Aini Li* and Wei Lai

Gender effects in Mandarin creaky voice evaluation: a matched-guise study

https://doi.org/10.1515/lingvan-2024-0237
Received November 22, 2024; accepted May 27, 2025; published online August 12, 2025

Abstract: This study investigates how Mandarin listeners evaluate creaky voice across different gendered voices using a matched-guise design. In contrast to findings in American English, where female creak is often evaluated negatively, Mandarin listeners did not display different evaluations of creaky voice based on speaker gender. Personality traits associated with creaky voice, such as attractiveness, competence, likeability, intelligence, and wealth, did not show systematic gender effects in Mandarin, and these dimensions were found to be highly correlated through principal component analysis. These results suggest that the gender bias against creaky female voices observed in American English does not generalize to Mandarin, offering a potential explanation for previously reported crosslinguistic differences in creak identification between American English and Mandarin Chinese. The absence of gender-based social bias in Mandarin may account for listeners' more phonetically driven identification of creak, making creak more identifiable in low-pitched male voices rather than high-pitched female voices. Taken together, the findings underscore the importance of incorporating social evaluation into speech perception studies and call for further social evaluation studies of creaky voice across various social dimensions and in crosslinguistic contexts.

Keywords: creaky voice; speaker gender; social evaluation; Mandarin Chinese

1 Introduction

Creaky voice, characterized by tightly adducted vocal folds and/or irregular vibrations (Gordon and Ladefoged 2001), is one type of non-modal phonation. Non-contrastive creak, the type that does not create phonemic distinctions, is widely known as a carrier of rich social meaning – such as perceptions of a speaker's social identity, status, or emotional tone – as evidenced by a bulk of sociophonetic studies on social associations of creaky voice (e.g., Esling 1978; Esposito 2016, 2017; Hildebrand-Edgar 2016; Mendoza-Denton 2011; Pittam 1987). For instance, it has been shown that non-contrastive creak can denote a speaker's social status (Esling 1978; Pittam 1987), emotions (Gobl and Chasaide 2003; Laver 1980), identity/persona (Esposito 2016, 2017; Hildebrand-Edgar 2016; Mendoza-Denton 2011), and many other associations.

Among these various social associations denoted by non-contrastive creak, one that has garnered significant interest among sociolinguists is the relationship between creak and gender. In speech production, the use of creaky voice in American English seems to be more of a prominent speech feature for young women (Abdelli-Beruh et al. 2014; Melvin 2015; Podesva 2013; Yuasa 2010). However, this gender-based difference is highly contingent on cultural and linguistic factors (Henton 1989; Kuang 2018). For instance, in Taiwan Mandarin, the use of creak does not seem to differ significantly in natural speech production between female and male speakers (Kuang 2018). When it comes to perception and evaluation, in certain varieties of English, such as in American English, creak is more stigmatized for female speakers than male speakers: a female voice that contains creaky voice is perceived as less competent and less hireable (Anderson et al. 2014), whereas a male creaky voice may sound "more authoritative, cool and attractive" (Greer and Winters 2015). Other personality traits associated with

***Corresponding author: Aini Li**, Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China, E-mail: ainili@cityu.edu.hk. https://orcid.org/0000-0002-5190-8065
Wei Lai, Department of Linguistics, California State University Long Beach, Long Beach, CA, USA. https://orcid.org/0000-0001-5250-724X

creak in previous literature include trustworthiness, education, desirability, authoritativeness, likeability, and intelligence (Anderson et al. 2014; Greer and Winters 2015; Ligon et al. 2019; Parker and Borrie 2018).

Crucially, gender-differentiated bias around creak appears to influence its identification in American English. In a creak identification experiment, Davidson (2019) found that listeners were more likely to identify creak when it was not in utterance-final position, and when the surrounding modal speech was low in pitch. That is, listeners often overlooked sentence-final creak when it predictably marked the end of a sentence as a prosodic cue. Additionally, *under certain conditions*, creak was identified more often in female voices: when the entire phrase was creaky, listeners showed a weak tendency to identify creak more frequently in female than male voices, potentially due to their bias against female creak.

In Mandarin, the pattern differs slightly. Li et al. (2023) adopted a similar design to that of Davidson (2019), except that their study tested participants with the same female voice, which was manipulated into a high-pitched female-sounding version and a low-pitched male-sounding counterpart. They found that, like English listeners, Mandarin listeners identified more creak when it appeared in nonfinal sentence position, due to the prosodic distribution of creak, and in the low-pitched male-sounding voice. This suggests that creak identification in tonal languages like Mandarin is similarly sensitive to phonetic environment. However, when utterances contained more extensive creak, Mandarin listeners still showed a preference for identifying more creak in the low-pitched male-sounding voice. Unlike in the English experiment, Mandarin listeners consistently identified more creak in the low-pitched male-sounding voice *across all conditions* (Li et al. 2023).

These differences in creak identification between Mandarin and American English listeners may be attributed to the absence of gender-differentiated social bias when it comes to creak in Mandarin. In English, evaluation bias may lead listeners to expect creak more in female voices, whereas the absence of such a bias in Mandarin could explain the lack of gender-based identification differences. However, little work has examined whether a gender-based evaluation bias around creaky voice exists in Mandarin. This study addresses this question by experimentally investigating whether creak has gender-differentiated social evaluation in Mandarin using a matched-guise experiment.

2 The present experiment

Against this background, the present study examines the influence of speaker gender on Mandarin creak evaluation using the matched-guise technique, which was pioneered by Lambert et al. (1960). The typical matched-guise method involves having listeners hear and give subjective evaluations of speech clips that differ minimally along a certain dimension of linguistic interest. If listeners diverge in their evaluations and judgments, this divergence can be attributed to the manipulated linguistic difference between the clips. Following this method, we presented listeners with pairs of voices from the same speaker, one with a high pitch range and higher vowel formants (resembling a female voice in perception) and the other with a low pitch range and lower vowel formants (resembling a male voice in perception). We then compared listeners' evaluations of highly controlled sentences with and without creak based on a number of social/personal metrics, as informed by previous research on creak evaluation (Anderson et al. 2014; Greer and Winters 2015; Ligon et al. 2019; Parker and Borrie 2018): likeability, competence, intelligence, attractiveness, richness, educatedness, friendliness, and trustworthiness.

An alternative approach to probing social evaluation of creaky voice is to use naturally occurring instances of vocal fry and normal voice from different speakers. We chose not to use this approach because variation in myriad acoustic features of different speakers' voices other than vocal fry could present difficulties in accounting for confounding effects induced by individual-level variabilities. In addition, it has been verified that social evaluation can be robust to speech style. Therefore, the use of read speech stimuli can be an appropriate methodological choice in matched-guise experiments (Tamminga 2017).

2.1 Experimental design

A block design was adopted in the current experiment to ensure that no guises from the same utterance appeared more than once within the same block. This approach prevented listeners from hearing multiple guises of the identical utterance consecutively. For example, if an utterance appeared in its high-pitched creak guise in the first block, it occurred in its other guises in other blocks, with filler items being interspersed among different blocks. Within each block, the critical and filler utterances were randomized, and were evaluated in groups of two. The block design was not made apparent to the participants.

2.2 Guise construction

A total of 64 distinct sentences were originally taken from Li et al. (2023): 32 sentences contained creak (2 Creak locality × 2 Prosodic position × 8 tones [4 lexical tones + 4 realizations of the Neutral tone]) and 32 were fully modal sentences. All of these sentences were declarative, each containing 12 syllables, as illustrated in (1).

(1) *Lǐ Ài zài gōngyuán sànbù pèngdào le Lǐ Āi.*
Lǐ Ài at park walk met ASPECTUAL.MARKER Lǐ Āi
'Lǐ Ài met Lǐ Āi while taking a walk in the park.'
(Li et al. 2023; the red highlight indicates a creak-containing target syllable)

In Li et al. (2023), creak locality distinguishes global creak and local creak, with the former referring to scenarios where the surrounding four or five syllables of the creaky syllable are creaky, and the latter representing cases where only the target syllable is creaky. Prosodic position distinguishes sentence-final position from sentence-nonfinal position. More details regarding the sentence construction can be found in Li et al. (2023). For the purpose of the current study, it was not obvious that listeners would evaluate different tones differently on the basis of the presence of creak. In particular, no evidence has been found yet in the literature about how Mandarin listeners might evaluate creaky Tone 1 and creaky Tone 2 differently. Therefore, we classified sentences into groups of two based on creak locality and prosodic position, overriding the distinction between different tones. In other words, sentence clips of the same type were grouped together such that within the same block, listeners evaluated sentences of the same type as a group. As a result, a total of 16 creak-containing sentences (out of the original 32 creak-containing sentences from Li et al. (2023)) were included as critical stimuli in the current study.

All the critical stimuli (N = 16) were read by a female native speaker of Mandarin. Each target syllable was produced first with creaky voice and then with modal voice, resulting in a total of 16 creaky-modal pairs. The recording session was conducted in a professional sound booth. Speech rate was controlled using an online metronome at 40 bpm. After recording, each sentence was subject to a trial-and-error process of cross-splicing to produce natural-sounding pairs that differed only in the presence of creak. For instance, if a sentence had two creaky syllables, both creaky syllables were replaced with their modal counterparts such that each pair differed only in the presence of creak. Likewise, for modal sentences, target syllables were replaced with the creaky production. This bidirectional cross-splicing ensured that all stimuli were edited to maintain consistency in the degree of manipulation applied.

The female-sounding high-pitched sentences were then converted to sentences with a low pitch range through the gender change function in Praat by lowering the formant (0.75), pitch median (110), and pitch range (0.7). The same cross-splicing process was repeated, this time targeting the low-pitched tokens. These low-pitched tokens served as the male-voice condition in the experiment.

In the end, the same critical sentence appeared in four different guises: female-sounding and creaky, female-sounding and modal, male-sounding and creaky, as well as male-sounding and modal. The experiment included four blocks (two blocks of creak-containing utterances and two blocks of modal utterances). The same 16 critical stimuli (in eight groups) showed in their different guises across the different blocks. Randomization was applied to both the presentation of blocks and the groups within each block. Within each block, there were 10 groups of sentences: eight critical groups and two filler groups. The filler groups were modal sentences selected from the 32

modal sentences in Li et al. (2023). As a result, there were a total of 40 sentence groups subject to listeners' evaluation (10 groups per block × 4 blocks), each with two audio clips of the same type, and listeners needed to give evaluation responses 40 times. In other words, listeners heard two sentences from the same group and rated them as a whole, rather than each sentence individually.

Table 1 showcases the makeup of a block with creak guises. For example, in Group 1, listeners heard two different utterances, each of which was in a female voice and also contained creak on multiple syllables in the sentence-final position.

Since the male-sounding voice was created by gender-shifting the natural stimuli produced by a female talker through acoustic means, it was essential to ensure that the resulting male voice sounded natural and typical. To achieve this, the two voices were evaluated for naturalness and gender typicality through a voice rating test. A total of 24 listeners were recruited for the test. They listened to 20 sentences, which were randomly sampled from the critical stimuli. For each sentence, listeners were asked to rate its naturalness and gender typicality on a 7-point Likert scale. Naturalness was rated from 1 ("very unnatural") to 7 ("very natural"), while gender typicality was rated from 1 ("typical male voice") to 7 ("typical female voice").

The results of this test showed that both voices were rated as natural, although the naturally produced female voice received a higher rating score (mean naturalness rating of the male voice = 5.15; mean naturalness rating for the female voice = 5.75). In addition, both voices were rated as gender typical (mean typicality rating for the male voice = 1; mean typicality rating for the female voice = 7).

2.3 Participants

A total of 40 Mandarin native speakers (self-reported male = 15, self-reported female = 23, self-reported other = 2) were recruited from mainland China to participate in the experiment in return for 30 Chinese yuan. Fourteen of them were under the age of 20, four of them were above 30, and the rest of them fell into the range of 21–30 years old. All participants spoke both standard Mandarin and another Mandarin dialect as their native language. Twenty-seven participants (68 %) self-reported that they had never heard of "vocal fry" or "creaky voice" prior to the study. No one reported having a hearing deficit. Note that none of these participants had taken part in the voice rating test.

2.4 Procedure

The experiment was administered online through the Qualtrics platform and was distributed through an anonymous link. Listeners were informed that they would hear a series of utterances in standard Mandarin and that they needed to evaluate the speaker who said those utterances along different evaluative dimensions.

Table 1: The makeup of a block with creak guises.

Group	Type of sentence	No. of sentences
Group 1	Sentence final + global creak + female (critical)	2
Group 2	Sentence final + global creak + male (critical)	2
Group 3	Sentence nonfinal + global creak + female (critical)	2
Group 4	Sentence nonfinal + global creak + male (critical)	2
Group 5	Sentence final + local creak + female (critical)	2
Group 6	Sentence final + local creak + male (critical)	2
Group 7	Sentence nonfinal + local creak + female (critical)	2
Group 8	Sentence nonfinal + local creak + male (critical)	2
Group 9	Fully modal (filler)	2
Group 10	Fully modal (filler)	2

Listeners were instructed to make their judgments based on their first impressions and to avoid focusing too much on the content of the utterances.

The experiment started with a practice session where listeners first heard an audio clip on one screen and were then asked to rate the voice that they just heard on a 7-point Likert scale using a total of eight different metrics: likeability, trustworthiness, friendliness, educatedness, wealthiness, competence, attractiveness, and intelligence. For example, after the audio was played, listeners were asked: "Based on the voice you just heard, how likeable does this person sound to you?". Listeners then needed to rate the likeability of the speaker on the 7-point Likert scale, with 7 being the most likeable and 1 being the least likeable. Similar questions were likewise asked for the other metrics.

After the practice session, listeners proceeded to the test session, which resembled the procedure of the practice session. However, instead of just one audio clip, listeners in the test session heard two audio clips of the same type before making evaluation judgments. During the test session, listeners could play and replay the stimulus audios one by one as many times as they wanted. No written cues were provided. Another change we made during the test phase was that, apart from rating different metrics, listeners were asked to type in at most three words that could best describe their impression of the speaker voice that they just heard. That is, listeners heard audio clips, then gave ratings of and commented on what they just heard. Note that the audio clips in the practice session were produced by a different Mandarin native speaker, distinct from the speaker voice that listeners heard during the test session. To ensure listeners paid enough attention during the rating task, two catch questions were interspersed within the critical rating items. For these questions, listeners were explicitly asked to choose a specific rating score rather than providing their own ratings.

After the survey, listeners were asked how many different voices they had heard and whether the voices were from the same speaker or different speakers. Since the male voice in the experiment was manipulated from its female counterpart, this question served as a sanity check to make sure that listeners perceived the low-pitched voice as a male speaker. After the sanity check, listeners' demographic information was collected. Listeners were asked about their age, gender (male/female/other/prefer not to reply), hometown, native languages and dialects, whether they had heard of creaky voice before the experiment, whether they had any hearing deficits, whether the sounds they heard in the experiment were clear, what headphones they used, and additional comments they might have about the experiment.

3 Analysis and results

The mean completion time for the experiment was 25 min. All listeners correctly answered the catch questions and indicated that they heard two gender voices (female and male), so no data points were excluded.

Figure 1 shows participants' overall mean ratings of the guises (creak vs. modal) across different voice genders. It is clear that listeners rated creak guises more harshly than modal guises. Other than this general pattern, modal guises in the male voice received lower overall ratings, compared to those in the female voice. Figure 2(a) further presents how listeners evaluated creak and modal guises for different gender voices on the basis of the eight different personality metrics (attractiveness, competence, educatedness, intelligence, likeability, wealthiness, and trustworthiness). Similar to the overall pattern as presented in Figure 1, in Figure 2(a), modal guises (the pair of bars on the right in each cell) were evaluated more favorably than creak guises (pairs of bars on the left) for each metric. In addition, guises read in the female voice (green) were rated more favorably than their counterparts read in the male voice (purple). However, we can also tell from Figure 2(a) that for both female and male voices, the ratings of different personality metrics have very similar patterns. In other words, these metrics are highly correlated with each other. For instance, listeners did not seem to have rated attractiveness differently from competence. Therefore, rather than examining the effects of voice gender and guise type on each individual metric, an integrative approach was adopted. Specifically, the highly correlated ratings of different personality metrics were analyzed using a principal component analysis (PCA) to identify the primary dimension that listeners relied on during their evaluations. As can be seen in Figure 2(b), all the personality metrics are converging toward the same dimension (Dim1), and Dim 1 – that is, the first principal component

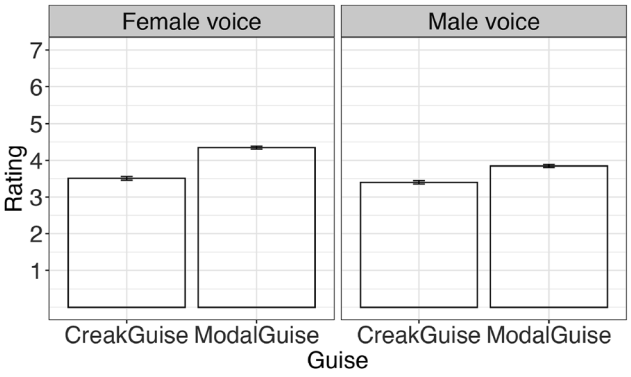

Figure 1: Overall average ratings for creak and modal guises across different gender voices: mean rating for female creak = 3.51, mean rating for male creak = 3.40, mean rating for female modal = 4.35, mean rating for male modal = 3.85.

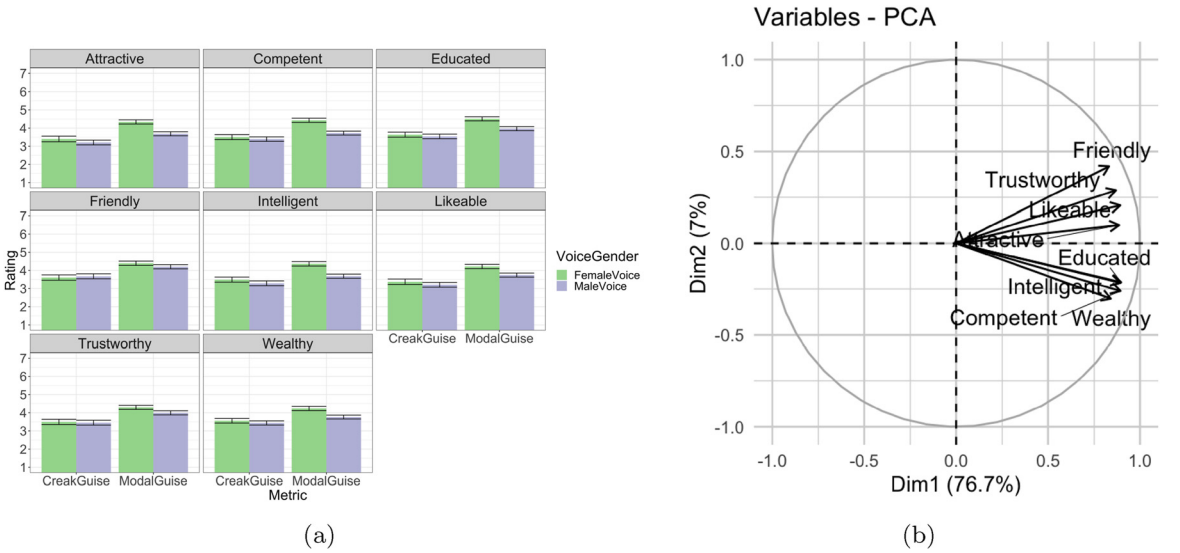

Figure 2: Principal component analysis of the rating space: (a) Ratings of voice gender and guise type in terms of different personality metrics. (b) The loadings for PC1 and PC2 of all the individual personality metrics: Dim1 on the *x*-axis stands for the loadings for the first principal component (i.e., PC1) and the *y*-axis represents the loadings for the second principal component (i.e., PC2).

(PC1) – captures 76.7 % of the variance in listeners' ratings across all the different personality metrics. The top five metrics that contribute most to PC1 are educatedness (36.2 %), intelligence (36.1 %), competence (36.1 %), likeability (36 %), and attractiveness (35.8 %). The top five metrics that contribute most to PC2 are friendliness (55.8 %), wealthiness (40.4 %), trustworthiness (38.5 %), competence (34.8 %), and intelligence (28.8 %). Based on the results of PCA, we thus gave a cover term "charisma" to the top five highly correlated dimensions that contribute to PC1, and treated PC1 (listener ratings of charisma) as the dependent variable for our main analysis.

As shown in Figure 3, the guise types (creak vs. modal guise) contrast primarily along PC1, compared to PC2, so do the two voice genders (male vs. female voices). This suggests that both the guise type and the voice gender differ in PC1 (charisma) ratings. PC2 appeared to capture the distinction between interpersonal likeability (e.g., friendly, trustworthy) and prestige-based likeability (e.g., wealthy, intelligent). However, since PC2 did not differentiate between creak and modal guises, it was excluded from further analysis.

To statistically test whether Mandarin listeners exhibit gender-differentiated bias around creaky voice, a linear mixed-effects regression model was conducted to predict listeners' charisma ratings (i.e., PC1) with guise, voice gender, and listener gender included in a three-way interaction as fixed effects and listener and pair as random intercepts. The random effects structure was initially configured in a maximal way according to Barr et al. (2013), but random slopes caused convergence failures. Therefore, the final model included listener and each

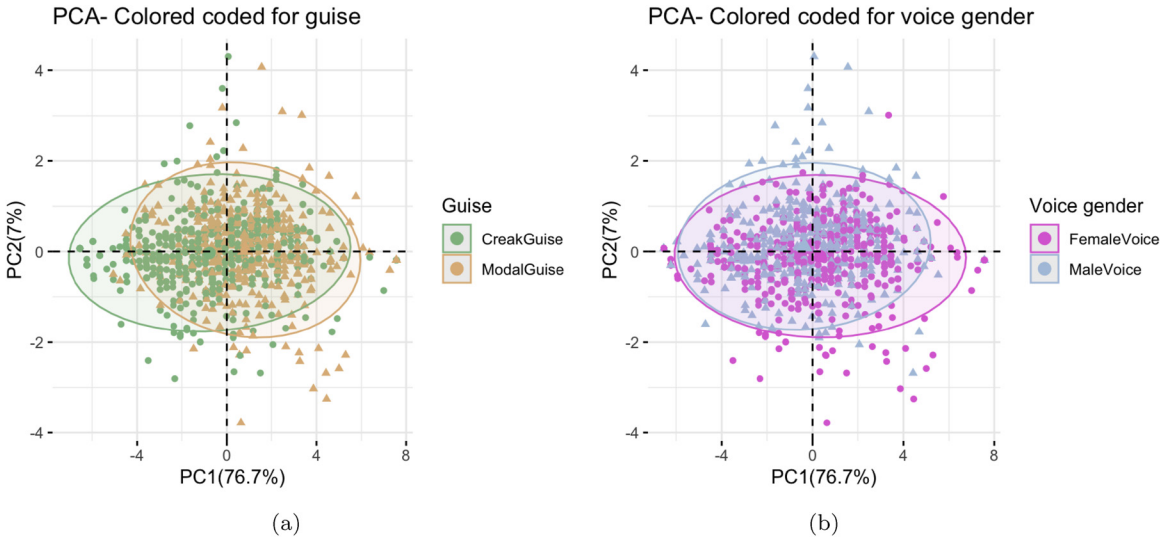

Figure 3: Principal component analysis of the rating space along guise and voice gender: (a) Color-coded for guise type. (b) Color-coded for voice gender. Concentration ellipse level = 0.95.

creak-modal stimulus pair as random intercepts. All the categorical variables were sum-coded. The model output (see Table 2) revealed that compared to the grand mean, listeners rated the creak guise significantly less favorably ($\beta = -0.77$, $p < 0.001$). In addition, there was a significant interaction between guise and voice gender, suggesting that there was a larger drop in listener rating from modal to creaky for the female voice compared to the male voice ($\beta = -0.25$, $p < 0.001$). No other predictors were tested significant.

The model above does not test whether PC1 ratings differ significantly within each voice gender condition in isolation. Post hoc pairwise comparisons were extracted from the fit model using `emmeans` while correcting for multiple comparisons. These results (see Table 3) further show that for the female voice, listeners gave significantly lower ratings to the creak guise than to the modal guise ($\beta = -2.03$, $p < 0.001$). The same is true for the male voice: lower ratings were given to the creak guise ($\beta = -1.04$, $p < 0.001$). Crucially, for the creak guise, the ratings of the female voice were not significantly different from those of the male voice ($\beta = 0.30$, $p = 0.44$). For the modal guise, the ratings of the female voice were significantly higher than those of its male counterpart ($\beta = 1.31$, $p = 0.01$). Crucially, the rating difference in PC1 between the creak guise and the modal guise for the female voice was significantly larger than that for the male voice ($\beta = -0.38$, $p < 0.001$).

Taken together, our analysis of PC1 shows that creak guises were indeed rated more harshly. However, within creak guises, listeners did not evaluate the female voice differently from the male voice. Notably, listeners tended to give lower ratings to modal guises in the male voice, which may be attributed to the fact that the male voice was

Table 2: Model output for creaky voice evaluation: PC1 ~ Guise × Voice gender × Listener gender + (1|Listener) + (1|Pair).

	Estimate	SE	*z* Value	Pr(>|z|)
(Intercept)	−0.07	0.27	−0.26	0.80
Guise (Creak)	−0.77	0.06	−13.39	<0.001
Voice gender (Female)	0.41	0.20	2.08	0.06
Listener gender (Female)	0.29	0.20	1.45	0.16
Guise (Creak): Voice gender (Female)	−0.25	0.06	−4.30	<0.001
Guise (Creak): Listener gender (Female)	0.08	0.06	1.50	0.14
Voice gender (Female): Listener gender (Female)	−0.09	0.06	−1.50	0.13
Guise (Creak): Voice gender (Female): Listener gender (Female)	−0.04	0.06	−0.72	0.47

Table 3: Post hoc comparisons based on the fitted model.

Contrast	Pair	Estimate	SE	*z* Ratio	*p* Value
Creak guise	Female voice – Male voice	0.32	0.41	0.79	0.44
Modal guise	Female voice – Male voice	1.31	0.41	3.19	0.01
Female voice	Creak guise – Modal guise	−2.03	0.16	−12.91	<0.001
Male voice	Creak guise – Modal guise	−1.04	0.18	−6.23	<0.001

originally manipulated from the female voice, leading to lower ratings based on perceptions of how natural the male voice sounded compared to the female voice.

4 Discussion and conclusion

The current study employed a matched-guise evaluation task to examine whether Mandarin listeners evaluate creaky voice differently based on voice gender. The results indicate that Mandarin listeners did not exhibit negative evaluations of female creak (creak in a high-pitched voice with higher vowel formants), compared to male creak (creak in a low-pitched voice with lower vowel formants). This suggests that, unlike American English, Mandarin listeners may not display explicit gender biases when it comes to creaky voice. In other words, the gender bias against women with creaky voice observed in American English does not seem to exist in Mandarin.

This brings us back to the main research question raised at the beginning: Can the differences in creak identification between American English and Mandarin listeners be attributed to the absence of gender-related social bias against creak in Mandarin? The answer seems to be yes, based on our current findings. In American English, gender bias in the evaluation of creaky voice can influence listeners' identification behaviors. In contrast, the absence of such bias in Mandarin leads to no observable gender-based differences in identification behaviors. This underscores the influence of social evaluation on speech perception, highlighting the need for future research to incorporate this in speech perception studies.

Notably, while Mandarin listeners did not evaluate female creak and male creak differently, they tended to downgrade the low-pitched modal voice. This may be attributed to the manipulation of sound files, as the low-pitched male voice was manipulated from the original high-pitched female voice. This manipulation could have influenced listeners' overall evaluations in ways that are not yet fully understood. Therefore, the results of this experiment should be treated with caution due to the potential influence of voice manipulation on listeners' evaluation. To have a better understanding of this manipulation effect, a follow-up experiment could be conducted using naturalistic stimuli from male speakers and generating female versions through manipulation.

In terms of personality metrics associated with creak such as "attractive", "competent", "likeable", "intelligent", and "wealthy", these did not show gender-based differences in Mandarin as in English. Most of these traits are highly correlated in listeners' rating dimensions, as evidenced by our PCA analysis, suggesting that creak may not carry the same social connotations in Mandarin. However, the limited set of stimuli tested may restrict the generalizability of the findings. Future work should explore the real-world social connotations of creaky voice using more naturalistic speech.

Despite potential limitations, this study confirmed that the different creak identification rates found between the high-pitched female-sounding voice and the low-pitched male-sounding voice in Li et al. (2023) was not underlyingly triggered by social evaluation biases in favor of male voices in Mandarin. The findings of this study highlight the importance of crosslinguistic comparisons of creaky voice and shed light on the relationships between speech evaluation and perception.

Acknowledgments: The authors would like to thank Nicole Holliday and Jianjing Kuang for their input during the early development of this project. We also appreciate the helpful suggestions from the two anonymous reviewers.

An earlier version of this work was presented at LSA 2023 and NWAV-AP7, and we thank the audience for their valuable feedback.

References

Abdelli-Beruh, Nassima B., Lesley Wolk & Dianne Slavin. 2014. Prevalence of vocal fry in young adult male American English speakers. *Journal of Voice* 28(2). 185–190.

Anderson, Rindy C., Casey A. Klofstad, William J. Mayew & Mohan Venkatachalam. 2014. Vocal fry may undermine the success of young women in the labor market. *PLoS One* 9(5). e97506.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Davidson, Lisa. 2019. The effects of pitch, gender, and prosodic context on the identification of creaky voice. *Phonetica* 76(4). 235–262.

Esling, John. 1978. The identification of features of voice quality in social groups. *Journal of the International Phonetic Association* 8(1–2). 18–23.

Esposito, Lewis. 2016. *I am a perpetual underdog: Lady Gaga's use of creaky voice in the construction of a sincere pop star persona*. Swarthmore: Swarthmore College undergraduate thesis.

Esposito, Lewis. 2017. "That's what it felt like, 'you're pathetic' ": Creaky voice, affective stance, and authentication in the speech of Lady Gaga. *Lifespans and Styles* 3(2). 2–12.

Gobl, Christer & Ailbhe Ní Chasaide. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40(1–2). 189–212.

Gordon, Matthew & Peter Ladefoged. 2001. Phonation types: A cross-linguistic overview. *Journal of Phonetics* 29(4). 383–406.

Greer, Sarah D. F. & Stephen J. Winters. 2015. The perception of coolness: Differences in evaluating voice quality in male and female speakers. In Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow. Available at: https://www.internationalphoneticassociation.org/icphs/icphs2015.

Henton, Caroline G. 1989. Sociophonetic aspects of creaky voice. *Journal of the Acoustical Society of America* 86(S1). S26.

Hildebrand-Edgar, Nicole. 2016. *Creaky voice: An interactional resource for indexing authority*. Victoria, BC: University of Victoria PhD thesis.

Kuang, Jianjing. 2018. The influence of tonal categories and prosodic boundaries on the creakiness in Mandarin. *Journal of the Acoustical Society of America* 143(6). EL509–EL515.

Laver, John. 1980. *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Lambert, Wallace E., Richard C. Hodgson, Robert C. Gardner & Samuel Fillenbaum. 1960. Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60(1). 44–51.

Li, Aini, Wei Lai & Jianjing Kuang. 2023. Creaky voice identification in Mandarin: The effects of prosodic position, tone, pitch range and creak locality. *Journal of the Acoustical Society of America* 154(1). 126–140.

Ligon, Claire, Carrie Rountrey, Noopur Vaidya Rank, Michael Hull & Aliaa Khidr. 2019. Perceived desirability of vocal fry among female speech communication disorders graduate students. *Journal of Voice* 33(5). 805.e21–805.e35.

Melvin, Shannon. 2015. *Gender variation in creaky voice and fundamental frequency*. Columbus, OH: Ohio State University PhD dissertation.

Mendoza-Denton, Norma. 2011. The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology* 21(2). 261–280.

Parker, Michelle A. & Stephanie A. Borrie. 2018. Judgments of intelligence and likability of young adult female speakers of American English: The influence of vocal fry and the surrounding acoustic-prosodic context. *Journal of Voice* 32(5). 538–545.

Pittam, Jeffery. 1987. Listeners' evaluations of voice quality in Australian English speakers. *Language and Speech* 30(2). 99–113.

Podesva, Robert J. 2013. Gender and the social meaning of non-modal phonation type. In *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, 427–448. Berkeley, CA: Berkeley Linguistics Society.

Tamminga, Meredith. 2017. Matched guise effects can be robust to speech style. *Journal of the Acoustical Society of America* 142(1). EL18–EL23.

Yuasa, Ikuko Patricia. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3). 315–337.