Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

How do listeners identify creak? The effects of pitch range, prosodic position and creak locality in Mandarin

Aini Li[1], *Wei Lai*[2], *Jianjing Kuang*[1]

[1] Department of Linguistics, University of Pennsylvania, USA
[2] Department of Psychology and Human Development, Vanderbilt University, USA

liaini@sas.upenn.edu, wei.lai@vanderbilt.edu, kuangj@ling.upenn.edu

Abstract

As a non-modal phonation, creak has been found to influence the perception of pitch range and prosodic boundary. However, few studies have examined how these factors could in turn affect listeners' perception of creak (e.g., Davidson 2019). This study examines the effects of pitch range, prosodic position, and creak locality on creak identification in Mandarin. 40 native Mandarin listeners listened to 128 auditory sentences online and identified whether and where (i.e., at which syllables) they heard creak. Sentences carrying target syllables were manipulated in terms of pitch range (high vs. low), the prosodic position of creak (final vs. non-final), and creak locality (global vs. local). Mixed-effects logistic regression was implemented to predict listeners' identification responses, with Creak locality, Pitch range, and Prosodic position as fixed effects and Item, Participant and Tone as random intercepts. The results revealed lower accuracy of creak identification at sentence-final positions than non-final positions. While low pitch facilitated creak identification, it also increased false-alarmed creak identification of modal speech. Global creak in the sentential context led to higher creak identification than local creak at specific syllables. These findings imply that creak perception is context dependent and reflects listeners' knowledge about its acoustic and linguistic distributions.

Index Terms: creaky voice, perception, Mandarin Chinese, pitch range, prosodic position

1. Introduction

Creaky voice, also known as 'creak', 'vocal fry', 'glottalization', refers to an aperiodic phonation that is often related to low pitch targets. Despite the fact that there does not exist only one single type of creaky voice, it is usually featured with multiple acoustic cues such as irregular pulses, low F_0, constricted glottis, damped pulses and presence of subharmonics [1]. It is well established that not only can creaky phonation signal phonemic contrast (in languages such as Zapotec), it can also function as an additional enhancement cue to the perception of low tone targets [2, 3], the low-end of pitch range [4, 5], and domain-final prosodic boundaries [6, 7, 8]. In some cultures such as the U.S., creaky voice is also associated with sociolinguistic variables such as social status [9], gender [10], as well as unjustified social bias [11, 12]. For instance, in the U.S. context specifically, female speakers are found to use creaky voice more frequently than male speakers do [13, 14, 15, 16]. However, it has been found that creaky voice is preferred for male voices but less so for female voices such that creak sounds more attractive for men, but not for women [17].

However, so far, few investigations have examined how these factors could in turn affect listeners' identification of creak. A recent study on creak identification has been done among English listeners. Driven by the fact that creaky female speech is subject to unjustified social biases in the U.S. context, [11] looked at the effects of pitch, gender and utterance type on the identification of creaky voice. Through manipulating the amount of creak (modal vs. partial creak vs. full creak) and the type of utterances (full sentence vs. sentence-final fragments) as well as speaker gender, this study showed that even though female speech in American English has been identified to be creakier than male speech, listeners' ability to recognize creak was less influenced by the gender of the speaker than it was by the environment in which the creak was produced. Notably, stimuli used in this study were drawn from naturally-produced podcasts, hence leaving segmental features and syntactic structures uncontrolled. Even though naturalistic speech would be useful to examine people's perception situated in real-life situations, it presents difficulties for teasing apart all sorts of linguistic and acoustic factors that might have influenced people's judgment. Even though all the utterances used in the study were declaratives and there were no extreme pitch changes, the stimuli were produced by four different individual speakers. The ways individual speakers produce creaky voice might differ, which could introduce further confounding factors. Last but not least, these two studies on creak perception had a focus on English, cross-linguistically speaking, whether the perception of creaky voice remains the same for speakers speaking different languages is still unknown.

Continuing along these lines, this study investigates the effects of pitch range, prosodic position, as well as creak locality on creak identification by Mandarin listeners using a fully controlled experiment. As reviewed above, creak presence in speech production is closely related to sentence-final positions, unstressed syllables, low pitch range as well as low tonal targets (as in tonal languages). It remains unclear if these factors are also important cues for listeners to identify creak. In this respect, Mandarin is a good test ground since it contains all these prosodic factors that could potentially trigger creaky voice. Moreover, in Mandarin, no social bias of creak associated with gender has been reported, making it a good comparison with the current findings in American English. Therefore, in our current study, we look at the effects of prosodic position (sentence-final vs. nonfinal), creak locality (global vs. local), pitch range (high vs. low) while controlling tonal effects on creak identification in Mandarin.

2. Method

2.1. Experimental design

A 8 (Tone) $\times$ 2 (Pitch) $\times$ 2 (Prosodic position) $\times$ 3 (Creak locality) within-subject design was implemented. There were three critical conditions: Pitch range condition, Prosodic posi-

10.21437/SpeechProsody.2022-98

tion condition as well as Creak locality condition.

2.2. Materials

2.2.1. Stimulus design

A total number of 64 simple declarative sentences were constructed. All the sentences were 12-syllable long. They were formed using the same syntactic structure (NP1-VP-NP2) but varying in terms of the exact content and lexical items. For each sentence, both NP1 and NP2 were disyllabic person names, of which only the tone of the second syllable was different (i.e., X Y1 vs. X Y2). Segments used to make up these names were all sonorants. The creak-containing syllables differed in terms of **Prosodic position** (either at sentence-final or non-final position), **Pitch range** (high-pitched vs. low-pitched: sentences were first read by a high-pitched female native speaker of Mandarin and were then manipulated into a low-pitched male voice) and **Creak locality** (global creak vs. local creak: global creak refers to scenarios where the surrounding 4-5 syllables of the creak-containing target were also creaky; local creak refers to cases where only the creak-containing target was creaky). Syllables in which creak were present were defined as creak-containing, regardless of its proportion and location in the syllable. For each target syllable, another two modal sentences were included to balance items with and without creak.

A schema of our manipulation of the creaky syllables is illustrated in Table 1 (M stands for modal and C stands for creak). Figure 1 further present examples of these creak types.

Table 1: *A schema of the experimental design*

Sentence	Prosodic Position	Creak Locality
MMMMMMMCCCCC	SentenceFinal	Global creak
MMMMMMMMMMC	SentenceFinal	Local creak
CCCCCMMMMMMM	SentenceNonfinal	Global creak
MCMMMMMMMMM	SentenceNonfinal	Local creak

2.2.2. Recording

The 64 sentences were naturally produced by a female native speaker of Mandarin in a professional recording booth using a high-quality BlueSnowball iCE microphone. Sentences were produced in equalized speech rates (at 40 bpm using an online metronome) with a sentence-final falling intonation as a way to further control the intonation patterns, sentence duration as well as speech rate. Recordings were digitized at a sampling rate of 44,100 kHz and 32-bit sample width. Recordings were saved in mp4 format and then configured for wav format using the software Audacity. The resulting audio data files were then analyzed using the software program Praat [18] and each sound file lasted for 2-3 seconds in duration.

The 64 sentences were then manipulated into low-pitched targets as if they were read by a low-pitched male voice to control speaker-induced variations (64 x 2= 128 sound files). To make sure the voice after manipulation sounds 'male', vowel formant frequencies and pitch range were also adjusted. In the end, the mean F_0 for the low-pitched recordings was 110 Hz whereas the mean F_0 for the original high-pitched recordings was 225 Hz. All the stimuli were normalized to an average intensity of 65 dB.

There were two recording deviants as one syllable was missed during sentence recording. Four audios were excluded from the final experiment due to unnaturalness (all of them were globally creaky). In the end, 124 experimental items were included in the final experiment with 122 having 12 syllables and two having 11 syllables.

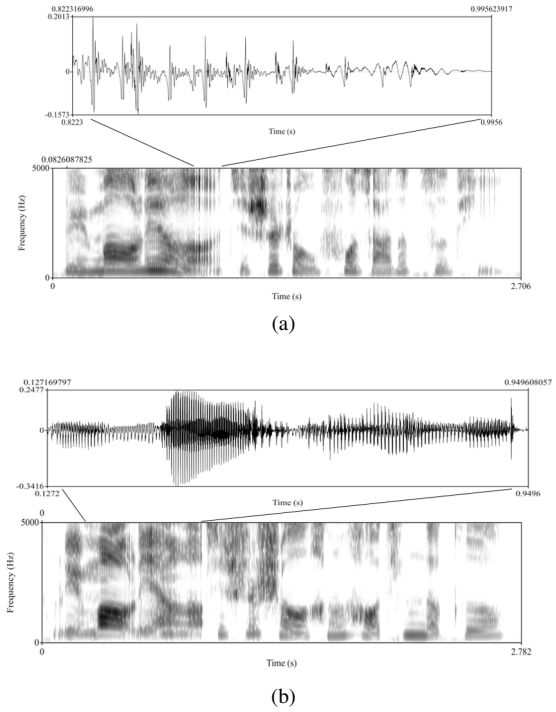

Figure 1: *Examples of creak locality. The waveform corresponds to the syllables that were produced creak (a): local creak (only one syllable was creaky); (b): global creak (multiple syllables were creaky*

2.3. Participants

A total number of 40 native speakers of Mandarin participated in this study online (self-reported: 8 men, 33 women). All of them were recruited from the mainland of China and were paid 20 RMB for their participation. Participants ranged from 19 to 36 years old (mean: 25.12). All the participants speak both standard Mandarin and another Mandarin dialect as their native language. 24 of them self-reported that they have never heard of "vocal fry" or "creaky voice" prior to the study. No one reported to have hearing deficits.

2.4. Procedure

The identification task was designed using the platform of Qualtrics and was conducted in Mandarin Chinese. After agreeing to participate, participants were provided a link to the experiment and were asked to follow the instructions provided. Participants were first informed of the purpose of the study and then were asked to identify for each sentence, **whether** and **where** creaky voice occurred. In other words, they needed to choose all the characters that they thought were creaky by clicking on the boxes below those characters. To familiarize participants with what creaky voice sounds like, a sample audio of creaky voice was provided during the introduction as a training phase and they were encouraged to listen to the sample audio as many

times as they wanted. Participants then went through four practice trials which resembled the format of test trials where they needed to choose all the characters where they heard creak for each auditory sentence, with feedback being provided immediately afterwards. After the practice trials, participants were then asked to listen to the audio carefully then choose, out of all the characters of the sentence, those that they thought were produced creaky. At the end of the identification task, participants were further asked for their demographic information. The whole experiment took around 25 minutes to finish.

3. Results

Statistical analyses were conducted using the R Statistical environment (Team, 2013); mixed-effects logistic regression was run using the lme4 library [19], and plots were created using ggplot [20]. Because creak locality only matters to creaky sentences but not to modal sentences, the analyses of target syllables in creaky and modal sentences were conducted separately.

3.1. Creaky syllables

The first analysis was of the creaky syllables. Figure 2 shows aggregated mean rates of creak identification for creaky syllables. Overall, the presence of creak could be reliably perceived, with identification rates being high above chance level across the board. In addition, the likelihood of perceiving creak was higher for creaky syllables at sentence non-final positions.

A mixed-effects logistic regression model was conducted to predict participants' responses, i.e., whether they perceived the presence of creak for creaky syllables or not, with Prosodic position, Creak locality and Pitch range (in a three-way interaction) as fixed effects (sum-coded) and Participant, syllable, and Tone as random intercepts. Model selection did not suggest excluding any one of these predictors.

The model output is summarized in Table 2. According to the results, there is a main effect of Pitch range, as listeners are less likely to perceive the creaky sound as creaky when it is high-pitched (β = -0.29, $p < 0.001$). The main effect of Prosodic position is also significant, suggesting that listeners are less likely to correctly identify creak in sentence-final positions (β = -0.28, $p < 0.001$). However, for high-pitched syllables, the probability of perceiving creak at sentence-final positions, becomes even significantly smaller, as suggested by the interaction between High Pitch range and Final Prosodic position (β = -0.10, $p = 0.03$). There is a main effect of Global creak, implying that the likelihood of correctly identifying a creaky syllable is significantly boosted when their surrounding syllables are also creaky ($\beta = 0.32$, $p < 0.001$). No significant effect is found for the interaction between High Pitch range and Global Creak locality. The interaction between Final Prosodic position and Global creak locality is significant, meaning that even though listeners are more likely to perceive creak when its surrounding syllables are creaky, this locality effect becomes less pronounced for sentence-final syllables (β = -0.22, $p < 0.001$). Finally, the significant three-way interaction between High Pitch range, Final Prosodic position and Global Creak locality further suggests that the creak locality effect reported above becomes even smaller for high-pitched syllable at sentence-final positions (β = -0.13, $p < 0.01$).

3.2. Modal syllables

Now, we turn to the analysis of how listeners perceived modal syllables. Figure 3 illustrates the aggregated rates of perceived

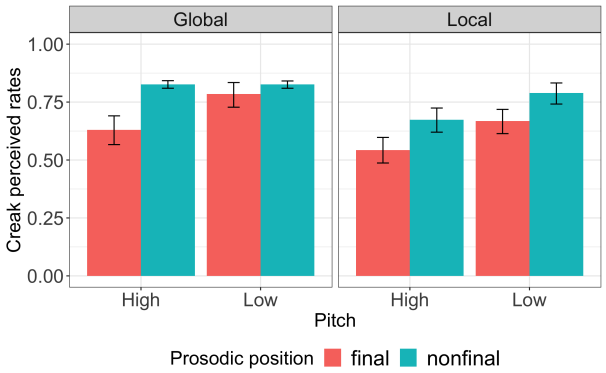

Figure 2: *The mean perceived creak rates for creaky syllables in different conditions*

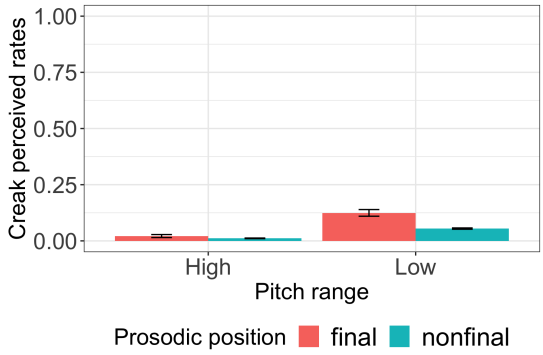

Figure 3: *The mean perceived creak rates for modal syllables in different conditions*

creak among modal syllables. It is shown that the amount of creak that was perceived by listeners among modal syllables decreased drastically overall, compared to the amount of creak that was perceived for creaky syllables, suggesting that listeners were in general accurate at identifying creak (c.f. Figure 2). For modal syllables specifically, listeners were inclined to perceive a modal syllable as creaky when it was at the sentence-final position and when it was low-pitched.

A similar mixed-effects logistic regression model was configured to predict participants' responses to modal syllables, i.e., whether they perceived creak for modal syllables or not, with Prosodic position and Pitch range (in a two-way interaction) as fixed effects (sum-coded) and Participant, syllable, and Tone as random intercepts. The model output, as shown in Table 3, reveal a main effect of Pitch range (β = -0.94, $p <$ 0.001) but no significant effect is found for Prosodic position (β = 0.06, p = 0.27) or the interaction between Pitch range and Prosodic position (β = -0.07, $p < 0.14$). Taken together, results for modal syllables suggest that listeners were likely to false-alarm on the low-pitched modal syllables.

4. Discussion and Conclusion

The current study used a controlled experiment to explore how Mandarin listeners identify creak. In particular, we tested the effects of Creak locality (global vs. local), Prosodic position (sentence-final vs. non-final), and Pitch range (high vs. low) on listeners' identification of creak. Our results suggest that all

Table 2: *Fixed effects of the model on the creak identification of creaky syllables*

Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	1.19	0.26	4.51	<0.001 ***
Pitch range (High)	-0.29	0.05	-6.44	<0.001 ***
Prosodic position (Final)	-0.28	0.06	-4.55	<0.001 ***
Creak locality (Global)	0.32	0.05	5.89	<0.001 ***
Pitch range (High) : Prosodic position (Final)	-0.10	0.05	-2.23	0.03 *
Pitch range (High) : Creak locality (Global)	0.08	0.05	1.74	0.08 .
Prosodic position (Final) : Creak locality (Global)	-0.22	0.05	-3.96	<0.001 ***
Pitch range (High) : Prosodic position (Final) : Creak locality (Global)	-0.13	0.05	-2.80	<0.01 **

Formula: Response $\sim$ Pitch range $\times$ Prosodic position $\times$ Creak locality + (1|Participant) + (1|Syllable) + (1|Tone)

Table 3: *Fixed effects of the model on the creak identification of modal syllables*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.68	0.19	-24.10	<0.001 ***
Pitch range (High)	-0.94	0.05	-20.00	<0.001 ***
Prosodic Position (Final)	0.06	0.06	1.10	0.27
Pitch range (High) : Prosodic Position (Final)	-0.07	0.05	-1.46	0.14

Formula: Response $\sim$ Pitch range $\times$ Prosodic position + (1|Participant) + (1|Syllable) + (1|Tone)

these factors are crucial cues in shaping how Mandarin listeners identify creaky voice.

For creaky syllables, the identification would be more accurate if the target syllables were in a context where their surrounding syllables were also creaky (i.e., global creak). In addition, creaky syllables at sentence non-final positions are easier to be perceived and identified. Creak identification is likely to be inhibited for creaky syllables located at sentence-final positions. Since sentence-final creak is a prosodic property signaling the end of declarative sentences in Mandarin, listeners may be more accustomed to creak at sentence-final positions and are less likely to notice their presence. The perception of sentence-final creak becomes harder when the syllables are read in a high-pitched voice. In a global creaky environment, the sentence-final creak is even less likely to be detected. Pitch range turns out to be another important cue for correctly identifying creak. High-pitched syllables are less likely to be perceived as creaky, giving rise to more misjudgment. Therefore listeners are less accurate and would not tend to perceive a high-pitched sound as creaky.

In terms of modal syllables, there are in fact two types of modal syllables based on our set-up: Some modal syllables are surrounded by creaky syllables in the same sentence, others are located in sentences completely comprised of modal syllables. In the latter case, modal syllables are less likely to be influenced by co-occurring creaky syllables and are more likely to trigger false alarms. For modal syllables in a creaky context, identification is more accurate when they are high-pitched. Sentence-final modal syllables are less likely to be perceived as creak, which suggests that listeners tend to perceive sentence-final pitch declination and creak distinctly. Interestingly, modal syllables are also subject to the locality of creaky syllables in the same sentence. When multiple creaky syllables are present in the sentence, listeners are less likely to perceive modal syllables as creaky because they are sound very differently from those creaky ones. But when there is only one creaky syllable in the sentence, listeners are more likely to perceive modal syllables as creaky as they are less certain about the location of creak. Finally, for modal syllables in modal sentences, listeners show a significant tendency to perceive low-pitched and sentence-final syllables creaky as they false alarm on low-pitched targets.

Notably, [11] found that there existed a weak tendency for English listeners to identify creak more often in females than males when the whole utterance was creaky, our results suggested the opposite: Mandarin listeners consistently identified creak more often in the male voice. [11] attributed her finding to how creaky voice is evaluated differently between female and male voices in the English context. Our future work will investigate whether this gender asymmetry in the social evaluation of creak also exists among Mandarin listeners. A comparison of linguistic evaluations based on these two languages will further shed light on how speech perception can be influenced or modulated by both linguistic contexts and social evaluation.

Our future work will also look at the effects of different tonal categories on creak identification, which is treated as a random effect in the current experiment and has not been fully addressed in our current results. Results on the influence of tone category will add to our understanding of how the perception of phonation type is shaped by language-specific characteristics.

Finally, our results show that creak co-occurring cues modulate listeners' sensitivity to creak in different ways, giving rise to a broader question of how contextual cues of different nature interact with speech perception. Intriguingly, although creak naturally co-occurs with sentence-final positions and low-pitched targets in speech production, these two factors play different roles in the identification of creak: Sentence-final positions can inhibit the identification of creak whereas low pitch always facilitates the identification of creak. These observations open up the possibility that covarying cues of different nature may interact with speech perception in different ways. The question of how these interactions vary depending on the nature of contextual cues still needs further inquiry.

5. Acknowledgements

The authors would like to thank members of the UPenn Phonetics Lab and Mark Garellek for their feedback.

6. References

[1] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice." in *ICPhS*, 2015.

[2] K. M. Yu and H. W. Lam, "The role of creaky voice in cantonese tonal perception," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1320–1333, 2014.

[3] Y. Huang, "Different attributes of creaky voice distinctly affect mandarin tonal perception," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1441–1458, 2020.

[4] J. Bishop and P. Keating, "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1100–1112, 2012.

[5] J. Kuang and M. Liberman, "Pitch-range perception: The dynamic interaction between voice quality and fundamental frequency." in *InterSpeech*, 2016, pp. 1350–1354.

[6] G. Kuo, "Perceived prosodic boundaries in taiwanese and their acoustic correlates," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] M. J. Crowhurst, "The joint influence of vowel duration and creak on the perception of internal phrase boundaries," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. EL147–EL153, 2018.

[8] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407–429, 2001.

[9] J. Esling, "The identification of features of voice quality in social groups," *Journal of the International Phonetic Association*, vol. 8, no. 1/2, pp. 18–23, 1978.

[10] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111–e116, 2012.

[11] L. Davidson, "The effects of pitch, gender, and prosodic context on the identification of creaky voice," *Phonetica*, vol. 76, no. 4, pp. 235–262, 2019.

[12] ——, "The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages," *Wiley Interdisciplinary Reviews: Cognitive Science*, p. e1547, 2020.

[13] I. P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[14] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 37, no. 1, 2011, pp. 427–448.

[15] G. Oliveira, A. Davidson, R. Holczer, S. Kaplan, and A. Paretzky, "A comparison of the use of glottal fry in the spontaneous speech of young and middle-aged american women," *Journal of Voice*, vol. 30, no. 6, pp. 684–687, 2016.

[16] N. B. Abdelli-Beruh, L. Wolk, and D. Slavin, "Prevalence of vocal fry in young adult male american english speakers," *Journal of Voice*, vol. 28, no. 2, pp. 185–190, 2014.

[17] S. D. Greer and S. J. Winters, "The perception of coolness: Differences in evaluating voice quality in male and female speakers." in *ICPhS*, 2015.

[18] P. Boersma, "Praat: doing phonetics by computer (version 5.0. 42)[computer program]," *http://www. praat. org/*, 2008.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[20] H. Wickham, "ggplot2," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 2, pp. 180–185, 2011.