Laryngealization and features for Chinese tonal recognition

Kristine M. Yu

Department of Linguistics, University of California Los Angeles

krisyu@ucla.edu

Abstract

It is well known that the lowest tone in Mandarin, a language without contrastive phonation, often co-occurs with laryngealization/creaky voice quality, and we provide evidence that this is also the case for the lowest tone in Cantonese. However, the effects of laryngealization on f0 feature extraction for tonal recognition, as well as the potential of laryngealization as a feature for improving tonal recognition, have not been well-discussed in the literature. We give evidence from a corpora of tonal production data for Cantonese and Mandarin that laryngealization is prevalent and significantly disturbs the extraction of f0 features, and suggest that laryngealization may in fact be a feature that could improve tonal recognition.

Index Terms: tone, tonal recognition, Chinese, Cantonese, Mandarin, laryngealization, voice quality, creak, glottalization, f0, irregular phonation, nonmodal phonation, vocal fry

1. Introduction

The purpose of this paper is to draw attention to the prevalence of irregular (non-modal) phonation in the phonetic realization of tone and discuss its implications for the acoustic front end in tonal recognition. In particular, we focus here on a particular subset of irregular phonation— laryngealization, also sometimes called creak, creaky voice, or glottalization—and we focus on two Chinese tonal languages, Cantonese and Mandarin.

1.1. Laryngealization in tonal languages

In tonal languages without contrastive phonation, there can nevertheless be interactions between tone and voice quality. In Mandarin, for instance, it is well known that the tone which occurs at the bottom of the pitch register, Tone 3, is typically realized with laryngealization [1, 2]. [2] found that Mandarin Tones 2 and 4 (rise and fall) also contained laryngealized regions, though much less often than Tone 3. In Cantonese, the lowest tone, Tone 4 (the mid-falling tone), has also been anecdotally noted to frequently co-occur with laryngealization [2]. Other tonal languages from Asia, e.g. Hmong, [3], as well as languages of Africa, such as Yoruba [4], have also been noted as having laryngealization in the lowest tone of their tonal inventory, either in isolation, or in sentence-medial positions as well as in isolation.

While it is unsurprising that tones with regions of low fundamental frequency (f0) co-occur with laryngealization due to physiological mechanisms of producing low f0 [5], there is also evidence that listeners use laryngealization as a cue in tonal perception. In a gating perceptual task, [2] compared the recognition point of Tone 3 in Mandarin for laryngealized instances of Tone 3 to that for non-laryngealized instances and found that the recognition point was earlier for laryngealized instances of Tone 3. Additionally, [6] found in a identification task for the six tones of Cantonese that listeners had both a higher percent correct and faster reaction time on creaky instances of Tone 4 than non-creaky instances (85% vs. 62% correct). (The Cantonese perception stimuli included the tones not occurring in checked syllables and were equalized for duration and normalized for amplitude.) That listeners can use laryngealization as a cue for tonal perception invites the possibility that laryngealization-based features could also be of use in automatic tonal recognition.

1.2. The interaction of laryngealization and f0

Even setting aside the potential of laryngealization-based features for improving tonal recognition, laryngealization presents a significant challenge for tonal recognizers. This is because f0 is not well-defined in laryngealized regions of speech. By "laryngealized", we mean the class of irregular phonation types shown by [5] to be perceptually equivalent for listeners. This class includes vocal fry and period-doubled phonation. Vocal fry is defined as "a train of discrete laryngeal excitations, or 'pulses', of extremely low frequency, with almost complete damping of the vocal tract between excitations, " with f0 ranging from 7 to 78 Hz [5]. In period-doubled phonation, the waveform shows cycles alternating in amplitude and/or frequency.

For both vocal fry and period-doubling, f0 is not well-defined. In vocal fry, glottal pulses are typically irregularly spaced, i.e. aperiodic, and in period-doubling, there is no unique value of f0, resulting in a "bitonal" percept.

Despite the interaction between the phonetic realization of tone and laryngealization, current tonal recognition systems do not discuss laryngealization-based features nor the effects of laryngealization on extraction of f0-related features. In the rest of the paper, we show that laryngealization is prevalent in the phonetic realization of tones with low f0 regions in Cantonese as well as Mandarin, particularly in Cantonese Tone 4 and Mandarin Tone 3. We also demonstrate that laryngealization poses a problem for f0 feature extraction. Finally, we discuss why laryngealization might not be currently discussed in work on automatic tonal recognition and propose that it should be.

2. The prevalence of laryngealization in Cantonese and Mandarin

In this section, we describe the prevalence and distribution of laryngealization in tonal production data from Cantonese and Mandarin. Because laryngealization in Mandarin has already been discussed to some extent in the literature, we primarily focus on Cantonese: laryngealization-tone interaction in Cantonese has been mentioned in [2] but never empirically described.

2.1. Materials and Methods

2.1.1. Corpus

We recorded a small corpora of read speech for both Cantonese and Mandarin. The corpora were designed to produce controlled contextual tonal variability following [7, 8], by including all possible bitones in the languages. The Cantonese corpus included the minimal set /lau/ over Tones 1-6 between mid-level tones (*lei5 yiu3 lau lau yaak8 cheung3/gaap3/sou3* 'you want lau-lau to eat sauce/duck/sauce'), with 5 repetitions of each item. The Mandarin corpus included the minimal set /ma/ over Tones 1-5 (including the neutral tone), uttered as sentence-initial, medial, and final bitones between high or low tones (e.g. *wo3 yao1/4 ma ma mai2/4 mao1* 'I invite/want mama to bury/buy a cat'), and we analyzed the sentence-medial bitones excluding Tone 5 (the neutral tone) for this paper.

Twelve native speakers of Cantonese from Hong Kong and Macau were recorded in a sound booth in Los Angeles, and twelve native speakers of Beijing Mandarin were recorded in Beijing, China. For both corpora, the speakers were asked to think of the target bitones as proper names, since the bitone combinations mostly 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 [7], which resulted in speech rates around 3.2 syllables/second, comparable to speech rates in [2]. The present analysis included four males and females from each language for a total of 1440 examples of each tone in Cantonese and 1280 in Mandarin. In Mandarin, we did not analyze the T3 tokens that were realized as T2 due to tone sandhi in the T3-T3 bitone.

2.1.2. Criteria for labelling laryngealization

After the recordings were segmented using Praat, we labeled each syllable in the target bitones as being laryngealized or not. We checked for irregular glottal pulses (vocal fry) or period doubling in the waveform, the appearance of subharmonics and/or loss of definition to harmonic structure in the narrow-band spectrogram, for f0 detection failure (with Praat's autocorrelation algorithm under appropriate settings), and for the auditory percept of laryngealization/creaky voice quality. These were the same criteria used to choose laryngealized vs. non-laryngealized instances of Tone 4 for the Cantonese tonal perception experiment in [6]. An example of using the criteria for performing the labeling is shown in Fig. 1. Note that the f0 detection failure occurs over the entire vowel.

2.2. Results

2.2.1. Distribution of laryngealization over Cantonese tones

We found that across the Cantonese speakers, there was laryngealization on Tone 4 about a quarter of the time. Moreover, laryngealization occurred much more frequently in the second syllable of the bitone than the first, as shown in Fig. 2. We suspect this was because the speakers thought of the bitone as a prosodic word, so that the laryngealization was more frequent at the end of a prosodic unit. There was great interspeaker variability in the proportion of laryngealized Tone 4s produced, ranging from 4% to 58%. Thus, while across all speakers, the majority of Tone 4 realizations were not laryngealized, in some speakers, more than half of Tone 4 instances were. Critically, laryngealization also rarely appeared on other tones. After Tone 4, it appeared most frequently in Tone 2, in the low f0 region at the onset of the rise. However, laryngealization was present

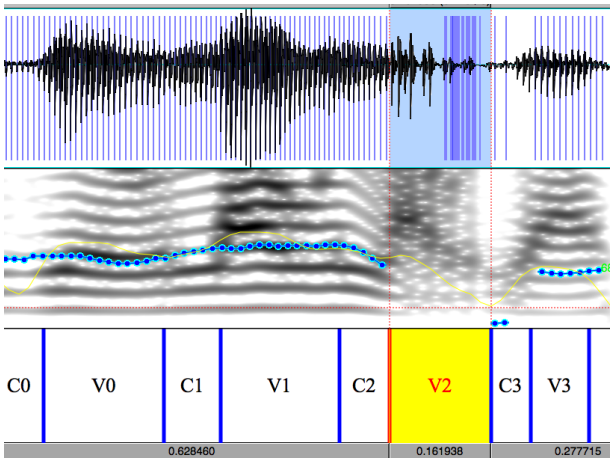

Figure 1: *Criteria for labeling segment as laryngealized. An example of vocal fry, with irregular, widely spaced glottal pulses, loss of definition in harmonic structure in the narrow-band spectrogram, and f0 detection failure.*

only 2% of the time for Tone 2.

Another characteristic of the laryngealized tone productions we found was that period doubling was not uncommon among female speakers. This is not unexpected since vocal fry is contingent on low absolute f0, but period doubling is a mechanism of laryngealization that is not contingent on any particular f0. An example of period doubling with alternating long and short cycles is in Fig. 3. Note that the laryngealization resulted in pitch halving in the f0 estimation.

2.2.2. Distribution of laryngealization over the four Mandarin basic tones

Our results for Beijing Mandarin were broadly similar to that of [2]. Laryngealization occurred most frequently for Tone 3, 68% across speakers, and also occurred in Tone 2 and Tone 4 and sporadically in Tone 1 (8% and 5%, 2% respectively). Thus, laryngealization is much more prevalent in Beijing Mandarin than in Cantonese, but also is comparatively less specific to a single tone in the inventory. Additionally, unlike in Cantonese, the distribution of laryngealization across bitones was closer to being equal, rather than concentrated in the second syllable of the bitone. As in Cantonese, the amount of creak across speakers in Mandarin was also variable: for Tone 3, the proportion of creak ranged in speakers from 0% to 100%.

Both period doubling and vocal fry were observed in female speakers, while vocal fry dominated in laryngealization in male speakers. For some speakers, the laryngealization produced resulted in such damped oscillations that the vowel was almost completely silent and contained only a single pulse.

3. Discussion

For both Cantonese and Mandarin, we found that laryngealization was prevalent in two senses: (i) a large proportion of instances of the lowest tones, Tone 3 in Mandarin, and Tone 4 in Cantonese, were laryngealized, and (ii) laryngealization was often present for nearly the entirety of the duration of the syllable nucleus.

Compared to Mandarin, the co-occurrence of laryngealization on the lowest tone in Cantonese was almost three

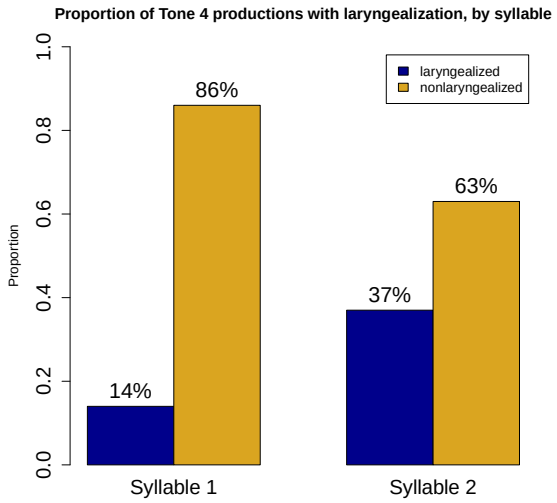

Figure 2: *Prevalence of laryngealization in Cantonese Tone 4 is greater in syllable 2 of sentence-medial bitone. Overall, laryngealization appeared in 25% of the Tone 4 realizations.*

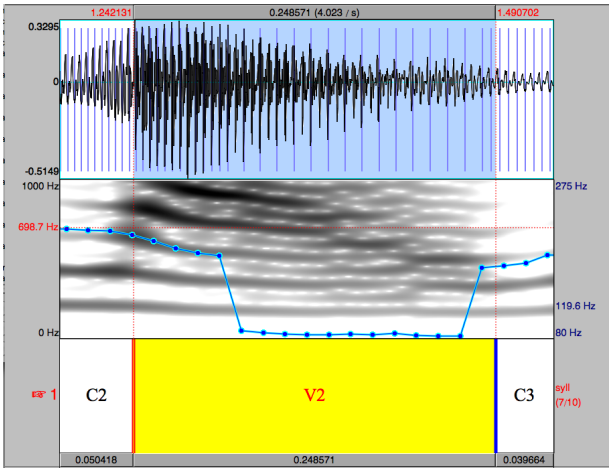

Figure 3: *Example of period doubling in Cantonese Tone 4 produced by a female speaker. Note the subharmonics in the narrow-band spectrogram and the pitch halving in the f0 track.*

times less frequent. However, the specificity of co-occurrence of laryngealization to Tone 4 in Cantonese suggests that laryngealization-based features could improve recognition of Tone 4. In the Cantonese perception study in [6], it was found that Tone 4 was most confusable with Tone 6, a mid-low level tone for non-laryngealized instances of Tone 4, but that the confusability between Tone 4 and 6 was much lower for laryngealized instances of Tone 4. The "supratone" model for Cantonese tonal recognition had 78% correct for Tone 4, with the majority of inaccuracies due to misidentification of Tone 4 as Tone 6 [12]; perhaps the introduction of laryngealization-based features could reduce these misidentifications, too.

In Mandarin, in comparison, there was significant laryngealization of low f0 regions of Tone 2 and Tone 4, in addition to laryngealization of Tone 3. However, Tone 3 laryngealization tended to be longer in duration, and even in females, typically vocal fry rather than period doubling. Thus, while the presence of laryngealization might not be specific to Tone 3, a more nuanced specification of laryngealization detailing mechanism and timing could be.

For both Cantonese and Mandarin, the duration of laryngealization over a large portion or even the entire syllable nucleus (examples in Fig. 1 and 3) poses a problem for extraction of f0 features. As can be seen in the figures, f0 detection failed. Thus, even if laryngealization-based features do not turn out to be useful for improving tonal recognition, the disturbance of f0 feature extraction due to laryngealization cannot be ignored.

3.1. Acoustic features in tonal recognizers and laryngealization

The majority of work on engines for tonal recognition has been for Mandarin and Cantonese. As f0 has always been considered the primary feature for tonal classification in these languages, tonal recognition for them is dominated by f0-based features: any tonal recognizer that has been built uses f0-based features.

For instance, in the acoustic front end for tone nucleus-based tone recognizers for Mandarin [9, 10, i.a.], mean log f0 and mean log f0 slope are extracted from three subsegments in the "tone nucleus" and from the preceding and following subsegments of neighboring tone nuclei. (The tone nucleus is the portion of the syllable segmented out by a discriminant function which "is a piece of F0 contour that represents pitch targets of the lexical tone, which contains the most critical information for tonality perception" [10].) In [11] for Mandarin tonal recognition, f0 features include values extracted from five evenly spaced points from the syllable final, as well as f0 max, standard deviation, and the slope of a linear fit to the f0 contour, and f0 features from neighboring syllable finals. In the "supratone" model for Cantonese tonal recognition in [12], average log f0 is extracted from three equally divided subsegments of the syllable final and its neighboring syllable finals.

The extensive use of f0 features for tonal recognizers and evidence that humans use laryngealization as a cue for tonal perception raises two issues with respect to laryngealization: (i) how come laryngealization doesn't seem to pose a problem for f0-related feature extraction?, and (ii) why haven't laryngealization-related features been considered? We address each question in turn.

Given that laryngealization has been suggested as frequently occurring in Mandarin Tone 3 in particular, and in Cantonese Tone 4, as discussed in §1.1, one would expect f0 detection problems to be a frequent occurrence for the acoustic front end in tonal recognizers. However, such problems have

not been highlighted in the tonal recognition literature. We infer that this is because of screening/manual correction of f0 tracks and smoothing/averaging. In the tone nucleus model work, [9] mentions manual correction of f0 contours for pitch halving or doubling, and [10] uses autocorrelation strength to segment out the window for feature extraction, the tone nucleus, a process which may filter out regions of f0 instability. Similarly, [11] states that f0 values were extracted only from "valid pitch tracked regions". Moreover, note in the list of typical f0-based features above that f0 based features are frequently averaged over subsegments, as in the supratone model in [12], as well as the tone nucleus model. Even in descriptive work in linguistics and phonetics on tone, f0 traces are typically hand-corrected and averaged over subsegments as in [7] and subsequent work. It is rare that f0 tracking difficulties are discussed, as in [13].

Finally, we note that for either manual correction or averaging, f0 detection problems in instances where laryngealization occurred across the duration of the syllable nucleus such as in our corpora would pose a problem: in such cases, there may be no region of the syllable where f0 is well-defined, and thus no f0 values to correct in a principled way, or to average over. For period-doubled regions, it is also not clear how one could define an f0 to manually correct or average over.

Additionally, filtering out, manually correcting, and averaging out effects of laryngealization on the f0 track may result in the discarding of acoustic information useful for tonal recognition, leading into the second question on why laryngealization-related features have not been considered in tonal recognition. There has in fact been work exploring the use of voice quality features in Mandarin [14]. In this work, voice quality features were parameterized using band energy and spectral features and indirectly relied on f0 features for the calculation of harmonics. Only band energy features (which do not rely on f0 detection) were shown to improve tonal recognition, and only for the neutral tone.

However, the dependence of any features involving harmonics on f0 feature extraction implies a large role for f0 detection in determining how useful such features may be for tonal recognition. In addition, in smaller-scale work with more controlled read speech, voice quality features involving harmonics (H1-A1) have been suggested as providing discriminatory power between Mandarin Tones 3 and 4 and other tones [2, 15]. Thus, we suggest that the utility of voice quality features for tonal recognition, including laryngealization-related ones, should not be ruled out without considering the effect of f0 detection problems in precisely those regions of the speech signal that may provide discriminatory power from irregular phonation such as laryngealization.

Finally, it may be that other parametrizations of voice quality features may show different results than those studied in [14]. The line of research on automatic detection of irregular phonation may provide useful alternatives [16, 17, 18].

4. Conclusion

Analysis of the proportion of laryngealized instances of tones in our Cantonese and Mandarin corpora suggests that laryngealization is widespread both in the sense that it can occur over the whole of the domain of f0 feature extraction for a syllable, thus severely disturbing f0 feature extraction, and also widespread in the sense that across speakers, laryngealization occurs frequently (and in some speakers, a majority of the time). The specificity of laryngealization to Tone 4 in Cantonese, and perhaps also to Tone 3 in Mandarin, coupled with results from human listeners that laryngealization is useful as a cue for these tones in tonal perception, suggests that laryngealization-based features should be considered in tonal recognition. Finally, even if laryngealization-based features may not be useful for tonal recognition, the disturbance of f0 tracking and thus extraction of f0 features due to laryngealization is problematic and should be addressed.

5. Acknowledgements

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship and NSF grant BCS-0720304.

6. References

[1] Davison, D.S., "An acoustic study of so-called creaky voice in Tianjin Mandarin", Working Papers in Phonetics, UCLA, 78: 50–77, 1991.

[2] Belotel-Grenie, A. and Grenie M,. "Types de phonation et tons en chinois standard", Cahiers de Linguistique - Asie Orientale, 26(2): 249–279, 1997.

[2] Vance, T.J., "Tonal distinctions in Cantonese", Phonetica, 34: 93–107, 1977.

[3] Huffman, M.K., "Measures of phonation type in Hmong", Working Papers in Phonetics, UCLA, 61: 1–25, 1985.

[4] Welmers, W.E., African Language Structures, University of California Press, 1973.

[5] Gerratt, B.R. and Kreiman, J., "Toward a taxonomy of nonmodal phonation", Journal of Phonetics, 29(4): 365–381, 2001.

[6] Lam, H.W. and Yu, K.M., "The role of creaky voice quality in Cantonese tonal perception, J. Acous. Soc. Amer., 127(3):2023–2023, 2010.

[7] Xu Y., "Contextual tonal variations in Mandarin", J. Phon., 25: 61–83, 1997.

[8] Wong, Y.W., "Contextual tonal variations and pitch targets in Cantonese", Proc. Speech Pros. 2006, 2006.

[9] Zhang, J. and Hirose, K., "Tone nucleus modeling of Chinese lexical tone recognition", Speech Comm., 42: 447–466, 2004.

[10] Wang, X., Hirose, K., Zhang, J., and Minematsu, N., "Tone recognition of continuous Mandarin speech based on tone nucleus model and neural network", IEICE Trans. Inf. Syst., E91–D(6): 1748–1755, 2008.

[11] Levow, G., "Unsupervised and semi-supervised learning of tone and pitch accent", Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, 224–231, 2006.

[12] Qian Y., Lee, T. and Soong F. K., "Tone recognition in continuous Cantonese speech using supratone models", J. Acous. Soc. Amer., 121(5): 2936–2945, 2007.

[13] Liu, S. and Samuel, A.G., "Perception of Mandarin lexical tones when f0 information is neutralized", Language and Speech, 47(2): 109–138, 2004.

[14] Surendran, D. and Levow, G., "Can voice quality improve Mandarin tonal recognition?", Proceedings of ICASSP, 4177–4180, 2008.

[15] Belotel-Grenie, A. and Grenie, M., "Phonation types analysis in Standard Chinese", Proceedings of ICSLP-1994, 343–346, 1994.

[16] Surana, K. and Slifka, J., "Acoustic cues for the classification of regular and irregular phonation", Proc. Interspeech 2006, 693–696, 2006.

[17] Vishnubhotla, S. and Espy-Wilson, C., "Automatic detection of irregular phonation in continuous speech", Proc. Interspeech 2006, 2006.

[18] Ishi, C.T., Sakakibara, K., Ishiguro, H., and Hagita, N., "A method for automatic detection of vocal fry", IEEE Trans. Aud., Speech, and Language, 16(1):47–56, 2008.