# The role of $F_0$ and phonation cues in Cantonese low tone perception

Yubin Zhang, and James Kirby

Citation: The Journal of the Acoustical Society of America **148**, EL40 (2020); doi: 10.1121/10.0001523
View online: https://doi.org/10.1121/10.0001523
View Table of Contents: https://asa.scitation.org/toc/jas/148/1
Published by the Acoustical Society of America

JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0001523

# The role of $F_0$ and phonation cues in Cantonese low tone perception

**Yubin Zhang**[1] **and James Kirby**[2,a)]

[1]*Department of Linguistics, University of Southern California, Los Angeles, California 90089, USA*
[2]*Department of Linguistics and English Language, University of Edinburgh, Edinburgh, EH8 9AD, United Kingdom*
*yubinzha@usc.edu, j.kirby@ed.ac.uk*

**Abstract:** For languages that primarily exploit $F_0$ to signal tonal contrast, the role of phonation cues in tonal perception remains controversial. This study revisits the use of $F_0$ and phonation cues in Cantonese low tone perception (tone 4, 21/tone 6, 22) using synthesized stimuli. In line with previous studies, $F_0$ contour and height were found to be the most salient cues, with $F_0$ height being more important. The effects of non-modal phonation (creaky and breathy voice) were relatively small. Non-modal phonation enhanced low tone perception only in the low $F_0$ range. The results are consistent with the *differential integration hypothesis* that the perceptual role of phonation is dependent on $F_0$ and that phonation cues integrate with $F_0$ differently depending on $F_0$ height. *© 2020 Acoustical Society of America*

[Editor: Martin Cooke] Pages: EL40–EL45

**Received:** 6 March 2020 **Accepted:** 17 June 2020 **Published Online:** 15 July 2020

## 1. Introduction

### *1.1 $F_0$ and phonation cues in tonal perception*

In speech perception, acoustic cues are distributed along multiple dimensions (Holt and Lotto, 2006). For lexical tone perception, it is generally agreed that aspects of $F_0$ realization, such as its height and contour, are the most important cues in a diverse array of tonal languages (Gandour, 1983), but phonation properties such as breathy and creaky voice are also germane to tonal perception. For a mixed $F_0$-phonation tonal system like Northern Vietnamese, phonation can be even more critical than $F_0$ for perceiving certain tones (Brunelle, 2009). Even for tonal languages where $F_0$ is clearly the primary perceptual dimension, non-modal phonation may be perceptually relevant. The perceptual relation between phonation and $F_0$ in this type of system can be predicted from the pitch-dependent phonation continuum proposed by Kuang (2013, 2017). In speakers' most comfortable pitch range, voice quality is modal. At one end of the continuum, as pitch is lowered, voice quality becomes tenser, which can finally lead to creaky voice (vocal fry) or breathy voice. At the other end of the continuum, as pitch is raised, voice quality also becomes tenser and can finally change into falsetto.

While this model is based primarily on acoustic results, there is also some evidence for the role of pitch-dependent phonation in tonal perception. For example, Yang (2015) found that Mandarin listeners relied on phonation cues in identifying the low-dipping tone 3 (T3, 214) compared to other tones. Yu and Lam (2014) found that creak biased listeners toward the lowest-pitched T4 in their study of Cantonese low tone identification. The presence of allophonic or redundant phonation does not seem to always entail its perceptual relevance, however. White Hmong also contains two low tones, low falling (*m*, 21) and low level (*s*, 22), and like Cantonese T4, the *m* tone can be realized with allophonic creak (Garellek *et al.*, 2013). Yet Garellek *et al.* (2013) reported that the presence of acoustic creak was not necessary for *m* tone identification.

While we cannot rule out the possibility that these are simply language-specific differences, methodological choices may also play a role. Creaky voice has multiple acoustic correlates including lower differences in the amplitudes of the $F_0$ harmonics (lower $H_1^* - H_2^*$), stronger higher-frequency harmonics (lower $H_1^* - A_n^*$),[1] and irregular pulsing (Keating *et al.*, 2015). Some types of creak can also be coproduced with a low $F_0$. It is worth considering the effects that the different manipulations used in previous studies may have on these cues, and the extent to which the individual contributions of the different properties can be assessed.

For example, the resynthesis method used by Garellek *et al.* (2013) might have had detrimental effects on certain acoustic properties of creak. In preparing their stimuli, they raised the

[a)]ORCID: 0000-0002-0502-5245.

 

https://doi.org/10.1121/10.0001523

$F_0$ of the non-creaky and creaky portions in a naturally produced token of White Hmong *m* tone using the pitch synchronous overlap and add (PSOLA) algorithm. Although Garellek *et al.* (2013) were careful to ensure that their manipulated stimuli retained the spectral tilt properties of creaky voice, the nature of the PSOLA process produces a fundamentally periodic signal (Moulines and Charpentier, 1990). Therefore, the null effect in their study could be because both low spectral tilt and irregular pulsing cues for creak are less important than $F_0$ cues. Alternatively, if $F_0$ irregularity contributes significantly to the percept of creak, the spectral tilt properties of creak in their stimuli may not have been sufficient to overcome the lack of irregular pulsing.

Conversely, the methods used by Yu and Lam (2014) may have confounded the acoustic cues to creak with $F_0$. In their Cantonese T4/T6 identification task (experiment 3), Yu and Lam (2014) spliced the naturally creaky /au$^4$/ onto the end of the disyllable /jiu lau/, which was originally a sequence of a mid level tone (T3, 33) followed by T6–/jiu$^3$ lau$^6$/. The target syllable /lau/ was assigned an ambiguous $F_0$ contour lying between T4 and T6 using the PSOLA algorithm, and the $F_0$ (semitone) in /jiu/ was manipulated to create a "pitch continuum"[2] in the target /lau/. For creak manipulation, they selected the type of double pulsing creak for cross-splicing, on the grounds that other types of creak, like vocal fry, typically co-occur with (extra-)low $F_0$ (Keating *et al.*, 2015). Double pulsing is a type of irregular pulsing defined as "pairs of vocal cycles alternating in period and/or amplitude" (Gerratt and Kreiman, 2001; Keating *et al.*, 2015). This kind of creaky voice has more than one $F_0$, typically one higher and one lower, yielding a percept of indeterminate pitch and roughness. Yu and Lam (2014) argued that their manipulation allowed them to tease apart the independent contributions of $F_0$ (pitch) and creak cues to T4/T6 perception. Yet, as their creaky fragments used for cross-splicing were naturally produced in the low $F_0$ region of T4, they might not belong to the canonical double pulsing creak, but instead could probably be classified as double pulsing with (extra-)low $F_0$, i.e., vocal fry with doublet cycle patterns [see Blomgren *et al.* (1998) and Gerratt and Kreiman (2001)]. In creaky portions of their stimuli, the distance between certain adjacent pulses was much longer than that in preceding non-creaky portions, which might lead to a predominant extra-low pitch percept at the end of the target syllable. Thus, it may be that this uncontrolled extra-low pitch percept, rather than the irregular pulsing or spectral tilt differences characteristic of creaky phonation, is responsible for the large main effect of creak found in Yu and Lam (2014).

### *1.2 The current study*

To gain finer control over these $F_0$-phonation interactions, we conducted a perceptual study of Cantonese T4/T6 perception using synthesized stimuli. We included not only creaky voice, but also tense and breathy voice, which have also been shown to covary with $F_0$ in production (Kuang, 2013, 2017). Both creak and tense phonations have pressed or constricted glottal configurations, which are assumed to be realized as lower spectral tilts like $H_1^* - H_2^*$ acoustically. Some work classifies tense phonation as a type of creak, i.e., pressed voice without irregularity (Keating *et al.*, 2015). Including a tense condition further allowed us to explore the differential contributions of spectral tilts and irregularity to tonal perception. For Cantonese, phonation difference between T4 and T6 may be described as tense versus modal (a lower versus higher spectral tilt). Garellek *et al.* (2013) argues for a similar tense/modal distinction between White Hmong *m* and *s* tones. Moreover, we included breathy voice, which can be adopted as another strategy for reaching low pitch targets (Kuang, 2017), as it has also been anecdotally mentioned as a correlate of Cantonese T4 (Rose, 2000; Yu and Lam, 2014).

This approach allows us to address two questions: (1) Do phonation cues play an independent role in T4/T6 perception? (2) How do phonation cues interact with $F_0$ cues? We hypothesized that non-modal phonation would play a role in T4/T6 perception, but when $F_0$ confounds are controlled for, the effect size might be smaller than that in Yu and Lam (2014). For the second question, we identify at least two possibilities. The first is that phonation serves as an independent cue, what we will call the *independent cue hypothesis* [see the discussion in Kuang (2017) and Yu and Lam (2014)]. Under this scenario, we would predict main effects of non-modal phonation on responses without interactions. However, non-modal phonation may be dependent on $F_0$ (Yang and Sundara, 2019), and lead to distinct percepts in different $F_0$ ranges (Gerratt and Kreiman, 2001). This *differential integration hypothesis* predicts patterns closely mirroring the covariation between $F_0$ and phonation in production (Kuang, 2013, 2017).[3] This leads us to predict an interaction: if, for example, double-pulsed creak leads to distinct percepts in the low versus middle or high $F_0$ range and only the integrated percept in the low $F_0$ range cues T4, we might find a T4 response bias only in the low $F_0$ range. Although Yu and Lam (2014) focused their discussion on the main effect of creak, their results actually revealed some evidence for an

interaction, especially with a female voice, but the creak effect was larger in the high-pitched conditions, e.g., negative contextual $F_0$ shift of −1.5 semitones in the preceding syllable /jiu/.

## 2. Methods

### *2.1 Materials*

We used the KLATTGRID synthesizer (Weenink, 2009) implemented in PRAAT (Boersma and Weenink, 2017) to generate speech stimuli that varied in $F_0$ mean, $F_0$ change, and phonation. A Cantonese-speaking male phonetician's productions of /wa$^6$/ in a modal voice and /wa$^4$/ in modal, syllable-final creak, and syllable-final breathy voices, were used to guide the synthesis. The vocal tract parameters were based on modal /wa$^6$/. The syllable was assigned a 500 ms duration and divided into three portions. The first 3/10 portion ($P_1$: 0–150 ms) modeled /w-a/ formant transitions. The middle 4/10 portion ($P_2$: 150–350 ms) was the steady-state /a/. The last 3/10 portion ($P_3$: 350–500 ms) was the vowel offset. As a result, $F_{1–4}$ began at 400, 760, 2600, and 3600 Hz and rose to 750, 1050, 3400, and 4300 Hz respectively at $P_1$ offset. They were held constant at these values in $P_{2,3}$. The intensity contour was adjusted to resemble that of modal /wa$^6$/ and the average intensity level was scaled to 70 dB.

The $F_0$ and phonation conditions were superimposed on /wa/ (see Fig. 1 for an illustration and supplementary materials[4] for audio files). For $F_0$ conditions, the lower half of the speaker's $F_0$ range was divided into five equally spaced $F_0$ levels (110, 100, 90, 80, and 70 Hz). We then generated 15 $F_0$ conditions (5 level and 10 linearly descending $F_0$ trajectories) by varying the degree of $F_0$ fall (0, 10, and 20 Hz) at each $F_0$ level. Four phonation types, i.e., modal, creaky, tense, and breathy, were synthesized by manipulating *open phase* (OP), *spectral tilt* (ST), *double pulsing* (DP), and *breathiness amplitude* (BA). The OP, defined as the ratio (% in KLATTGRID) of glottal opening time to the whole glottal cycle, is correlated with $H_1^* - H_2^*$ (Esposito, 2012). The ST in KLATTGRID specifies extra dB down at 3000 Hz and thus affects various spectral tilt measures like $H_1^* - A_n^*$ used in previous studies. OP and ST measures, especially $H_1^* - H_2^*$, have been found to distinguish phonation types in a variety of languages (Keating *et al.*, 2010). Breathy phonation typically exhibits a higher OP and ST, while creaky and tense phonations have lower ones. We assigned a steady 60% OP and a steady 10 dB ST to modal tokens. For creaky tokens, OP and ST began at 60% and 10 dB and fell to 30% and 0 dB, respectively, at the end of $P_{1,2}$, and they were held constant at 30% and 0 dB in $P_3$. The ST of tense tokens had the same profile as that of creaky tokens, but the OP of tense tokens began at 60% and fell to 25% at the end of $P_{1,2}$. This lower final OP rendered the tense tokens an even "tenser" percept than the creaky tokens.[5] The OP and ST of breathy tokens also started at 60% and 10 dB, but were then increased to 70% and 20 dB, respectively. The exact values were determined to ensure that the resulting spectral measures of synthesized phonations roughly matched those in the phonetician's non-modal tokens of /wa$^4$/. The DP, which modifies glottal pulses in a pair by delaying the timing and attenuating the amplitude of the first pulse,[6] was also adjusted to render double pulsing creak in $P_3$. The DP has been used to mimic utterance-final creak (Klatt and Klatt, 1990) and to approximate "fry" or "rough" quality in clinical settings (Yiu *et al.*, 2002). A steady 40% DP, selected based on Yiu *et al.* (2002), was added to $P_3$ for creaky tokens. The BA, which adds noise amplitude during the open phase, was increased from 0 to 60 dB in $P_{1,2}$ and held constant at 60 dB in $P_3$ to approximate the extra breathiness amplitude in breathy voice (Klatt and Klatt, 1990).

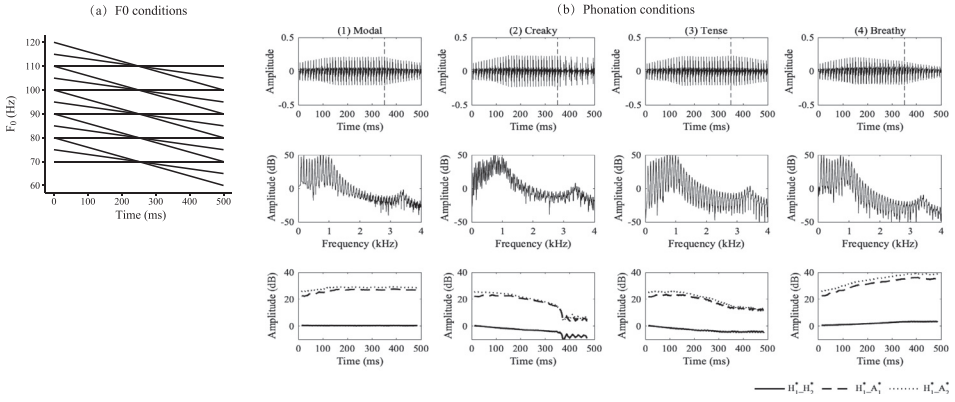


Fig. 1. Illustration of 15 $F_0$ conditions (a) and 4 phonation conditions (b). For (b), the top, middle, and bottom panels show synthesized waveforms, *fft* spectra (reference 20 $\mu$Pa) of the final 3/10 portion $P_3$ (350 to 500 ms), and resulting spectral tilt changes obtained from VOICESAUCE (Shue *et al.*, 2011), respectively.

https://doi.org/10.1121/10.0001523

*2.2 Participants and procedures*

A total of 31 Hong Kong Cantonese speakers—14 males (age 22.43 ± 2.58 years) and 17 females (age 22.76 ± 2.94 years)—were recruited from the student population at Hong Kong Polytechnic University. They carried out a combined identification and goodness rating task in a sound-treated booth. Goodness rating was administered to recognize particularly artificial-sounding tokens, as they tend to receive low ratings (Brunelle, 2009). The speech stimuli were presented in PRAAT over a headset at a comfortable volume. Their task was to first identify the tone by clicking on one of the two boxes representing Chinese characters for T4 (華, wa$^4$) and T6 (語, wa$^6$). Then, they rated the goodness of the tone on a scale of 1 (very bad)–5 (very good). The 240 stimuli (4 repetitions × 15 $F_0$ trajectories × 4 phonations) were pseudo-randomly presented in 4 blocks.

## 3. Results

A generalized linear mixed-effects model (Bates *et al.*, 2015) and two cumulative link mixed models (Christensen, 2015) were fit to analyze the identification responses and rating responses of T4 and T6, respectively. We included $F_0$ mean, $F_0$ change, phonation (reference level "modal"), $F_0$ mean: $F_0$ change, and $F_0$ mean:phonation as fixed effects. $F_0$ mean and $F_0$ change were $z$-standardized. The random-effects structure justified by the likelihood ratio tests included by-participant intercepts and slopes for $F_0$ mean and $F_0$ change.

For identification (see Fig. 2), both $F_0$ mean ($B = -3.27$, $p < 0.001$) and $F_0$ change ($B = 0.36$, $p < 0.001$) significantly predicted the probability of a T4 response, but $F_0$ mean had a larger effect than $F_0$ change. The main effects of phonation did not reach significance (breathy: $B = -0.07$; creaky: $B = -0.15$; tense: $B = 0.07$, $ps > 0.05$). Significant interactions between $F_0$ mean and breathy voice ($B = -0.24$, $p < 0.05$), and between $F_0$ mean and creaky voice ($B = -0.26$, $p < 0.05$) were found. Breathy voice and creaky voice slightly increased the probability of T4 responses in the low $F_0$ range, but this effect seemed to be diminished or somewhat reversed in higher $F_0$ range.

For T4 rating ($mean = 3.62$, $sd = 1.31$), $F_0$ mean and $F_0$ change significantly predicted its rating responses. As $F_0$ mean increased, the probability of a T4 response being rated in higher categories on the rating scale decreased ($B = -2.27$, $p < 0.001$). Furthermore, as $F_0$ change increased, the probability of a T4 response being rated in higher categories went upwards ($B = 0.37$, $p < 0.001$). This effect was modulated by its interaction with $F_0$ mean—as $F_0$ mean went higher, the positive effect of $F_0$ change on the probability of a T4 response being rated in higher categories decreased ($B = -0.15$, $p < 0.01$). No significant effects of phonation types were found, suggesting no extremely unnatural T4 tokens in our manipulation. For T6 rating ($mean = 3.27$, $sd = 1.25$), as $F_0$ mean increased, the probability of a T6 response being rated in higher categories significantly decreased ($B = -0.80$, $p < 0.001$).

## 4. Discussion

In line with previous studies (e.g., Gandour, 1983), our results reaffirm that both $F_0$ height and contour are the primary perceptual cues to Cantonese T4/T6 perception, with $F_0$ height being more critical than $F_0$ contour. More importantly, we found that the effects of non-modal phonation on Cantonese lexical tone perception are relatively small, and non-modal phonation interacts with $F_0$. In our study, the presence of creaky and breathy voice increased T4 responses in the low $F_0$ range, but the effects were diminished or somewhat reversed in higher $F_0$ range.

In Sec. 1.2, we sketched two predictions about how phonation cues might interact with $F_0$. The presence of an interaction effect rather than a main effect of creaky and breathy voice agrees with the *differential integration hypothesis* rather than the *independent cue hypothesis*. This is because only the former hypothesis predicts the perceptual dependency of phonation cues on $F_0$ cues (Yang and Sundara, 2019), and differential perceptual integration and categorization of

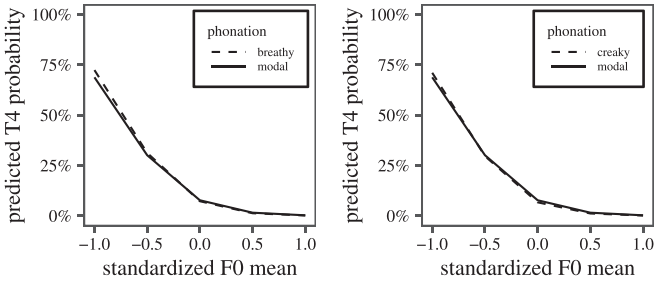


Fig. 2. Predicted effects (fixed effects) of breathy voice and creaky voice on the probability of a T4 response.

these two phonation cues in different $F_0$ ranges (Gerratt and Kreiman, 2001). One possible underlying perceptual mechanism for the differential cue integration is that the percept induced by non-modal phonation varies according to $F_0$ height. Double-pulsed creak could cause a "vocal-fry-like" percept in the low $F_0$ range, while it might be associated with a "rough" percept in relatively higher $F_0$ ranges. Listeners may merely employ the integrated vocal-fry-like percept/mixture as a cue for low tone perception, resulting in more T4 responses in the low $F_0$ range. A further possibility is that non-modal phonation interacts with $F_0$ by altering the pitch percept. Kuang and Liberman (2016) found that vocal fry, characterized by low and irregular $F_0$ (jitter), could lead to a lower pitch percept in a pitch classification task. However, the enhancement may still be dependent on pitch ranges, as only low $F_0$ ranges (baselines of 50, 70, and 90 Hz) typical of vocal fry were included in that study.

The present findings extend the results of Yu and Lam (2014) and Garellek *et al.* (2013) in finding that the effect of creak on T4/T6 perception is small and dependent on $F_0$. In Yu and Lam (2014), the large main effect of creak might be confounded by low $F_0$. Interestingly, the experiment of Yu and Lam (2014) also revealed an interaction between creak and $F_0$, but they found a larger creak effect in high-pitched conditions. This difference might be due to the $F_0$ confounds in their study or their manipulation of contextual $F_0$, rather than $F_0$ in the target syllable. In Garellek *et al.* (2013), since the $F_0$-creak interaction was not explicitly tested, it remains unclear whether a similar interaction would also be predicted to hold for the low tone contrast in White Hmong, or more generally in other low tone contrasts. Note that we are not disputing the validity of the findings of Yu and Lam (2014). Their creaky tokens, which mixed (extra-)low $F_0$ and double pulsing creak, did greatly enhance T4 identification. Nevertheless, the interaction effect in our study suggests that the observed enhancement may be due to listeners' enhanced low pitch percept or integration of low $F_0$ and irregular pulsing in the vocal-fry-like mixture, rather than attention to creak cues themselves. Furthermore, the difference could also be because perceptual cues are richer in natural stimuli or the extent of creakiness was larger in their study. However, since it is yet unclear what other cues are relevant in T4/T6 contrast and how to quantify the effects of different cues of creakiness in natural tokens, we still suspect the assessment of the independent effect of creak in Yu and Lam (2014) could be overestimated.

Positive evidence for creak and breathiness in low tone perception is consistent with the pitch-dependent phonation continuum (Kuang, 2013, 2017). However, our study did not detect any effects of the tense voice manipulation. One possibility is that tense voice may be a less effective cue to the extent that pitch-dependent tense voice can occur in both low and high $F_0$ ranges. Since tenseness is also a property of creaky voice, this result indicates that for the two types of creak cues, lower spectral tilts may be less critical in cuing low tones than irregular pulsing. However, it is also likely that the integration of double pulsing and tense properties plays a more crucial role in perception. Future studies can further include a condition with the manipulation of irregular pulsing only to test this possibility.

A major limitation of the current study is that the stimuli were all based on a single monosyllable /wa/ spoken by a male voice. This may hinder the extent to which we can generalize the present findings to other situations of $F_0$-phonation weighting, as cue weighting can be adaptive or flexible (Broersma, 2008; Yang and Sundara, 2019). Future studies can employ a design with a wider range of stimuli presented in various listening conditions.

## 5. Conclusion

This study reaffirms that $F_0$ height and $F_0$ contour are the most critical cues for Cantonese T4/T6 perception, with pitch height being a more important cue than $F_0$ contour. Creaky and breathy phonation types can also play a role in the perception of this contrast, but their effects are relatively small and mainly exerted through their interactions with $F_0$ height, consistent with the *differential integration hypothesis*. Models of the perceptual weighting and integration of phonation cues to tonal contrasts should take into account the pitch ranges in which they are produced.

## Acknowledgments

This project was supported by funds from the Department of Linguistics and English Language, University of Edinburgh, and by ERC Grant No. 758605 to Dr. James Kirby. We thank Dr. Yao Yao, Dr. Kristine Yu, Dr. Xuefeng Gao, and two anonymous reviewers for constructive comments.

## References and links

[1] $H_1^* - H_2^*$ represents the amplitude differential between the first and second harmonics and $H_1^* - A_n^*$ denotes the difference between the amplitude of the first harmonic and that of the most prominent harmonic of the nth formant. The asterisk indicates that the measures are corrected for formant effects. See Iseli and Alwan (2004) for more details.

https://doi.org/10.1121/10.0001523

[2]We refer to this as a "pitch continuum" as the $F_0$ of the target syllable /lau/ was not directly manipulated.
[3]Note, however, that perceptual integrality is distinct from marginal distributional distinctiveness and/or overlap. As a result, perceptual integration (Repp, 1988) or auditory enhancement (Kingston and Diehl, 1994) cannot always be predicted purely from the distributional properties of the acoustic cues involved.
[4]The supplementary materials including the audio files of the stimuli, response data, and code for reproducing the statistical analysis and figures are available online (Zhang and Kirby, 2020).
[5]We hypothesized that the absence of effects of tense voice in our pilot study might be due to the non-saliency of tense voice in our synthesis. Therefore, we further decreased the OP to 25% in the current study to make it sound tenser.
[6]The degree of delay and attenuation is specified in percentage (%). For example, a value of 50% causes the first pulse in a pair to be reduced by 50% in amplitude and to migrate half-way toward the second pulse. Note that the synthesis of vocal fry in Kuang and Liberman (2016) used the jitter parameter to render irregular $F_0$. Also, their $F_0$ baselines were 50, 70, and 90 Hz, which are common for a vocal fry register. The percept of double pulsing creak and canonical vocal fry might be different.


Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.
Blomgren, M., Chen, Y., Ng, M. L., and Gilbert, H. R. (**1998**). "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," J. Acoust. Soc. Am. **103**(5), 2649–2658.
Boersma, P., and Weenink, D. (**2017**). "Praat: Doing phonetics by computer," http://www.fon.hum.uva.nl/praat/ (Last viewed 7/2/2020).
Broersma, M. (**2008**). "Flexible cue use in nonnative phonetic categorization," J. Acoust. Soc. Am. **124**(2), 712–715.
Brunelle, M. (**2009**). "Tone perception in Northern and Southern Vietnamese," J. Phon. **37**(1), 79–96.
Christensen, R. H. B. (**2015**). "Ordinal—Regression models for ordinal data," R package version 2015.6-28.
Esposito, C. M. (**2012**). "An acoustic and electroglottographic study of White Hmong tone and phonation," J. Phon. **40**(3), 466–476.
Gandour, J. (**1983**). "Tone perception in Far Eastern languages," J. Phon. **11**(2), 149–175.
Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (**2013**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**(2), 1078–1089.
Gerratt, B. R., and Kreiman, J. (**2001**). "Toward a taxonomy of nonmodal phonation," J. Phon. **29**(4), 365–381.
Holt, L. L., and Lotto, A. J. (**2006**). "Cue weighting in auditory categorization: Implications for first and second language acquisition," J. Acoust. Soc. Am. **119**(5), 3059–3071.
Iseli, M., and Alwan, A. (**2004**). "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proceedings of ICASSP 2004*, IEEE, Vol. 1, pp. I–669.
Keating, P., Esposito, C., Garellek, M., Khan, S., and Kuang, J. (**2010**). "Phonation contrasts across languages," in *Working Papers in Phonetics*, Department of Linguistics, UCLA, Vol. 108, pp. 188–202.
Keating, P., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the International Congress of Phonetic Sciences XVIII*, Glasgow.
Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**(3), 419–454.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**(2), 820–857.
Kuang, J. (**2013**). "The tonal space of contrastive five level tones," Phonetica **70**(1-2), 1–23.
Kuang, J. (**2017**). "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," J. Acoust. Soc. Am. **142**, 1693–1706.
Kuang, J., and Liberman, M. (**2016**). "The effect of vocal fry on pitch perception," in *Proceedings of ICASSP 2016*, IEEE, pp. 5260–5264.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**(5-6), 453–467.
Repp, B. H. (**1988**). "Integration and segregation in speech perception," Lang. Speech **31**(3), 239–271.
Rose, P. (**2000**). "Hong Kong Cantonese citation tone acoustics: A linguistic tonetic study," in *Proceedings of the 8th Australian International Conference on Speech Science & Technology*, pp. 198–203.
Shue, Y., Keating, P., Vicenik, C., and Yu, K. (**2011**). "Voicesauce: A program for voice analysis," in *Proceedings of the International Congress of Phonetic Sciences XVII*, pp. 1846–1849.
Weenink, D. (**2009**). "The Klattgrid speech synthesizer," in *Interspeech 10*, pp. 2059–2062.
Yang, M., and Sundara, M. (**2019**). "Cue-shifting between acoustic cues: Evidence for directional asymmetry," J. Phon. **75**, 27–42.
Yang, R. (**2015**). "The role of phonation cues in Mandarin tonal perception," J. Chinese Ling. **43**(1B), 453–472.
Yiu, E. M., Murdoch, B., Hird, K., and Lau, P. (**2002**). "Perception of synthesized voice quality in connected speech by Cantonese speakers," J. Acoust. Soc. Am. **112**(3), 1091–1101.
Yu, K. M., and Lam, H. W. (**2014**). "The role of creaky voice in Cantonese tonal perception," J. Acoust. Soc. Am. **136**(3), 1320–1333.
Zhang, Y., and Kirby, J. (**2020**). "The role of F0 and phonation cues in Cantonese low tone perception," https://osf.io/wp8mv/ (Last viewed 7/2/2020).