JASA ARTICLE

Creaky voice identification in Mandarin: The effects of prosodic position, tone, pitch range and creak locality

Aini Li,[1,a)] Wei Lai,[2] and Jianjing Kuang[1]

[1]*Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

[2]*Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee 37203, USA*

ABSTRACT:

Creaky voice, a non-modal aperiodic phonation that is often associated with low pitch targets, has been found to not only correlate linguistically with prosodic boundary, tonal categories, and pitch range, but also socially with age, gender, and social status. However, it is still not clear whether co-varying factors such as prosodic boundary, pitch range, and tone could, in turn, affect listeners' identification of creak. To fill this gap, this current study examines how creaky voice is identified in Mandarin through experimental data, aiming to enhance our understanding of cross-linguistic perception of creaky voice and, more broadly, speech perception in multi-variable contexts. Our results reveal that in Mandarin, creak identification is context-dependent: factors including prosodic position, tone, pitch range, and the amount of creak all affect how Mandarin listeners identify creak. This reflects listeners' knowledge about the distribution of creak in linguistically universal (e.g., prosodic boundary) and language-specific (e.g., lexical tone) environments. © *2023 Acoustical Society of America*. https://doi.org/10.1121/10.0019941

(Received 28 February 2023; revised 6 June 2023; accepted 12 June 2023; published online 11 July 2023)

[Editor: Jody Kreiman] Pages: 126–140

I. INTRODUCTION

As a non-modal phonation, creaky voice is typically defined as vocal folds being tightly adducted and/or vibrating irregularly (Gordon and Ladefoged, 2001). From the articulatory point of view, creaky voice can be realized differently, and thus exhibit different acoustic features. Research on creaky voice has so far established a basic common ground that there is not one kind of creaky voice, as no single defining property is shared by all kinds (Keating *et al.*, 2015). The term "creaky voice" (also known as "vocal fry" and "glottalization"), therefore serves as a cover term for several types of phonation. Similar to the term "rhoticity" where a bottom-level $F3$ is a widespread sign but cannot serve as a unique, defining acoustic feature for the whole range of phenomena under the notion "rhoticity," the existence of proposed categories such as "whispery creaky falsetto" makes it difficult to equate creak simply with low $F0$. This being said, the prototypical creaky voice is often associated with irregular vocal fold vibration, constricted glottis, and low pitch.

Creaky voice can be used both contrastively and non-contrastively depending on specific languages. In languages where creak is non-contrastive, it is often associated with domain-final prosodic boundaries, unstressed syllables, and low pitch and tonal targets (Bishop and Keating, 2012; Crowhurst, 2018; Huang, 2020; Kuang and Liberman, 2016b; Kuo, 2012; Yu and Lam, 2014, *inter alia*). In addition, non-contrastive creak bears various social meanings, also depending on different cultures. For instance, in the English-speaking world, creaky voice is enriched with sociolinguistic connotations, including but not limited to: social status (Esling, 1978), age, and gender (Dallaston and Docherty, 2020; Wolk *et al.*, 2012). In the U.S. context specifically, creaky voice can even invoke unjustified social bias (Davidson, 2019, 2020): Women are found to use creaky voice more frequently than men in their production (Abdelli-Beruh *et al.*, 2014; Oliveira *et al.*, 2016; Podesva, 2011; Yuasa, 2010). However, creaky voice is not preferred for male voices but even more so for female voices in evaluation (Greer and Winters, 2015). On the contrary, in Mandarin, no apparent gender-differentiated creak usage in production has been found (Kuang, 2018), and Mandarin appears to have no gender-differentiated social bias toward creak (Li and Lai, 2023). Despite these established (socio)-linguistic implications creaky voice is imbued with, few studies have examined how factors like prosodic position, pitch range (or gender), and tone influence listeners' identification of creak, as answers to this question could further enhance our understanding of speech perception under multi-variable contexts. This forms the main goal of our current study.

Previous studies based on English have demonstrated that the perception of creak is subject to influences from prosodic position, pitch range, creak locality, and gender (Davidson, 2019). Crucially, although creaky female speech is susceptible to unjustified social biases in the U.S. context, the rates of identifying creak in male and female speakers are similar for English listeners (Davidson, 2019). Therefore, listeners' ability to recognize creak is more determined by the environment in which the creak is produced instead of its social connotations. For languages such as

[a)] Electronic mail: liaini@sas.upenn.edu

Mandarin, non-contrastive creak also interacts with tone as the creakiness of different tones can be realized differently (Kuang, 2018). The different linguistic and social profiles between Mandarin and English as mentioned previously regarding how creak is used and evaluated differently further call into question how creak is identified cross-linguistically. Specifically, it remains unclear as of yet how tone, together with its interaction with prosodic position and pitch range, further shapes the identification of creaky voice. On top of this, how production influences creak perception has also been frequently overlooked in previous studies.

Continuing along these lines, this study probes how creaky voice is identified among Mandarin listeners through a fully controlled experiment. In particular, we examine the effects of prosodic position (final vs non-final), pitch range (high vs low), tone (lexical and neutral tones), as well as creak locality (global vs local) on creak identification by Mandarin listeners, with the ultimate goal of fostering a holistic view of how creaky voice is identified in speech perception. More importantly, given that creaky voice plays a vital role in determining linguistic meanings and conveying real-world social connotations, the identification of creak at a prelexical level would feed into the recognition of relevant linguistic units at interplay. Results of the current study will help disentangle the complicated relationships between creak and its acoustic, linguistic, and social correlates, as well as delineate their relative contributions when they are put into interaction under various contexts in speech communication. This represents an important step in developing a better understanding of speech perception in a multi-variable environment.

II. BACKGROUND

Creaky voice as one type of non-modal phonation can be used in both contrastive and non-contrastive ways. While contrastive, phonemic creak is specified in phonology, non-contrastive creak is more driven by mechanical reasons (e.g., pitch range and aerodynamics) that are more commonly attested among languages (Kuang, 2013, 2017). In languages such as Jalapa Mazatec, Chong, and Southern Yi, where creak is a contrastive property and can be used to distinguish between different words, it is contrastive independently of tones (DiCanio, 2009; Garellek and Keating, 2011; Kuang, 2011). However, for other languages such as Green Mong, Northern Vietnamese, and White Hmong, albeit being contrastive, creaky voice (in broad terms) covaries with certain tonal categories (Andruski, 2006; Brunelle, 2009; Esposito, 2012) (see Brunelle and Kirby, 2016; Gordon and Ladefoged, 2001 for more systematic reviews). More universally, non-contrastive creaky voice functions as an important prosodic cue (Kuang, 2018).

A. Non-contrastive creaky voice as a prosodic cue

One well-established finding about non-contrastive creaky voice in the literature is that the presence of creaky voice is closely related to prosodic positions (i.e., positions in prosodic domains). Cross-linguistically, creaky voice is often found at sentence-final positions. This has been attested in languages such as English, Mandarin, German, Greek, and many more (Abdelli-Beruh *et al.*, 2016; Garellek, 2015; Henton, 1989; Lehiste *et al.*, 1975; Pierrehumbert, 1979; Redi and Shattuck-Hufnagel, 2001, *inter alia*). To illustrate, Henton (1989) found that in British English, compared to other positions in an utterance, creak occurred more often in syllables at the end of utterances. Similarly, Abdelli-Beruh *et al.* (2016) analyzed the distribution of different acoustic patterns of vocal fry in American English and found that the frequencies of different patterns of vocal fry overall were greatest at the end of a sentence, regardless of the length of the sentence. Crucially, not only does sentence-final creak have a demarcation function of marking prosodic boundaries, but also it can be modulated by the size of the boundary, and the larger the prosodic boundary, the higher amount of creak (e.g., Garellek, 2015; Kuang, 2018; Redi and Shattuck-Hufnagel, 2001).

Other than marking prosodic positions and boundaries across multiple words within an utterance, creaky voice can be used to cue turn-taking in running discourse with utterances at larger scales (Heldner *et al.*, 2019; Ogden, 2002). For instance, Belotel-Grenié and Grenié (2004) found that in standard Mandarin TV newscasts, creaky voice often appears at the end of paragraphs. Recent studies further suggested that the amount of creak is conditioned by different speech contexts: creakiness occurs far more often in dialogues than in monologues (Aare *et al.*, 2014). With respect to dialogue context, Lee (2015) discovered that creaky voice can in fact mark the detachment of an utterance from the surrounding discourse or even the speaker, suggesting further that creaky voice plays an important role in discourse.

An example of this is that creaky voice facilitates the resolution of sentence disambiguation, as it cues certain prosodic boundaries (Kuo, 2012). In her study on how Taiwanese listeners use prosodic phrasing to interpret ambiguous sentences, Kuo (2018) pointed out that listeners made use of duration (i.e., final lengthening) and $F0$ measures (i.e., pitch declination) to detect disambiguation points during sentence gating; on top of that, they also relied on voice quality, specifically creaky voice, to judge whether they detected an early boundary or a late boundary. Similarly, Crowhurst (2018) showed that during sentence disambiguation, English listeners relied on both duration and creak-based cues to identify utterance-internal prosodic boundaries: listeners associated both duration and creak with a following prosodic phrase boundary and these effects were additive when duration and creak provided redundant information. However, when these two cues were in conflict, the effect of creak was subtractive such that listeners became less likely to report perceiving a prosodic boundary after a long word when the short word was creaky. In this case, creak can turn a syntactic boundary within a prosodic phrase into a boundary between two prosodic phrases.

All these studies have demonstrated clearly that creaky voice is associated with prosodic position in both speech perception and production. Yet, the question of how prosodic position influences creak identification by listeners has received little attention. This is a crucial issue to address, as creak serves as an important acoustic cue for marking prosodic boundaries. Therefore, understanding the factors that affect creak identification could have broader implications for prosodic boundary perception.

B. Pitch-driven voice quality variation

Voice quality naturally covaries with pitch, which, as an auditory concept, has mostly been operationalized as fundamental frequency ($F0$). In speech production, while a tense voice is associated with the highest pitch range (because tense voice happens naturally when pitch is close to its highest limit without changing into falsetto), creaky voice is always associated with the lowest pitch range (Blomgren *et al*., 1998; Hollien and Michel, 1968; Kuang, 2017). In more concrete terms, Blomgren *et al.* (1998) for instance, reported that the typical pitch range for creaky voice is about 20–70 Hz. This covariation between voice quality and pitch range, and more specifically, between creaky voice (despite not being the same type of creaky voice) and low pitch targets, has been observed and verified across multiple languages such as English, Mandarin, Javanese, and many others (e.g., Abramson *et al*., 2004; Andruski and Ratliff, 2000; DiCanio, 2009; Esposito, 2012; Gobl and Chasaide, 2010; Thurgood, 2004).

In speech perception, it has also been well established that voice quality cues are able to shift the way listeners perceive pitch range. For instance, Kuang and Liberman (2015) provided perceptual evidence using non-speech signals to show that listeners' classification of pitch can be significantly shifted by different spectral cues representing different voice qualities. Listeners tended to perceive higher pitch with a tenser voice quality (as implemented with spectral balance tilted towards higher frequency), and vice versa. Later, similar patterns were replicated by them in a follow-up study where speech cues were used instead of non-speech signals (Kuang and Liberman, 2016a). On a related note, Kuang and Liberman (2016b) further showed that voice quality can introduce significant shifts in the classification of pitch, and listeners were more likely to interpret boosted spectrum as creaky voice when $F0$ stimuli were on the lower side of the speaker's range. These findings suggested that in general, voice quality is a useful cue for pitch perception (Kuang and Liberman, 2018). The fact that creaky voice is associated with low pitch range in speech production and spectrum cues of creaky voice influence the perception of pitch range further speaks to the strong mutually dependent relationship between voice quality and pitch range.

Importantly, pitch has often been considered the primary cue of tonal contrasts for tonal languages. In tonal languages where creak phonation is not used contrastively, the presence of creak also interacts with tone, with non-contrastive creaky voice often being associated with low pitch tones (Kuang, 2017). Take Mandarin as an example: according to Zhu (2012), creak serves as an accompanying feature of low tones in "hundreds of local varieties of Chinese and other tonal languages in China." Specifically, creaky voice is highly correlated with full tones like Tone 3 and Tone 4 (Belotel-Grenié and Grenié, 1994, 2004). Through a systematic examination of the distribution of Mandarin creaky voice in a large-scale corpus, Kuang (2018) pointed out that apart from Tone 3, creaky voice was also likely to occur in the neutral tone ($T0$). Similarly, in Cantonese, Yu and Lam (2014) collected a corpus of Cantonese read speech of eight native speakers (4 M, 4 F) and found that there was systematically a higher tendency for creak to be produced for Cantonese low fall Tone 4 (24.2%) than other tones.

Not only has creaky voice been found to co-occur with certain tones, it has been found to be a salient cue to facilitate low tone perception. For example, In a Mandarin tone identification experiment, Belotel-Grenié and Grenié (1997) found that Tone 3 stimuli produced with creaky voice was recognized significantly faster than Tone 3 stimuli produced without creaky voice. This was later confirmed by Yang (2011) with a similar experiment on a larger scale. Huang (2020) found that when the extra-low $F0$ cue was absent, creak attributes such as period doubling and irregular $F0$ did become perceptually prominent in low tone perception. Similar effects were reported for Cantonese. Yu and Lam (2014) demonstrated that Cantonese listeners were reported to have a higher accuracy rate when the low tone (i.e., Tone 4) was realized with creak than when it was not. Meanwhile, the presence of creak could bias the listeners' perception towards the low tone (Tone 4) instead of the low level tone (Tone 6).

According to these studies, it is clear that creaky voice is sensitive to pitch range and further interacts with tonal categories (especially those with low-pitched targets). Now, a natural question that arises is whether and how pitch range as well as tone can influence the identification or perception of creaky voice, a better understanding of which could further unveil the underlying mechanisms behind pitch range and tone perception.

C. Creaky voice and society

Other than the previously-mentioned linguistic correlates, a substantial body of work has suggested that creaky voice is sociolinguistically meaningful in various cultures. One of the earliest pieces of evidence comes from Esling (1978). In her study, Esling compared the acoustic correlates of voice quality across different regional and social groups in the Edinburgh speech community. She found that creaky voice was judged to show up predominately among speakers with higher social status, suggesting that creaky voice can be indicative of people's regional and social profiles. Similarly, in Australian English, it has been documented that creaky voice is associated with high social status (Pittam, 1987). Other than these macro social categories, Creaky voice can be used to express certain emotions in speech such as boredom, sadness, and suppressed rage (Gobl and Chasaide, 2003; Laver, 1980), and to mitigate face threatening acts (e.g., Butler, 2017; Cullen *et al.*, 2013; Zetterholm, 1998). For example, Butler (2017) explored the use of creaky voice in various types of speech acts such as commands, requests, disagreements, suggestions, and jokes, and found that the use of creaky voice can help to mitigate the face-threatening effect of the act. Moreover, creaky voice has been viewed as an interactional resource for persona and identity construction (Esposito, 2016, 2017; Hildebrand-Edgar, 2016; Mendoza-Denton, 2011). When speakers felt less intimately connected to their interlocutors during conversations, they were inclined to resort to creaky voice for epistemic stance-taking (Hildebrand-Edgar, 2016). A more concrete example, illustrated in Mendoza-Denton (2011), demonstrated that creaky voice became enregistered within an early narrative context and, catapulted by centrifugal media forces, was then taken as part of a constellation of features that cluster around the persona of "hardcore Chicano gangster."

Most importantly, in the context of American English, there exists a widespread belief that creaky voice is gender-differentiated, such that creaky voice is increasingly prevalent among young American women (Abdelli-Beruh *et al.*, 2014; Podesva, 2011; Wolk *et al.*, 2012; Yuasa, 2010). For instance, Yuasa (2010) examined the occurrence of creaky voice in natural conversations among relatively young educated American and Japanese speakers. She found that California women used creaky voice much more frequently than their counterpart American men and also comparable Japanese female speakers. In read speech, Melvin (2015) found that among Midwestern college students, female speakers produced more creak than their male counterparts with a small difference (18% vs 12%). Further, the gender-differentiated use of creaky voice is susceptible to cultural differences and can differ cross-linguistically. For instance, Henton (1989) conducted a large-scale quantitative study of the sociocultural usage of creaky voice in two dialects of British English (Received Pronunciation and Modified Northern). They discovered that for both dialects, male speakers used much more (between around three and ten times) creaky voice than female speakers. Therefore, while being a female speech marker in the U.S. context, creak marks male speech in Britain. In Mandarin (in this case, Taiwan Mandarin), Kuang (2018) failed to find significant differences between female and male speakers in how they used creaky voice linguistically in speech production.

In the U.S. context, these quantitative differences in creak usage between women and men can lead to substantial perceptual bias. Studies have shown that listeners generally prefer creaky voice for male voices but less so for female voices (Greer and Winters, 2015). As mentioned in Davidson (2019), female speech containing creaky voice was likely to be rated as less competent and less hirable by listeners, rendering creak an important aspect of speech that could have a negative impact on certain groups of speakers. With that being said, the role of creaky voice as a marker of gender-differentiated speech can be more pronounced in some cultures than others. However, in non-English speaking cultures, documentation of how creaky voice is evaluated is much less available. In Mandarin, for instance, it has been reported that the gender bias against women speaking with creak has not been found (Li and Lai, 2023).

III. THE CURRENT STUDY

Based on the background, we have learned that creaky voice in speech production is closely associated with prosodic position, pitch range, and various social factors. However, it remains unclear if these factors are also important cues for listeners to identify creaky voice in perception.

So far, there still exist very limited empirical studies on creak identification. One prior study on English examined the influence of acoustic contexts (i.e., prosodic position, pitch range, and utterance type/creak amount) and social factors (i.e., gender) on creak identification (Davidson, 2019). The goal of the Davidson (2019) study was to probe whether creak identification was driven by pitch range differences, or by the social bias against female creak that exists in American English. However, the results of this study did not provide a clear answer to this question: Rates of creak identification in male and female speakers were similar, implying that listeners' ability to recognize creak was less influenced by speaker gender, whereas there existed a weak tendency for listeners to identify more creak in female speech, indicating that gender bias might still play a role.

One limitation Davidson (2019) is that stimuli were drawn from naturally-produced podcasts, hence leaving segmental features and syntactic structures uncontrolled. Even though naturalistic speech is useful for examining how language is perceived in real-world scenarios, it presents difficulties to tease apart all of the linguistic and acoustic factors that might have contributed to listeners' identification of creak. Moreover, although all the utterances used in her study were declaratives and there were no extreme pitch changes, the utterances were produced by four different speakers. Because voice production is highly variable and voice quality involves a huge amount of individual variation, the ways individual speakers produce creaky voice might differ in various ways such that listeners might not even be listening to the same kind of creaky voice. This could potentially further confound the results. Notably, despite these limitations, Davidson (2019) indeed found some robust effects of prosodic position (i.e., sentence-final creak is harder to identify) and creak amount (i.e., creak identification was easier when the whole utterance was creaky) on English creak identification, suggesting that the identification of creak appeared to be primarily driven by acoustics rather than gender bias.

In light of these established findings, Mandarin provides a better test ground for developing a comprehensive understanding of creak identification and perception. For one, there is no reported gender-differentiated bias toward creaky voice, which makes it easier to tease apart the effect of gender and pitch range on creak identification. For another, testing creak identification with Mandarin allows us to examine all the previously mentioned phonetic factors (prosodic position, pitch range, and also tone) that could trigger creaky voice simultaneously. To further control for individual variation in voice production, instead of using naturalistic speech as in Davidson (2019), we chose to use resynthesized stimuli as a methodological improvement. Therefore, the goal of this current study is to systematically examine the effects of prosodic position (sentence final vs non-final), creak locality (global vs local), pitch range (high vs low), and tone (lexical tones and neutral tone) on creak identification in Mandarin using a fully controlled experiment.

IV. METHOD

A. Experimental design

This experiment employed an 8 (Tone) × 2 (Prosodic position) × 2 (Pitch range) × 2 (Creak locality) within-subjects design. The creak-containing syllables, i.e., syllables that were produced with creaky voice regardless of the proportion and location of creak in the syllable, varied in Tone, Pitch range, Prosodic position, and Creak locality. The four critical conditions are described in detail in the following subsections. Note that creak in creaky syllables was deliberately produced for each stimulus to control for where it occurred.

1. Tone condition

The tone condition refers to what tonal category was carried by creak-containing syllables. In Mandarin, there are four full lexical tones and one neutral tone. The four full lexical tones, including High level (Tone 1), Rising (Tone 2), Dipping/low (Tone 3), and Falling (Tone 4), have different pitch heights and contour shapes (see Fig. 1). The neutral tone, also called the “fifth tone” or Tone 0, is only carried by weak syllables. Syllables with the neutral tone in Mandarin show greater susceptibility to weakening and lenition and often appear in word-final positions; therefore they are usually regarded as unstressed (Duanmu, 2007). Unlike full lexical tones, neutral tone is phonologically “underspecified” as it does not have its own underlying tonal targets. Therefore, how neutral tone is realized in pitch implementation depends on the tone preceding it (see Fig. 2). In this sense, the phonetic realization of neutral tone following Tone 1, Tone 2, Tone 3, and Tone 4 differs. Our study included these four different cases and treated them differently. We refer to instances of neutral tone preceded by Tone 1 as “Neutral 1,” neutral tone preceded by Tone 2 as “Neutral 2,” and so forth. To obtain a systematic pattern of how different tonal categories and their realizations influence creak identification, a total number of eight individual levels for tonal condition hence was included (four lexical tones + four realizations of the neutral tone = eight levels).

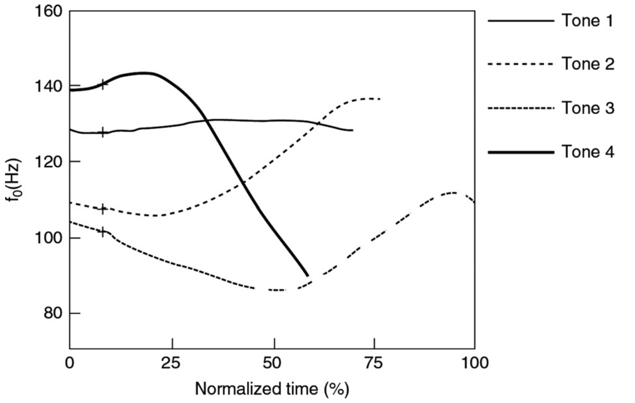

FIG. 1. Pitch contour of Mandarin lexical tones (Tones 1–4), sourced from Xu (1997).

2. Prosodic position condition

As for prosodic position, the creak-containing target syllables were in either sentence-final or sentence non-final positions to capture the boundary effects on creak identification.

3. Creak locality condition

With respect to the creak locality conditions, syllables were either locally creaky or globally creaky. Global creak in this case refers to scenarios where the surrounding 4–5 syllables, including the target syllables, were also creaky. Local creak refers to cases where a single creaky syllable (i.e., the target) was embedded in a modal sentential context.

4. Pitch range condition

As for the pitch range conditions, stimuli were manipulated into either a high-pitched voice or a low-pitched voice (see Sec. IV B for more details).

B. Stimuli construction

Sixty-four sentences were constructed. Each sentence was 12 syllables in length, approximately the average sentence length in Mandarin (Huang, 2018). All the sentences

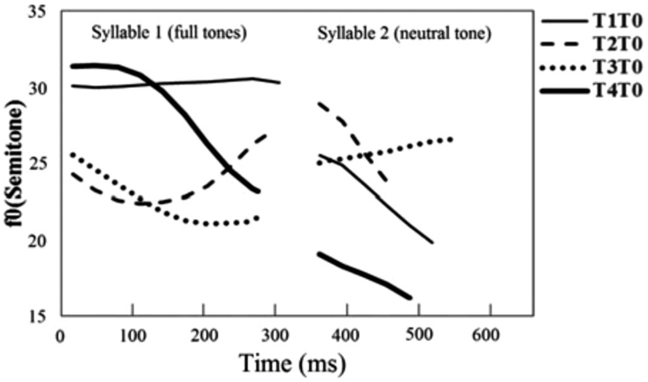

FIG. 2. Pitch contour of Mandarin neutral tones (T0) in different tonal contexts (following Tones 1–4), sourced from Tang (2014).

were simple declarative sentences and had sentence-final pitch declination. As for the syntactic structure, all the sentences shared the same structural frame of *NP1* + *TP* + *NP2*, conveying a general meaning of "who doing what sees/meets/knows whom." Notably, in order to avoid priming effects, each sentence differed in terms of its exact content so that listeners did not hear the same sentence multiple times during the experiment. Thus, listeners had to pay attention to the whole sentence instead of just the portion that varies.

Here, in the sentence frame mentioned previously, NP1 and NP2 are places where syllables were manipulated in terms of their tone targets. Both NP1 and NP2 were person names that shared the same final syllable, controlling for the word context where each target syllable appeared. In other words, for the two person names (NP1 and NP2) in each sentence, each target syllable shared the same preceding syllable and the same segmental features except for tone. Crucially, creaky target syllables were in either NP1 or NP2 positions, with the other one being "modal" to counterbalance the distribution of different tones in different prosodic positions. For instance, in the example that follows, for the two constructed names "Li3, Ai4" and "Li3, Ai1," only the two syllables (i.e., "Ai") differ in tones, one with Tone 4 and one with Tone 1, with the Tone 4 syllable being creaky and the Tone 1 syllable being modal. Note that the lexical frame where target syllables were situated was also controlled for such that the target syllables were preceded by the same character/syllable. Although these syllables rhymed in terms of their phonotactic structures, their orthographic forms were different. Syllables used to makeup these names contained only sonorants. Person names were adopted in utterance construction essentially because, in Mandarin, tones are not equally distributed. Particularly, syllables sharing the same phonotactic structure while allowing for the full set of tonal entries (i.e., the four lexical full tones) are extremely few and are not necessarily real words.

(1) **Li3 Ai4** zai gongyuan sanbu pengdao le **Li3 Ai1**.
Li3Ai4 at park walk meet ASP Li3Ai1
"Li3 Ai4 met Li3 Ai1 while taking a walk in the park."

All the sentences were produced by a female native speaker of Mandarin (the second author) and recorded in a professional recording booth at the University of Pennsylvania.[1] Sentences were read at 40 bpm using an online metronome to further control the speech rate. Crucially, one of the key manipulations in our current study is that the 64 sentences were manipulated into low-pitched counterparts as if they were read by a natural-sounding low-pitched male voice (64 × 2 = 128 sound files), as a way to maximally control for speaker-induced variation in voice production. We did not choose to record multiple different speakers because different speakers may differ in the way creak is produced. That means the single-speaker alternative could maximally decrease speaker-induced creak quality differences such that the nature of produced creak would not change even if the pitch range changes.

Manipulation was implemented on the female voice through the gender change function in Praat by lowering the formant (0.75), pitch median (110), and pitch range (0.7). Each sound file lasts around 2.5 s in duration. There were two recording deviants as one syllable was missed during sentence recording. Four audio files were excluded from the final experiment due to unnaturalness based on the authors' judgment (native speakers of Mandarin) (all of them were globally creaky). The excluded recordings were not re-recorded out of the concern that different recording sessions may influence the sound quality in various ways, which should be crucial since voice quality is the focus of our current study. More importantly, these excluded files were not in the important conditions. The exclusion of them would not result in missing a critical condition in the experiment and the final results would not thus be influenced as there were multiple stimuli within each condition. In the end, 124 experimental items were included in the final experiment with 122 having 12 syllables and two having 11 syllables. All the stimuli were normalized to an average intensity of 65 dB. All the stimuli are available at https://osf.io/zdhka/.

A schema of the experimental design implemented in our current study is provided in Table I. The letter C (red) stands for creak-containing syllables in which natural creaky voice was produced by the speaker. Letter C with an underscore represents the target syllable in that particular stimulus. The letter M (black) stands for modal syllables where no creak was produced by the same speaker. In short, Table I showcases how each tone (here: Tone 1) was manipulated to occur in different contexts. Figure 3 further presents examples of local creak [Fig. 3(a)] and global creak [Fig. 3(b)]. Following Davidson (2019), we also included in our design sentences that were fully modal, i.e., no creak was produced for all the syllables. An equal number of modal items was included to have an overall better balance of items with and without creak that would be heard by the listeners.

C. Participants

A total number of 41 native speakers of Mandarin participated in this study (8 male, 33 female). All of them were recruited from the mainland of China and were paid 20 RMB for their participation. Participants ranged in age from 19 to 36 (average = 25.12 years). All the participants speak both standard Mandarin and another Mandarin dialect as their native language. Twenty-four of them (59%) self-reported that they have never heard of "vocal fry" or "creaky voice" prior to the study. No one reported having a hearing deficit.

TABLE I. A schema of the experimental design.

Sentence	Tone	Prosodic position	Creak locality
MMMMMMMCCCC$\underline{C}$	Tone1	SentenceFinal	Global creak
MMMMMMMMMMM$\underline{C}$	Tone1	SentenceFinal	Local creak
C$\underline{C}$CCCMMMMMMM	Tone1	SentenceNonfinal	Global creak
M$\underline{C}$MMMMMMMMMM	Tone1	SentenceNonfinal	Local creak

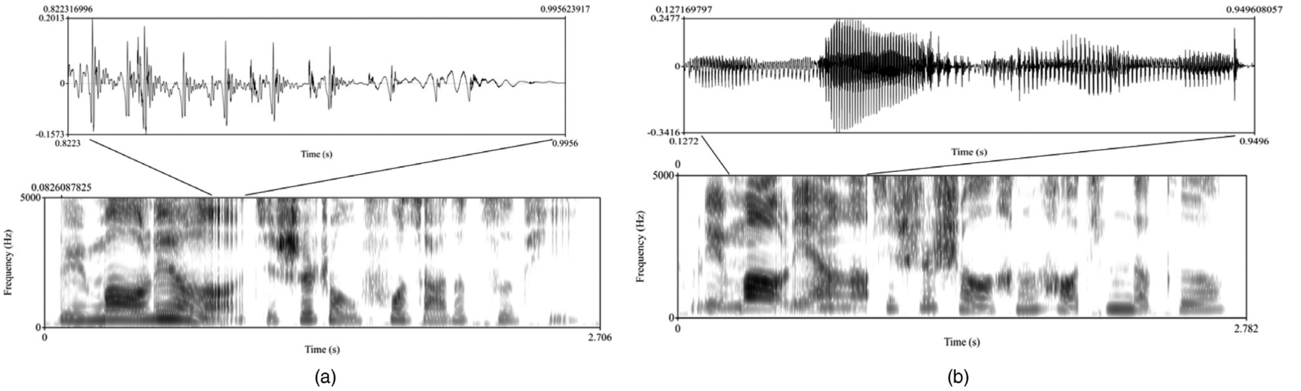

FIG. 3. Examples of creak locality. The waveform corresponds to the syllables that are produced with creak (a), local creak (only one syllable is creaky); (b), global creak (multiple syllables are creaky).

D. Procedure

The identification task was implemented using the online platform Qualtrics and was conducted in Mandarin Chinese. Participants were informed beforehand that headphones were needed throughout the whole experiment and they needed to finish the task in one sitting and in a quiet environment. With respect to the survey procedure, participants were first informed of the purpose of our study as follows: "This study aims to investigate how Mandarin listeners recognize creaky voice. Creaky voice conveys the sense as if vocal folds of the speaker are tightened and constricted. In this experiment, you will hear sentences that may or may not contain creaky voice. Your task is to decide whether the sentence that you hear contains any creaky voice. If you do not hear any creaky voice, you can just check the box saying "no creak is hear"; if you think the sentence does have creak, please try to choose all the characters that you think are creaky by clicking on the boxes below them." To make sure that participants were familiarized what creaky voice sounds like, a sample audio of creaky voice produced by a different female speaker was provided during the introduction and they were encouraged to listen to the sample audio as many times as they wanted before proceeding.

After reading the instructions and becoming familiarized with creaky voice, participants were presented with four practice audio clips: one case of global creak in high-pitched voice, one case of local creak in low-pitched voice, and two instances of fully modal sentences, one in high-pitched voice and the other in low-pitched voice. These practice trials resembled the format of the critical trials. Participants needed to choose all the Chinese characters where they heard creak for each auditory sentence, with feedback being provided immediately afterward. Participants were also allowed to listen to these exercise audio files as many times as they wanted before moving on so that they were able to become familiarized with what creaky voice sounded like and what the final task looked like. The speaker who produced the practice utterances was different from the speaker who produced the critical stimuli in the final test phase.

During the test phase, participants were presented with 124 stimuli sound files. For each sound file, they were asked to listen to the audio carefully and then choose, out of all the characters of the sentence, those that they thought were produced with creaky voice. For each question, participants had access to both the audio of the sentence and its written form. The characters varied in each trial with the context of the sentence so that participants' judgments were unlikely to be influenced by the orthography of each individual character.

After the identification task, participants were further asked for their age, sex (male; female; other; prefer not to say), where they are from, their native languages and dialects (if any), whether they have heard of creaky voice before the experiment, what they think of the sounds of the experiment, how they made their judgments and any other comments they might have about the experiment in open-ended responses. It took participants 25 min on average to complete the whole experiment. Crucially, based on their feedback, none of the participants perceived the stimuli as unnatural or artificial.

V. RESULTS

Here, we lay out three separate analyses in service to our research question of how Mandarin listeners identify creaky voice. Our first analysis focused on the identification of *creaky* syllables, investigating the effects of prosodic position, creak locality, and pitch range on creak identification. Our second analysis focused on the identification results of *modal* syllables, examining the effect of previous factors on how listeners false alarmed in modal syllables. Finally, we zoomed in to examine the identification results of the target creaky syllables in the *local* creak condition, which allowed us to further examine the effect of tone, since tone was only intentionally manipulated and balanced for creaky syllables in the local creak condition.

Statistical analyses were conducted using the *R* Statistical environment version 4.0.5 (R Core Team, 2013); mixed-effects logistic regression was run using the *lme4* library version 1.1–27.1 (Bates *et al.*, 2014), and plots were created using *ggplot* version 3.3.5 (Wickham, 2011). Model comparison was implemented for all the models reported here using log-likelihood ratio tests to diagnose non-significant factors and find the model with the optimal fit.

Model selection began from a maximal model and proceeded in a backward stepwise manner by removing insignificant factors one at a time and comparing the reduced model with the superset model at each step. A Chi-square test was used to assess the significance of the difference in log-likelihood between the two models. The superset model is chosen if $p < 0.05$ and the subset model otherwise.

A. Creaky syllables: The effects of creak locality, pitch range and prosodic position

In this subsection, we report results from our first analysis that probed the effects of creaky locality, pitch range, and prosodic position on creak identification for creaky syllables. Figure 4 shows listeners' rates of perceived creak among creaky syllables as conditioned by creak locality, prosodic position, and pitch range. At first blush, the likelihood of creak perception was overall higher when multiple syllables in a sentence were creaky (cf. Global vs Local). In addition, the likelihood of perceived creak seemed higher for creaky syllables at sentence non-final positions (blue bars). Compared with high pitch range, low pitch range also seemed to facilitate creak identification.

To confirm this pattern, a mixed-effects logistic regression model was implemented using *lme4* to predict listeners' responses, i.e., whether they identified the presence of creak in creaky syllables or not, with CreakLocality (global vs local), PitchRange (high vs low), and syllable's ProsodicPosition (sentence final vs non-final) as fixed effects (in a three-way interaction). All the categorical predictors were sum-coded, which allowed us to compare rates of perceived creak under specific conditions with the grand mean. Participant and syllable were included as random intercepts to account for baselines of by-item and by-participant variance. Model selection did not suggest excluding any one of these predictors. In the end, the optimal model took ProsodicPosition, PitchRange, and CreakLocality in a three-way interaction as fixed effects and Participant and Syllable as random intercepts.

The model output is summarized in Table II. According to the model results, there was a main effect of PitchRange. Compared to the grand mean, listeners were less likely to identify creak when the utterance was produced with a high pitch range ($\beta = -0.29$, $p < 0.001$). There also existed a main effect of ProsodicPosition, which suggests that listeners were less likely to identify creak when creaky syllables were in sentence-final positions ($\beta = -0.31$, $p < 0.001$). There was a main effect of Global creak, implying that the likelihood of detecting a creaky syllable was significantly boosted when the surrounding syllables were also creaky ($\beta = 0.34$, $p < 0.001$). For high-pitched utterances, the low probability of detecting creak in sentence-final positions became even smaller ($\beta = -0.10$, $p < 0.03$), as indicated by the significant interaction between PitchRange (High) and ProsodicPosition (Final). No significant effect was found for the interaction between PitchRange (High) and CreakLocality (Global). The interaction between Final Prosodic position and Global creak locality was significant, implying that even though listeners were more likely to perceive creak when the surrounding syllables were creaky, this locality effect became significantly less pronounced for when target syllables were at sentence-final positions ($\beta = -0.24$, $p < 0.001$). Finally, the significant three-way interaction between PitchRange(High), ProsodicPosition(Final), and CreakLocality(Global) further suggested that the creak locality effect reported previously became even smaller for high-pitched syllables at sentence-final positions ($\beta = -0.13$, $p < 0.01$).

In short, the overall analysis of creaky syllables showed that the creak identification was easier when target syllables were situated in a context where their surrounding syllables were also creaky (i.e., global creak). In addition, low pitch range facilitated creak identification. Moreover, creaky syllables in sentence-final positions were harder to identify. In other words, for syllables containing creaky voice, sentence-final position inhibited creak identification.

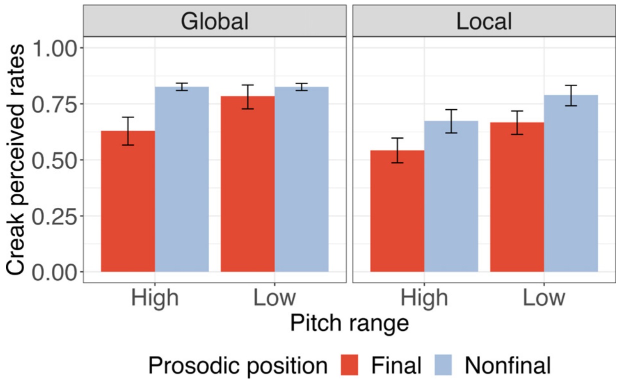

FIG. 4. (Color online) Creak perceived among creaky syllables: the effects of pitch range, prosodic position and creak locality).

B. Modal syllables: The effects of pitch range and prosodic position

Next, we turn to the analysis of how modal syllables were perceived by listeners. Figure 5 illustrates the

TABLE II. Model output of identification accuracy for creaky syllables: Response ∼ PitchRange * CreakLocality * ProsodicPosition + (1|Subject) +(1|Syllable).

Fixed Effects	Estimate	Standard error	t value	Pr(>|t|)
(Intercept)	1.16	0.23	5.12	$<0.001^{***}$
PitchRange (High)	−0.29	0.06	−6.44	$<0.001^{***}$
ProsodicPosition(Final)	−0.31	0.06	−5.13	$<0.001^{***}$
CreakLocality(Global)	0.34	0.05	6.44	$<0.001^{***}$
PitchRange(High) × ProsodicPosition(Final)	−0.10	0.05	−2.32	0.03^{*}
PitchRange(High) × CreakLocality(Global)	0.08	0.05	1.78	0.08
ProsodicPosition(Final) × CreakLocality(Global)	−0.24	0.05	−4.53	$<0.001^{***}$
PitchRange(High) × ProsodicPosition(Final) × CreakLocality(Global)	−0.13	0.05	−2.81	0.01^{**}

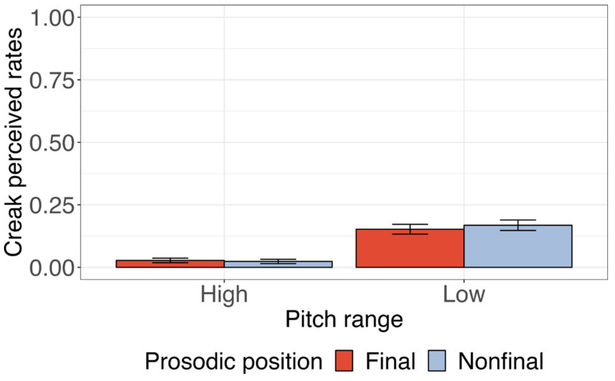

FIG. 5. (Color online) The mean perceived creak rates in modal syllables.

aggregated rates of perceived creak among modal syllables. It is clear that the amount of creak that was perceived by listeners among modal syllables was drastically lower, compared with the amount of creak that was perceived for creaky syllables, suggesting that listeners were in general accurate at identifying creak. However, listeners were also inclined to false-alarm and tended to perceive modal syllables as creaky when they were low-pitched.

A similar mixed-effects logistic regression model was configured to predict listeners' responses to modal syllables, i.e., whether they perceived creak for modal syllables or not, with ProsodicPosition and PitchRange (in a two-way interaction) as fixed effects (sum-coded). Participant and Syllable were included as random intercepts. The model output, as shown in Table III, revealed a main effect of PitchRange ($\beta = -1.08$, $p < 0.001$), implying that listeners were significantly more likely to have false-alarms on the low-pitched modal syllables. No significant effect was found for ProsodicPosition ($\beta = 0.01$, $p = 0.89$) or the interaction between PitchRange and ProsodicPosition ($\beta = 0.08$, $p = 0.28$).

Taken together, even though prosodic position did not generally influence listeners' identification of creak among modal syllables, listeners did show a strong tendency towards false alarm on low-pitched targets.

C. Analysis of local creaky syllables

Finally, we report our analysis based on locally creaky syllables, where we were able to focus on the effects of tone, prosodic position, and pitch range on the identification of creaky voice. As mentioned previously, four audio files were excluded from the final experiment due to unnaturalness and all of them were globally creaky. This presented difficulties in examining the tone effect under the global locality condition since some tone targets were missing. In addition, as each syllable was a tone-bearing unit, the effects of tone can be better captured only when we zoomed in to examine tone targets at local contexts (i.e., local creak). Therefore, here we conducted our analysis by focusing on syllables that are locally creaky.

TABLE III. Model output for creak identification of modal syllables: Response ~ PitchRange × ProsodicPosition + (1|Participant) + (1|Syllable).

Fixed Effects	Estimate	Standard error	z value	Pr(>|z|)
(Intercept)	−2.98	0.19	−15.50	<0.001***
PitchRange (High)	−1.08	0.07	−15.18	<0.001***
ProsodicPosition (Final)	0.01	0.07	0.18	0.86
PitchRange (High): Prosodic Position (Final)	0.07	0.07	1.08	0.28

1. Quantification of local creak in production

To better understand what kind of creak was identified, here we provide descriptively the durational properties of the produced creak in our stimuli before we dive into detailed results of creak identification for locally creaky syllables. Since only one speaker was included in our current study, instead of extracting acoustic measurements to capture individual differences in creak production, we chose to measure the proportion of creak produced in the creaky stimuli. Figure 6 presents produced creak proportion for target syllables in each tone category and prosodic position. Creak proportion was calculated as the duration of creak divided by the duration of the entire syllable. Across both prosodic positions, it is clear that Tone 3 target syllables had the highest amount of creak in production, followed by Tone 1 and then Tone 2 and Tone 4. In terms of the four different realizations of neutral tone, overall, they also seemed to contain higher proportions of creak, compared with lexical tones, with those following Tone 2 (i.e., Neutral 2) containing the smallest amount of creak.

2. The effects of tone, pitch range and prosodic position on perception

Figure 7 presents the creak identification results for creaky syllables with a full lexical tone, and Fig. 8 presents

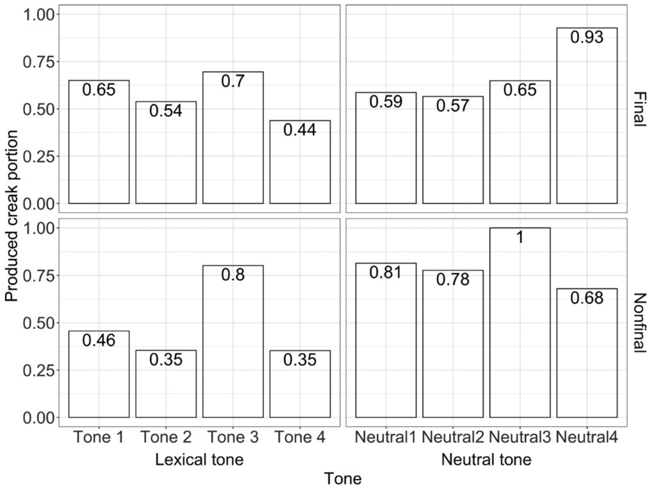

FIG. 6. Average creak proportion produced within each tone category and prosodic position (Neutral 1 stands for Neutral tone being preceded by Tone 1; Neutral 2 stands for Neutral tone being preceded by Tone 2, and so forth.).

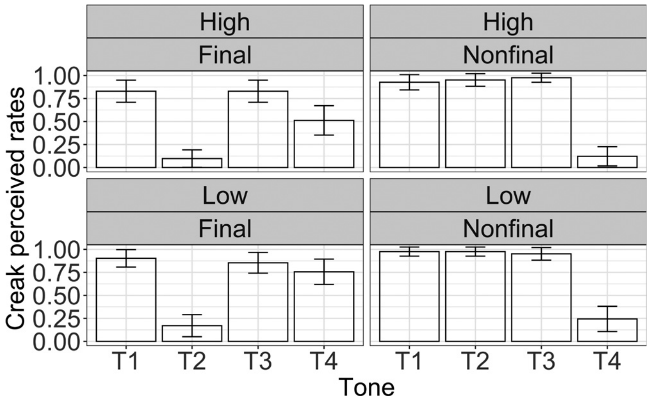

FIG. 7. Creak identification rates for creaky syllables with a lexical tone. “T1” stands for Tone 1, “T2” stands for Tone 2, and so forth. Error bars are based on 95% confidence interval.

the results for creaky syllables with a neutral tone. Our analysis focused on the effects of tone, prosodic position, and pitch range. For both kinds of tones, it seemed that lexical tones in sentence non-final positions were easier to identify. Among lexical tones, creak in Tone 1 and Tone 3 displayed higher rates of identification compared to that in Tone 2 and Tone 4 when in the sentence-final position. For neutral tones, the interaction between tone and prosodic position was not as consistent. For instance, the identification rate was higher in sentence non-final positions than in sentence-final positions for Neutral1, but not for Neutral4.

A mixed-effects model taking Tone, ProsodicPosition and PitchRange in a three-way interaction (sum-coded) was performed to predict listeners' responses (whether they perceived creak or not). Model selection suggested removing the three-way interaction of *PitchRange* × *ProsodicPosition* × *Tone* [χ^2 (7)= 5.65, $p = 0.58$], and the two-way interaction of *PitchRange* × *Tone* [χ^2 (7)= 0.56, $p = 0.48$]. The optimal model took Tone, ProsodicPosition, PitchRange, as well as the interaction between PitchRange and ProsodicPosition, and the interaction between ProsodicPosition and Tone as fixed effects. Participant and Syllable were included as random intercepts.

As shown in Table IV, the model output revealed a main effect of PitchRange. All else being equal, high-pitched syllables were significantly less likely to be identified as creak-containing ($\beta = –0.49$, $p < 0.001$). The effect of ProsodicPosition also turned out to be significant, suggesting that creaky syllables in sentence-final positions significantly dampened the identification of creak ($\beta = -1.33$, $p < 0.001$). None of the Neutral tones displayed significant differences in terms of perceived creak responses compared to the grand mean. For lexical tones, there was a significant main effect of Tone 1, which indicated that listeners were significantly more likely to report hearing creak when the syllable was the high level tone ($\beta = 2.43$, $p < 0.001$). The main effect of Tone 3 also demonstrated a similar pattern ($\beta = 2.47$, $p < 0.001$): low Tone was likely to elicit higher rates of creak identification, which for creaky syllables also indicated a higher accuracy rate. The effect of Tone 2 turned out to be not statistically significant ($\beta = 0.18$, $p = 0.66$). Conversely, there was a main effect of Tone 4, but the direction of its influence on creak identification was reversed ($\beta = -1.28$, $p < 0.001$), suggesting that listeners were significantly less likely to identify creak when it was a Tone 4 syllable.

The interaction between PitchRange (High) and ProsodicPosition (Final) was not significant ($\beta = 0.06$, $p = 0.47$). There were significant interactions between different Neutral tone categories and ProsodicPosition (Final), as suggested by the interaction between Neutral 2, Neutral 3, Neutral 4, and ProsodicPosition (Final). These interactions further suggested that although overall sentence-final position hindered creak identification, specific prosodic effects varied depending on different Neutral tone

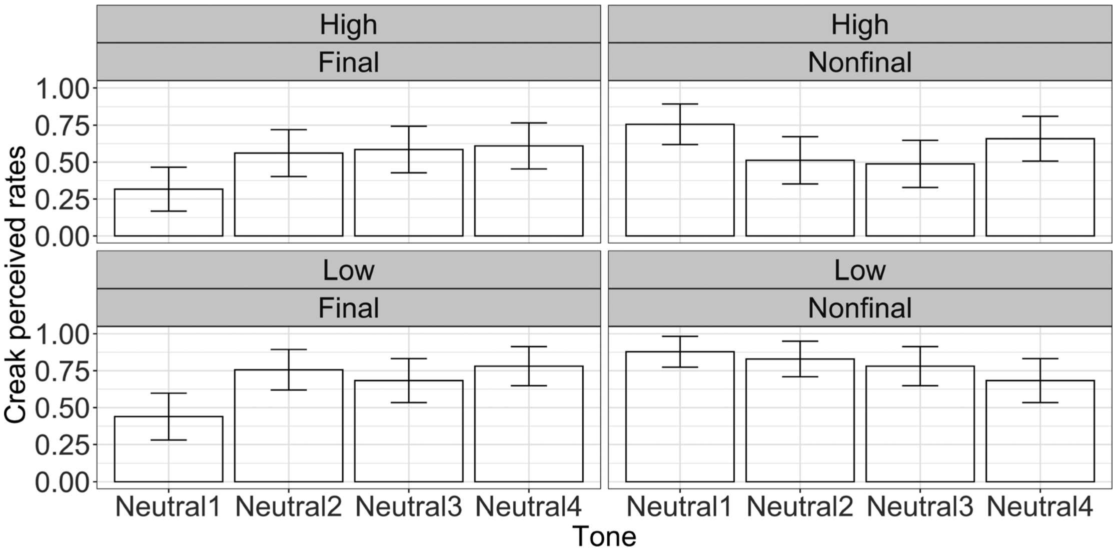

FIG. 8. Creak identification rates for creaky syllables with a neutral tone. “Neutral1” stands for Neutral tone following Tone 1, “Neutral2” stands for Neutral tone following Tone 2, and so forth. Error bars are based on 95% confidence interval.

TABLE IV. Model output for local creaky syllable identification: *Response ∼ PitchRange + ProsodicPosition + Tone + PitchRange * ProsodicPosition + ProsodicPosition * Tone + (1|Subject)+(1|Syllable).*

Fixed effects	Estimate	Standard error	z value	Pr(>|z|)
(Intercept)	0.65	0.28	2.30	0.02*
PitchRange (High)	−0.49	0.08	−5.99	<0.001***
ProsodicPosition(Final)	−1.33	0.21	−6.31	<0.001***
Neutral2	0.26	0.28	0.93	0.35
Neutral3	0.08	0.28	0.29	0.78
Neutral4	0.37	0.28	1.31	0.19
Tone1	2.43	0.39	6.22	<0.001**
Tone2	0.18	0.41	0.44	0.66
Tone3	2.47	0.41	6.00	<0.001***
Tone4	−1.28	0.30	−4.31	<0.001***
PitchRange(High) × ProsodicPosition(Final)	0.06	0.08	0.73	0.47
Netural2 × ProsodicPosition(Final)	1.30	0.29	4.54	<0.001***
Netural3 × ProsodicPosition(Final)	1.33	0.28	4.71	<0.001***
Netural4 × ProsodicPosition(Final)	1.41	0.29	4.90	<0.001***
Tone1 × ProsodicPosition(Final)	0.68	0.38	1.77	0.08
Tone2 × ProsodicPosition(Final)	−1.89	0.41	−4.60	<0.001***
Tone3 × ProsodicPosition(Final)	0.40	0.41	0.99	0.33
Tone4 × ProsodicPosition(Final)	2.70	0.31	8.78	<0.001***

categories. For lexical tones, the picture was more complex. The interaction effect between Tone 1 and ProsodicPosition (Final) was not significant ($\beta = 0.68$, $p = 0.08$), neither was the interaction between Tone 3 and ProsodicPosition (Final) ($\beta = 0.40$, $p = 0.33$). There was a significant interaction effect between Tone 2 and ProsodicPosition (Final), suggesting that even though listeners did not have an overall preference for identifying Tone 2-bearing syllables as creaky, the likelihood of them detecting creaky syllables with Tone 2 became significantly smaller in sentence-final positions ($\beta = –1.89\, p < 0.001$). On the contrary, although listeners were less likely to perceive Tone 4 syllables as creak-containing syllables, they tended to be significantly more likely to identify these syllables as creak-containing when these syllables showed up in sentence-final positions.

Based on our examination of locally creaky syllables, it seemed that overall listeners were less likely to perceive syllables as creaky when they were high-pitched or in sentence-final positions. However, sensitivity to prosodic position differed variably depending on different tonal realizations, with identification rates being boosted by some tonal categories (such as Neutral tone), while being diminished by others (such as Tone 2 and Tone 4). In particular, high perceived creak rates in certain tones like Tone 1 and Tone 3 were partially driven by the produced creak in production. Interestingly, the contour shape, together with prosodic position, further modulated creak identification: for sentence non-final positions, creak was more likely to be identified when embedded in tones with rising contours such as Tone 2, compared to Tone 4 with a falling contour. On the contrary, for sentence-final positions, the pattern was reversed: more creak identified in Tone 4 (falling contour) than in Tone 2 (rising contour). Unlike lexical tones, neutral tone interacted with prosodic position such that there was more perceived creak when the neutral tone was at sentence-final positions. That further suggested that creak in unstressed syllables like neutral tone were more easily identified when it was in sentence-final positions.

VI. DISCUSSION

This study investigated in a laboratory setting how Mandarin listeners identify non-contrastive creaky voice. Specifically, we tested the effects of prosodic position (sentence final vs non-final), pitch range (high vs low), tone (lexical tones and different realizations of neutral tone), and creak locality (global vs local) on listeners' identification of creak. Our results showed that all these factors were crucial cues for Mandarin creak identification. All else being equal, creak was difficult to identify in sentence-final positions, compared to non-final positions. While low pitch range facilitated creak identification for creaky syllables, it also caused false alarms in modal speech. Creak identification was easier when syllables were globally rather than locally creaky. For local creak identification in particular, the amount of produced creak and tonal contour both mattered. Although listeners tended to identify more creak on low-tone targets such as Tone 3, the large amount of creak produced in high-tone targets such as Tone 1 could still lead to high creak identification rates. In the following subsections, we discuss in more detail how different factors and their interactions worked together to influence creak identification.

A. Sentence-final position inhibits creak identification

One robust finding we have established so far is that sentence-final position inhibited creak identification for creaky syllables. In other words, creaky syllables in sentence non-final positions were more easily perceived and identified. Since sentence-final creak is a prosodic cue signaling the end of a declarative sentence in Mandarin, listeners may be more accustomed to hearing creak at sentence-final positions and thus were less likely to notice its presence. This is consistent with what has been reported in English creak identification (Davidson, 2019).

Other than this general inhibition effect, sentence-final position also interacted with creak locality and pitch range. For instance, sentence-final creak was even less likely to be detected in a globally creaky environment. Recall that for creaky syllables the identification rates were higher if the target syllables were situated in a context where the surrounding syllables were also creaky (i.e., global creak). The strength of the inhibition of sentence-final creak may be further enhanced, hence increasing the threshold for identifying creak. Therefore, although global creak made creak identification easier, identifying sentence-final creak under such circumstances became even harder. Another interesting interaction that we saw is that the perception of sentence-final creak also became significantly harder when the syllables were read in a high-pitched voice. On the one hand, listeners were less likely to perceive creak under high pitch range, since creak is often associated with low pitch targets. On the other hand, the sentence-final position inhibited creak identification. Therefore, this interaction may be driven by the additive inhibition effects of both high pitch range and sentence-final position. As a result, creak identification became harder when sentence-final creak was embedded in a high pitch range.

However, although sentence-final position inhibited creak identification for creaky syllables, it caused false alarms of creak presence among modal syllables as sentence-final position cued the existence of low-pitched targets, which were further associated with creak. In addition, modal syllables in sentence-final position were not perceived as different from sentence non-final ones, which suggested that listeners did not tend to perceive sentence-final pitch declination and creak distinctly for modal speech. In the next section, we further discuss the relationship between pitch range and creaky voice.

B. Low pitch facilitates creak identification

In addition to prosodic position, pitch range is another important cue in Mandarin creak identification. Overall, high-pitched creaky syllables were less likely to be perceived as creaky as creak is often associated with a low pitch range. Listeners' expectation of hearing creak in low pitch may make it more difficult for them to notice creaky syllables in that environment. Nevertheless, for modal syllables, creak identification rates were significantly higher when the utterance was read in a low-pitched voice. These findings further suggested that low pitch facilitates creak identification in Mandarin. It is also worth pointing out that low pitch in our experiment was manipulated through both lowering the pitch range and vowel formants. How different pitch ranges within the same gender would influence Mandarin creak identification still needs further inquiry. In addition, as one of the reviewers pointed out, in future research, it may be worthwhile to include a male speaker for "natural" low-pitched stimuli and re-synthesize to a higher pitch, better separating the effects of pitch range and gender.

Our results further provided insights into understanding the relative contributions of gender/social bias in creak identification. In English, female creaky voice is less preferred than male creaky voice. When it comes to creak identification, however, no strong gender effects have been found, indicating that English listeners are not simply biased to recognize creak in female voices as opposed to male voices (Davidson, 2019). However, Davidson (2019) indeed found that there existed a weak tendency for English listeners to identify creak more often in female voices than male voices when the whole utterance was creaky. This finding has been attributed to how creaky voice is evaluated differently between female and male voices in the English context. Our results based on Mandarin suggest the opposite: Mandarin listeners consistently identified creak more often in the low-pitched male voice and it did not significantly interact with creak locality (i.e., creak amount). In addition, to the best of our knowledge, gender asymmetry involved in the social evaluation of creak among Mandarin listeners has not been found (Li and Lai, 2023). These findings based on these two languages further suggest that gender or social bias may still play a role in English creak identification whereas in Mandarin, creak identification is more acoustically-driven.

C. Lexical level information matters in creak identification

Apart from prosodic position and pitch range, another two important factors that we found to be crucial in Mandarin creak identification are tone and stress. When we zoomed in on the effects of tone for locally creaky syllables, we found that listeners were inclined to identify creak in both Tone 1 and Tone 3 syllables while they tended not to perceive Tone 4 syllables as creak-containing. This appeared to partially contradict what has been found before based on corpus work: in Mandarin, creaky voice often occurs in low-pitched targets such as Tone 3 and Tone 4, but not Tone 1 (Belotel-Grenié and Grenié, 1994, 2004; Kuang, 2018). We suspect this contradiction is potentially driven by the amount of creak that was produced in Tone 1-bearing syllables (cf. Fig. 6). How listeners identify creak can also be influenced by the amount of creak originally produced by the speaker. According to the data on the speaker's produced creak amount, in this set of stimuli, Tone 3 and Tone 1 have the highest amount of produced creak proportion, compared with Tone 2 and Tone 4. In broad terms, this speaks to the fact that listeners do pay attention to the acoustic details in the signal. In the literature, speech perception is often assumed to be influenced by production (Pardo and Remez, 2021). While some accounts argue for a tight linkage between perception and production such as Motor Theory (Liberman and Mattingly, 1985), others think that the connection is minimal as neither perception nor production is an automatic consequence of the other (Mitterer and Ernestus, 2008). Based on our current results, the relationship between production and perception is rather complicated. Produced creak influenced creak perception in some tones but not others, suggesting that perception and production coordinate with each other, but when and how much acoustic details are recruited in perception depends on the contexts. Alternatively, as one of the reviewers suggested, it could be listeners' expectations based on their previous experience (in both production and perception), that caused the high identification rates in Tone 1 and Tone 3. Since it is unusual or unexpected to hear creak in Tone 1, listeners became more ready to identify them whenever creak showed up in these tones. In future work, a potential approach to teasing apart the perception experience from the production experience would involve controlling the amount of creak in each Tone and see if the same pattern still holds.

In addition, while Tone 1 and Tone 3 did not significantly interact with prosodic position, Tone 2 and Tone 4 both did. It is interesting that Tone 2 and Tone 4 interacted with prosodic position in opposite directions (cf. Fig. 7), despite the fact that Tone 2 and Tone 4 did not seem to differ in the amount of creak produced in production (cf. Fig. 8). While Tone 2-carrying creaky syllables in sentence-final position significantly inhibited creak identification, Tone 4-carrying ones at the same positions facilitated it. Creaky syllables with Tone 4 (falling tone) were more likely to be perceived as creaky at the end of the sentence, while the ones with Tone 2 (rising tone) were more likely to be perceived as creaky at sentence medial position. Therefore, the effect of prosodic position on creak identification was modulated by pitch contours. Note that even though Tone 3 in isolation is realized with a rising contour, it is essentially a low tone, and its contour is rarely realized in connected speech (Kuang, 2018), hence rendering the contour a less important cue in creak identification for Tone 3.

Here, we propose some possible explanations to account for the (non)-interaction between tone and prosodic position. To start with, if we assume the effect of Tone 3 on creak identification functioned the same as the effect of sentence-final position, when Tone 3 was situated at the end of a sentence listeners in principle should be less likely to identify creak. However, Tone 3 did not significantly interact with sentence-final prosodic position. This suggested that in Mandarin, Tone 3 with creak and sentence-final creak may be perceived distinctly, similar to the contrast of /t/-glottalization and phrase-final creak in English (Garellek, 2015). As for the interaction between Tone 4 and sentence-final position (i.e., creak identification was significantly easier for Tone 4 syllables at the end of sentence), it is possible that the falling pitch contour on Tone 4 indicated low pitch at the end of the contour, thus facilitating creak identification. When the falling contour was in sentence-final position, it canceled the inhibition effect on creak identification from sentence-final declination, thus giving rise to more creak identification. When different levels of contexts/processing co-occur and compete, influence from the lexical context is much stronger. Similarly, since Tone 2 bears a rising pitch contour, the high-pitched targets at the end of the contour might inhibit creak identification. Given that sentence-final declination already inhibited creak identification, when tone- and boundary-based cues provide redundant information, their effects were additive, further inhibiting creak identification. As suggested before in Crowhurst (2018), the effects of different types of contextual clues in speech perception can be either additive or subtractive based on whether these cues are present in either redundant or conflicting arrangements. However, a seemingly plausible pattern is that cues at the lexical level tend to win over cues at the sentence level when it comes to creak identification.

With respect to neutral tone, different realizations of the neutral tone did not elicit an obvious difference on the production side, and none of them showed up as individually influencing listeners' identification of creak. However, we did see that neutral tone interacted with prosodic position such that when the neutral tone was in sentence-final position, listeners were more likely to be able to perceive creak. Given that creak in Mandarin is also associated with unstressed syllables, while sentence-final position inhibited creak identification, this speaks to the possibility that unstressed syllables are more heavily weighted in speech perception.

VII. CONCLUSION

Based on the experiment conducted in this study, we conclude that creak perception is context-dependent and reflects listeners' knowledge about its acoustic and linguistic distributions. Crucially, we established that cues co-occurring with creak modulate listeners' sensitivity to creak in different ways. 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 position can inhibit the identification of creak, whereas low pitch consistently facilitates the identification of creak. This gives rise to a broader question of how contextual cues of a different nature interact with speech perception. In addition, while in speech production Mandarin Tone 1, as a high level tone, is not a tone target where creaky voice has often been found (Belotel-Grenié and Grenié, 1994; Kuang, 2018), in creak identification, the rates of identified creak for Tone 1 were relatively high, due to the high proportion of creak produced in Tone 1 targets. These observations open up the possibility that first, speech production may mirror speech perception (Pardo, 2012). Second, co-varying cues of a different nature may interact with speech perception in different ways. The question of how these interactions vary depending on the nature of contextual cues needs further inquiry.

One limitation of the current study is that speech from only one speaker was used as the vocal stimulus. This potentially invited the possibility listeners' perceptions were heavily influenced by the idiosyncratic nature of that voice, making it hard to generalize across different voices. Conversely, any studies that have examined how creaky voice is perceived often present listeners with voices from different speakers talking in naturalistic speech settings with one in a normal tone of voice and the other manipulated digitally to generate vocal fry. On the one hand, there does not seem to exist a "best" method to generate such a manipulation (Klofstad *et al.*, 2012). On the other hand, even if manipulation succeeds, these studies are always susceptible to critiques related to individual differences. Because different speakers may produce creaky voice differently, it always raises the question of which "creaky-sounding" phonation (sub)type has been identified or perceived. This paper is unable to provide an optimal solution to this dilemma. However, it is worth pointing out that naturalistic speech should be used to, for one, confirm the findings of the current study; and for another, helps address how generalizable these findings are when the stimuli come from multiple different speakers when less experimental control is given.

Taken together, this study aims to develop a better understanding of Mandarin creaky voice by integrating how creaky voice is identified and evaluated in a single-speaker study. Results indicate that in Mandarin, where no obvious gender-differentiated social bias of creaky voice exists, listeners' ability to identify creak is mostly influenced by the environment in which the creak is produced. Creaky voice can be easily recognized when there are multiple creaky syllables in a sentence. Sentence-final creak is hard to identify. A high tone can also elicit a high proportion of creak identification if creak is highly present in the production. While low-pitched voice triggers higher rates of identification of creaky voice, it also causes false alarms for the presence of creak in modal syllables. Further, this preference for creak identification in a low-pitched environment is unlikely to be driven by biases in social evaluation. This study sheds light on cross-linguistic research on creaky voice and further contributes to current studies in speech perception.

[1] We are aware of the fact that the speaker was not unaware of creak as a phonetic category since the speaker is one of the co-authors, and hence not in the best position to read sentences without self-awareness of phonation awareness. However, we wanted to make sure all the sentences in intended conditions can be reproduced. It is hard to achieve this by recruiting a speaker who simply has little knowledge about phonetics.

Aare, K., Lippus, P., and Šimko, J. (**2014**). "Creaky voice in spontaneous spoken Estonian," in *XXVII Fonetiikan Päivät*, edited by K. Jähi and L. Taimi. Turku, 25–26 October 2013, pp. 27–35.

Abdelli-Beruh, N. B., Drugman, T., and Owl, R. R. (**2016**). "Occurrence frequencies of acoustic patterns of vocal fry in American English speakers," J. Voice **30**(6), 759.E711–759.E720.

Abdelli-Beruh, N. B., Wolk, L., and Slavin, D. (**2014**). "Prevalence of vocal fry in young adult male american english speakers," J. Voice **28**(2), 185–190.

Abramson, A. S., Theraphan, L., and Nye, P. W. (**2004**). "Voice register in Suai (Kuai): An analysis of perceptual and acoustic data," Phonetica **61**(2-3), 147–171.

Andruski, J. E. (**2006**). "Tone clarity in mixed pitch/phonation-type tones," J. Phon. **34**(3), 388–404.

Andruski, J. E., and Ratliff, M. (**2000**). "Phonation types in production of phonological tone: The case of Green Mong," J. Int. Phon. Assoc. **30**(1–2), 37–61.

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

Belotel-Grenié, A., and Grenié, M. (**1994**). "Phonation types analysis in standard Chinese," in *Proceedings of the ICSLP Conference*, September 18–22, Yokohama, Japan.

Belotel-Grenié, A., and Grenié, M. (**1997**). "Types de phonation et tons en chinois standard" ("Phonation types and tones in standard Chinese"), Cahiers de Clao **26**(2), 249–279.

Belotel-Grenié, A., and Grenié, M. (**2004**). "The creaky voice phonation and the organisation of chinese discourse," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, March 28–31, Beijing, China.

Bishop, J., and Keating, P. (**2012**). "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," J. Acoust. Soc. Am. **132**(2), 1100–1112.

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.

Brunelle, M. (**2009**). "Tone perception in Northern and Southern Vietnamese," J. Phon. **37**(1), 79–96.

Brunelle, M., and Kirby, J. (**2016**). "Tone and phonation in Southeast Asian languages," Lang. Ling. Compass **10**(4), 191–207.

Butler, E. (**2017**). "The use of creaky voice in mitigating face threatening acts," in *Student Research Submissions* (University of Mary Washington, Fredericksburg, VA), Vol. 164.

Crowhurst, M. J. (**2018**). "The joint influence of vowel duration and creak on the perception of internal phrase boundaries," J. Acoust. Soc. Am. **143**(3), EL147–EL153.

Cullen, A., Kane, J., Drugman, T., and Harte, N. (**2013**). "Creaky voice and the classification of affect," in *Proceedings of WASSS*, August 22–23, Grenoble, France.

Dallaston, K., and Docherty, G. (**2020**). "The quantitative prevalence of creaky voice (vocal fry) in varieties of English: A systematic review of the literature," PLoS One **15**(3), e0229960.

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

Davidson, L. (**2020**). "The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages," in *Wiley Interdisciplinary Reviews: Cognitive Science* (Wiley, New York), p. e1547.

DiCanio, C. T. (**2009**). "The phonetics of register in Takhian Thong Chong," J. Int. Phonetic Assoc. **39**(2), 162–188.

Duanmu, S. (**2007**). *The Phonology of Standard Chinese* (Oxford University Press, Oxford, UK).

Esling, J. (**1978**). "The identification of features of voice quality in social groups," J. Int. Phonetic Assoc. **8**(1/2), 18–23.

Esposito, C. M. (**2012**). "An acoustic and electroglottographic study of White Hmong tone and phonation," J. Phon. **40**(3), 466–476.

Esposito, L. (**2016**). "'I am a perpetual underdog': Lady Gaga's use of creaky voice in the construction of a sincere pop star persona," Undergraduate thesis, Swarthmore College, Swarthmore, PA.

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

Garellek, M. (**2015**). "Perception of glottalization and phrase-final creak," J. Acoust. Soc. Am. **137**(2), 822–831.

Garellek, M., and Keating, P. (**2011**). "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec," J. Int. Phonetic Assoc. **41**(2), 185–205.

Gobl, C., and Chasaide, A. N. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**(1-2), 189–212.

Gobl, C., and Chasaide, A. (**2010**). "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences* (Wiley, New York), Vol. 50, p. 378.

Gordon, M., and Ladefoged, P. (**2001**). "Phonation types: A cross-linguistic overview," J. Phon. **29**(4), 383–406.

Greer, S. D., and Winters, S. J. (**2015**). "The perception of coolness: Differences in evaluating voice quality in male and female speakers," in *Proceedings of the ICPhS 2015*, August 10–14, Glasgow, Scotland.

Heldner, M., Wlodarczak, M., Beňuš, Š., and Gravano, A. (**2019**). "Voice quality as a turn-taking cue," in *Interspeech 2019*, September 15–19, Graz, Austria, pp. 4165–4169.

Henton, C. G. (**1989**). "Sociophonetic aspects of creaky voice," J. Acoust. Soc. Am. **86**(S1), S26–S26.

Hildebrand-Edgar, N. (**2016**). "Creaky voice: An interactional resource for indexing authority," Ph.D. thesis, University of Victoria, Victoria, Canada.

Hollien, H., and Michel, J. F. (**1968**). "Vocal fry as a phonational register," J. Speech Hearing Res. **11**(3), 600–604.

Huang, Y. (**2020**). "Different attributes of creaky voice distinctly affect mandarin tonal perception," J. Acoust. Soc. Am. **147**(3), 1441–1458.

Huang, Z. (**2018**). "Yi 'zi' wei danwei de hanyu pingjun juchang yu juchang fenbu yanjiu" ("A study on the mean length and distribution of sentences in Chinese"), J. Qiqihar Univ. (Philos. Social Science Ed.) **1**, 133–138.

Keating, P. A., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the ICPhS 2015*, August 10–14, Glasgow, Scotland.

Klofstad, C. A., Anderson, R. C., and Peters, S. (**2012**). "Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women," Proc. R. Soc. B **279**(1738), 2698–2704.

Kuang, J. (**2011**). "Production and perception of the phonation contrast in Yi," Ph.D. thesis, University of California, Los Angeles, CA.

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**(3), 1693–1706.

Kuang, J. (**2018**). "The influence of tonal categories and prosodic boundaries on the creakiness in mandarin," J. Acoust. Soc. Am. **143**(6), EL509–EL515.

Kuang, J., and Liberman, M. (**2015**). "Influence of spectral cues on the perception of pitch height," in," in *Proceedings of the ICPhS 2015*, August 10–14, Glasgow, Scotland.

Kuang, J., and Liberman, M. (**2016a**). "The effect of vocal fry on pitch perception," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 20–25, Shanghai, China, pp. 5260–5264.

Kuang, J., and Liberman, M. (**2016b**). "Pitch-range perception: The dynamic interaction between voice quality and fundamental frequency," in *Proceedings of InterSpeech*, September 8–12, San Francisco, CA, pp. 1350–1354.

Kuang, J., and Liberman, M. (**2018**). "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," Front. Psychol. **9**, 2147.

Kuo, G. (**2012**). "Perceived prosodic boundaries in taiwanese and their acoustic correlates," in *Thirteenth Annual Conference of the International Speech Communication Association*, September 9–13, Portland, OR.

Kuo, G. (**2018**). "Prosodic disambiguation of syntactically ambiguous sentences," in *Proceedings of TAL2018*, June 18–20, Berlin, Germany.

Laver, J. (**1980**). "The phonetic description of voice quality," Cambridge Stud. Ling. Lond. **31**, 1–186.

Lee, S. (**2015**). "Creaky voice as a phonational device marking parenthetical segments in talk," J. Socioling. **19**(3), 275–302.

Lehiste, I., Cohen, A., and Nooteboom, S. (**1975**). "Structure and process in speech perception," in *Proceedings of the Symposium on Dynamic Aspects of Speech Perception*, August 4–6, Eindhoven, the Netherlands, pp. 195–203.

Li, A., and Lai, W. (**2023**). "How do listeners evaluate creak: A matched-guise study in mandarin chinese," *paper presented at the Linguistics Society of America 2023 Annual Meeting*, June 5–8, Denver, CO.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**(1), 1–36.

Melvin, S. (**2015**). "Gender variation in creaky voice and fundamental frequency," Ph.D. thesis, The Ohio State University, Columbus, OH.

Mendoza-Denton, N. (**2011**). "The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a chicana/o gang persona," J. Ling. Anthropol. **21**(2), 261–280.

Mitterer, H., and Ernestus, M. (**2008**). "The link between speech perception and production is phonological and abstract: Evidence from the shadowing task," Cognition **109**(1), 168–173.

Ogden, R. (**2002**). "Creaky voice and turn-taking in Finnish," in *Colloquium British Association of Audiological Physicians*, March 25–27, Newcastle, UK.

Oliveira, G., Davidson, A., Holczer, R., Kaplan, S., and Paretzky, A. (**2016**). "A comparison of the use of glottal fry in the spontaneous speech of young and middle-aged American women," J. Voice **30**(6), 684–687.

Pardo, J. S. (**2012**). "Reflections on phonetic convergence: Speech perception does not mirror speech production," Lang. Ling. Compass **6**(12), 753–767.

Pardo, J. S., and Remez, R. E. (**2021**). "On the relation between speech perception and speech production," in *The Handbook of Speech Perception* (Wiley Online Library, New York), pp. 632–655.

Pierrehumbert, J. (**1979**). "The perception of fundamental frequency declination," J. Acoust. Soc. Am. **66**(2), 363–369.

Pittam, J. (**1987**). "Listeners' evaluations of voice quality in Australian English speakers," Lang. Speech **30**(2), 99–113.

Podesva, R. J. (**2011**). "Gender and the social meaning of non-modal phonation types," in *Proceedings of the 37th Berkeley Linguistics Society*, February 12–13, Berekely, CA, pp. 427–448.

R Core Team (**2013**). "R: A language and environment for statistical computing," http://www.R-project.org/ (Last viewed April 26, 2013).

Redi, L., and Shattuck-Hufnagel, S. (**2001**). "Variation in the realization of glottalization in normal speakers," J. Phon. **29**(4), 407–429.

Tang, P. (**2014**). "A study of prosodic errors of chinese neutral tone by advanced Japanese students (Chinese version)," TCSOL Stud. **56**(4), 39–47.

Thurgood, E. (**2004**). "Phonation types in Javanese," Oceanic Ling. **43**, 277–295.

Wickham, H. (**2011**). "ggplot2," WIREs. Comp. Stat. **3**(2), 180–185.

Wolk, L., Abdelli-Beruh, N. B., and Slavin, D. (**2012**). "Habitual use of vocal fry in young adult female speakers," J. Voice **26**(3), e111–e116.

Xu, Y. (**1997**). "Contextual tonal variations in mandarin," J. Phon. **25**(1), 61–83.

Yang, R.-X. (**2011**). "The phonation factor in the categorical perception of mandarin tones," in *Proceedings of the 2011 ICPhS*, August 17–21, Hong Kong, pp. 2204–2207.

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.

Yuasa, I. P. (**2010**). "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?," Am. Speech **85**(3), 315–337.

Zetterholm, E. (**1998**). "Prosody and voice quality in the expression of emotions," in *Proceedings of the 5th ICSLP*, November 30–December 4, Sydney, Australia.

Zhu, X. (**2012**). "Multiregisters and four levels: A new tonal model," J. Chin. Ling. **40**(1), 1–17, available at https://www.jstor.org/stable/23754196.