Perceiving the social meanings of creaky voice in Mandarin Chinese

Yao Yao[1], *Meixian Li*[1], *Shiyue Li*[1], *Charles B. Chang*[2]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]Boston University, USA

y.yao@polyu.edu.hk, meixian.li@connect.polyu.hk, shiyue.li@connect.polyu.hk, cc@bu.edu

Abstract

While there is a growing literature on the social meanings of nonmodal voice qualities, most of the existing studies focus on English and use either naturally produced speech stimuli (which are hard to control acoustically) or a small set of fully synthesized stimuli. This paper reports a perceptual study of the social meanings of creaky voice in Mandarin Chinese, using a large set of resynthesized stimuli featuring 38 talkers (19F) and 6–10 pairs of sentences per talker that differed only in voice quality (creaky vs. modal). Sixty listeners (33F) answered 4 questions about the talker's demographic profile (age, gender, sexuality, education) and gave 19 ratings of personality traits (e.g., confident, professional, charismatic) and interactive potential (e.g., engagingness). Using factor analysis and mixed-effects modeling, our results showed that for male listeners, creaky voice significantly decreased the perceived warmth of male talkers but increased the perceived warmth of female talkers; creaky voice also led to more gender identification errors on female talkers by female listeners and made male talkers sound older. These findings point toward multifaceted social meanings of creaky voice in Mandarin, which extend beyond talker attractiveness and are closely linked to gender, both the talker's and the listener's.

Index Terms: social perception, creaky voice, voice quality, phonation type, Mandarin Chinese

1. Introduction

1.1. Social perception of nonmodal voice qualities

In daily use of spoken language, listeners are often exposed to nonmodal voice qualities such as creaky voice (i.e., "vocal fry", involving low F0 and semiregular/irregular glottal pulsing) and breathy voice (involving incomplete vocal fold closure during the closed phase of vibration). The perception and production of nonmodal voice qualities have been of growing interest to both speech-language pathologists and linguists in the past decade: while speech pathologists are mainly concerned with the effects of aging, smoking, fatigue, and other physiological or lifestyle factors on voice quality (e.g., [1], [2]), linguists are interested in how voice quality interacts with language, cognition and communication.

Of particular interest to us is the research on the socio-indexical functions of nonmodal voice qualities, which links creaky voice—the focus of this paper—and breathy voice with the perception of various social properties of the talker. The existing literature documents a wide range of socio-indexicalities associated with creaky voice, mostly based on studies of (North American) English (see [3] and [4] for recent reviews). Notably, the social meanings of creaky voice are often gendered. Although creaky voice in English was traditionally tied to male speech and masculinity, partly due to the concomitant low F0, since the 2000s creaky voice has been increasingly associated with young female speakers (in North America); however, it remains contested whether young women indeed produce more creaky voice than other groups [5]. Regardless, existing research (e.g., [6]–[9]) reports overall more negative social perceptions of vocal fry for female speech (e.g., as "vain", "bored", "sleepy", "less competent/educated /attractive") than for male speech (cp. "authoritative", "cool", "attractive"), although, at least for some listeners, creaky voice in young female speech could index positive social traits (e.g., "educated"/"upward mobility" [9]; "chill", "sexy", "cool" [8]).

In addition to gender, creaky voice has been associated with sexuality and gender identity [10], social class [11], [12], ethnicity [13], and social personae (e.g., gangster persona [14]; "chilled" adolescent [15]). Apart from mainstream North American speech, this line of research draws evidence from other varieties of English (e.g., UK English, Chicano English, Maori English in New Zealand). By comparison, much less is known about the socio-indexicalities of creaky voice in other languages, including Chinese languages.

1.2. Creaky voice in Mandarin Chinese

In tonal languages such as Mandarin and Cantonese, creaky voice is known to have a close relationship with the perception and production of lexical tones, especially the low-pitch tonal targets, given the association between creaky voice and low pitch [16]–[18]. That is, unlike in non-tonal languages, creaky voice often plays a functional role in distinguishing lexical tones and, by extension, different words.

This typological difference vis-à-vis non-tonal languages has potential implications for the indexical functions of creaky voice in a language like Mandarin, yet, to the best of our knowledge, there has not been any systematic investigation of the socio-indexical meanings of creaky voice in Mandarin. Some preliminary findings by [19] suggest that Mandarin listeners overall disfavor creaky voice, while [20] reported the use of creaky voice to portray dangerous female sexuality in Chinese TV shows. More recently, [21] and [22] attempted to examine the social perception of creaky voice in Mandarin, but with a small set of stimuli, these studies reported overall lower talker attractiveness for creaky voice than modal voice, which is consistent with previous research, while finding no gendered patterns of creaky voice perception.

1.3. Current study

In the current study, we applied a novel method of resynthesizing modal and creaky utterances from naturally produced modal utterances to the investigation of creaky voice perception in Mandarin. The resynthesized stimuli were used in

a social perception experiment where listeners evaluated talker properties based on the (resynthesized) speech they heard. With this approach, we were able to conduct a social perception experiment with a sizable group of talkers and listeners and well-controlled speech stimuli.

2. Methods

2.1. Participants

The experiment enrolled 60 adult listeners (33F, 27M; ages 18–35) from a university in Hong Kong. All the participants were native Mandarin Chinese speakers born and raised in Mainland China. They exhibited normal or corrected-to-normal hearing and vision and had no known speech or language disorders. None of the participants had previously studied linguistics or psychology.

A separate group of 40 Mandarin speakers (20F, 20M; ages 20–35) with similar linguistic and demographic backgrounds as the listeners were recruited as talkers. In addition, one female Mandarin speaker with speech-language pathology expertise served as an expert rater of stimuli naturalness and intelligibility, and a separate group of 22 Mandarin speakers (11F, 11M) with no prior training in phonetics served as nonexpert raters.

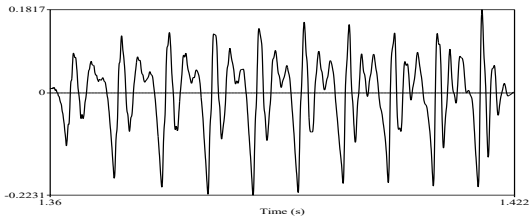
2.2. Stimuli

The stimuli for the experiment were constructed based on speech samples from the talkers, comprising 120 emotion-neutral Mandarin sentences averaging ten characters (syllables) in length (see https://osf.io/6hc7n/ for a full list of the sentences, other study materials, data, and model outputs). Due to Covid restrictions on in-person testing, the stimulus recording was done over an online recording platform using the talker's own device at a quiet location of the talker's choice. The utterances were then segmented in Praat [23] into consecutive voiced portions (e.g., vowels, sonorant consonants) and voiceless portions (e.g., voiceless consonants).

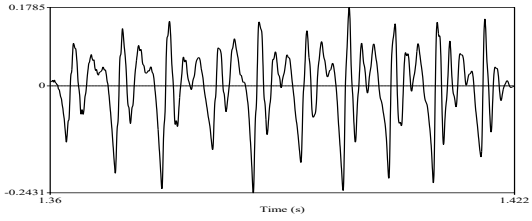
After the segmenting, a resynthesis process was applied to the speech samples to create two resynthesized versions (modal and creaky) of each utterance. Previously annotated voiced portions were resynthesized with a Klatt synthesizer that tracked the pitch, formants, and intensity profiles of the original production and then concatenated the resynthesized portions with neighboring (voiceless) portions with smoothed boundaries. The only difference between the modal and creaky versions was that the creaky resynthesis inserted double pulsing points in **all** the voiced portions, resulting in doubly-pulsed creak throughout the creaky utterance ([24]; see Figure 1). Thus, the two versions of a given utterance differed solely in voice quality and were otherwise acoustically identical.

To obtain a subset of resynthesized tokens with good quality, a two-stage selection process was applied. After members of the research team excluded stimuli with significant flaws (e.g., low intelligibility, background noise, or substantial distortion), the remaining tokens (N = 3768) were evaluated by one expert rater and a group of nonexpert raters for intelligibility and naturalness (separately) on a 5-point scale. The combined expert (80%) and nonexpert (20%) ratings were used to select the best 6–10 utterance pairs (modal vs. creaky) per talker, excluding two talkers (1F, 1M) whose resynthesis quality was exceptionally low due to fast speech rates.

The final set of stimuli for the social perception experiment consisted of 38 talkers (19F, 19 M) and 328 utterance pairs (on average 8.63 pairs per talker, SD = 1.15). Modal/creaky utterances of the same talker were concatenated into one sound file, so each talker had a collection of modal utterances and a collection of creaky utterances that were maximally similar. Acoustic analysis confirmed that the creaky utterances had significantly lower H1*–H2* and HNR (harmonic-to-noise ratio in the 0–3500 Hz spectrum) values than the corresponding modal stimuli (see Table 1), consistent with previously documented acoustic differences between modal and creaky productions [24].



(a)



(b)

Figure 1: *Modal (a) and creaky (b) resynthesis of an [i] vowel produced by a female talker.*

Table 1: *Acoustic properties of creaky and modal stimuli. Acoustic measures obtained with VoiceSauce* [25].

Voice quality	H1*–H2* (dB)	HNR (dB)
Creaky	M = -6.83 (SD = 6.04)	M = 29.19 (SD = 4.01)
Modal	M = 4.42 (SD = 3.41)	M = 43.63 (SD = 6.20)

2.3. Procedure

The social perception experiment comprised a single laboratory session of approximately one hour. After providing informed consent, participants completed a perception task on the Gorilla platform, using a mouse and keyboard while seated before a 24-inch monitor equipped with a headset. The experiment consisted of two parts: a social perception task followed by a post-task questionnaire.

In the social perception task, each participant listened to the resynthesized auditory stimuli of 14 talkers (balanced across talker gender and voice quality) and provided judgments. Each participant heard one voice quality version (modal or creaky) of a given talker, and each talker was evaluated by at least 10 listeners in each voice quality. Participants were told that the voices they would hear were synthesized by speech engineers and were asked to provide subjective evaluations of the imaginary talkers. Participants assessed four demographic characteristics (age, gender, sexuality, education level) and

rated 15 personality traits (e.g., confidence, professionalism, charisma) and four aspects of interactive potential (e.g., engagingness, friendliness) of the talker on a 5-point scale.

In the post-task questionnaire, participants answered questions concerning gender attitudes, where they rated degree of agreement with statements describing gender stereotypes (or gender equality) and homosexual tolerance (or intolerance).

In this paper, we focus on the analysis of data from the social perception task.

2.4. Analysis

Ratings of personality traits and interactive potential were combined and submitted to a factor analysis in R (v3.4.1; [26]). Mixed-effects models with by-talker and by-listener random intercepts and slopes were built to examine the effects of voice quality, talker gender, and listener gender on talker evaluations. All the mixed-effects models were built with the lmerTest R package (v3.1.3; [27]). Models were initially built with maximal fixed- and random-effect structures, and then underwent backward elimination to eliminate non-significant fixed predictors and to ensure model convergence.

3. Results

3.1. Factor analysis of perceived talker personality traits and interactive potential

Informed by the results of the scree test and parallel analysis [28], we conducted a factor analysis of talker personality traits and interactive potential, and yielded three factors, which we name as COMPETENCE, LIKEABILITY, and WARMTH. The COMPETENCE factor is mainly loaded by ratings of the talker sounding professional, formal, confident, smart, hard-working, authoritative, and convincing, and **not** lazy, casual, or hesitant. The WARMTH factor is mainly loaded by ratings of the talker sounding gentle, genuine, and hard-working, but **not** aggressive or pretentious, as well as ratings of the listener wanting to talk more with and befriend the talker. The LIKEABILITY factor is loaded by both competence-related (e.g., talker sounding confident, convincing, charismatic, and authoritative) and warmth-related (listener wanting to talk more with and befriend the talker) items.

3.2. Effects of voice quality, talker gender, and listener gender on the perception of personality traits and interactive potential

Mixed-effects models showed that both talker gender ($p = .001$) and the interaction of talker gender and listener gender ($p = .026$) had significant effects on the perception of talker COMPETENCE. Female talkers were higher in COMPETENCE scores than male talkers, and the cross-gender differences were greater for female listeners than male listeners. No significant effects of voice quality or interactions of voice quality and talker/listener gender were observed.

A marginal effect of the interaction between talker gender and listener gender ($p = .073$) was found on LIKEABILITY: male listeners gave female talkers slightly higher LIKEABILITY-associated ratings than they did male talkers, but this difference was not present for female listeners. No significant effects of voice quality or interactions of voice quality and talker/listener gender were observed.

As for WARMTH, however, the models revealed significant two-way interactions between voice quality and talker gender ($p = .002$) and between talker gender and listener gender ($p = .003$) and a significant three-way interaction between voice quality, talker gender, and listener gender ($p = .004$). Post-hoc analyses suggested that the effects of voice quality were mainly driven by male listeners. Specifically, male listeners perceived female talkers as **higher** in warmth when speaking in creaky than in modal voice; in contrast, they perceived male talkers as **lower** in warmth when speaking in creaky than in modal voice (see Figure 2; in all figures, error bars indicate 95% confidence intervals and MOD = modal, CRK = creaky). In other words, for male listeners, creaky voice increased the perceived warmth of female talkers but decreased the perceived warmth of male talkers. Female listeners, on the other hand, seemed to be unaffected by voice quality in their perception of talker warmth.

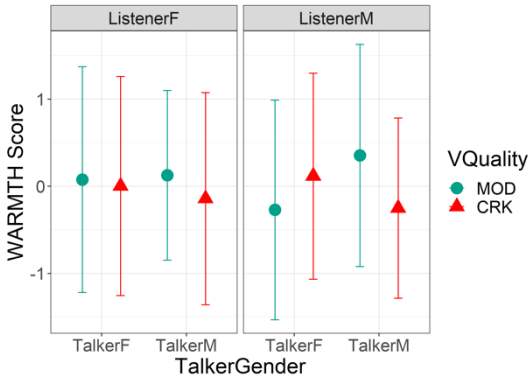


Figure 2: *Average WARMTH score by voice quality, talker gender, and listener gender.*

3.3. Effects of voice quality, talker gender, and listener gender on the perception of age, gender, and education

Listeners' estimation of talker age was overall quite accurate. As shown in Figure 3, the estimated age was around "3" on a 6-point scale, corresponding to the age range of 21–25, which coincides with the age range of the majority of the talkers at the time of speech recording.

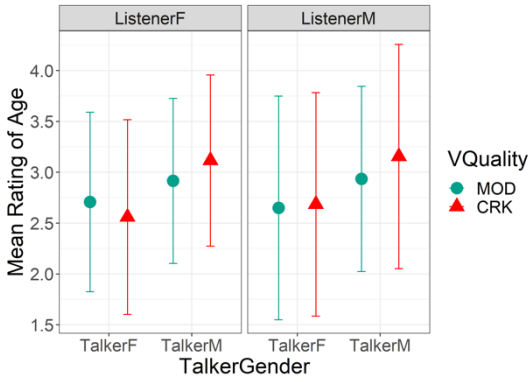


Figure 3: *Average rating of talker age (on a 6-point scale) by voice quality, talker gender, and listener gender. On the rating scale: 1 = under age 16; 2 = 16–20; 3 = 21–25; 4 = 26–30; 5 = 31–35; 6 = 36–40.*

The models revealed marginal effects of talker gender ($p = .076$) and the interaction between voice quality and talker gender ($p = .061$) on the perception of talker age: male talkers

were overall perceived as older than female talkers, and furthermore, creaky voice **increased** the estimated age of male talkers but had no effect on that of female talkers.

Consistent with previous literature reporting swift and accurate talker sex assignment [29], listeners in the current study were overall highly accurate in identifying talker gender, yielding a total accuracy rate of over 96%. However, analysis of gender identification errors (N = 29) suggested that female talkers (accounting for 27 out of 29 cases) and female listeners (accounting for 20 out of 29 cases) were far more prone to errors than male talkers and listeners. On top of that, creaky voice seemed to exacerbate the confusion of gender perception observed for female talkers and listeners: about two-thirds (14 out of 20) of identification errors with female talkers and female listeners occurred with creaky voice, whereas only one-third occurred with modal voice. It should be noted that gender identification errors, albeit infrequent in general, affected a sizable group of unique talkers (N = 9) and listeners (N = 21), and thus cannot be easily attributed to idiosyncratic features of talkers/listeners (e.g., some talkers' voices being naturally gender-ambiguous or some listeners being particularly inaccurate with voice gender perception).

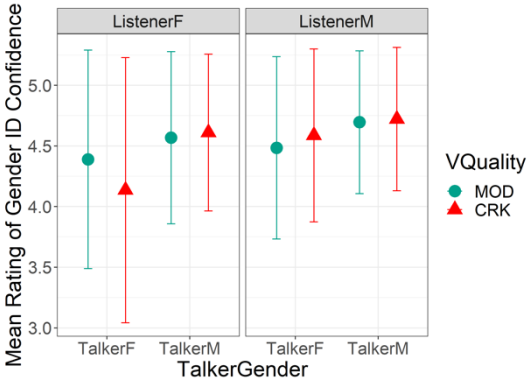


Figure 4: *Average confidence level of talker gender identification by voice quality, talker gender, and listener gender.*

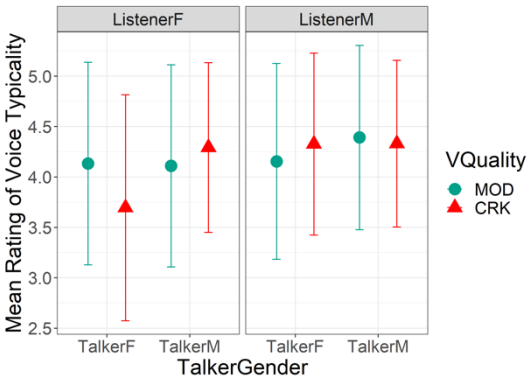


Figure 5: *Average rating of voice gender typicality by voice quality, talker gender, and listener gender.*

The patterns of talker gender identification reported above are supported by listeners' reports of their confidence level of talker gender identification and their ratings of the talkers' voice gender typicality (both on a 5-point scale). As shown in Figures 4 and 5), female listeners hearing female talkers in creaky voice reported the lowest confidence level of talker gender identification and the lowest rating of voice gender typicality compared to all other conditions, and the differences were statistically significant (p's < .04) as shown in mixed-effects models.

No significant effects on the perception of education level were observed.

4. Discussion

In this paper, we report a social perception study of creaky voice (as compared to modal voice) in Mandarin Chinese. By using resynthesized speech samples, the modal and creaky stimuli were strictly controlled to only vary in voice quality but not any other acoustic properties. Our results from 38 talkers and 60 listeners revealed significant, interacting effects of voice quality and gender (both talker gender and listener gender) on the perception of multiple talker properties (age, gender, and personality traits). We showed that when creaky voice was present, female talkers were perceived as higher in warmth (by male listeners) and were more likely to be misidentified as male (by female listeners), whereas male talkers were perceived as lower in warmth (by male listeners) and older in age (by both male and female listeners).

Some of these effects, such as increased age estimation for male talkers and increased gender identification errors for female talkers, could be explained via the effects of creaky voice on lowering perceived pitch [30], [31]. Furthermore, the creaky stimuli used in the current study evinced doubly-pulsed creak, which superimposes an additional low F0 on the original F0 that could lead to an overall lower perceived pitch for the creaky than the modal stimuli. Thus, lower pitch perception could be responsible for creaky male talkers sounding older (i.e., more mature) and creaky female talkers sounding more gender-ambiguous (i.e., less stereotypically female-sounding), although it is not clear why the latter result was only observed in female listeners. The finding of creaky male talkers being perceived as lower in warmth by male listeners could be explained in a similar vein: since the talkers and listeners are similar in age in real life, more mature-sounding male talkers might be perceived as more distant and less approachable by their peers, especially those of the same sex.

The effect of creaky voice on the perception of warmth in female talkers by male listeners is compatible with a previous report of creaky voice being connected with female sexuality in Chinese TV shows [20]. It is possible that creaky voice increases the sexual appeal of female talkers for male listeners; if true, this would explain why the effect was not observed for other combinations of talker and listener gender. However, more research is needed to verify the validity of this account.

To summarize, results of this study reveal several patterns of gender-related social perception of creaky voice in Mandarin. Not only did the social meanings of creaky voice vary by talker gender, the perception of such meanings was also sensitive to listener gender. These findings highlight the multifaceted nature of the socio-indexicalities of creaky voice, and call for further research on the perception of nonmodal voice qualities in diverse languages.

5. Acknowledgements

This research is supported by grant no. 15611322 awarded to PI Yao Yao by the Research Grants Council (General Research Fund) of Hong Kong.

6. References

[1] 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," *J. Voice*, vol. 30, no. 6, pp. 684–687, 2016.

[2] D. Pinar, H. Cincik, E. Erkul, and A. Gungor, "Investigating the effects of smoking on young adult male voice by using multidimensional methods," *J. Voice*, vol. 30, no. 6, pp. 721–725, 2016.

[3] L. Davidson, "The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 12, no. 3, pp. 1–18, 2021.

[4] R. J. Podesva and P. Callier, "Voice quality and identity," *Annu. Rev. Appl. Linguist.*, vol. 35, pp. 173–194, 2015.

[5] K. Dallaston and G. Docherty, "The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature," *PLoS One*, vol. 15, no. 3, p. e0229960, 2020.

[6] R. C. Anderson, C. A. Klofstad, W. J. Mayew, and M. Venkatachalam, "Vocal fry may undermine the success of young women in the labor market," *PLoS One*, vol. 9, no. 5, p. e97506, 2014.

[7] S. D. F. Greer and S. J. Winters, "The perception of coolness: Differences in evaluating voice quality in male and female speakers," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 2015, paper 0833.

[8] C. Ligon, C. Rountrey, N. V. Rank, M. Hull, and A. Khidr, "Perceived desirability of vocal fry among female speech communication disorders graduate students," *J. Voice*, vol. 33, no. 5, pp. 805.e21–805.e35, 2019.

[9] I. P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?," *Am. Speech*, vol. 85, no. 3, pp. 315–337, 2019.

[10] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," *Annu. Meet. Berkeley Linguist. Soc.*, vol. 37, no. 1, p. 427–448, 2011.

[11] J. Stuart-Smith, "Glasgow: Accent and voice quality," in *Urban voices: Accent studies in the British Isles*, P. Foulkes and G. Docherty, Eds. Leeds, UK: Arnold, 1999, pp. 201–222.

[12] G. Knowles, "The nature of phonological variables in Scouse," in *Sociolinguistic patterns in British English*, P. Trudgill, Ed. London, UK: Edward Arnold, 1978, pp. 129–149.

[13] A. Szakay, "Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception," *J. Socioling.*, vol. 16, no. 3, pp. 382–397, 2012.

[14] N. Mendoza-Denton, "The semiotic hitchhiker's guide to creaky voice: Circulation and gendered jardcore in a Chicana/o gang persona," *J. Linguist. Anthropol.*, vol. 21, no. 2, pp. 261–280, 2011.

[15] T. Pratt, "Affective sociolinguistic style: An ethnography of embodied linguistic variation in an arts high school," Stanford University Ph.D. dissertation, 2018.

[16] A. Li, W. Lai, and J. Kuang, "Creaky voice identification in Mandarin: The effects of prosodic position, tone, pitch range and creak locality," *J. Acoust. Soc. Am.*, vol. 154, no. 1, pp. 126–140, 2023.

[17] J. Kuang, "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," *J. Acoust. Soc. Am.*, vol. 142, pp. 1693–1706, 2017.

[18] K. M. Yu and H. W. Lam, "The role of creaky voice in Cantonese tonal perception," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1320–1333, 2014.

[19] A. Xu and A. Lee, "Perception of vocal attractiveness by Mandarin native listeners," in *Proceedings of the 9th International Conference on Speech Prosody*, 2018, pp. 344–348.

[20] P. Callier, "Voice quality, rhythm and valorized femininities," paper presented at *Sociolinguistics Symposium 18*, Southampton, UK, November 1–4, 2010.

[21] Q. Li and P. Mok, "A perception study on voice quality and stance in Mandarin Chinese," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2023, pp. 1786–1790.

[22] A. Li and W. Lai, "How do listeners evaluate creak: A matched-guise study in Mandarin Chinese," paper presented at the *Linguistic Society of America 2023 Annual Meeting*, June 5–8, Denver, CO, 2023.

[23] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." 2022.

[24] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 2015, paper 0821.

[25] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, 2011, pp. 1846–1849.

[26] R Development Core Team, "R: A language and environment for statistical computing." Vienna, Austria, 2023.

[27] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, 2017.

[28] J. K. Sakaluk and S. D. Short, "A methodological review of exploratory factor analysis in sexuality research: Used practices, best practices, and data analysis resources," *J. Sex Res.*, vol. 54, no. 1, pp. 1–9, 2017.

[29] M. J. Owren, M. Berkowitz, and J. O. A. Bachorowski, "Listeners judge talker sex more efficiently from male than from female vowels," *Percept. Psychophys.*, vol. 69, no. 6, pp. 930–941, 2007.

[30] L. Davidson, "Contributions of modal and creaky voice to the perception of habitual pitch," *Language*, vol. 96, no. 1, pp. e22–e37, 2020.

[31] J. Kuang and M. Liberman, "Influence of spectral cues on the perception of pitch height," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 2015, paper 0435.