JASA ARTICLE

# Perception of Xiapu Min checked syllables and tones in citation and sandhi forms[a)]

Yuan Chai[1,b)] and Shihong Ye[2]
[1]Department of Linguistics, University of Washington, Seattle, Washington 98195, USA
[2]Second High School Attached to Beijing Normal University, Beijing 100088, China

**ABSTRACT:**
This study investigates the acoustic cues for listeners to differentiate checked syllables and tones from unchecked ones. In Xiapu Min, checked and unchecked syllables and tones differ in *f*0, glottalization, and duration, whereas these differences are reduced in their sandhi forms. In citation forms, listeners utilize all three cues while relying on duration the most. The results indicate that duration is an independent perceptual cue for checked syllables and tones, rather than a peripheral cue resulting from the syllable structure of /CVʔ/. In sandhi forms, where checked and unchecked syllables and tones are phonologically neutralized, the duration and *f*0 still influence listeners' perception of checked constituents significantly. Data from Xiapu Min, along with other languages, illustrate that cues consistently found in the production of checked syllables and tones are likely to be utilized in their perception. © *2025 Acoustical Society of America*. https://doi.org/10.1121/10.0034785

(Received 4 June 2024; revised 18 November 2024; accepted 5 December 2024; published online 9 January 2025)
[Editor: Mark A. Hasegawa-Johnson] Pages: 70–83

## I. INTRODUCTION

In Chinese languages, checked syllables are syllables closed by a voiceless obstruent, whereas unchecked syllables are open syllables or syllables closed by sonorant. Checked tones are tones borne by checked syllables (Chai, 2022). Checked syllables and tones (or *Rù* "entering" syllables 入声韵 and tones 入声调) have been extensively studied in terms of their origin in Old Chinese and Middle Chinese (Dong, 2020; Haudricourt, 1954; Zhengzhang, 2003) and their acoustic and articulatory properties in production (e.g., Taiwanese Min, see Pan, 2017; Pan *et al.*, 2016; Shanghainese Wu, see Chen and Gussenhoven, 2015; Gao and Kuang, 2022; Meixian Hakka, see Shao, 2012; Hong Kong Cantonese, see Bauer and Matthews, 2017). The production studies of checked syllables and tones illustrate their multi-dimensional phonetic properties. However, limited research has addressed what phonetic cues are involved for listeners to perceive checked syllables and tones. The current study aims to contribute more original empirical data to the perception mechanism of checked syllables and tones by investigating the independent perceptual roles of three acoustic cues in a low-resourced language: Xiapu Min.

Checked syllables and checked tones are two phonological concepts instantiated on the same phonetic constituent (i.e., the nucleus vowel). Checked syllable refers to a syllable structure. Checked tone refers to a tonal melody. In the production of checked syllables and tones, we frequently observe their nucleus vowels have a short duration, distinct pitch contours, and glottalization,[1] compared with unchecked syllables and tones (Shao, 2012; Tang, 2014; Wu, 2018). The short duration could be a result of the syllable structure (i.e., vowels in closed syllables are shorter than those in open syllables), or an inherent property of the checked tone. The glottalization on the vowel is likely due to the coarticulation of the glottal/glottalized stop coda in the checked syllable. The distinct pitch contour is the property of the checked tone. Although these three acoustic properties likely come from different phonological structures, they are realized on the same nucleus. Thus, we integrate the acoustic cues resulting from checked syllable and tone in the same perception experiment to investigate what acoustic cues listeners attend to when identifying a word with checked syllable and checked tone.

In the following, we review studies that have tested the perception of checked syllables and tones, or phonetic constituents that have similar phonetic properties to checked syllables and tones. The three phonetic properties of checked syllables and tones, *f*0, duration, and glottalization, have all been found to be perceptually salient for the identification of checked-like syllables and tones in some languages.

The perceptual weight of each acoustic cue varies by language. Several languages found glottalization to be a less important cue than duration and/or *f*0. Danish "stød" is a phonation that is realized with a high *f*0 onset and creaky offset (Peña, 2022, 2023). When Danish listeners identify Danish stød, glottalization plays a less important role than high *f*0. White Hmong has a creaky tone resembling checked tones phonetically, such that it is realized with low-falling *f*0, short duration, and vowel-final glottalization (Peña, 2023). For White Hmong listeners, glottalization is not used by listeners

[a)]This paper is part of the special issue on Acoustic Cue–Based Perception and Production of Speech by Humans and Machines.
[b)]Email: yuanchai@uw.edu

 

https://doi.org/10.1121/10.0034785

to identify the creaky tone, whereas *f*0 and duration have a significant influence on listeners' perception (Garellek *et al.*, 2013). Gao (2004) conducted identification experiments on the tones of Shanghainese Wu, a Wu variety that has checked syllables closed with a glottal stop and two checked tones (55 and 12 in Chao numeral notation) borne by checked syllables (Chen and Gussenhoven, 2015). Gao (2004) found that shortening a naturally produced unchecked without vowel-final glottalization to 56 ms led listeners to consistently identify it as a checked tone; whereas lengthening a naturally produced checked tone with vowel-final glottalization to twice its original length led listeners to identify it as an unchecked tone 100% of the time. This suggests that in Shanghainese Wu, vowel-final glottalization is not necessary for checked tone perception when the duration is sufficiently short, and that the presence of glottalization does not guarantee a checked tone percept.

In contrast, glottalization can also play an essential role in other languages. Burmese "creaky" and "glottalized" tones are characterized by high *f*0, vowel-final glottalization, and shorter duration. Stimuli with glottalization predominantly elicited "creaky" and "glottalized" responses regardless of the *f*0 and duration conditions (Gruber, 2011). Northern Vietnamese B2 (22 nặng) tone is realized with low *f*0 and a full glottal stop closure or strong glottalization at the end of the vowel. Listeners perceived stimuli with final-glottalization predominantly as the B2 (22) tone, regardless of the *f*0 condition (Brunelle, 2009).

We also observe cases where the combination of multiple cues leads to a checked-like tone percept. Sgaw Karen "glottalized" tone (T4) has a mid-falling *f*0 contour, strong vowel-final glottalization, and the shortest duration among the six tones in the language. Brunelle and Finkeldey (2011) reported that Sgaw Karen listeners identified the stimuli as having the "glottalized" T4 when they exhibited mid- or final-glottalization and short duration, whereas *f*0 was not a relevant cue for identifying the "glottalized" tone. Cues can also interact with each other in their effects on eliciting checked tones. Taiwanese Southern Min has both mid- and high-registered checked tones borne by /Vp,t,k,ʔ/ syllables. The study by Zhang and Lu (2023) showed that the listeners utilized the duration cue when distinguishing the mid-registered tones (T3 and T33) but not the high-registered tones (T5 and T55). Tang (2017) and Tang and Li (2018) explored the effects of duration and *f*0 on the perception of checked tones in Yangzhou Jianghuai Mandarin, which has checked syllables closed by a glottal stop and one checked tone (T5). They found that lengthening the duration of checked tones elicited more unchecked tone responses while shortening the duration of unchecked tones increased the percentage of checked tone responses. Notably, the duration threshold for a categorical shift from unchecked to checked tone responses was longer for checked tone-based tokens than for unchecked ones. Thus, the authors argued that while glottalization is not a necessary cue for checked tone perception when the duration is sufficiently short, it still plays a facilitative role. When the duration is ambiguous between checked and unchecked tones, the presence of glottalization biases the listeners towards a checked tone percept. They also found that varying the *f*0 contours of a checked or unchecked token did not significantly change listeners' perception of the tone, indicating that the listeners do not rely on *f*0 cue to differentiate between checked and unchecked tones in Yangzhou Jianghuai Mandarin.

The current study investigates the perceptual cues used by listeners of Xiapu Min through two identification experiments to contribute to the rich perceptual variations for checked(-like) constituents. Xiapu Min is a Min variety spoken in Xiapu County, Fujian, China. There were 475 936 residents in Xiapu County in the year 2020 (Xiapu County Bureau of Statistics, 2021). Based on the field research by the authors, Mandarin Chinese is the language of government, work, and school in Xiapu County. Xiapu Min is used mostly at home between family members. There is a tendency for the younger generation to speak Mandarin exclusively but are passive listeners of Xiapu Min. Thus, the number of Xiapu Min speakers is smaller than the number of residents in Xiapu County. Xiapu Min contains checked syllables ending in glottal stop and checked tones T21 and T53 (in Chao numerals; hereafter referred to as T2 and T5). It also has five unchecked tones T33, T11, T24, T13, and T42, borne by open or nasal-closed syllables. A seven-way minimal set of the seven tones is presented in Table I. Their pitch track is in Fig. 1. Checked syllables and tones always co-occur with each other: checked syllables carry either a high-falling or a low-falling tone; the high-falling and low-falling are always realized on checked syllables. The purpose of this study is to test what acoustic cues lead to listeners' perceptions of checked syllables and tones: whether the vowel has to be short, whether the vowel has to be glottalized, and whether the vowel has to carry a particular pitch. These three acoustic properties have been observed in Xiapu Min checked syllables and tones (Chai and Ye 2022).

In Xiapu Min, checked syllables and tones occur in both citation and sandhi forms. Chai and Ye (2022) observed partial phonetic neutralization between the sandhi forms of checked and unchecked syllables and tones in compound words. Here, we follow the definition of neutralization as underlying phonological contrast being lost in the surface form in particular contexts (Dinnsen, 1985; Winter and Roettger, 2011), resembling the tapping/flapping neutralization of English /t/ and /d/ in intervocalic position (Braver, 2014). In Xiapu Min, tokens with a citation tone of T2, T13, and T33 are neutralized into a 33 *f*0 contour in a compound-initial position. Tokens with a citation tone of T5, T24, and

TABLE I. Minimal set of seven tones in Xiapu Min.

| Tone | Middle Chinese correspondence | Word | Gloss |
|---|---|---|---|
| T33 | 阴平 Yinping | /θi 33/ | 诗 "poem" |
| T11 | 阳平 Yangping | /θi 11/ | 时 "time" |
| T42 | 上 Shang | /θi 42/ | 死 "die" |
| T24 | 阴去 Yinqu | /θi 24/ | 四 "four" |
| T13 | 阳去 Yangqu | /θi 13/ | 是 "to be" |
| T5 | 阴入 Yinru | /θiʔ5/ | 湿 "wet" |
| T2 | 阳入 Yangru | /θiʔ2/ | 实 "concrete" |

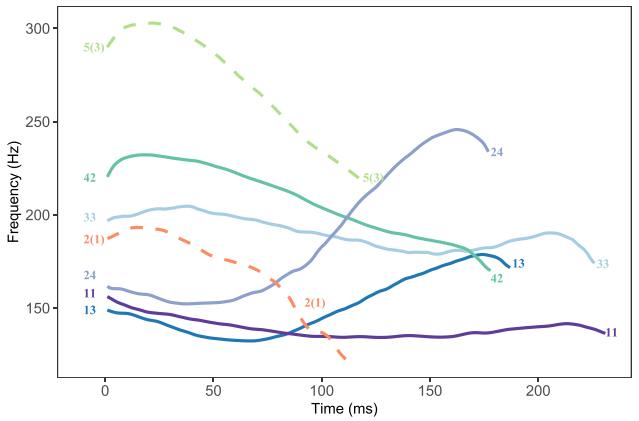


FIG. 1. (Color online) *F*0 track of /θi/ in seven tones by a female speaker. Adapted from Fig. 2 from Chai and Ye (2022), Languages **7**(1), 47. Copyright 2022 Author(s), licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

T42 all carry a 55 *f*0 contour in sandhi forms. T11 does not go through sandhi in compound-initial position. Checked syllables lose their glottal stop coda in sandhi forms. The formal representation of the sandhi rules and examples are in Table II.

Neutralizations can be either complete or incomplete on the phonetic level (Dinnsen 1985). On the phonetic level, the completeness of neutralization needs to be evaluated from the perspectives of production and perception. A complete neutralization in production means there are no phonetic differences between the realizations of neutralized contrasts. An incomplete one in production means there are still consistent phonetic differences between their realizations. A complete neutralization in perception means the listeners are not able to tell one contrast from the other when they are neutralized. An incomplete neutralization in perception means that the listeners are still able to differentiate the neutralized forms of the contrasts above chance. For instance, Taiwanese Min has a neutralization between T2 and T5 in sandhi forms. Their sandhi forms do not differ in *f*0 or duration (Chien and Jongman, 2019). The listeners cannot differentiate their sandhi forms above chance (Kuo, 2013). These suggest a complete neutralization between T2 and T5 in Taiwanese Min in both production and perception. In contrast, in German, voiced obstruents are devoiced in word-final position, and consequently neutralized with voiceless obstruents in word-final position (Dinnsen, 1985). Roettger *et al.* (2014) found that the vowels before devoiced stops are significantly longer than those before voiceless stops. The listeners were able to differentiate minimal pairs with neutralized devoiced stops from voiceless stops significantly above chance. This indicates that in German, the phonological neutralization between voiced and devoiced stops is phonetically incomplete in terms of both production and perception.

TABLE II. Tone sandhi rules in Xiapu Min.

| Sandhi rule | Example | Gloss |
|---|---|---|
| /T2, T13, T33/ → [T33] / ____$_{\sigma}$ + $_{\sigma}$ X | /koʔ 2 tioŋ 42/ → [ko 33 tioŋ 42] | 局长 "governor" |
| | /to 13 k$^{h}$eu 42/ → [to 33 k$^{h}$eu 42] | 路口 "intersection" |
| | /ko 33 p$^{h}$o 42/ → [ko 33 p$^{h}$o 42] | 歌谱 "song sheet" |
| /T5, T24, T42/ → [T55] / ____$_{\sigma}$ + $_{\sigma}$ X | /t$^{h}$eʔ 5 to 13/ → [t$^{h}$e 55 to 13] | 铁路 "railroad" |
| | /t$^{h}$e 24 tsoi 13/ → [t$^{h}$e 55 tsoi 13] | 替罪 "scapegoat" |
| | /t$^{h}$e 42 tain 13/ → [t$^{h}$e 55 tain 13] | 体重 "body weight" |
| /T11/ → [T11] / ____$_{\sigma}$ + $_{\sigma}$ X | /θi 11 kein 33/ → [θi 11 kein 33] | 时间 "time" |

For Xiapu Min, Chai and Ye (2022) conducted a statistical analysis of the neutralization among tones in sandhi forms in production. We found that the neutralization between checked and unchecked syllables and tones in sandhi forms was incomplete in production, such that checked syllables and tones consistently had a shorter duration than unchecked syllables and tones in their neutralized sandhi form. For the neutralization between T2 and T33, we also found differences in *f*0 and voice quality, such that T2 in sandhi forms has higher *f*0 and less degree of glottal constriction (higher H1$^{*}$–H2$^{*}$). For the neutralization between T5 and T24, and between T5 and T42, we also found differences in voice quality, such that T5 is noisier (lower HNR). While we know that there are still phonetic differences between the neutralized checked syllables and tones in sandhi forms in production from Chai and Ye (2022), our following question is, are these phonetic differences found in production meaningful to the listeners in perception? Will the listeners be able to differentiate among T2, T13, and T33, and/or among T5, T24, and T42 in their sandhi forms in identification tests? In the current study, we investigate whether the neutralization between checked and unchecked syllables and tones is phonetically complete in perception.

To answer the research questions of (1) what acoustic cues listeners use when identifying checked syllables and tones in citation form; (2) whether the phonological neutralization between checked and unchecked syllables and tone in sandhi form is phonetically complete, we designed an experiment for each question, respectively. Resynthesized stimuli are used in Experiment I, because it is most likely that the listeners are able to correctly identify the tones and syllable types in their citation forms in the language, and our goal is to determine what acoustic cues the listeners use for citation form identification. Experiment II used natural stimuli because we still do not know whether the listeners are able to distinguish the neutralized tones and syllable types. We need to answer that question before asking what acoustic cues the listeners use to distinguish the neutralized contrasts. The experiment web page, stimuli, data, and data analysis script are available at https://doi.org/10.17605/OSF.IO/94FVT.

## II. EXPERIMENT I: PERCEPTION OF CHECKED TONES IN CITATION FORMS

In Experiment I, we aim to understand how *f*0, glottalization, and duration affect native Xiapu Min listeners'

https://doi.org/10.1121/10.0034785

perception of checked syllables and tones in citation forms. To this end, we resynthesized a token /θi/ with five *f*0 contours, modal or glottalized phonations, and short or long durations, and performed a word identification task on native listeners of Xiapu Min.

### A. Stimuli, participants, and procedure

In order to test the independent effect of *f*0, phonation, and duration on checked syllable and tone identification, we resynthesized stimuli varying in *f*0, phonation, and duration orthogonally. The base token of the resynthesis is a token of /θi/ in a mid-level tone (33) produced by a female Xiapu Min speaker (hereafter referred as Speaker #1).[2] We created five conditions for *f*0, they are low-falling (21), high-falling (53), low-level (11), mid-level (33), and mid-falling (42). The *f*0 contour values are based on another female speaker (hereafter referred to as Speaker #2[3]), who is evaluated as having a prototypical production of the seven tones by the second author of this article, a native speaker of Xiapu Min. The *f*0 values for the five *f*0 conditions are in Table VII in the Appendix. These contours were chosen to introduce ambiguity between checked and unchecked tones in the *f*0 space. For instance, the low-falling (21), low-level (11), and mid-level (33) *f*0 could be interpreted as the *f*0 of T2, while the high-falling (53), mid-falling (42), and mid-level (33) *f*0 could resemble the *f*0 of T5. In terms of the duration conditions, the short condition has a vowel 115 ms. The long condition has 235 ms. These values are based on the duration of the shortest and the longest tones among the seven tones realized on /θi/ and /θiʔ/ produced by Speaker #2. The vowel duration of the natural production by Speaker #2 for the seven words in the word options are: T2: 111 ms; T5: 133 ms; T42: 183 ms; T13: 186 ms; T24: 190 ms; T33: 225 ms; T11: 230 ms. The short duration condition 115 ms is close to the duration of the shortest tone—T2—among the seven tones; while the long duration condition 235 ms is close to the longest tone—T11. The recordings of the seven word tokens produced by Speaker #2 are available in the supplementary material at https://doi.org/10.17605/OSF.IO/94FVT.

In terms of the two voice quality conditions, the base token has modal phonation. The glottalized phonation was created by lowering and jittering the second half of the vowel. The *f*0 values and corresponding time points of each pitch point are in Table VIII in the Appendix. Glottalization was added to the second half of the vowel because, in the citation forms of the checked syllables produced by the nine speakers in the production experiment by Chai and Ye (2022), the average proportion of glottalization is 50% of the vowel, occurring at the end of the vowel. We simulated the glottalization by lowering and jittering the *f*0, because a prototypical creaky voice is produced with low and irregular *f*0, as well as a constricted glottis (Keating *et al.*, 2015). Previous studies have also used this method to simulate creaky voice, and have proved its effectiveness in eliciting a perception of creaky voice. Frazier (2009) simulated the glottalization in glottalized vowels [V̰V̰] in Yucatec Maya by inserting a pitch point with an extra-low *f*0 value (35 Hz) in the middle of the vowel. Such stimuli elicited more responses for glottalized vowels from the Yucatec Maya listeners. Huang (2020) resynthesized Mandarin tones with extra-low *f*0 and found that extra-low *f*0 shortened the reaction time for identifying the creaky Mandarin tone (T3) and lengthened the reaction time of identifying non-creaky Mandarin tones (T1, T2, T4).

We resynthesized the stimuli with the conditions previously noted in three steps. Starting with a base token of /θi/ in a mid-level tone (33), we first adjusted the duration to create short (115 ms) and long (235 ms) variants. Following this, we resynthesized these tokens into five f0 contours: low-falling (21), high-falling (53), low-level (11), mid-level (33), and mid-falling (42). In the final resynthesis step, we altered the *f*0 of the latter half of each vowel, making its value low and irregular, aiming to produce a glottalized percept. This process resulted in twenty conditions (short/long duration * 5 *f*0 value sets * modal/glottalized phonation). The pitch tracks of these stimuli are illustrated in Fig. 2. Figure 3 presents the spectrograms of sample stimuli with high-falling *f*0 varying in the four conditions of duration and glottalization; high-falling *f*0 long modal; high-falling *f*0 short modal; high-falling *f*0 long glottalized; high-falling *f*0 short glottalized. The sound files of the stimuli are available at https://doi.org/10.17605/OSF.IO/94FVT.

Twenty-nine people participated in the experiment (17 women; 12 men; average age = 48). Their age distribution is in Fig. 4(a). All participants self-identified themselves as Xiapu Min native speakers, with eleven also considering Mandarin as their other native language. None reported fluency in languages other than Xiapu Min and Mandarin.

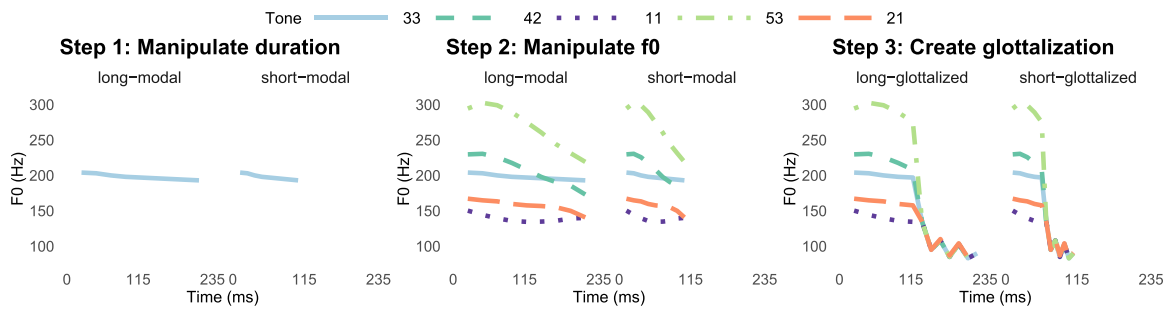


FIG. 2. (Color online) Pitch track of Experiment I stimuli.

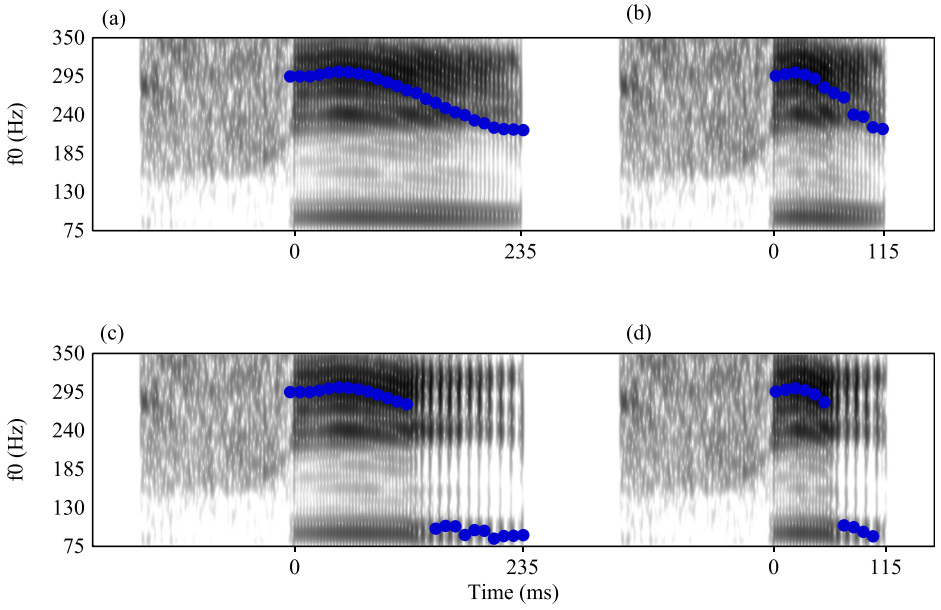


FIG. 3. (Color online) Spectrograms (0–5000 Hz) of the four conditions of duration and glottalization with High-falling *f*0. (a) High-falling *f*0 long modal; (b) high-falling *f*0 short modal; (c) high-falling *f*0 long glottalized; (d) high-falling *f*0 short glottalized. The blue dots represent the *f*0.

The participants engaged in a word identification task. The experiment was presented through an HTML webpage on a Microsoft Surface Pro 6 laptop (Microsoft, Redmond, WA) with a sound card of Realtek High Definition Audio (SST). The participants listened to the stimuli through SONY MDR-ZX110AP headsets (Sony, Tokyo, Japan), and selected the word they heard out of the seven options contrasting in tone and syllable structure (presented in Chinese characters) (in Table I). The task consisted of two blocks, each block presenting 40 stimuli tokens (20 test tokens + 20 fillers) in a randomized order unique to each participant. The participants can listen to each token as many times as desired. The experiment is available online at https://yuanchaiyc.github.io/xmperception/.

At the end of Experiment I, participants were instructed to produce the seven tonal contrasts from the test trials to verify their familiarity with the target words and tones. Twelve participants produced one or more of the seven words differently than anticipated, either using a different tone or different phonemes. Notably, the word 实 "concrete" /θiʔ 2/ was frequently produced with an unexpected tone, possibly because it commonly occurs in compounds but rarely in isolation in Xiapu Min. We believe that excluding these participants enhances the accuracy and reliability of our findings on checked tone perception, presenting a more conservative approach that allows for more confidence in the results. The data from participants who have different pronunciations of the segments and/or tones of the target words introduce uncertainty unrelated to the perceptual cues under investigation. For example, one excluded participant produced 实 "concrete" with a rising tone T13 instead of the expected T2. Consequently, when this participant selected the word 实 "concrete" in the perception experiment, we were not sure whether their responses correspond to T13 or T2, making the interpretation of the results inaccurate. Another participant produced the word 湿 "wet" /θiʔ 5/ as [tan 11]. In the test, they never selected the word 湿 "wet," but it is most likely because their pronunciation of word 湿 "wet" differs from the expected pronunciation in segmental structure, rather than because the

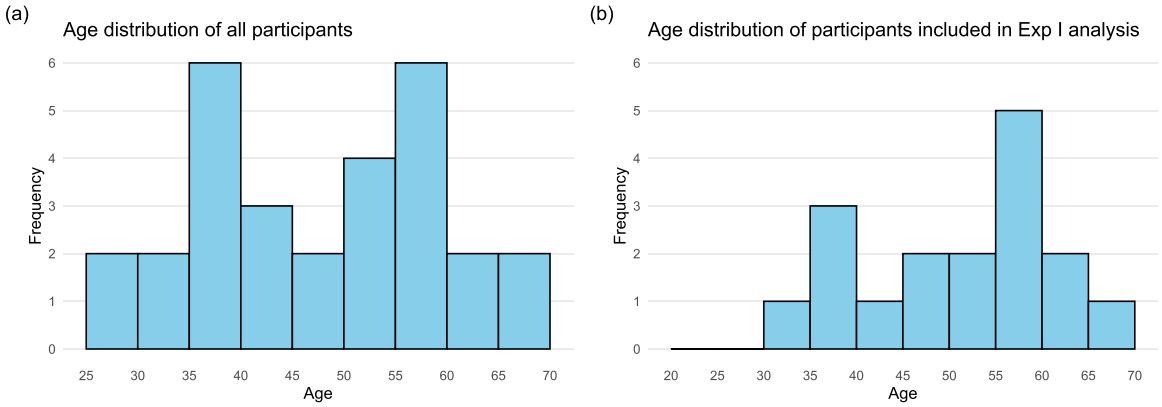


FIG. 4. (Color online) Age distribution of participants. (a) Age distribution of all participants; (b) age distribution of participants whose data are used for Experiment I data analysis.

https://doi.org/10.1121/10.0034785

TABLE III. Percentage of checked tone (overall and T2 and T5 separately) by fixed effects.

| | f0 | | | | | Duration | | Glottalization | |
|---|---|---|---|---|---|---|---|---|---|
| | 21 | 11 | 33 | 42 | 53 | Long | Short | Modal | Glottalized |
| Checked | 57.4 | 55.1 | 38.2 | 46.3 | 73.2 | 37.1 | 70.9 | 43.5 | 64.5 |
| T2 | 53.7 | 54.4 | 9.6 | 1.5 | 2.3 | 17.4 | 31.2 | 21.2 | 27.4 |
| T5 | 3.7 | 0.7 | 28.7 | 44.9 | 70.9 | 19.7 | 39.8 | 22.4 | 37.1 |

*f*0, duration, or voice quality cues do not meet their expectation of checked tone T5. As a result, we excluded the responses from these twelve participants with production deviations from Experiment I data analysis. The final dataset includes seventeen listeners (10 women, 7 men; average age = 51), whose age distribution is shown in Fig. 4(b). This yielded a total of 679 data points for analysis (20 test tokens × 2 repetitions × 17 participants, minus 1 response error). To address concerns about the impact of these exclusions on our results, we conducted the same statistical analyses including all participants. The patterns observed were consistent with those reported in the main text, supporting the robustness of our findings. These additional results are provided in the supplementary material at https://osf.io/zgp5f.

## B. Results

We calculated the percentages of overall checked syllable and tone responses, as well as checked T2 and T5 tone responses separately, under varying conditions of *f*0, duration, and glottalization, as shown in Table III. The percentages of checked tone (T2 and T5), T2, and T5 responses under each of the 20 conditions are in Fig. 5. On average, checked syllable and tone responses (T2 and T5) were more frequent in the short duration condition than in the long duration condition; and more in the glottalized condition than the modal condition. The effect of *f*0 varied by which checked tone was elicited. The low checked T2 tone was predominantly elicited by low *f*0 conditions, specifically the low-falling (21) and low-level (11) contours. Conversely, the high checked T5 tone was primarily elicited under mid to high *f*0 conditions, including the mid-level (33), mid-falling (42), and high-falling (53) contours. The mid-level *f*0 condition was the most ambiguous, eliciting the lowest percentage (38.2%) of checked tone responses among the five *f*0 conditions.

We analyzed the statistical significance of the effect of each variable using a logistic regression, with listeners' responses as the dependent variable, and *f*0, duration, and glottalization as the independent variables. The responses were categorized as either "checked" (for T2 and T5

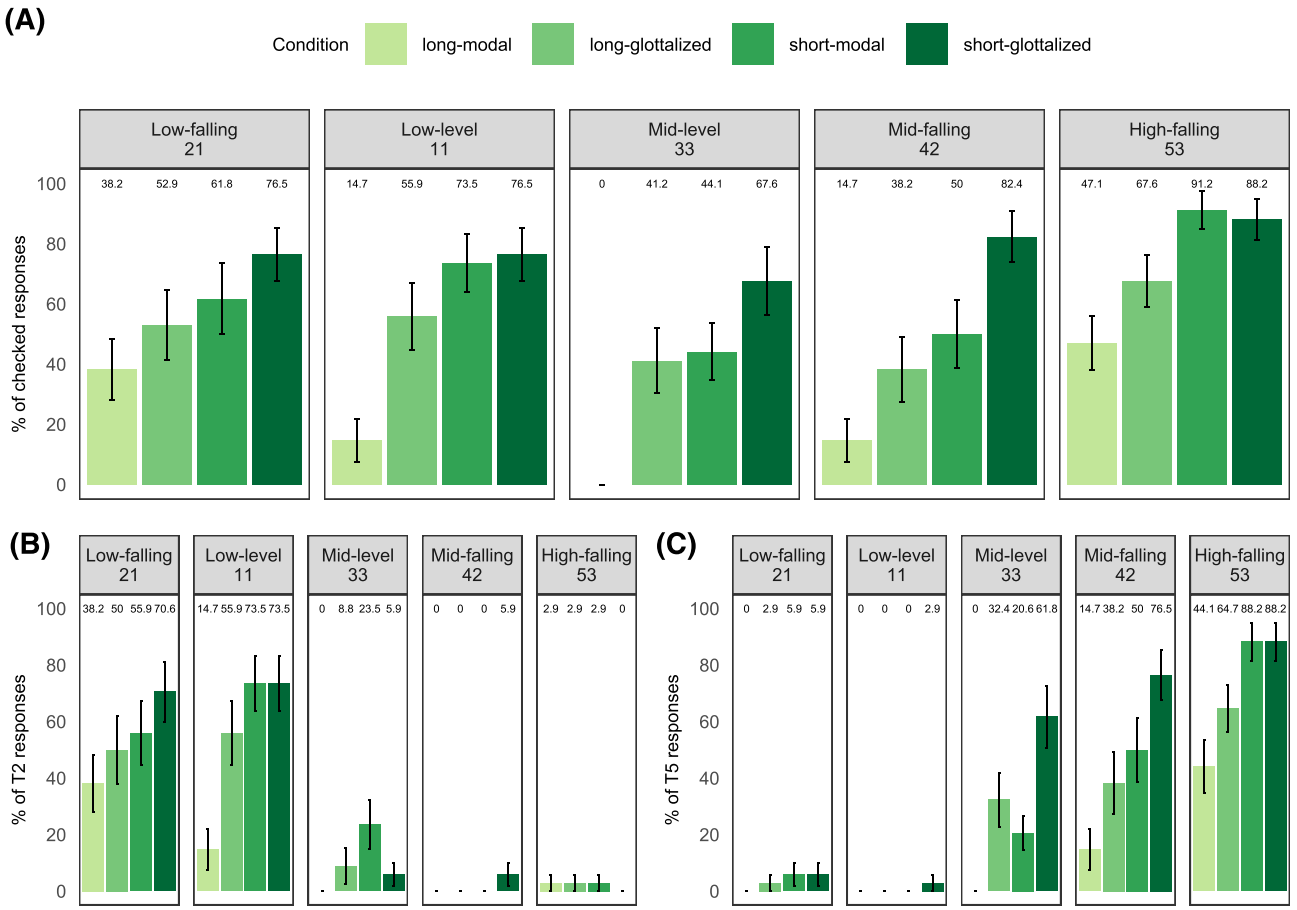


FIG. 5. (Color online) Percentage of checked tones (A), checked T2 (B), and checked T5 (C) by each condition. The error bars represent the standard errors of the percentages in each condition.

responses) or "unchecked" (for T11, T13, T33, T24, and T42 responses). We used orthogonal coding to code the categorical variables in order to test the significance of the differences between categorical groups. The five-leveled *f*0 conditions were coded by four variables: Variable 1 (*f*53_42_33_*vs*_21_11) represents the difference in responses between the average of 53 and 42 and 33 vs the average of 21 and 11; Variable 2 (*f*21_*vs*_11) represents the difference in responses between 21 and 11; Variable 3 (*f*53_*vs*_42_33) represents the difference in responses between 53 vs the average of 42 and 33; Variable 4 (*f*42_*vs*_33) represents the difference in responses between 42 vs 33. The rationale of choosing these four variables comes from our observation in the descriptive data in Figs. 5(B) and 5(C). In our data, *f*0 conditions 53, 42, and 33 elicit predominantly high-checked tone T5 while the *f*0 conditions 21 and 11 predominantly elicit low-checked T2. Thus, we divide the five *f*0 conditions into two groups, a high *f*0 contour group and a low *f*0 contour group. Variable 1 (*f*53_42_33_*vs*_21_11) demonstrates, on average, whether high *f*0 contour group elicits more checked tone than low *f*0 contour group. In other words, it tests whether high-checked tones were elicited more frequently than low-checked tones. Variables 2, 3, and 4 examine whether there are significant differences *within* the low-*f*0 group and high-*f*0 group. Variable 2 tests whether the low-falling *f*0 contour (21) of checked tone T2 elicits more checked responses than the unchecked tone contour 11; Variable 3 tests whether the high-falling *f*0 contour (53) of checked tone T5 elicits more checked responses than unchecked tone contours 42 and 33; Variable 4 tests whether there is a difference between the two unchecked tone contours 42 and 33. There are another two orthogonal coded variables: *dur* and *gl*. Variable *dur* represents the difference between short and long conditions; Variable *gl*, represents the difference between glottalized and modal conditions. Interaction between each two variables is also included in the model. The output of the logistic regression is in Table IV. We also included two sociolinguistic variables, *age* and *gender*, to test whether the listeners' age and gender affect their perception of syllables and tone. The data and data analysis script are both available at https://doi.org/10.17605/OSF.IO/94FVT.

We observed high-falling (53) *f*0 elicited more checked responses compared to mid-falling (42) and mid-level (11) *f*0 (*f*53_42_33_*vs*_21_11: $b = 2.395$, $p < 0.001$). Mid-falling (42) *f*0 elicited more checked responses than mid-level (33) *f*0 marginally (*f*42_*vs*_33: $b = 0.671$, $p = 0.058$). More checked syllables and tones were elicited by shorter duration (*dur*: $b = 2.452$, $p < 0.001$) and glottalized phonation (*gl*: $b = 1.531$, $p < 0.001$) compared with longer duration and modal phonation. We also observed significant interactions between duration and glottalization, between duration and *f*0, and between *f*0 and glottalization. First, the effect of glottalization interacts with duration. Adding glottalization to short tokens is less effective than adding to long tokens for eliciting checked responses (*dur* : *gl*: $b = -1.235$, $p = 0.007$). Second, we observe that the effect of duration also varies with *f*0. Shortening duration of low-falling *f*0 (21) is less effective than shortening low-level *f*0 (11) (*dur* : *f*21_*vs*_11: $b = -1.166$, $p = 0.071$). Last, we find significant interaction between the effect of glottalization and *f*0. Adding glottalization to high-falling (53) tokens was less effective than adding it to mid-falling (42) or mid-level (33) tones (*gl* : *f*53_*vs*_42_33; $b = -1.906$, $p = 0.003$). These results indicate that listeners employ *f*0, duration, and glottalization to identify checked syllables and tones in Xiapu Min. Meanwhile, their effects are not simply additive. The presence of one checked tone-conducive factor (e.g., short duration) diminishes the impact of introducing another conducive factor (e.g., glottalization and checked tone *f*0). The age ($p = 0.437$) and gender ($p = 0.118$) of the listeners did not have a significant effect on the probability of perceiving a checked tone.

TABLE IV. Statistical output of the logistic regression for Experiment I. Significance is indicated as: ".": for $0.05 < p < 0.1$, "*" for $0.001 < p < 0.05$, "**" for $0.001 < p < 0.01$, "***" for $p < 0.001$.

| | Estimate | Std. Error | *z* value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.110 | 0.380 | 0.290 | 0.771 |
| *dur* | 2.452 | 0.250 | 9.805 | <0.001*** |
| *gl* | 1.531 | 0.230 | 6.666 | <0.001*** |
| *f*53_42_33_*vs*_21_11 | −0.284 | 0.219 | −1.293 | 0.196 |
| *f*53_*vs*_42_33 | 2.395 | 0.341 | 7.028 | <0.001*** |
| *f*42_*vs*_33 | 0.671 | 0.354 | 1.895 | 0.058. |
| *f*21_11 | 0.133 | 0.321 | 0.414 | 0.679 |
| *gender* | −1.199 | 0.767 | −1.562 | 0.118 |
| *age* | 0.029 | 0.037 | 0.777 | 0.437 |
| *dur* : *f*53_42_33_*vs*_21_11 | 0.650 | 0.443 | 1.467 | 0.142 |
| *dur* : *f*53_*vs*_42_33 | −0.308 | 0.660 | −0.466 | 0.641 |
| *dur* : *f*42_*vs*_33 | −0.022 | 0.725 | −0.031 | 0.975 |
| *dur* : *f*21_*vs*_11 | −1.166 | 0.647 | −1.803 | 0.071. |
| *gl* : *f*53_42_33_*vs*_21_11 | 0.512 | 0.437 | 1.171 | 0.242 |
| *gl* : *f*53_*vs*_42_33 | −1.901 | 0.643 | −2.956 | 0.003** |
| *gl* : *f*42_*vs*_33 | −0.546 | 0.725 | −0.753 | 0.451 |
| *gl* : *t*21_*vs*_11 | −0.617 | 0.645 | −0.957 | 0.339 |
| *dur* : *gl* | −1.235 | 0.455 | −2.712 | 0.007** |

To evaluate the overall importance of *f*0, duration, and glottalization in eliciting checked syllables and tones in Xiapu Min, we employed Random Forest models (Ali *et al.*, 2012; Prajwala, 2015) with the checked syllable and tone responses as the independent variable and the stimuli conditions as the predictors. The dataset was randomly split into an 80% training set and a 20% test set. We trained the Random Forest model with the training set, applying hyperparameter settings from Strobl *et al.* (2007) to mitigate bias towards factors that have more categorical levels. The Random Forest model output an importance score for each predictor based on its weight in the model. We split the training set and test set and repeated the Random Forest analysis 100 times with different random seeds and found that duration was consistently ranked highest, while voice quality scored higher than *f*0 in 65 instances. Duration has the highest importance score (the mean score out of the 100 repetitions is 0.095). The importance between *f*0

https://doi.org/10.1121/10.0034785

(mean = 0.031) and voice quality (mean = 0.037) is similar to each other.

### C. Discussion

Experiment I demonstrates that all three acoustic cues —duration, *f*0, and glottalization—significantly influence Xiapu Min listeners' identification of checked syllables and tones. On average, shorter duration elicits significantly more checked responses than longer duration. In the descriptive data [Fig. 5(A)], we observe that shortening duration consistently increases the percentage of checked responses across all *f*0 and glottalization conditions. However, the effect size of duration decreases significantly when other checked response-inducing variables (i.e., checked tone *f*0 contours and glottalization) are present.

In terms of *f*0, the low-checked tone T2 is predominantly elicited by low *f*0 conditions: low-falling (21) and low-level (11) contours [Fig. 5(B)]; the high-checked tone T5 is primarily elicited by high *f*0 conditions: high-falling (53), mid-falling (42), and mid-level (33) [Fig. 5(C)]. Within the high *f*0 conditions favoring checked tone T5 responses, we also observe gradations: the checked tone T5 contour (53) elicits a significantly higher probability of checked responses than the unchecked tone contours 42 and 33. In the descriptive data [Fig. 5(A)], we observe that the high-falling contour (53) consistently elicits a higher percentage of checked responses than the mid-falling (42) and mid-level (33) contours across all duration and voice quality conditions, illustrating a robust effect of high-falling contour 53 in eliciting checked responses.

Within the low *f*0 group, we did not observe a significant difference between the low-level (11) and low-falling (21) contours in eliciting checked tones on average. However, the descriptive data [Fig. 5(A)] show that when duration is long and voice quality is modal, the checked tone T2 contour 21 elicits twice as high a percentage of checked tone responses than the low-level contour (11) (38.2% vs 14.7%). In contrast, in long and glottalized and short and modal conditions, the differences between contours 21 and 11 are reversed: the low-falling contour (21) elicits slightly *lower* percentages of checked tone than the low-level contour (11) (52.9% vs 55.9%; 61.8% vs 73.5%). In short and glottalized condition, the 21 and the 11 contours elicit the same percentage of checked tone responses (76.5%). This pattern reflects that when duration and voice quality do not favor checked responses, checked tone T2 contour (21) facilitates the perception of checked tone. However, when duration and/or voice quality already favor a checked percept, the effect of checked tone T2 *f*0 contour (21) diminishes. We believe this is likely because low-falling (21) and low-level (11) contours are very similar in the *f*0 space. When other checked tone-inducing factors are present, the effect of checked tone T2 contour (21) is diminished.

In terms of glottalization, we see that on average, adding glottalization to modal tokens significantly increases the probability of eliciting a checked tone response. In addition, the effect of glottalization interacts with duration and *f*0. As illustrated in Figs. 5(B) and 5(C), adding glottalization to short low-level (11) stimuli and short high-falling (53) stimuli does not increase the percentage of checked tone T2 and T5 responses, respectively. This suggests that when the conditions of *f*0 and duration already strongly favor checked tone perception, the listeners are not sensitive to the additional glottalization in the stimuli.

Overall, our data demonstrate that while all three cues contribute to the perception of checked tones, their effects are not simply additive. The influence of *f*0 and glottalization may disappear when other checked-tone-inducing factors are present, whereas the effect of duration remains robust across conditions. The random forest analysis corroborates these findings, assigning the highest importance score to duration in determining listeners' responses.

While the current study provides strong evidence confirming the independent effects of *f*0, duration, and glottalization, future studies can explore more fine-grained effects by introducing additional levels to each acoustic parameter. For instance, creating evenly spaced *f*0 conditions that systematically vary in both height and contour shape could demonstrate whether *f*0 height and slope have linear effects on checked tone perception. Similarly, adding more gradations to the duration variable could reveal whether the listeners are sensitive to more subtle changes in duration. In terms of glottalization, this study synthesized a "prototypical" creaky voice characterized by low f0, irregularity, and a high degree of glottal constriction (Keating *et al.*, 2015). Future studies can examine the effects of other types of creaky voice—such as vocal fry, tense, irregular phonation, or pitch doubling—on checked tone perception, as explored by Huang (2020). Additionally, while the glottalization condition in this study occupied the latter 50% of the vowel, future research can vary the proportion of glottalization within the vowel to assess whether listeners are sensitive to the relative duration of glottalization in the vowel.

## III. EXPERIMENT II: PERCEPTION OF CHECKED TONES IN SANDHI FORMS

Experiment I established that *f*0, duration, and voice quality influence the perception of the citation form of checked syllables and tones in Xiapu Min. Experiment II aims to determine whether listeners can still distinguish checked syllables and tones from unchecked ones when these factors are largely neutralized in sandhi forms.

### A. Stimuli, participants, and procedure

As presented in Table II, phonological neutralization happens among T2, T13, and T33, and among T5, T24, and T42. Thus, there are six pair-wise neutralizations: T2-T13, T2-T33, T13-T33, T5–T24, T5–T42, and T24–T42. We tested five neutralization pairs in Experiment II: T2–T33, T13–T33, T5–T24, T5–T42, and T24–T42. The contrast T2–T13 was not included due to the lack of suitable

TABLE V. Stimuli for testing the perceptual neutralization between checked and unchecked tones in sandhi forms.

| Neutralized contrast | Underlying segment | | Underlying compound | | Surface compound |
|---|---|---|---|---|---|
| T13–T33 | /to 13/ | 路 | /to 13 keu 42/ | 路口 | [to 33 keu 42] |
| | /to 33/ | 刀 | /to 33 keu 42/ | 刀口 | [to 33 keu 42] |
| T2–T33 | /tsaʔ 2/ | 杂 | /tsaʔ 2 ki 33/ | 杂技 | [tsa 33 ki 33] |
| | /tsa 33/ | 查 | /tsa 33 kaŋ33/ | 查岗 | [tsa 33 kaŋ 33] |
| T5–T24 | /$t^h$eʔ 5/ | 铁 | /$t^h$eʔ 5 pain 42/ | 铁板 | [$t^h$e 55 pain 42] |
| | /$t^h$e 24/ | 替 | /$t^h$e 24 po 42/ | 替补 | [$t^h$e 55 po 42] |
| T5–T42 | /θiʔ 5/ | 湿 | /θiʔ 5 ti 24/ | 湿地 | [θi 55 ti 24] |
| | /θi 42/ | 死 | /θi 42 tsui 24/ | 死罪 | [θi 55 tsui 24] |
| T24–T42 | /ka 24/ | 价 | /ka 24 kai 42/ | 价格 | [ka 55 kai 42] |
| | /ka 42/ | 假 | /ka 42 θe 42/ | 假设 | [ka 55 θe 42] |

minimal pairs. The target words are presented in Table V. In each compound word pair, the target syllables are the first syllable, with identical segmental structure and surface tone but different underlying tones. The second syllables of the compounds are identical in tone, onset place of articulation, and sonority to minimize their coarticulatory effects on the target syllable. An exception is the contrast for T24–T42, where the onset of the second syllable's onset differs in the place of articulation (/k/ vs /θ/). We allowed this exception because the formant values in the target syllables /ka 24/ and /ka 42/ were similar in the tokens selected for stimuli.

The stimuli were developed from the natural productions by two female Xiapu Min speakers. We selected two natural tokens for each target word in Table V as the base tokens.[4] In total, there are 40 tokens for the compound identification task (five contrasts * two words per contrast * two speakers * two tokens). Because the second syllables of the compounds differed segmentally (except in the T13–T33 contrast), we replaced the second syllables with 342 ms of pink noise generated in Audacity (Audacity Team, 2022). This noise was adjusted to 60% of the amplitude of the first segment. After the noise concatenation, the amplitude of the entire compound was normalized to 70 dB. Additionally, a 50 ms silence was padded at the beginning and the end of each compound. Sample stimuli tokens are in Fig. 6.

We measured the *f*0, duration, H1*–H2*, H1*–A3*, and Harmonic-to-Noise Ratio (HNR) of the stimuli using VoiceSauce (Shue *et al.*, 2011). VoiceSauce output a value of each acoustic measure for every millisecond of the vowel in the words. We averaged the values of each acoustic measure over the whole duration of each word. Then we calculated the mean value of these acoustic measures for each tone category in each neutralization pair. The data are presented in Fig. 7 and Table VI. *F*0 is transformed into semitones, with the mean value of *f*0 of all tokens in the stimuli as the baseline, because semi-tone is closer to listener's perception of pitch. The descriptive statistics of the stimuli show that the *f*0, H1*–H2*, and H1*–A3* of the tones within each neutralization pair are similar. In terms of duration, T2 and T13 have a shorter duration than T33. T5 has a shorter duration than T24 and T42. In terms of HNR, T5 has a lower HNR than T24. The acoustics of the same tone differ in

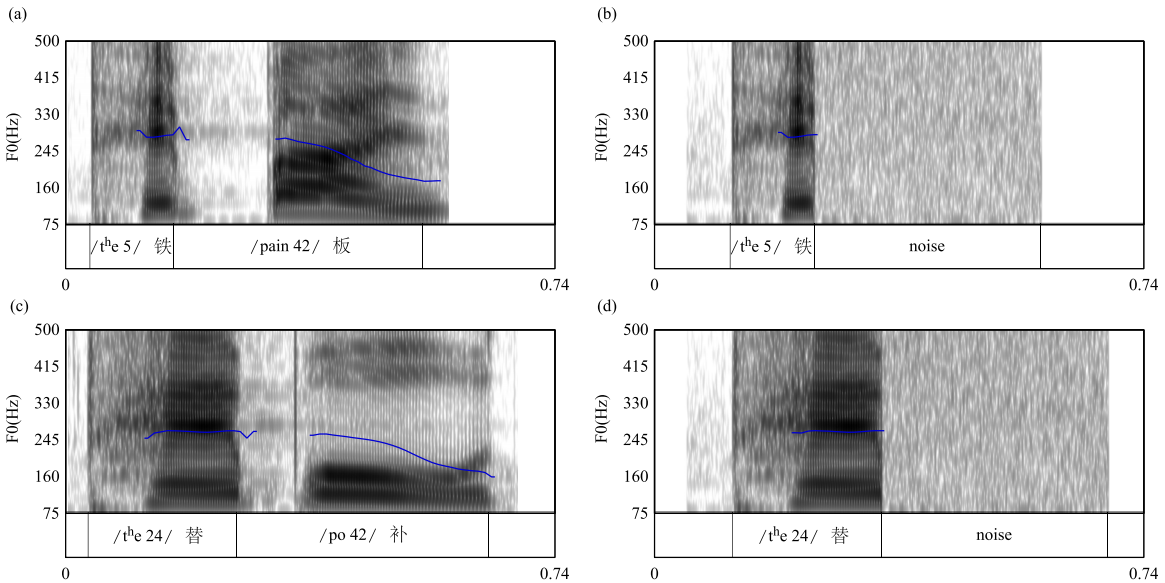


FIG. 6. (Color online) Sample tokens for T5–T24 contrast in Experiment II stimuli. Here, (a) and (c) are the original tokens; (b) and (d) are noise-masked tokens.

https://doi.org/10.1121/10.0034785

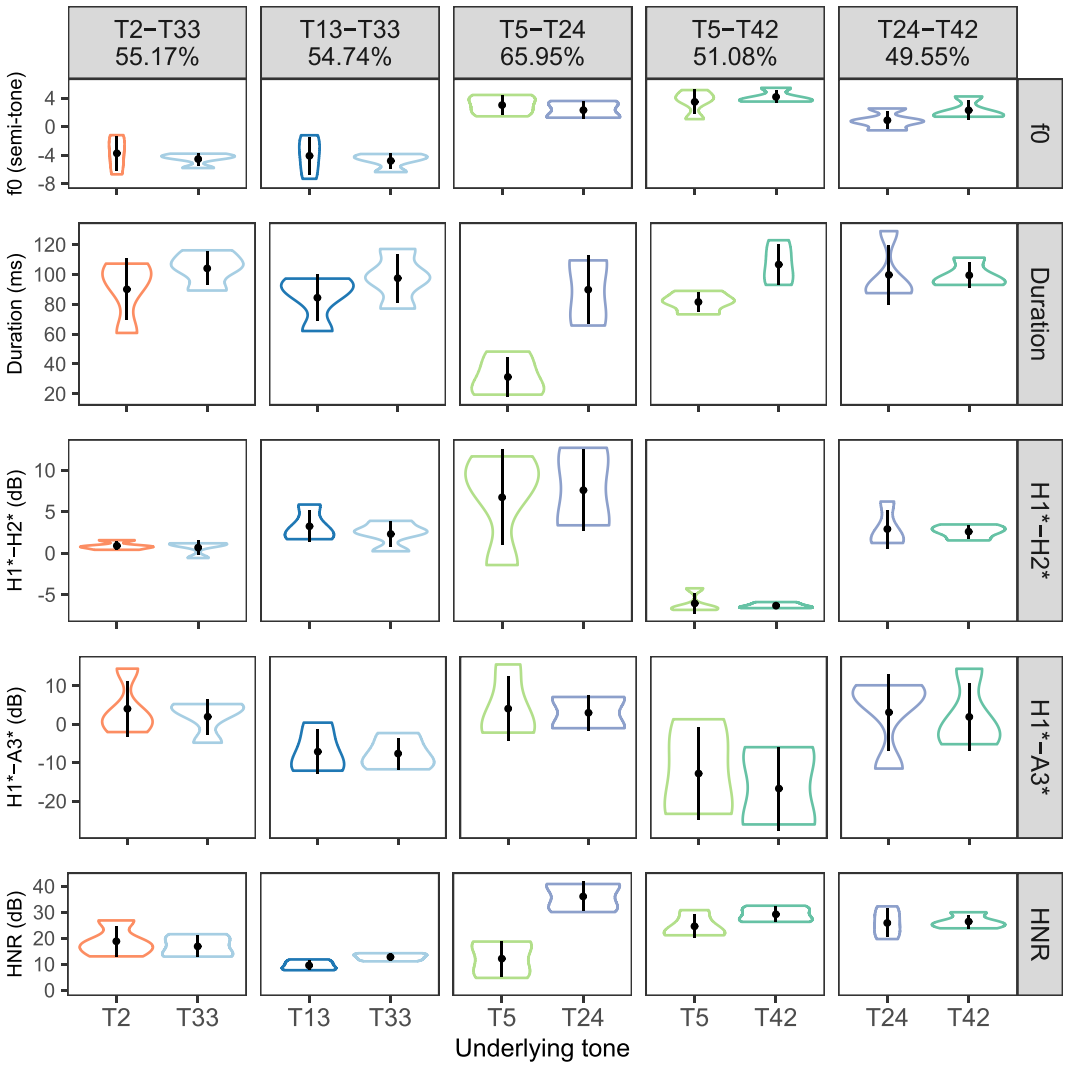


FIG. 7. (Color online) Acoustics of the stimuli in Experiment II. The dot represents the mean. The upper and lower whiskers extend to the largest and smallest values in the data, but not further than the 1.5 times the box length.

different pairs because they have different onsets and are influenced by the coarticulation of the onset.

The participants and the equipment were the same as in Experiment I. The 40 stimuli tokens, divided into two blocks by the speaker's identity, were presented in a randomized order for each participant. Before the start of the test trials, four practice trials with the same task structure but different stimuli were conducted. The participants engaged in a compound word identification task, being informed that the second word of each compound was masked by noise. Their task was to identify the compound based on the first word's recording. Each trial was a forced choice between the two

TABLE VI. Mean *f*0, H1*–H2*, HNR, and duration of the stimuli of the neutralization experiment.

| Contrast | Underlying tone | *F*0 (Hz) | *F*0 (semitone) | H1*–H2* (dB) | H1*–A3* (dB) | HNR | Duration (ms) |
|---|---|---|---|---|---|---|---|
| T2–T33 | T2 | 199.594 | –3.761 | 0.891 | 4.005 | 18.802 | 90.007 |
| | T33 | 189.324 | –4.569 | 0.668 | 1.931 | 16.849 | 104.050 |
| T13–T33 | T13 | 195.862 | –4.109 | 3.238 | –7.153 | 9.555 | 84.390 |
| | T33 | 186.645 | –4.821 | 2.304 | –7.646 | 12.762 | 97.458 |
| T5–T24 | T5 | 293.499 | 3.008 | 6.727 | 4.039 | 12.133 | 31.262 |
| | T24 | 281.536 | 2.296 | 7.584 | 2.967 | 36.057 | 89.758 |
| T5–T42 | T5 | 301.742 | 3.459 | –6.057 | –12.835 | 24.578 | 81.522 |
| | T42 | 313.442 | 4.162 | –6.332 | –16.744 | 29.182 | 106.626 |
| T24–T42 | T24 | 259.697 | 0.876 | 2.904 | 3.080 | 25.864 | 99.687 |
| | T42 | 281.533 | 2.278 | 2.591 | 1.901 | 26.393 | 99.427 |

compound options within the neutralization pairs in Table V. The options were presented in Mandarin characters on the computer screen with their order randomized for each trial and participant. The participants can listen to each token as many times as desired. In total, 1151 data points (40 tokens * 29 participants − 9 errors) were included for data analysis.

## B. Results

The percentages of responses for each underlying tone within each neutralization pair are illustrated in Fig. 8. Using binomial tests, we found that only the neutralization pair of T5–T24 had a significant above-chance accuracy rate (65.95%). All the other four pairs have an accuracy at chance (T2–T33: 55.17%; T13–T33: 54.74%; T5–T42: 51.08%; T24–T42: 49.55%). For the remaining pairs, we observe a bias towards one of the two options: a preference towards T2 in the T2–T33 pair; T13 in the T13–T33 pair; T5 in the T5–T42 pair, and T24 in the T24–T42 pair.

To investigate the acoustic parameters used by listeners to identify checked syllables and tones in sandhi forms, we correlated their responses (checked syllable/tone or not) with the acoustic parameters of *f*0, duration, $H1^*$–$H2^*$, and Harmonic-to-Noise Ratio (HNR). $H1^*$–$H2^*$ and HNR are two commonly-used acoustic measures for voice quality, correlated with the degree of vocal folds constriction and noisiness (Garellek, 2019; Klatt and Klatt, 1990). $H1^*$–$A3^*$ is another measure of spectral tilt. Lower $H1^*$–$A3^*$ is correlated with a higher degree of vocal folds constriction. It has been found to differentiate checked from unchecked tones in Taiwanese Min (Pan, 2017; Pan *et al.*, 2016). We used *f*0 in semi-tones for a closer representation of pitch perception. The left-skewed distribution of the duration was transformed into a normal distribution by subtracting each duration from the maximum duration of all stimuli and then adding one, followed by a square-root transformation. F0, duration, $H1^*$–$H2^*$, $H1^*$–$A3^*$, and HNR were transformed into z-scores for scale uniformity. Given that the data size within each contrast pair is relatively small, we conducted model comparisons between complex models including these four predictors and simpler ones, dropping one predictor at a time using analysis of variance (ANOVA). Predictors that did not significantly improve the model accuracy were dropped.

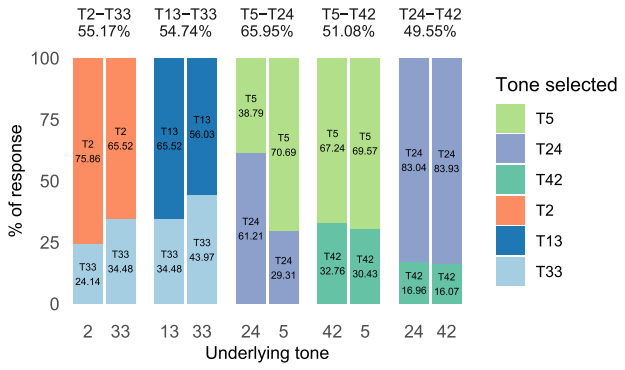


FIG. 8. (Color online) Responses of compound identification task. The percentage number under the neutralization pair text is the identification accuracy of each pair. The percentage in the stacked bar plot is the percentage of responses of each tone in each neutralization pair.

For the T2–T33 contrast, we compared the complex model $T2\_or\_not \sim f0 + duration + H1^* - H2^* + H1^* - A3^* + HNR + (1|subject)$ with a simpler model $T2\_or\_not \sim f0 + duration + (1|subject)$, finding that the complex model does not explain the variances significantly better than the simpler model ($p = 0.346$). Thus, we judged the predictors of $H1^*$–$H2^*$, $H1^*$–$A3^*$, and HNR did not improve the model significantly and dropped them from the model. Lower *f*0 ($p = 0.016$) and shorter duration ($p = 0.010$) led to a higher likelihood of checked T2 responses.

For the T5–T24 contrast, we compared the complex model $T5\_or\_not \sim f0 + duration + H1^* - H2^* + H1^* - A3^* + HNR + (1|subject)$ with a simpler model $T5\_or\_not \sim duration + (1|subject)$, finding that the complex model does not explain the variances significantly better than the simpler model ($p = 0.822$). This indicates the predictors of *f*0, $H1^*$–$H2^*$, $H1^*$–$A3^*$, and HNR did not improve the model significantly. As a result, we dropped these four predictors from the model. A shorter duration ($p < 0.001$) led to a higher likelihood of checked T5 responses.

For the T5–T42 contrast, we compared the complex model $T5\_or\_not \sim f0 + duration + H1^* - H2^* + H1^* - A3^* + HNR + (1|subject)$ with a model that only predicted the intercept—$T5\_or\_not \sim 1 + (1|subject)$. We found that the model with all the predictors does not explain the variances in the data significantly than a model that only predicts intercept ($p = 0.389$). This means that none of the acoustic variables had a significant effect on listeners' responses. There is a high bias toward the word "湿地" /θiʔ 5 ti 24/ in the contrast. This observed bias in their responses is not attributable to the acoustic variables included in the model but has to be a lexical bias.

## C. Discussion

In Experiment II, we tested whether the phonologically neutralized checked and unchecked syllables and tones were still perceptually discernible. Using natural stimuli, we found that listeners of Xiapu Min were only able to identify the sandhi forms of T5 and T24 with an above-chance accuracy. There was a bias towards one specific tonal category in the other contrasts attested (T2 in T2–T33; T13 in T13–T33; T5 in T5–T42, and T24 in T24–T42). We hypothesize that these biases are due to the imbalance in word frequency between the compounds in each pair. Several participants reported that the word "假设" /ka 42 θe 42/ in the T24–T42 pair was rarely used in daily speech. The alternative hypothesis is the listeners are biased towards T2, T13, T5, and T24 specifically. The ambiguity is a limitation in the current study, stemming from the absence of a comprehensive Xiapu Min corpus for word frequency reference. Future studies could address this by conducting a word frequency survey with Xiapu Min speakers. With the word-frequency survey results, we can assign more frequent

https://doi.org/10.1121/10.0034785

words to the less favored tone options in the current experiment, to determine whether the bias observed in the current experiment is due to word frequency or specific tone.

Regarding the acoustic correlates of checked syllables and tones in sandhi forms, as anticipated, shorter duration was a key factor in eliciting checked tones T2 and T5. Checked T2, against T33, was also elicited by lower *f*0. This suggests that listeners still rely on the acoustic cues of the underlying tones when processing the sandhi forms in perception. Although in sandhi forms, the checked syllables and tones are largely neutralized with unchecked syllables and tones phonetically, listeners appear to associate the acoustic characteristics of checked syllables and tones in citation forms with their sandhi forms.

The significant effect of duration on checked syllable and tone response explains why T5–T24 is the only pair that received an above-chance identification accuracy. The difference in duration between T5 and T24 in the natural stimuli is larger than the difference between T2 and T33, and T5 and T42 (T24 – T5 = 58 ms; T33 – T2 = 14 ms; T42 – T5 = 25 ms) (as shown in Table VI). Future studies can systematically manipulate the duration in evenly-spaced steps and examine whether a decrease in duration leads to more checked tone responses for all checked–unchecked tone neutralization pairs.

## IV. GENERAL DISCUSSION

What cues do listeners use to identify a checked syllable and tone? This study answered this question by testing Xiapu Min listeners' perception of Xiapu Min checked syllables and tones in citation and sandhi forms. In citation forms, duration, glottalization, and *f*0 all play a part, though duration has a more salient role in it. In sandhi forms, listeners still attend to duration and *f*0 cues to identify a checked tone, despite their phonetic differences being largely neutralized in the surface form compared with the citation form.

What do the results of this study speak for the nature of checked syllables and tone? Chai and Ye (2022) described the checked syllables and tones in Xiapu Min as having short duration and vowel-final glottalization. One of the questions raised by Chai and Ye (2022) was whether the short duration is a by-product of the closed syllable V?, or an independent articulatory target. Chai and Ye (2022) argued that the short duration of the checked syllable and tone is an independent articulatory goal because when the glottal stop coda in checked syllables was lost in sandhi forms, the vowel in checked syllables was still shorter than the vowel in unchecked syllables. Our results argue for the independence of short duration from closed syllable structure from a perceptual perceptive. In citation forms, duration has a higher weight than glottalization in checked tones elicitation. In sandhi forms, duration has a significant correlation with T2 and T5 responses while voice quality does not. Short duration can elicit checked syllable and tone perception even when a glottalization cue is absent.

How is Xiapu Min's perception results in comparison to the phonetically "checked-like" constituents in other languages? Xiapu Min checked syllables and tones share similarities with Shanghainese Wu (Gao and Kuang, 2022) and Yangzhou Jianghuai Mandarin (Tang, 2017; Tang and Li, 2018), such that listeners are sensitive to the short duration cue, independent of glottalization. If the duration is short enough, the listeners can still perceive a checked syllable and tone in the absence of glottalization. However, Xiapu Min also differs from Yangzhou Jianghuai Mandarin in the effect of *f*0. While the changes in *f*0 does not affect listeners' perception of the checked syllable and tone in Yangzhou Jianghuai Mandarin, listeners of Xiapu Min take advantage of the checked tone *f*0 contours to identify checked syllables and tones. Xiapu Min checked syllables and tones also behave similarly to Taiwanese Min low checked tone T3, such that shortening of duration leads to checked tone responses, but diverges from Taiwanese Min checked tone T5, for which duration is not an effective cue (Zhang and Lu, 2023). The divergence between Taiwanese Min and Xiapu Min checked tones is likely due to the difference in the checked tone development in the two languages. Checked tones have a tendency of "opening" in Chinese languages. This sound change process results in checked syllables losing their syllable codas and becoming open syllables; and checked tones losing their distinct contour and merging with the tonal melody of unchecked tones. Taiwanese Min is in a more advanced stage of checked syllable opening compared with Xiapu Min. As suggested by Zhang and Lu (2023), Taiwanese Min high-checked tone,

TABLE VII. *F*0 of the modal tokens in the stimuli. The *f*0 values are assigned to the relative time point in the vowel. P1 to P9 refer to the first to the ninth *f*0 point, located at the 1/10 to 9/10 of the duration of the vowel.

| *f*0 point | Time point | Low-falling (21) | High-falling (53) | Low-level (11) | Mid-falling (42) | Mid-level (33) |
|---|---|---|---|---|---|---|
| P1 | 1/10 | 167 | 294 | 150 | 230 | 204 |
| P2 | 2/10 | 165 | 303 | 144 | 231 | 203 |
| P3 | 3/10 | 163 | 299 | 140 | 226 | 200 |
| P4 | 4/10 | 160 | 289 | 136 | 218 | 198 |
| P5 | 5/10 | 158 | 275 | 134 | 208 | 197 |
| P6 | 6/10 | 157 | 261 | 135 | 198 | 196 |
| P7 | 7/10 | 155 | 244 | 136 | 191 | 195 |
| P8 | 8/10 | 150 | 231 | 139 | 186 | 194 |
| P9 | 9/10 | 141 | 219 | 140 | 173 | 193 |

TABLE VIII. *F*0 of the glottalized tokens in the stimuli. The *f*0 values are assigned to the relative time point in the vowel. P1 to P12 refers to the first to the 12th *f*0 point. P1 to P5 are at the 1/10 to the 5/10 of the duration of the vowel. P6 to P12 are at the 9/16 to the 15/16 of the duration of the vowel.

| *f*0 point | Time point | Low-falling (21) | High-falling (53) | Low-level (11) | Mid-falling (42) | Mid-level (33) |
|---|---|---|---|---|---|---|
| P1 | 1/10 | 167 | 294 | 150 | 230 | 204 |
| P2 | 2/10 | 165 | 303 | 144 | 231 | 203 |
| P3 | 3/10 | 163 | 299 | 140 | 226 | 200 |
| P4 | 4/10 | 160 | 289 | 136 | 218 | 198 |
| P5 | 5/10 | 158 | 275 | 134 | 208 | 197 |
| P6 | 9/16 | 135 | 135 | 135 | 135 | 135 |
| P7 | 10/16 | 95 | 95 | 95 | 95 | 95 |
| P8 | 11/16 | 110 | 110 | 110 | 110 | 110 |
| P9 | 12/16 | 87 | 87 | 87 | 87 | 87 |
| P10 | 13/16 | 104 | 104 | 104 | 104 | 104 |
| P11 | 14/16 | 84 | 84 | 84 | 84 | 84 |
| P12 | 15/16 | 91 | 91 | 91 | 91 | 91 |

T5, is undergoing vowel lengthening in production. Their perception test results illustrate that the listeners are adapted to the vowel lengthening also in perception. In contrast, the production study of Xiapu Min by Chai and Ye (2022) maintains that the two checked tones are still shorter than the unchecked tones in Xiapu Min. Vowel lengthening is not observed in Xiapu Min checked tones. It is reasonable that listeners still use short duration for both checked tones in Xiapu Min.

Xiapu Min's checked syllables and tones also resemble White Hmong's creaky tones (Garellek *et al.*, 2013), such that *f*0 and duration both play an important part in their identification. However, they also differ in the effect of glottalization. Vowel-final flottalization facilitates the identification of checked syllables and tones for Xiapu Min listeners, but not for White Hmong listeners. The divergence is in accordance with the production of White Hmong. Garellek and Esposito (2021) found that in production, the low-falling creaky tone (-m) was not consistently creakier than the low modal tone (-s). Glottalization is not a consistent cue for the creaky tone in White Hmong in production. In contrast, Chai and Ye (2022) consistently observed glottalization at vowel-final position of the checked syllables and tones in Xiapu Min. It is reasonable that White Hmong listeners ignore glottalization cue while Xiapu Min listeners rely on that cue. The comparisons among the checked constituents in Xiapu Min, Taiwanese Min, and White Hmong illustrate a robust correlation between checked constituent production and the cues used for its perception.

## ACKNOWLEDGMENTS

Thanks to Professor Marc Garellek, Professor Gabriela Caballero, Professor Sarah Creel, Professor Sharon Rose, and Professor Will Styler for their insightful guidance and feedback on the study design, data interpretation, and manuscript. We also thank Professor Gabriela Caballero for funding this field research. We greatly appreciate the constructive suggestions provided by the four reviewers, and the valuable feedback from the audience at PhonCo at UCSD and ASA Fall 2021.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Ethical Approval

Consent forms have been obtained from all participants in this study. The study was approved by the Institutional Review Board of the University of California San Diego (Protocol Code No. 190,550; date of approval, April 29, 2019).

## DATA AVAILABILITY

The data used for analysis and the programming scripts are available in Open Science Foundation at https://doi.org/10.17605/OSF.IO/94FVT.

## APPENDIX

See Tables VII and VIII for the parameter values for creating the modal tokens and glottalized tokens, respectively, in the stimuli for Experiment I.

[1]In this study, we follow the definition of glottalization in Garellek (2013), which refers to "the articulatory and acoustic effects on targets of either glottal stop or laryngealization." The term glottalization is chosen to describe the phonetic effect of the glottal stop coda of a checked syllable on the syllable nucleus.

[2]We used only female voices in the stimuli based on evidence that their acoustic properties are representative of Xiapu Min tone production across genders. Analyzing the production data from Chai and Ye (2022), which comprised 629 unique words produced by four male and four female speakers (with at least 55 words per tone), we conducted a mixed-effects model with duration as the dependent variable and gender and tone as independent variables. The analysis revealed that gender was not a significant predictor of duration, indicating no significant difference in tone duration between male and female speakers. Regarding *f*0, the *f*0 contour shapes were consistent across genders. The male speakers exhibited a lower overall *f*0 and a slightly compressed *f*0 range. Therefore, we hypothesize that similar perceptual results would be obtained using male voices; however, this requires empirical validation in future research.

[3]We did not use the production by Speaker #2 as the base token for resynthesis because her recording contained background noise.

[4]One exception is that, for the word /t$^{h}$e 24 po 42/ 替补 "substitute" in the T5–T24 contrast, one speaker only produced a single token. In order to ensure a balance of the data, I duplicated that single token in the stimuli.

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (**2012**). "Random forests and decision trees," Int. J. Comput. Sci. Issues **9**(5), 272–278.
Audacity Team (**2022**). "Audacity (version 3.1.3) [computer program]" http://audacityteam.org/ (Last viewed 25 June 2022).
Bauer, R. S., and Matthews, S. (**2017**). "Cantonese," in *The Sino-Tibetan Languages*, 2nd ed., edited by G. Thurgood and R. J. LaPolla (Routledge, London), pp. 169–184.
Braver, A. (**2014**). "Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping," Lingua **152**, 24–44.
Brunelle, M. (**2009**). "Tone perception in Northern and Southern Vietnamese," J. Phon. **37**(1), 79–96.
Brunelle, M., and Finkeldey, J. (**2011**). "Tone perception in Sgaw Karen," in *Proceedings of the ICPhS XVII 2011*, August 17–21, Hong Kong, pp. 372–375.
Chai, Y. (**2022**). "Phonetics and phonology of checked phonation, syllables, and tones," Ph.D. thesis, University of California San Diego, San Diego, CA.
Chai, Y., and Ye, S. (**2022**). "Checked syllables, checked tones, and tone sandhi in Xiapu Min," Languages **7**(1), 47.
Chen, Y., and Gussenhoven, C. (**2015**). "Shanghai Chinese," J. Int. Phonetic Assoc. **45**(3), 321–337.
Chien, Y.-F., and Jongman, A. (**2019**). "Tonal neutralization of Taiwanese checked and smooth syllables: An acoustic study," Lang. Speech **62**(3), 452–474.
Dinnsen, D. A. (**1985**). "A re-examination of phonological neutralization," J. Ling. **21**(2), 265–279.
Dong, H. (**2020**). *A History of the Chinese Language*, 2nd ed. (Routledge, New York).
Frazier, M. (**2009**). "The production and perception of pitch and glottalization in Yucatec Maya," Ph.D. thesis, The University of North Carolina at Chapel Hill, Chapel Hill, NC.
Gao, X., and Kuang, J. (**2022**). "Phonation variation as a function of checked syllables and prosodic boundaries," Languages **7**(3), 171.
Gao, Y. (**2004**). "Shengdiao ganzhi yanjiu 声调感知研究" ("Study on tone perception"), Ph.D. Dissertation, Shanghai Normal University, Shanghai, China.
Garellek, M. (**2013**). "Production and perception of glottal stops," Ph.D. thesis, University of California, Los Angeles, Los Angeles, CA.
Garellek, M. (**2019**). "The phonetics of voice," in *Routledge Handbook of Phonetics*, edited by W. Katz and P. Assmann (Routledge, Oxford, UK), pp. 75–106.
Garellek, M., and Esposito, C. M. (**2021**). "Phonetics of White Hmong vowel and tonal contrasts," J. Int. Phonetic Assoc. **53**, 213–232.
Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (**2013**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**(2), 1078–1089.
Gruber, J. F. (**2011**). "An articulatory, acoustic, and auditory study of Burmese tone," Ph.D. thesis, Georgetown University, Washington, DC.
Haudricourt, A. G. (**1954**). "De l'origine des tons en Vietnamien" ("The origin of tones in Vietnamese"), J. Asiatique **242**, 69–82.
Huang, Y. (**2020**). "Different attributes of creaky voice distinctly affect Mandarin tonal perception," J. Acoust. Soc. Am. **147**(3), 1441–1458.
Keating, P., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**(2), 820–857.
Kuo, C.-H. (**2013**). "Perception and acoustic correlates of the Taiwanese tone sandhi group," Ph.D. thesis, University of California, Los Angeles, CA.
Pan, H.-H. (**2017**). "Glottalization of Taiwan Min checked tones," J. Int. Phonetic Assoc. **47**(1), 37–63.
Pan, H.-H., Huang, H.-T., and Lyu, S.-R. (**2016**). "Coda stop and Taiwan Min checked Tone sound changes," in *Proceedings of Interspeech 2016*, September 8–12, San Francisco, CA, pp. 1011–1015.
Peña, J. M. (**2022**). "Stød timing and domain in Danish," Languages **7**(1), 50.
Peña, J. M. (**2023**). "Effects of fundamental frequency and harmonics-to-noise ratio on the perception of Danish laryngealized phonation," in *Proceedings of the 20th ICPhS 2023*, Prague, Czech Republic, August 7–11, pp. 1736–1740.
Prajwala, T. (**2015**). "A comparative study on decision tree and random forest using R tool," Int. J. Adv. Res. Comput. Commun. Eng. **4**, 196–199.
Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., and Grice, M. (**2014**). "Assessing incomplete neutralization of final devoicing in German," J. Phon. **43**, 11–25.
Shao, D. (**2012**). "Jiyu EGG de meixian, fuzhou, changsha fangyan shengdiao shiyan yanjiu 基于 EGG 的梅县、福州、长沙方言声调实验研究" ("Acoustic experimental analysis of Meixian, Fuzhou, and Changsha dialect based on EGG data"), Master's thesis, Nanjing Normal University, Nanjing, China.
Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. M. (**2011**). "VoiceSauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetic Science*, August 17–21, Hong Kong, pp. 1846–1849.
Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (**2007**). "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinf. **8**(1), 25.
Tang, Z. (**2014**). "Wanshu jianghuai guanhua rusheng shiyan yanjiu 皖属江淮官话入声 实验研究" ("Experimental analysis of checked tone in Anhui Jianghuai Mandarin"), Master's thesis, Nanjing Normal University, Nanjing, China.
Tang, Z. (**2017**). "Jianghuai guanhua rusheng shengxue-shengli-ganzhi shiyan yanjiu 江淮官话入声声学-生理-感知实验研究" ("Acoustic, articulatory, and perceptual experimental studies of jianghuai mandarin rushing"), Ph.D. thesis, Nanjing Normal University, Nanjing, China.
Tang, Z., and Li, S. (**2018**). "Perceptual studies on the distinctive features of Rusheng入声 tone in Yangzhou dialect, Jiangsu province 扬州方言入声区别性特征的感知研究," Dialect方言 **4**, 411–420.
Winter, B., and Roettger, T. (**2011**). "The nature of incomplete neutralization in German: Implications for laboratory phonology," Grazer Linguistische Studien **76**, 55–74.
Wu, B. (**2018**). "An acoustic analysis of vowels in checked syllables in Chinese," Chin. J. Acoust. **37**(4), 491–502, accessible outside China at https://link.oversea.cnki.net/doi/10.15949/j.cnki.0217-9776.2018.04.009; accessible inside China at https://link.cnki.net/doi/10.15949/j.cnki.0217-9776.2018.04.009.
Xiapu County Bureau of Statistics (**2021**). "Xiapu xian diqici quanguo renkou pucha gongbao" ("Xiapu County seventh national census report"), http://www.xiapu.gov.cn/zwgk/zfxxgkzdgz/tjxx/tjgb/202106/t20210628_1491098.htm (Last viewed 14 November 2024).
Zhang, W., and Lu, Y.-A. (**2023**). "The role of duration in the perception of checked versus unchecked tones in Taiwanese Southern Min," in *Proceedings of 20th International Congress of Phonetics Science*, edited by R. Skarnitzl and J. Volín (GUARANT International, Prague, Czech Republic), pp. 226–230.
Zhengzhang, S. (**2003**). *Shanggu Yinxi* 上古音系 *(Old Chinese Phonology)*, 1st ed. (Shanghai Educational Publishing House, Shanghai, China).