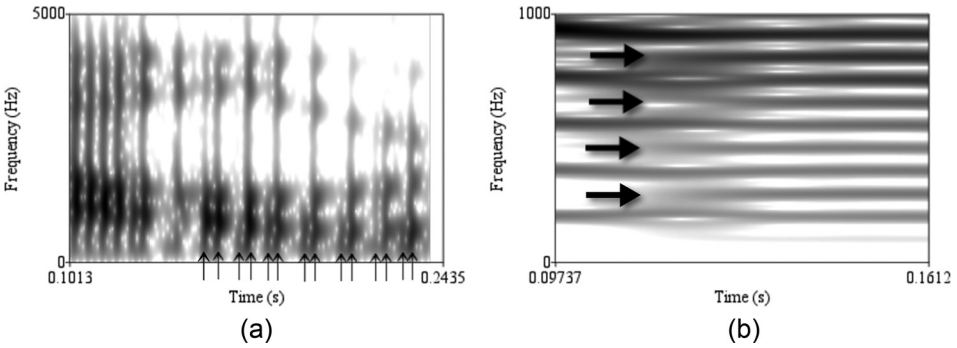

FIG. 2. Spectrograms of female disyllabic stimuli for Experiment 3. Wide-band spectrogram of /au/ from narrow width stimulus, showing doubled pulses, indicated by arrows (left). Narrow-band spectrogram of pitched stimulus, showing strong sub-harmonics, indicated by arrows (right).

we measured the duration of the rime, i.e., the vowel, since pre-boundary lengthening has been observed in the rime (Wightman *et al.*, 1992). However, the mean duration of the vowel was not significantly different between the two syllables [$t(7) = -2.3$, $p = 0.056$]; in fact, there was a mean decrease in length between the first and second vowels by 9.4 ms (SD 11.6 ms). Thus, there was no durational evidence of a prosodic edge following the second /lau/.

In sum, the corpus results suggest that the presence of creak might be informative in the identification of Cantonese T4 since there was a higher incidence of creak in T4 than in other tones. The next two sections present perceptual experiments showing that listeners do indeed use creak in identifying T4.

III. EXPERIMENT 2: CANTONESE TONAL IDENTIFICATION AND CREAKY VOICE IN TONE 4

Experiment 2 tested whether Cantonese listeners benefited from the presence of creak in T4 in naturally produced tones. We designed the tonal identification task to be very difficult to ensure listeners were not at ceiling. The stimuli were monosyllables sampled from the target bitones in the corpus described in Sec. II A 1. To represent a spectrum of naturalistic variability in tonal realization, they came from multiple speakers and varied in which tones had flanked the stimulus in the original utterance, as well as in their prosodic position in the original utterance. The contextual information available to the listener was minimal since the stimuli were monosyllables; the presentation of stimuli was randomized, and the identification task required choosing among all six tones of Cantonese.

TABLE I. Experiment 1 incidence of creak in T4s by speaker and prosodic position.

		Percent of creaky T4s (%)		
Speaker	Sex	Syllable 1	Syllable 2	Overall
1	F	17.8	43.3	30.6
2	F	1.1	6.7	3.9
3	F	15.6	44.2	29.5
4	F	18.9	46.7	32.8
5	M	0	8.9	4.5
6	M	35.6	76.7	56.1
7	M	4.4	47.8	26.1
8	M	6.7	13.5	10.1

A. Methods

1. Materials

The stimuli were 576 tokens of sentence-medial /lau/ syllables drawn from the corpus described in Sec. II A 1. An additional four speakers were recorded following the corpus study for a total of 12 speakers in the expanded corpus. From this corpus, 96 tokens were drawn for each of the six tones, 12 tokens per tone per speaker, from four males and four females who were chosen to be widely distributed in pitch range. The bitone sequence (e.g., T1-T1, T1-T2,…, T6-T6) and position of the syllable in the bitone (1st or 2nd) were balanced among the tokens from each speaker. For the T4 tokens, half of the tokens were chosen to be creaky and the other half non-creaky, following the criteria for defining a token to be creaky from Experiment 1. By "non-creaky," we do not necessarily mean "modal." Non-creaky tokens may have had a region of relatively low amplitude or breathiness. Because of great inter-speaker variability in the prevalence of creaky T4s, it was not possible to fully balance the presence of creak in T4s within a speaker, although this was balanced within sex. See Table III for the breakdown of the number of creaky and non-creaky T4s by pitch range and sex of speaker.

An expert in voice quality listened to the creaky/non-creaky subsets and confirmed that the tokens had/did not have the percept of creaky voice. Duration and amplitude were then controlled because this study was designed to single out the contribution of creaky voice as a cue. All tokens were resynthesized using Pitch Synchronous Overlap and Add (PSOLA) in Praat to have a duration of 313 ms, the grand mean of token durations, and the average intensity of each token was scaled to a constant value, 78 dB (relative to the auditory threshold).[6] Sample perceptual stimuli are available at http://www.krisyu.org/blog/posts/2014/06/supp-material-cantonese-creak-perception/#exp2.

TABLE II. Logistic model for probability of occurrence of creak in Experiment 1.

Factor	Coefficient β	SE(β)	z-score	p-value
Tone 4	3.802	0.726	5.236	$1.6e^{-7}$
Male speaker	0.301	0.933	0.322	0.747
Second syllable (S2)	0.575	0.215	2.671	0.008
Tone 4 × male speaker	−0.696	1.417	−0.491	0.624
Tone 4 × S2	1.339	0.301	4.455	$8.4e^{-6}$
Male speaker × S2	0.822	0.431	1.910	0.056
Tone 4 × S2 × male	−0.207	0.601	−0.344	0.730

TABLE III. Experiment 2 stimuli. Pitch range was estimated from monosyllabic tones.

Speaker	Sex	Pitch range (Hz)	Number of creaky/non-creaky T4s
1	F	175–266	6/6
2	F	134–255	6/6
3	F	122–263	6/6
4	F	147–263	6/6
5	M	95–159	2/6
6	M	80–140	6/6
7	M	90–130	7/7
8	M	82–119	9/5

2. Participants

The participants were 16 native Cantonese speakers recruited from the student population at the University of California, Los Angeles; all spoke Cantonese on a daily basis. They received cash compensation. All were born in Hong Kong except one born in Macau, and their mean time of stay in the U.S. was 4.0 ± 1.8 yr. There were 11 males (age 20.6 ± 1.6 yr) and five females (age 21.2 ± 0.8 yr). Two other participants were excluded from analysis due to equipment failure.

3. Procedure

Participants were tested in a sound-attenuated booth. The perception experiment was run in Matlab using Psychophysics Toolbox extensions (Brainard, 1997). Stimuli were played from an Echo Indigo IO sound card on a laptop over Sony MDR-7506 studio monitor headphones at a comfortable volume fixed across subjects, and the responses and reaction times of the subjects measured from the onset of the stimulus were recorded. Reaction time results were not found to be significant and are not further reported. The interstimulus interval was 3 s.

The task of the participants was to identify each stimulus by a keyboard press of one of six keys labeled with the characters for the minimal tonal set over *lau*. The participants were told that the stimuli were extracted from sentences *lei*$^{25/35}$ *jiu*^{33} *lau lau yak*^{3} *sou*^{33} and that the sentences were read by multiple speakers. They were asked to respond as quickly and accurately as possible and told they would be timed. The order of the stimulus presentation as well as which key was labeled with which word was randomized across participants, and participants received three short breaks during the experiment, which took about 45 min.

4. Data analysis

The overall confusion matrix for all stimuli was calculated and further analysis was performed on the T4 stimulus subset. Data were excluded from analysis for one T4 stimulus which yielded long reaction times that were outliers and sounded highly unnatural after resynthesis. Correctness of identification in the T4 stimulus subset was analyzed using mixed effects logistic regression following the procedures outlined in Sec. II A 3.

For modeling the probability of a correct T4 response, the logistic model included the following fixed effects: *presence of creak* (present, absent), *prosodic position* in target bitone (second syllable, not second syllable), *speaker sex* (male, not male), and their interactions.

B. Results

Overall, identification accuracy for T4 was high (70.51%, SE = 10.93%) compared to that of other tones (see the overall confusion matrix in Table IV), and the tone most confusable with T4 was T6.[7] Breaking down the 70.51% identification accuracy for T4 by phonation, identification accuracy for T4 was 82.03% (SE = 2.27%) for creaky T4s but only 58.98% (SE = 3.57%) for non-creaky T4s. Confusion of creaky T4s with T6 was also 13% lower than confusion of non-creaky T4s with T6. Broken down by speaker sex and the prosodic position the stimulus was drawn from (syllable 1 or 2), T4 identification accuracy jumped from 50.5%–53.6% for non-creaky T4s to 80.1%–82.8% for both male and female stimuli from syllable 2, as well as for male stimuli from syllable 1. However, for female stimuli from syllable 1, T4 identification accuracy was already 79.2% for non-creaky T4s, compared to 87.0% for creaky T4s.

The results of the logistic regression are shown in Table V. There was a significant effect of *creak*, as well as a significant interaction *creak* × *prosodic position* × *male speaker*. To unpack this interaction and to check if the effect of creak generalized across the whole data set, additional logistic regressions for *creak* × *male speaker* were performed for data subsets partitioned by prosodic position. In logistic models for both the first and second syllables, the presence of creak had a positive regression coefficient and reached significance at the 0.05 level (first syllable: $\beta = 1.38$, $p = 6.0\text{e-}6$, second syllable: $\beta = 1.84$, $p = 0.03$). There were no other significant effects for the second syllable subset model. But the first syllable subset model also showed a trend for male speaker ($\beta = -0.67$, $p = 0.06$) and a significant effect for male speaker × creak ($\beta = 1.27$, $p = 0.02$). Logistic models for subsets partitioned by speaker sex within the first syllable showed a significant effect of creak for male stimuli ($\beta = 1.91$, $p = 8.2e^{-5}$), but not female stimuli ($\beta = 0.53$, $p = 0.37$).

Thus, the logistic regressions reflect that creaky T4s drawn from the second syllable of the target bitone were identified with higher accuracy than non-creaky T4s,

TABLE IV. Confusion matrix for tones in Experiment 2.

	Tonal identification response (%)					
Actual	T1	T2	T3	T4	T5	T6
T1	**85.94**	0.85	8.53	0.07	1.04	3.58
T2	1.30	**35.09**	3.65	1.50	55.99	2.47
T3	27.28	1.04	**32.23**	3.39	3.12	32.94
T4	1.89	3.19	2.93	**70.51**	8.07	13.41
T5	2.34	9.77	7.68	5.34	**63.93**	10.94
T6	8.46	1.76	20.25	15.56	4.56	**49.41**

TABLE V. Logistic model for probability of correctness in identifying T4 in Experiment 2.

Factor	Coefficient β	SE(β)	z-score	p-value
Second syllable in bitone	−0.3785	0.5232	−0.723	0.4694
Male speaker	−0.4228	0.2714	−1.558	0.1193
Presence of creak	1.7300	0.3308	5.229	$1.7e^{-7}$
Second syllable × male speaker	0.9309	0.4977	1.871	0.0614
Creak × second syllable	0.6668	0.6382	1.045	0.2961
Creak × male speaker	0.2055	0.3848	0.534	0.5934
Creak × second syllable × sex	−1.7383	0.7673	−2.266	0.0235

regardless of speaker sex. For T4s from the first syllable, creaky T4s were identified with higher accuracy for the male but not the female stimuli.

C. Discussion

Experiment 2 showed that the presence of creak in the utterance of a Cantonese T4, the lowest tone in Cantonese, increased the probability of it being identified correctly in native tonal perception. This is positive evidence for listeners' use of phonation cues in a tone language where phonation cues are secondary to pitch cues, as well as the first evidence we are aware of that phonation cues can improve tonal identification accuracy. This result builds on Belotel-Grenié and Grenié's (1997) results showing an earlier isolation point for Mandarin T3 when creak was present. While listeners in Belotel-Grenié and Grenié (1997) were already at ceiling in their perceptual task, thus limiting the potential beneficial role of phonation cues, our task was designed to be difficult through the choice of a language with a large tonal inventory and the multiple sources of variability in the stimuli. The low accuracies in the confusion matrix in Table IV reflect that listeners in our task were indeed not at ceiling overall.

The one subpart of the stimulus set where creaky T4s were not identified significantly more accurately than non-creaky T4s was for stimuli drawn from the first syllable in the target bitone from female speakers. Identification accuracy of creaky T4s for these stimuli was higher than for non-creaky T4s, 87% vs 79%, though not significantly different. Accuracy was already close to ceiling for these non-creaky T4s, leaving less room for the benefit of creak. (For the rest of the stimulus set, non-creaky T4s were identified with 50% accuracy and creaky T4s with 80% accuracy.) The high accuracy for this particular subset of non-creaky T4s may have been driven by the fact that it happened to contain very few tokens with ambiguous pitch cues for T4 such as a flat or slight rise in the pitch contour. Such tokens were more numerous in the other non-creaky T4 stimuli. This brings up the possibility that, in general, in the naturally variable realizations of T4, creaky T4s were also instances where pitch was low and/or falling in non-creaky portions of the stimulus, while non-creaky T4s had more ambiguous pitch cues.

Experiment 2 showed that the presence of creak in T4 could aid its correct identification in the face of natural variability in tonal realization. However, like any other study using naturally produced stimuli [e.g. Belotel-Grenié and Grenié (1997)], it could not isolate the role of creaky voice from the role of concomitant pitch cues. Experiment 3 was designed to complement the results of the naturalistic study and to better isolate the role of creaky voice as a cue in T4 perception.

IV. EXPERIMENT 3: RESYNTHESIS OF F0 AND CREAKY VOICE QUALITY

In Experiment 3, we tested if the presence of creak biased listeners toward identifying a tone as T4 even when concomitant f0 cues were controlled.

For this purpose, we resynthesized and cross-spliced speech materials from the production corpus described in Sec. III A 1 to generate stimuli with controlled contextual pitch and creaky voice cues. These included both monosyllabic and disyllabic stimuli. The monosyllabic stimuli allowed us to test for listeners' sensitivity to creak in T4 perception in the absence of absolute low pitch cues and pitch movement cues. (Suppose pitch at a point in an utterance is so low that it could only be at the lower bound of any person's pitch range, i.e., absolute f0 at this point is near the lower bound of the human pitch range in speech. Then, listeners might use this absolutely low pitch as a cue for T4, since T4 canonically has the lowest pitch target among Cantonese tones. Thus, if a creaky T4 that is identified as T4 with high accuracy is also produced with absolute low pitch, it is unclear that it is specifically the creak that is influencing T4 perception, rather than the pitch. Similarly, if a T4 production that is identified as T4 with high accuracy is produced with a falling pitch, it is unclear that it is specifically the creak rather than the falling pitch that is influencing tonal perception, since T4 is canonically a fall.) The disyllabic stimuli allowed us to probe listeners' sensitivity to creak while we systematically controlled listeners' percepts of different pitch levels in the target syllable by manipulating f0 in the preceding syllable [following Wong and Diehl (2003)].

A. Methods

1. Materials

Productions of the Cantonese syllable /lau/ for T4 and T6 were selected from one male and one female speaker with high pitch ranges out of the production corpus described in Experiment 2 (Sec. III A 1) to preclude the availability of absolute low pitch as a cue for T4. For each speaker, the utterance with the lowest level f0 contour instance of lau^{22} immediately following the sentence frame $lei^{25/35}$ jiu^{33} was selected. Inspection of the production corpus indicated that T6 as well as T4 occurred with level f0 variants, and we did not want an f0 fall over the target syllable because that would introduce f0 information that might bias the listener toward a T4 response.

Three additional utterances were selected for each speaker to exhibit a controlled range of variation in creaky realizations of T4 immediately following $lei^{25/35}$ jiu^{33}. The range of variation was chosen to explore ways to further tease apart the role of creaky voice and pitch cues. Because

the vocal fry mechanism is contingent on a low f0 (Gerratt and Kreiman, 2001), we selected instances of period doubling and avoided instances of vocal fry. Period doubling is nonmodal phonation with "pairs of vocal cycles alternating in period and/or amplitude" in which a pitch percept is ill-defined due to bitonality (Gerratt and Kreiman, 2001), e.g., see doubled pulses in the wide-band spectrogram in Fig. 2, left. A stimulus was considered period doubled based on narrow-band spectrographic evidence of subharmonics, see Fig. 2, right. Because a pitch in period doubled speech might still be detected based on the pulse width between glottal pulses, we included period doubled stimuli with varying pulse widths and strength of pitch percept. One had a wider pulse width ("wide," Fig. 1, bottom row), another a narrower width ("narrow," Fig. 1, top row), and one had a very clear and audible pitch percept and its f0 was trackable by Praat's autocorrelation algorithm ("pitched," Fig. 2, right) (Boersma and Weenink, 2010). For the male speaker, no "pitched" stimulus could be found, so a "pitched" stimulus from another male speaker was used.

The utterances were processed and resynthesized in Praat. The disyllable *jiu lau* was extracted for each utterance, and the f0 of the utterance was resynthesized using PSOLA as follows: (1) the absolute f0 of the diphthong /au/ was resynthesized to a constant value ambiguous between T4/6 for high pitch range males/females in the production corpus (180 Hz for the female; 107 Hz for the male); (2) the f0 of /jiu/ preceding /lau/ was resynthesized to be 31 Hz higher than the f0 of /au/, a relative f0 difference ambiguous between the f0 drop from T3 to T4/T6 in the production corpus for high pitch range speakers, and then incremented upward and downward in half-semitone steps from 1.5 semitones below to 2 semitones above that (the value 31 Hz above the f0 in /au/) in an eight-step continuum; (3) f0 was linearly interpolated in Hz over /l/ between the offset of /jiu/ and onset of /au/.

The creaky T4 /au/s were cross-spliced with the T6 in /jiu lau/ utterances. For each creaky /au/ token, three splice points were chosen to manipulate creak proportion: "heavy," "medium," and "light," where "light" included the last three pulses from the creaky region of /au/; "heavy" included the maximal amount of material from the /au/ that was creaky; "medium" was set at the approximate midpoint between the other two (see Fig. 1). While all other stimuli included creaky regions persisting through the end of the syllable, the end of the /au/ for the male "narrow" stimulus included a few regular pulses at the end of the syllable which were not counted as part of the creaky region of the /au/. Therefore, the splice point for "light" proportion for this stimulus included more pulses than for other "light" stimuli. Both authors and another trained phonetician checked that creak was audible for all creaky stimuli. All splice points were taken at the nearest zero crossing in a low amplitude region.

Durations of /jiu/, /l/ and /au/ were equalized to their averages between all utterances to facilitate naturalistic cross-splicing and also to standardize durations for reaction time measurements. The average intensity of /jiu lau/ for the two non-creaky T6 utterances was resynthesized to 78 dB (relative to the auditory threshold). The average amplitude of the /au/s extracted from the creaky T4 utterances was set to be slightly lower at 75 dB because otherwise the creaky portion of the utterance sounded unnaturally loud so that there was not a continuous percept across the splice boundary.

The disyllable stimulus set included 3 repetitions of each modal stimulus, and 2 *male speaker* (male, not male) $\times$ 8 *contextual f0 shift* (8 f0 levels) $\times$ 12 different qualities of creak, i.e., 3 *creak type* (wide, narrow, pitched) $\times$ 3 *creak proportion* (heavy, medium, light) + 3 repetitions of modal stimuli, for 192 stimuli in total. The monosyllable stimulus set was balanced between creaky and modal stimuli: 2 *male speaker* $\times$ (3 *creak type* $\times$ 3 *creak proportion* + 9 modal stimuli) for 36 stimuli in total. Monosyllable /lau/ stimuli were extracted from the /jiu lau/ stimuli. While the creaky stimuli were drawn from the "0 st" f0 shift level, the 9 modal stimuli were drawn across f0 shift levels from 2 semitones (st) above the 0 st level to 2 st below so that listeners were not exposed to only a single modal stimulus token 9 times. Thus, there was some variability in the fall over the /l/ in the modal stimuli not present in the creaky stimuli. This variability showed no significant effect when modeled in statistical analyses (see Sec. IV B 1).

Sample perceptual stimuli are available at http://www.krisyu.org/blog/posts/2014/06/supp-material-cantonese-creak-perception/#exp3.

2. Participants

The participants were 20 native Cantonese speakers who were born and raised in Hong Kong and currently living there. There were 10 males (age 20.3 $\pm$ 1.9 yr) and 10 females (age 21.8 $\pm$ 1.7 yr). They were recruited from the local Hong Kong university student population and received cash compensation.

3. Procedure

Participants were tested in a sound-attenuated booth. The perception experiment was run as in Experiment 2, except that the task of the participants was to identify each stimulus by a keyboard press of either a key labeled with the character for lau^{21} (a common family name) or one labeled with lau^{22} "drip." Participants were asked to respond as quickly and accurately as possible and told they would be timed. Each participant heard two replicates of the stimulus set. The order of the stimulus set within each replicate, as well as which key was labeled with which word, was randomized across participants. Participants received a short break between stimulus sets. Testing took about 30–40 min.

Participants were told that the stimuli were extracted from sentences $lei^{25/35}$ jiu^{33} zi^{22} "You want word" in the discourse context of looking up a word in a dictionary and a sheet with the sentence was placed before them during the experiment. Participants were also told that there was more than one speaker, that the speakers were asked to say the sentences in different pitch ranges, and that the relative proportions of the two different words played was randomized (so they would not know what proportion to expect).

4. Data analysis

Statistical analyses involved mixed effects regression, following the procedures outlined in Sec. II A 3. Contrasts for *creak proportion* were coded with forward difference coding (light vs medium, medium vs heavy) and *creak proportion* with treatment contrasts with "wide" creak as the reference level. Following Wong and Diehl (2003, p. 417), *(contextual) f0 shift* was treated as a continuous, interval-scale variable because it was based on the semitone scale. *Replicate* was included as a noninteracted fixed effect covariate (because each stimulus set was presented twice). Control predictors justified by experimental design such as *replicate* were included in final models (Barr *et al.*, 2013). Because listeners showed systematically different patterns of behavior for the male and female stimulus sets, models were typically fitted to each of the two stimulus sets separately.

B. Results

Results are presented organized by research question.

1. Does the presence of creak bias listeners toward T4 responses?

Analysis of the monosyllable data indicated that the presence of creak biased for T4 responses: (1) in the absence of absolute low pitch cues either from f0 or from vocal fry, and (2) in the absence of pitch movement cues which may have been present in unequal amounts between the creaky and non-creaky stimuli in the naturalistic stimulus set in Experiment 2. Figure 3 shows that the proportion of T4 responses was higher in the presence of creak than in the absence of creak for both the male and female stimuli sets.

The logistic model for the effect of creak on the probability of a T4 response for the monosyllable stimuli is given in Table VI. There was a significant effect for the presence of creak ($\beta = 2.78$, $p = 6.6e^{-7}$). There were no other significant effects; in particular, the variability in the fall over the /l/ in the non-creaky stimulus (f0 shift nested in the presence of creak) showed no significant effect.

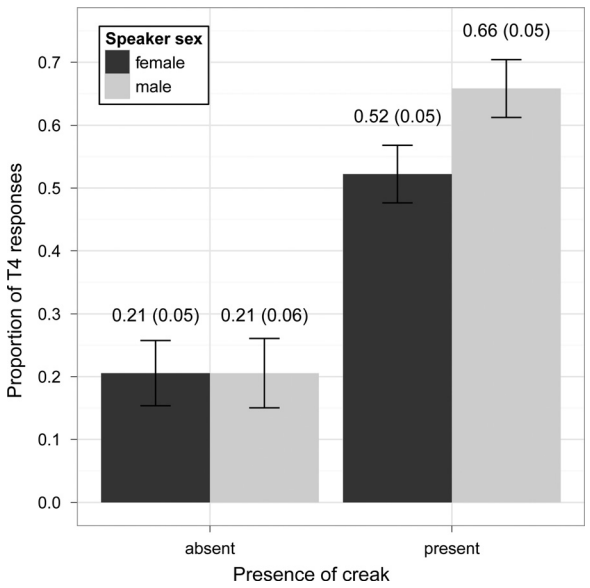

FIG. 3. Overall proportion of T4 responses conditioned on presence of creak and stimuli speaker sex for monosyllabic stimuli in Experiment 3. A higher proportion of creaky stimuli than non-creaky stimuli was identified as T4 for the male and female stimuli sets. Error bars show ±1SE.

TABLE VI. Summary of mixed logit model for creaky monosyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
Male speaker	0.41	0.57	0.7	0.47
Presence of creak	2.77	0.56	5.0	$7.1e^{-7}$
Replicate	−0.02	0.23	−0.1	0.92
f0 shift in creak	−0.13	0.44	−0.3	0.77

As a more stringent test for the effect of creak, we also checked if the effect that the presence of creak biased listeners toward T4 responses held even for only the "light" stimuli—those with the shortest duration of creak in the stimulus set. With only these creaky stimuli included, a logistic model with an identical structure to the one in Table VI showed a significant effect for creak ($\beta = 1.45$, $p = 0.002$), with no other significant effects. Thus, the presence of even a few pulses of creak biased listeners toward T4 responses for both the male and female speaker stimuli.

2. Does contextual f0 information bias responses?

Before we could study how listeners weighed relative f0 cues and creaky voice cues, we needed to establish that we successfully replicated Wong and Diehl's (2003) effect of contextual f0 on tone perception. To do this, we analyzed the non-creaky disyllabic stimulus subset. Figure 4 plots the

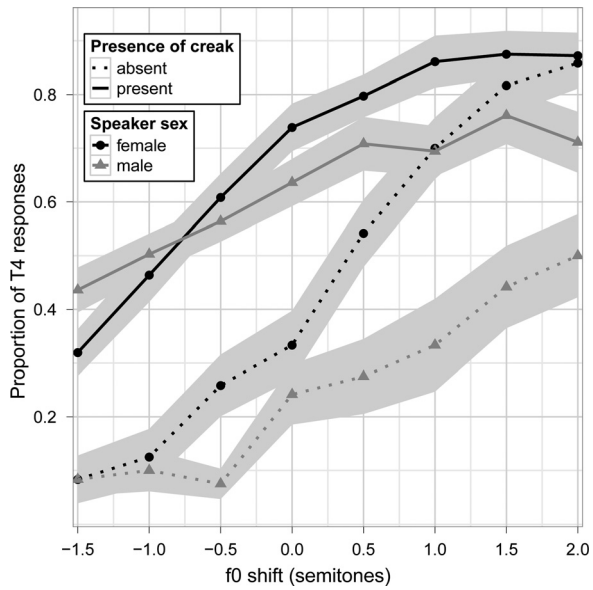

FIG. 4. Overall proportion of T4 responses as a function of contextual f0 shift conditioned on the presence of creak and speaker sex for disyllabic stimuli, from Experiment 3 aggregated across listeners. The 0 point for contextual f0 shift indicates the base resynthesized f0 level (31 Hz higher than fixed f0 in the target syllable), from which the contextual f0 shift continuum was created in increments of half-semitones. Ribbons show ±1 SE. For the non-creaky stimuli, the response curve for female stimuli is much steeper than for male stimuli; for both the male and female stimuli, the response curve is less steep for the creaky stimuli and globally shifted upward from the response curve for the non-creaky stimuli.

TABLE VII. Summary of mixed logit model for non-creaky male disyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	1.16	0.22	5.2	$2.27e^{-7}$
Replicate	−0.74	0.26	−2.8	0.0044

proportion of T4 responses as a function of f0 shift. The positive slope of T4 response curves for the non-creaky stimuli (indicated by dashed lines) in the plot indicates that as f0 on the preceding syllable became higher, pitch on the target syllable was perceived to be lower, driving listeners toward T4 responses.

For both the male and female stimulus subsets, logistic models with *f0 shift* and *replicate* as fixed effects showed a significant effect for f0 shift (male: $\beta_{\text{f0shift}} = 1.16$, $p = 2.27e^{-7}$; female: $\beta_{\text{f0shift}} = 1.92$, $p = 8.57e^{-16}$), see Tables VII and VIII. Thus, the probability of a T4 response increased with f0 shift in the non-creaky disyllabic stimuli. These results replicate and extend Wong and Diehl's (2003) findings that higher f0 on the preceding syllable biased listeners toward perceiving lower Cantonese level tones.

For the male stimuli, there was also a significant effect for *replicate* ($\beta = -0.74$, $p = 0.004$) indicating a lower probability of T4 response in the second replicate. Perhaps some subjects became sensitized to the presence of creak and this caused them to weigh the presence of creak more heavily in deciding on a T4 response.

3. How do listeners weigh creaky voice cues together with relative pitch cues?

The presence of creak biased listeners toward T4 even in the presence of relative pitch cues, and the effect of creak varied across the f0 shift continuum. As shown in Fig. 4, the response curves for the creaky stimuli were both globally shifted upward from and flatter than those for the non-creaky stimuli. For the female stimuli, at the endpoint of the f0 shift continuum biasing for T4, the effect of creak had little effect beyond the biasing effect of f0 shift. It appears that the listeners were already at a saturation point from the f0 shift bias (with about 90% T4 responses), leaving little room for an additional effect of creak. For the male stimuli, the response curves for creaky vs non-creaky stimuli across the f0 shift continuum are closer to being parallel. At the endpoint of the f0 shift continuum biasing for T4 in male stimuli, listeners were not at a saturation point (with about 50% T4 responses), and the presence of creak further biased listeners toward T4 responses.

Logistic models for the full disyllabic stimulus set (Tables IX and X) showed significant effects for the presence of creak (male: $\beta_{\text{creak}} = 2.65$, $p = 2.14e^{-9}$; female: $\beta_{\text{creak}} = 1.83$, $p = 6.8e^{-6}$) in addition to significant effects for f0 shift. There was also a significant interaction of f0 shift and creak for the female stimuli ($\beta_{\text{f0}\times\text{creak}} = -0.83$, $p = 2.8e^{-4}$), and a trend for an interaction in the male stimuli ($\beta_{\text{f0}\times\text{creak}} = -0.53$, $p = 0.052$). The significant interaction for female stimuli reflects the limited effect of creak at the endpoint of the f0 shift continuum biasing for T4. The non-significant trend for an interaction is reflected in the nearly parallel response curves for creaky vs non-creaky male stimuli.

TABLE VIII. Summary of mixed logit model for non-creaky female disyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	1.92	0.24	8.0	$8.57e^{-16}$
Replicate	0.03	0.25	0.1	0.91

TABLE IX. Summary of mixed logit model for f0 shift × creak interaction, male disyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	0.78	0.19	4.2	$3.16e^{-5}$
Presence of creak	2.65	0.44	6.0	$2.14e^{-9}$
Replicate	−0.10	0.17	−0.6	0.56
f0 shift × presence of creak	−0.53	0.27	−1.9	0.052

To further check how listeners pit relative pitch cues vs creaky voice cues, we analyzed responses at the endpoint of the f0 shift continuum biasing for T6 (f0 shift = −1.5, −1.0 st) (Tables XI and XII). Here, the pitch and phonation cues were in conflict. Logistic models for this region of the f0 shift continuum show a significant effect for creak (male: $\beta_{\text{creak}} = 3.08$, $p = 1.72e^{-4}$, female: $\beta_{\text{creak}} = 3.18$, $p = 2.07e^{-5}$). Within this region of the f0 shift continuum, the proportion of T4 responses jumped from 10.4% to 39.2% for female stimuli and from 9.2% to 46.9% for male stimuli when creak was present. This shows that as the speech signal unfolded, creaky voice cues could shift listener's responses in the opposite direction from that of previous pitch cues.

4. Do listeners show sensitivity to glottal pulse width and duration of nonmodal phonation?

The controlled variation in creak quality in the stimulus set also presented the opportunity to check for effects of details of creak quality on biasing listeners toward T4. We explored the effects of creak quality in the monosyllabic subset. The effect of *creak proportion* on T4 responses in monosyllables is shown in Fig. 5. The figure shows that the proportion of T4 responses increased as the proportion of creak in the stimulus increased.

Logistic models for *creak proportion* and *creak type* showed significant effects for *creak proportion* for both male and female stimuli, and an effect for *creak type* for the male stimuli, see Tables XIII and XIV. (Backward

TABLE X. Summary of mixed logit model for f0 shift × creak interaction, female disyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	1.36	0.16	8.3	$<2e^{-16}$
Presence of creak	1.83	0.41	4.5	$6.8e^{-6}$
Replicate	0.08	0.13	0.6	0.57
f0 shift × presence of creak	−0.83	0.23	−3.6	$2.8e^{-4}$

TABLE XI. Summary of mixed logit model for the effect of creak, for f0 shift of −1.5 to −1 st, male stimuli.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	0.85	1.20	0.7	0.48
Presence of creak	3.08	0.82	3.8	$1.72e^{-4}$
Replicate	0.33	0.21	1.6	0.10

elimination did not support the inclusion of an interaction between *creak type* and *proportion*.) For the male stimuli, the probability of a T4 response significantly increased from light to medium ($\beta = 1.74$, $p = 8.27e^{-6}$) and from medium to heavy creak proportion ($\beta = 1.12$, $p = 8.1e^{-3}$). For the female stimuli, the probability of a T4 response significantly increased only from light to medium creak proportion ($\beta = 1.77$, $p = 1.2e^{-4}$). There were no significant effects for *creak type* for the female stimuli. However, for the male stimuli, the probability of a T4 response significantly increased for the narrow pulse width stimulus relative to the wide pulse width stimulus and for the wide and narrow pulse widths relative to the "pitched" stimulus.

C. Discussion

In Experiment 3, we built on the result from Experiment 2. We showed that listeners were sensitive to creak in Cantonese tonal perception with a more controlled experiment where listeners were tasked to choose between T4 and its most confusable tone, T6. We controlled for f0 preceding the region of nonmodal phonation, creating an eight-step continuum of f0 on the syllable preceding the syllable to be identified, and interpolated from that f0 through the onset consonant to a constant f0 at the onset of the target vowel, in an f0 range ambiguous between T4 and T6. In the target vowel, we used cross-splicing to control the duration of the nonmodal region, as well as characteristics of the glottal pulse train. The nonmodal regions were period doubled and thus did not have a clear low absolute f0 percept.

Under these controlled conditions, we found, first, that the presence of creak biases listeners toward T4 responses in the absence of absolute low pitch and pitch movement cues that may have been present in the naturalistic T4 stimuli in Experiment 2. This bias was present even for the subset of creaky stimuli with the shortest duration of nonmodal phonation, the stimuli with a "light" proportion of creak.

We also replicated and extended results of the effect of f0 in preceding context in tone perception. Wong and Diehl (2003) previously showed that preceding f0 strongly biases listener responses for the Cantonese level tones, T1, T3, and T6. Huang and Holt (2009) showed that preceding f0 also affected perception of the T2 rise in Mandarin, demonstrating an effect of preceding context on contour tone perception. Here, the proportion of T4 responses increased as f0 on the preceding syllable increased. As described in Sec. IV A 1, we exploited allophonic variation which can cause T4 and T6 both to be realized with level f0 contours, although T4 is considered a fall based on its citation form, while T6 is considered a level tone. Thus, we showed that preceding context affects contour tone perception in Cantonese, although we tested listeners on variants of T4 which were level over the rime.

We also found that the presence of creak could shift listeners in the opposite direction in tonal identification from relative pitch cues in the preceding syllable. When the f0 on the preceding syllable was at its lowest in the continuum, it biased listeners toward a T6 response. However, if creak was also present, listeners sometimes changed their response to T4 instead. Overall, the presence of creak diminished the effect of the preceding f0 context on tonal perception: the T4 response curves for creaky stimuli as a function of preceding f0 were flatter than those for the non-creaky stimuli. The

TABLE XII. Summary of mixed logit model for the effect of creak, for f0 shift of −1.5 to −1 st, female stimuli.

Factor	Coefficient β	SE(β)	z-score	p-value
f0 shift	1.64	0.93	1.8	0.078
Presence of creak	3.18	0.75	4.3	$2.07e^{-5}$
Replicate	0.15	0.19	0.8	0.41

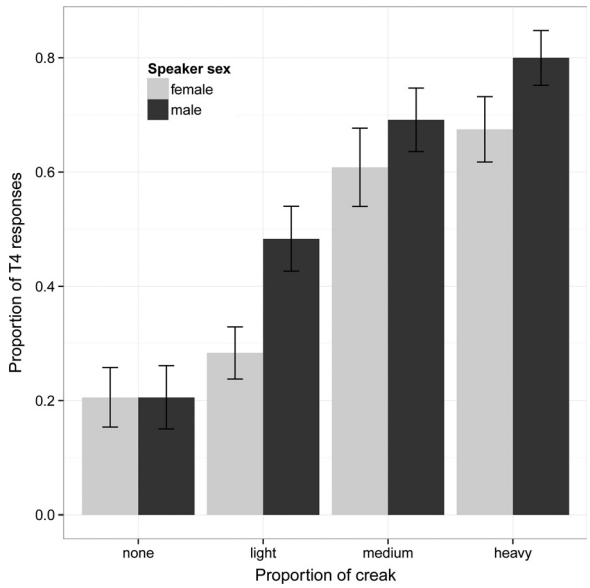

FIG. 5. Overall proportion of T4 responses conditioned on the presence of creak type for male and female creaky monosyllabic stimuli, aggregated across listeners. Error bars show ±1SE. The proportion of T4 responses is significantly higher for "light" proportion vs no creak and "medium" vs "light" proportion for both the male and female stimuli. Only in the male stimuli is the proportion of T4 responses significantly higher in the "heavy" proportion than the "medium" proportion.

TABLE XIII. Summary of mixed logit model for creak quality in creaky male monosyllables.

Factor	Coefficient β	SE(β)	z-sore	p-value
Creak proportion (light-medium)	1.74	0.39	4.5	$8.27e^{-6}$
Creak proportion (medium-heavy)	1.12	0.42	2.6	$8.1e^{-3}$
Creak type (pitched-wide)	−2.31	0.38	−6.1	$1.18e^{-9}$
Creak type (narrow-wide)	1.94	0.44	4.4	$1.22e^{-5}$
Replicate	−0.06	0.32	−0.2	0.85

presence of creak did not simply produce flat response curves with responses to T4 at ceiling over the f0 shift continuum. This suggests that the presence of creak was not interpreted simply as absolute low f0, which should have strongly biased listeners toward T4 responses.

Finally, Experiment 3 provided evidence that listeners are sensitive to the duration of the region of nonmodal phonation in a creaky tone. For both the female and male monosyllable stimulus sets, listeners showed a greater probability of a T4 response with longer durations of the region of nonmodal phonation. One interpretation of this is that longer duration may make listeners more confident that the creak is linguistically meaningful, rather than being due purely to something about the particular physiological state of the speaker. To our knowledge, this is the first result showing sensitivity of listeners to details of creak quality.

The evidence for an effect of details of the pulse train such as pulse width are not as clear. We expected widely spaced glottal pulse stimuli to provide a relatively lower pitch percept than the narrowly spaced glottal pulse stimuli and thus favor a T4 response. But in the male stimuli, the narrow width stimuli actually yielded a higher T4 response proportion than the wide width stimuli. This may have been because the male narrow width stimuli had the longest durations of material cross-spliced from the creaky T4. In the male stimuli, there was also a lower T4 response proportion for the "pitched" stimuli than the other two stimuli types. But this may have been because we cross-spliced speech produced by another vocal tract to create these stimuli. There were no effects of *creak type* in the female monosyllabic stimuli.

V. GENERAL DISCUSSION

In Experiment 1, we ascertained that creak occurs more often on Cantonese T4 than other tones in a multi-speaker corpus of lab speech, suggesting that creak might provide a useful cue in tonal perception. We then showed that listeners are sensitive to the presence of creak in native lexical tone perception in Cantonese, a tone language where pitch is the primary dimension of tonal contrast. Experiment 2 demonstrated that naturally produced creaky T4s were identified with 20% more accuracy than non-creaky T4s. Previous experiments on the presence of creak in Mandarin tonal perception could not demonstrate improvement in tonal identification accuracy because listeners were already at ceiling in the task. Experiment 3 showed that the presence of creak biased Cantonese listeners toward perceiving tones as T4 even in the absence of low absolute f0 cues within the syllable, and when any falling pitch movement in the syllable was accounted for in the analysis. Furthermore, when relative pitch cues from the preceding syllable biased listeners toward perceiving T6, the most confusable tone with T4, the presence of creak could still shift listeners toward perceiving T4. Finally, listeners were more likely to perceive T4 when the region of nonmodal phonation in a creaky stimulus was longer.

TABLE XIV. Summary of mixed logit model for creak quality in creaky female monosyllables.

Factor	Coefficient β	SE(β)	z-score	p-value
Creak proportion (light-medium)	1.77	0.46	3.8	$1.2e^{-4}$
Creak proportion (medium-heavy)	0.26	0.35	0.7	0.46
Creak type (pitched-wide)	−0.04	0.22	−0.2	0.85
Creak type (narrow-wide)	−0.20	0.25	−0.8	0.42
Replicate	0.13	0.25	0.5	0.60

All together, our results from Experiment 2 and 3 conclusively demonstrate that Cantonese listeners are sensitive to the presence of creak in tone perception. This is positive evidence for the use of phonation cues in tonal perception when pitch and phonation cues are correlated, adding to a handful of case studies which thus far have been split between positive and negative evidence for the use of phonation cues. Moreover, this is evidence for the use of creak in tonal perception when pitch and phonation cues are correlated, in contrast to the negative evidence of Garellek *et al.* (2013) for the use of creak in identifying the lowest tone (also 21, ˩) in White Hmong. Finally, this is evidence for the use of phonation cues in a tone language where phonation cues are secondary to pitch cues, as well as the first evidence we are aware of that suggests that phonation cues can improve tonal identification accuracy. Our demonstration of the sensitivity of listeners to creak in tonal perception in a tone language where pitch is the primary dimension of contrast further supports that register and tone languages lie along a continuum of co-variance between phonation and pitch cues as dimensions of perceptual contrast.

One open question remaining from our study is the ultimate source of the creak in Cantonese T4. One hypothesis is that creak is produced as an enhancement of cues specifically for T4. Another hypothesis is that creak occurs as a side-effect of producing low f0 in general, and thus would occur for any tone—not just T4—uttered in a speaker's low f0 range. In Mandarin, Belotel-Grenié and Grenié (1997) and Belotel-Grenié and Grenié (2004) found that creak occurred more often on T3 than other tones, but that it also occurred quite frequently in other tones as well. Thus, it seems unlikely that creak is a cue for a specific tone in Mandarin. However, our study found that the occurrence of creak was much more asymmetrically distributed across tones in Cantonese: creak on other tones than T4 was rare. Without further work investigating the co-occurrence of creak and phonetic details of pitch realization across Cantonese tones, or manipulations of pitch range in tonal production, it thus is unclear whether creak in Cantonese might specifically cue T4. Finally, until we have results from more case studies on the use of phonation cues in tonal perception, the underlying reasons for the positive result for the use of creak in tonal perception in Cantonese but the negative result in Hmong remain murky.

Another open question remaining from our study is whether creaky phonation and f0 are perceptually integrated in the sense that they are dimensions that are not processed independently. Our result showing that listeners weight relative pitch cues from preceding context together with creaky

phonation in tone perception is consistent with creak being perceived or processed as low f0. Brunelle (2012) showed that F1, f0, and breathiness/spectral tilt show perceptual integration in this sense in Cham. To show this for creaky phonation and f0, one would need to define a dimension of gradient variability within creaky phonation and show that perception of variation among this dimension interacts with perception of variation in f0. The manipulation of glottal pulse width and duration of the creaky region in our study was an initial attempt to investigate potential dimensions of gradient variability within creaky phonation. The idea of gradience within creaky phonation remains unexplored.

In conclusion, the results of this study underscore that it is necessary to consider voice quality-related parameters beyond f0 and pitch for understanding human tonal perception and production in a potentially wide range of tonal languages—even tone languages in which pitch is the primary dimension of contrast. We hope that studies like ours draw attention to the role of phonation cues in tone languages and increase work incorporating phonation in automatic tonal recognition and in clinical applications for tone production and perception.

ACKNOWLEDGMENTS

We wish to acknowledge Patricia Keating, John Kingston, Jody Kreiman, Edward Stabler, Megha Sundara, and Kie Zuraw for illuminating discussions, and Eric Zee for allowing us to test participants at the City University of Hong Kong, Henry Tehrani for help with programming the experimental task, and Katherine Hill for editorial assistance. Experiment 2 was conducted for the second author's UCLA bachelor's honors thesis, and this paper is based in part on a chapter of the first author's UCLA dissertation. This work was supported by a NSF graduate fellowship to the first author and by NSF grant BCS-0720304 to Patricia Keating, Abeer Alwan, and Jody Kreiman. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1]The literature on voice quality is fraught with inconsistencies in terminology (Gerratt and Kreiman, 2001; Surana and Slifka, 2006). We use *phonation* as a term for a specific class of *voice quality* which can be divided into two classes: *modal* (default, baseline) and *nonmodal phonation*. The class of nonmodal phonations discussed in this paper is sometimes called *creaky* in referring to the acoustic signal, and we use *creaky*/*creak* to describe both period doubling and vocal fry. We reserve the term *creaky voice* to refer to the percept associated with *creak* in the acoustic signal.

[2]Throughout the paper, IPA tonal transcriptions are included in the first mention of a tone, but tones are referred to with abbreviated names after that, e.g., "T3" for Tone 3 or "T3s" for plural.

[3]Some descriptions also distinguish these tones from the shorter "entering" tones (high, mid, and low level) which occur in syllables with unreleased stop codas.

[4]Reduplicated expressions, e.g., when the bitone is composed of a sequence of identical morphemes, can be the target of tonal change in certain morphological/semantic contexts in Cantonese, such as for marking vocatives or intensification in adjectives (Matthews and Yip, 1994), but there were no tonal changes in the recordings.

[5]Silence threshold = 0.03, voicing threshold = 0.45, octave cost = 0.01, octave-jump cost = 0.35, voiced/unvoiced cost = 0.14.

[6]The particular value of 78 dB is not at all critical; what is critical is that the average intensity of a token was a constant value across tokens. Stimuli were played to participants over headphones at a comfortable volume fixed across subjects.

[7]The identification accuracy of the other tones is not the focus of this paper, but we note the following: first, perhaps because T3 (*lau*^{33}) is uncommon, its identification accuracy was the poorest (32.23%). Perhaps because of that and also because most of the Cantonese level tones are toward the bottom of the pitch range, T1 was identified most accurately (85.94%). Also, there was a bias for T5 responses: 55.99% of T2 stimuli were identified as T5s, replicating the pattern of results for sentence-medial tones in Ma *et al.* (2005). We speculate that T2 rises uttered sentence-medially may have had relatively small pitch excursions, making them particularly confusable with T5 when presented in isolation.

Abramson, A. S., L-Thongkum, T., and Nye, P. W. (**2004**). "Voice register in Suai (Kuai): An analysis of perceptual and acoustic data," Phonetica **61**, 147–171.

Abramson, A. S., and Luangthongkum, T. (**2009**). "A fuzzy boundary between tone languages and voice-register languages," in *Frontiers in Phonetics and Speech Science*, edited by G. Fant, H. Fujisaki, and J. Shen (The Commercial Press, Beijing, China), pp. 149–155.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**, 255–278.

Barry, J. G., and Blamey, P. J. (**2004**). "The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese," J. Acoust. Soc. Am. **116**, 1739–1748.

Bates, D., and Maechler, M. (**2010**). lme4: Linear mixed-effects models using S4 classes, URL http://lme4.r-forge.r-project.org/, R package version 0.999375-37 (Last viewed 21 September 2010).

Belotel-Grenie, A., and Grenie, M. (**1994**). "Phonation types analysis in standard Chinese," in *The 3rd International Conference on Spoken Language Processing, ICSLP 1994*, September 18–22, Yokohama, Japan, pp. 343–346.

Belotel-Grenié, A., and Grenié, M. (**1997**). "Types de phonation et tons en chinois standard" ("Phonation types and tones in standard Chinese"), Cah. Ling. – Asie Orient. **26**, 249–279.

Belotel-Grenié, A., and Grenié, M. (**2004**). "The creaky voice phonation and the organisation of Chinese discourse," *TAL-2004*, pp. 5–8.

Boersma, P., and Weenink, D. (**2010**). "Praat: Doing phonetics by computer (version 5.1.32) [computer program]," http://www.praat.org (Last viewed 21 September 2010).

Brainard, D. H. (**1997**). "The psychophysics toolbox," Spatial Vision **10**, 433–436.

Brunelle, M. (**2009**). "Tone perception in Northern and Southern Vietnamese," J. Phonet. **37**, 79–96.

Brunelle, M. (**2012**). "Dialect experience and perceptual integrality in phonological registers: Fundamental frequency, voice quality and the first formant in Cham," J. Acoust. Soc. Am. **131**, 3088–3102.

Brunelle, M., and Finkeldey, J. (**2011**). "Tone perception in Sgaw Karen," in *Proceedings of ICPhS XVII*, pp. 372–375.

Davison, D. S. (**1991**). "An acoustic study of so-called creaky voice in Tianjin Mandarin," Work. Pap. Phonet., Depart. Ling., UCLA **78**, 50–57.

DiCanio, C. T. (**2009**). "The phonetics of register in Takhian Thong Chong," J. Int. Phonet. Assoc. **39**, 162–188.

DiCanio, C. T. (**2012**). "Coarticulation between tone and glottal consonants in Itunyoso Trique," J. Phonet. **40**, 162–176.

Fok, C. (**1974**). "A perceptual study of tones in Cantonese," No. 18 in Occasional Papers and Monographs (University of Hong Kong, Centre of Asian Studies, Hong Kong).

Gårding, E., Kratochvil, P., and Svantesson, J.-O. (**1986**). "Tone 4 and Tone 3 discrimination in modern Standard Chinese," Lang. Speech **29**, 281–293.

Garellek, M., and Keating, P. (**2011**). "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec," J. Int. Phonet. Assoc. **41**, 185–205.

Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (**2013**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**, 1078–1089.

Gerratt, B. R., and Kreiman, J. (**2001**). "Toward a taxonomy of nonmodal phonation," J. Phonet. **29**, 365–381.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**, 3983–3994.

Khouw, E., and Ciocca, V. (**2007**). "Perceptual correlates of Cantonese tones," J. Phonet. **35**, 104–117.
Kong, J. (**2001**). "Study on dynamic glottis through high-speed digital imaging," Ph.D. thesis, City University of Hong Kong.
Kuang, J. (**2013**). "The tonal space of contrastive five level tones," Phonetica **70**, 1–23.
Lee, K. Y., van Hasselt, C., Chiu, S., and Cheung, D. M. (**2002**). "Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children," Int. J. Pediatr. Otorhinolaryngol. **63**, 137–147.
Ma, J. K.-Y., Ciocca, V., and Whitehill, T. (**2005**). "Contextual effect on perception of lexical tones in Cantonese," *INTERSPEECH-2005*, pp. 401–404.
Matthews, S., and Yip, V. (**1994**). *Cantonese: A Comprehensive Grammar* (Routledge, New York), pp. 1–432.
R Development Core Team (**2010**). "R: A language and environment for statistical computing," http://www.R-project.org ISBN 3-900051-07-0 (Last viewed 21 September 2010).
Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "Voicesauce: A program for voice analysis," *Proceedings of ICPhS XVI*.
Silverman, D., Blankenship, B., Kirk, P., and Ladefoged, P. (**1995**). "Phonetic structures in Jalapa Mazatec," Anthrolopolog. Linguist. **37**, 70–88.
Surana, K., and Slifka, J. (**2006**). "Acoustic cues for the classification of regular and irregular phonation," in *Proceedings of INTERSPEECH-2006*, 693–696.
Vance, T. J. (**1977**). "Tonal distinctions in Cantonese," Phonetica **34**, 93–107.
Wang, M., Wen, M., Hirose, K., and Minematsu, N. (**2010**). "Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model," *Proceedings of INTERSPEECH-2010*, pp. 2166–2169.
Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**, 25–47.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717.
Wong, P. C. M., and Diehl, R. L. (**2003**). "Perceptual normalization for inter- and intratalker variation in Cantonese level tones," J. Speech, Lang. Hear. Res. **46**, 413–421.
Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phonet. **25**, 61–83.
Yu, K. M. (**2010**). "Laryngealization and features for Chinese tonal recognition," *INTERSPEECH-2010*, pp. 1529–1532.
Zhu, X. (**2012**). "Multiregisters and four levels: a new tonal model," J. Chin. Linguist. **40**, 1–17.