Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

The prosodic characteristics of Standard Chinese rhetorical questions in naturalistic settings

Shuwen Chen[1], *Qingke Sun*[1], *Yue Huang*[1], *Yingyi Luo*[1]

[1]Institute of Linguistics, Chinese Academy of Social Sciences
chensw@cass.org.cn, Hekuang_e2@163.com, huangyue_ucr2021@163.com, luoyingyi@cass.org.cn

Abstract

The current study investigated the prosodic features of rhetorical questions in Standard Chinese within naturalistic settings, utilizing data collected from 103 native Mandarin speakers. String-identical information-seeking questions (ISQs) and rhetorical questions (RQs) were recorded through an online platform, ensuring a diverse and representative sample. The results revealed that, when freely reading aloud to the phone, speakers tended to convey rhetorical meaning by either shifting prominence toward or enhancing the prominence of the verb or modal verb. Acoustically, the prominent verbs were characterized by higher pitch and longer duration. In other positions, phonetic markers varied depending on the sentence structure. Our large-sample findings highlight the interface of prosody and syntax in conveying communicative intention. This research also addresses a critical gap by moving beyond laboratory-based data to examine speech phenomena in more ecologically valid contexts.

Index Terms: rhetorical questions, natural setting, prosodic characteristics, Standard Chinese

1. Introduction

Questions can be used to serve distinct communicative functions: they may be used to seek information from the interlocutor (referred to as information-seeking questions, or ISQs) or to make assertions, express dissatisfaction, or convey rhetorical meaning (referred to as rhetorical questions, or RQs). Notably, prosody plays a critical role in differentiating communicative meanings in speech, particularly when the compositional semantics of the utterances are identical [1, 2, 4, 6]. While the importance of prosody is well established, the specific prosodic characteristics distinguishing RQs from ISQs remain controversial. Previous studies have yielded inconsistent findings, largely due to insufficient attention to two key factors: the influence of syntactic structure and individual variability. To address these limitations, the current study examines the prosodic marking of rhetorical expression in Standard Chinese in a more ecologically valid context, by using daily-used polar-questions with different structures and recruiting a larger and more diverse sample of speakers.

Previous research across languages has observed phonetic differences between ISQs and RQs, despite the diverse acoustic profiles involved. With respect to pitch, evidence from stress languages reveal that RQs and ISQs differ in the type and the position of nuclear pitch, as well as the types of boundary tones [1]. Japanese, a pitch accent language, marks rhetorical meanings with sentence-initial *f0* lowering, sentence-final *f0* raising [2]. For tonal languages, the use of pitch to mark rhetorical meanings is more complex due to the dual role of pitch in distinguishing lexical tones and conveying intonation. A study focusing on the sentence-final particle (SFP) in Cantonese *wh*-questions found an increased *f0* rising for positive RQs (where the answer is known to both parties and specific) compared to ISQs, but not for negative RQs (where the answer is supposed to be "nobody"), which yet had a numerically lower pitch [3]. In Standard Chinese, a relatively consistent finding is that RQs tend to end with a falling contour [4, 5]. However, specific prosodic patterns remain controversial. Zahner-Ritter et al.[4] observed that both polar questions started with *yǒurén* ("anyone") and *wh*-questions with *shéi* ("who") are produced with an overall lower *f0* for RQs. However, Lo and Kiss [5] found that RQs started with *yǒushéi* ("who") are marked by a higher *f0* at the beginning and a lower *f0* in the SFP. In addition, duration has consistently been reported to be globally longer for RQs across various language types (stress languages: English [1] and German [6]; pitch accent languages: Japanese [2]; tonal languages: Cantonese [3] and Standard Chinese [4]). Specifically, evidence from Standard Chinese highlighted that RQs are globally lengthened compared to ISQs, except at the SFP [4, 5].

To interpret the convergence and discrepancies among the findings, we consider that two issues should be revisited. Firstly, questions that use interrogative words followed by a Verb and an Object Noun are intensively concerned [1, 3, 4, 5], but such constructions inherently have their focused element placed on the preceding part of the sentence. Given the apparent impact of grammatical structure on intonation, diverse syntactic structure should be employed to test the generalization of prosodic characteristics for RQs. Therefore, in the current study, two types of polar questions, containing no interrogative words, were utilized. Secondly, most studies have relied on speech data collected from a limited number of participants in highly controlled laboratory settings, where the individual difference and the lab-induced emotional states may confound the results. Therefore, for enhanced ecological validity, an online data collection method was adopted in the current study, wherein participants used their own smartphones for sound recording. This approach has been shown to produce speech data comparable to laboratory recordings for pitch-related analyses [7]. It not only reduces the likelihood of performative tones by simulating real-life voice messaging but also facilitates participation from a larger and more diverse speaker sample.

In summary, this research aims to examine the prosodic characteristics of ISQs and RQs in Standard Chinese speakers, with different sentence structures examined in a more naturalistic setting context and with a larger number of participants.

10.21437/Interspeech.2025-863

2. Method

2.1. Participants

A total of 116 native Mandarin speakers were recruited. To ensure the representativeness of the speakers and include diverse dialectal backgrounds, we did not place any restrictions on variables such as age, gender, geographical region, educational background, or other demographic factors. The data from thirteen participants were later excluded due to either not following instructions or suboptimal recording quality. Consequently, speech data from 103 participants (24 males and 79 females) were used in the analysis.

2.2. Materials

The materials analyzed in the current study constitute a subset of recordings from a larger production project involving a set of 84 polar questions ending with the neutral-toned interrogative SFP *ma* ("吗"). Two types of polar questions are targeted to our interest (Table 1): those with the predicate/modal verb being followed by a content word which acts as a complement (verb-complement, VC) and those without (null-complement, NC). With syllable count controlled, there are 5 NC questions of 4 syllables, 8 NC of 5 syllables, 9 VC of 5 syllables, and 7 VC of 6 syllables.

Table 1: *Sentence sequence types and exemplars. Subj: subject; PM:perfective marker; SFP: sentence-final particle; syl: syllable*

Null-Complement Structure (NC)					**Verb-Complement Structure (VC)**					
shū	kàn	le	ma		tā	huì	tiào	wǔ	ma	
书	看	了	吗		他	会	跳	舞	吗	
book	read				he	can		dance		
Topic	Verb	PM	SFP		Subj.	Model Verb		Complement	SFP	
Have (you) read the book?					Can he dance?					
xí	tí	zuò	le	ma	tā	xué	guò	shēng	wù	ma
习	题	做	了	吗	他	学	过	生	物	吗
exercises		do			he	learn		biology		
Topic		Verb	PM	SFP	Subj.	Verb	PM	Complement		SFP
Have (you) completed the exercises?					Has he learned biology?					

For each target question, a pair of two distinct scenarios were designed to convey different communicative intentions: one involving neutral events, where questions were straightforward inquiries to obtain information, and the other involving negative events, where the speaker expressed dissatisfaction toward the other interlocutor, a third party, or an object, using the questions in rhetorical manner (Table 2). For example, for the target sentence "Have you completed the exercises?", the context that induced ISQs was "The mother wanted to check her son's homework, so she asked." The context that induced RQs was "Xiao Ming did not do well on the exam, but he is still spending all his time on TikTok instead of doing his homework. So his mother said." Notably, none of the words in the target questions are default focused element like *wh*-word. Also, the content in the target question did not differ between the two scenarios in terms of referencing new/old information.

In total, 58 items (29 questions × 2 scenarios) were created. These items, along with other sentences from a larger project, were divided into four lists (each containing 42 items), ensuring that the two scenarios of the same target question did not appear in the same list. Each participant was assigned one of the four lists, and the items within each list were presented in a randomized order. A total number of 1433 sentences were analyzed in this paper.

2.3. Procedures

The experiment was conducted using the online survey platform Wenjuanxing (https://www.wjx.cn). Participants were directed to scan a QR code and complete the experiment on their smartphones in a quiet environment. While the use of headphones for recording was recommended, there were no strict restrictions on the type of mobile devices or recording equipment used. Participants were first asked to read the provided contextual information and then press the microphone icon on the screen to record the target sentences. They were instructed to fully engage with the scenario, and imagine themselves as the individual posing the question in the conversation. If any errors occurred during recording, such as mispronunciations, omissions, or added words, participants were permitted to re-record their responses.

2.4. Data analysis

The recordings were segmented and annotated using a custom-written forced-alignment tool, and then manually adjusted by two research assistants. The analysis includes auditory evaluations of prominence and acoustic measurements. For the perceptual judgments, a native Mandarin speaker with linguistic training, who was blinded to the experimental conditions of each sentence, was instructed to listen to all sentences and annotate perceived prominence based on auditory perception. If two levels of prominence were identified, the most prominent word(s) were labeled as 2, and the less prominent word(s) as 0. If three levels of prominence were detected, the most prominent word(s) were labeled as 2, the second most prominent word(s) as 1, and the least prominent word(s) as 0. Multinomial logistic regression was then used to examine the effect of context (ISQs vs. RQs) on the level of prominence (three categories: 0, 1, 2).

Pitch was first automatically tracked and subsequently manually checked and corrected using Praat [8]. For each syllable, *f0* values were extracted at ten equidistant points, and the time-normalized *f0* contours were compared across question types (ISQs vs. RQs). Generalized additive mixed models (GAMMs) were employed to examine differences in time-normalized *f0* trajectories between ISQs and RQs. Model fitting was performed using the R package *mgcv* [9], and the package *itsadug* [10] was utilized to visualize the model results. The number of basis functions (k) was adjusted as needed, and the best-fitting models were corrected for autocorrelation using a correlation parameter determined by the *acf_resid()* function.

To compare the duration of ISQs and RQs, the Shapiro-Wilk test was first conducted to assess normality. The results indicated that the duration data for both ISQs and RQs were not normally distributed. Therefore, the non-parametric Wilcoxon rank-sum test (also known as the Mann-Whitney U test) was applied to evaluate differences between the two groups.

3. Results

3.1. Perceptual prominence

The perceptual prominence for NC sentences is summarized in Figure 1. The verb, positioned at the last third syllable before the two neutral-toned particles, was most likely to carry primary prominence (the blue bar) compared to other elements, regardless of sentence type. The percentage of primary prominence on verb was higher in RQs than that in ISQs, but the difference was not statistically significant. Additionally, the interrogative particle *ma* in RQs was assigned more secondary prominence (denoted as *stress_1* in the figure) compared to that in ISQs (4-syllable NC: β = 0.395, z = 1.107, p < 0.001; 5-syllable NC: β = 1.089, z = 2.827, p = 0.005).

For VC sentences, the pattern of verbs being the most perceptually prominent was observed only in RQs. In contrast, ISQs exhibited the highest proportion of prominence on the post-verb complement, which is the last content word before the sentence-final particle (SFP) (Figure 2). This observation was further supported by statistical results: the probability of primary prominence on the verb was significantly higher in RQs than in ISQs (5-syllable VC: β = 1.866, z = 5.524, p < 0.001; 6-syllable VC: β = 2.682, z = 5.205, p < 0.001). The interrogative particle in RQs was also assigned more secondary prominence (5-syllable VC: β = 0.750, z = 2.163, p = 0.030; 6-syllable VC: β = 0.749, z = 2.163, p = 0.030).

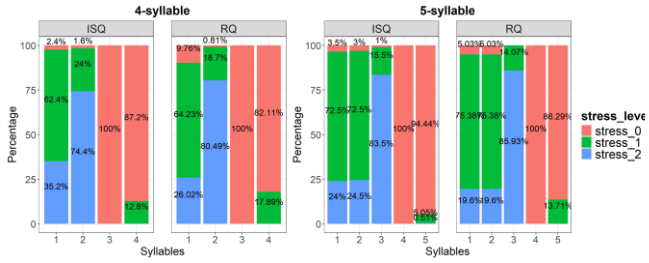

Figure 1: *Stress patterns for 4-syllable and 5-syllable NC sentences*

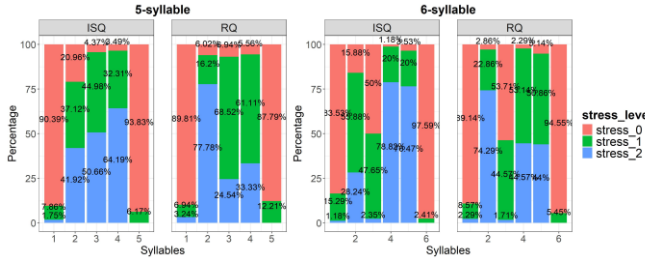

Figure 2: *Stress patterns for 5-syllable and 6-syllable VC sentences*

3.2. Pitch

The best GAMM models included sentence type (ISQ vs. RQ) as a parametric effect (fixed effect), along with a random smooth for speakers over normalized time. Figures 3 and 4 provide visualizations of the time-normalized *f0* trajectories for ISQs and RQs, as well as the predicted differences in *f0* between the two sentence sequence types.

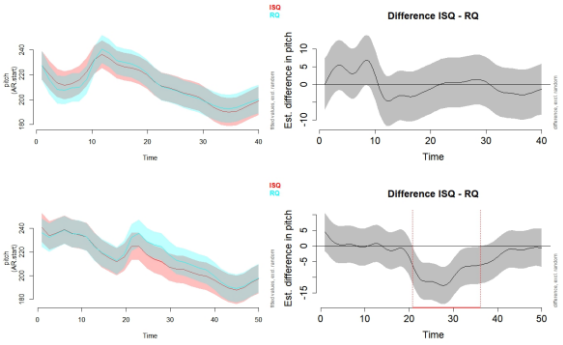

Figure 3: *Time normalized average f0 trajectory for ISQs (red) and RQs (blue) NC sentences (left panel) and the predicted difference in f0 (right panel) for 4-(upper panel) and 5-syllable (lower panel) sentences. The grey shading displays 95% confidence interval (CI) of the predicted mean difference. The difference in f0 is significant if zero is not included in the 95% CI. Significant differences are delimited by the vertical red lines.*

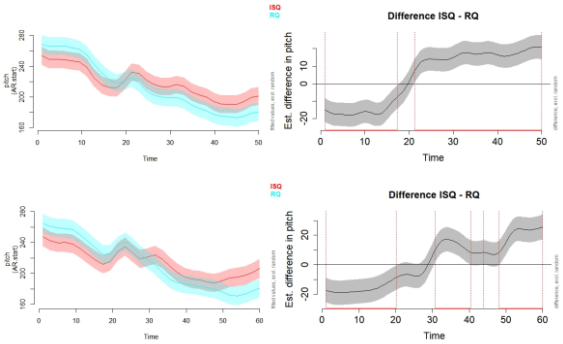

Figure 4: *Time normalized average f0 trajectory for ISQs (red) and RQs (blue) VC sentences (left panel) and the predicted difference in f0 (right panel) for 5-(upper panel) and 6-syllable (lower panel) sentences.*

Regarding NC sentences (Figure 3), for those with five syllables, RQs exhibited a significantly higher pitch during the verb and the first half of the perfective marker *le*. No significant differences were found in four-syllable NC sentences, although a trend of higher pitch in the verb for RQs could still be observed. As for VC sentences (Figure 4), both 5-syllable and 6-syllable questions showed a reversal in the pitch pattern before and after the verb/modal verb. Specifically, RQs displayed significantly higher pitch than ISQs from the sentence beginning to the verb/modal verb; however, the pitch was significantly lower for RQs than for ISQs in the subsequent portion of the sentence.

3.3. Duration

The duration of each syllable was compared between ISQs and RQs. No significant durational difference was observed in 4-syllable NC sentences. For five-syllable NC sentences, the Wilcoxon rank-sum test showed that the verb in RQs was significantly longer than that in ISQs (W = 17377, p = 0.029). In five-syllable VC sentences, the subject, modal verb, and main verb in RQs were all significantly longer compared to those in ISQs (first syllable: W = 17377, p < 0.001; second

syllable: W = 13132, $p < 0.001$; third syllable: W = 19238, $p < 0.001$; fourth syllable: W = 21248, $p = 0.010$). For six-syllable VC sentences, the duration of the verb (second syllable) was significantly longer in RQs (W = 7007.5, $p < 0.001$), while the subject (first syllable) was significantly shorter (W = 17621, $p = 0.003$).

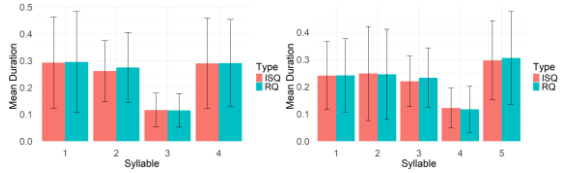

Figure 5: *Averaged word duration for four-syllable NC sentences (left) and five-syllable NC sentences (right)*

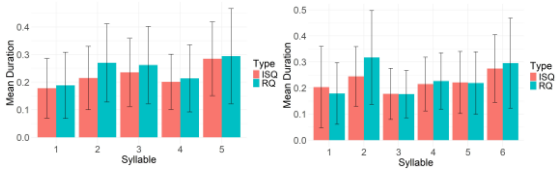

Figure 6: *Averaged word duration for five-syllable VC sentences (left) and six-syllable VC sentences (right)*

3.4. Results summary

In summary, conveying rhetorical meanings enhances the prominence of the verb (NC sentences) or shifts prominence toward the verb or modal verb (VC sentences), while the sentence-final particle receives more secondary prominence. Acoustically, the verb or modal verb in RQs exhibited significantly higher *f0* and longer duration in both NC and VC sentences. For VC sentences, the subject of RQs also displayed an elevated pitch, while the subsequent parts were characterized by a lower *f0*.

4. Discussion

Our data showed that the marking of RQs in naturalistic settings is primarily associated with the prosody of the verb. The verb in RQs was generally perceived more prominent and exhibited a higher *f0* and a longer duration than that in ISQs. Since these features are predominantly linked to focus marking [11, 12], we consider our results as reflect in focus reassignment driven by the demand of conveying rhetorical meanings. Particularly, two sentence sequence types depicted distinct verb-rooted dynamics. In the case of VC structures, the default focus determined by the syntax and the lexical semantics of the sentence is positioned at the complement, as clearly reflected in ISQs' measures: the complement took the most prominence and the longest duration per syllable. However, rhetorical expressions shift the focus to the verb; as a result, not only the verb is accented and lengthened, but also the *f0* rises up to the verb and has a sharper drop right after the verb. On the other hand, if the verb has been assigned the default focus in the ISQ, such as in the NC type, its prominence is further enhanced perceptually and acoustically, to achieve a stronger emphatic effect for rhetorical marking. The remaining part of the NC sentence, other than the verb, is less affected in intonation due to the stable focus position.

The verb was focused in our RQs, presumably because of the rhetorical meanings conveyed within the context. For instance, for the question 'Have (you) completed the exercises', the rhetorical context, which is known to both parties, implies a negation to this presupposition ('have NOT completed'), thus causing the act of asking to conflict with the fact. The verb in the question constitutes a focused element in our study, similarly to the *wh*-word in previous studies [1, 4]. In addition, the differences between sentence sequence types (Null-Complement VS Verb-Complement) suggest that syntactic construction plays a role in the phonetic realization of rhetorical questions.

The results of this study partially align with those of [4], but differences are also evident. While [4] reported global lengthening of duration, the lengthened duration observed in our study is localized mainly on the verb. Additionally, [4] found that RQs consistently exhibited significantly lower *f0* at the *wh*-word as well as the post-verb part, whereas in our study, lower f0 was observed only after the focused verb in VC sentences. One possible reason is the lexical tone for the focused element, which is a rising tone (who "*shéi*") in [4] whereas our study involves diverse tones because different verbs were used. Another factor that should be taken into account is whether performative speech is measured. Rather than solely in a lab environment with a limited number of participants from a specific population, this study collected data in naturalistic settings with speakers with diverse backgrounds. The inclusion of participants with diverse dialectal backgrounds enables identifying common patterns in rhetorical meaning signaling in Standard Chinese across dialects. These findings can help reveal generalizable patterns, offering valuable insights for speech synthesis applications.

Besides, we note that the short, 4-syllable NC questions were prosodically marked in a weaker manner compared with the longer NC questions. It may be attributed to the limited utterance length, which may affect the adjustment of global intonation, especially in a tonal language [13]. In addition to pitch and duration, voice quality also warrants consideration as a potential marker of rhetorical meaning. Previous research has established that non-modal voice quality serves as a phonetic correlate of rhetorical meaning in both tonal and non-tonal languages [1, 3, 4, 6]. It is plausible that differences in voice quality exist between information-seeking questions (ISQs) and rhetorical questions (RQs) in 4-syllable NC questions. Therefore, in subsequent studies, we plan to conduct a more in-depth analysis of voice quality to achieve a more comprehensive understanding of this issue.

5. Conclusions

This study shows that in naturalistic settings, rhetorical meaning in Standard Chinese is prosodically marked by enhanced prominence on the verb/modal verb. Acoustically, this prominence is characterized by a higher *f0* and longer duration of the verb/modal verb. The findings further reveal that the prosodic realization of rhetorical expressions is systematically influenced by syntactic structure. These results support a model in which focus reassignment, driven by rhetorical intent, interacts with syntactic structure, phonological factors, and utterance length to shape prosodic patterns. These insights hold potential applications in speech synthesis and cross-dialectal studies. Future investigations should focus on elucidating the cognitive and linguistic mechanisms underlying these observed patterns.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC 62276185) and Key Laboratory of Linguistics, Chinese Academy of Social Sciences (Project #2024SYZH001).

7. References

[1] N. Dehé and B. Braun, "The prosody of rhetorical questions in English," *English Language and Linguistics*, vol. 24, pp. 607 – 635, 2020.

[2] I. Miura and N. Hara, "Production and perception of rhetorical questions in Osaka Japanese," *Journal of Phonetics*, vol. 2, pp. 291 – 303, 1995.

[3] R.-Y.-H. Lo, A. Kiss, and M. Tulling, "The prosodic properties of the Cantonese sentence-final particles aa1 and aa3 in rhetorical wh-questions," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, Melbourne, Australia, 2019, pp. 502 – 506.

[4] K. Zahner-Ritter, Y. Chen, N. Dehé, and B. Braun, "The prosodic marking of rhetorical questions in Standard Chinese," *Journal of Phonetics*, vol. 95, p. 101190, 2020.

[5] R.-Y.-H. Lo and A. Kiss, "Durational and pitch marking of rhetorical wh-questions in Mandarin," in *Proceedings of the 10th International Conference on Speech Prosody (Speech Prosody 2020)*, Tokyo, Japan, 2020.

[6] B. Braun, N. Dehé, J. Neitsch, D. Wochner, and K. Zahner, "The prosody of rhetorical and information-seeking questions in German," *Language and Speech*, vol. 62, pp. 779 – 807, 2019.

[7] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements," in P*roceedings of Interspeech 2021*, Prague, Czech, 2021, pp. 245 – 249.

[8] P. Boerma, "Praat, a system for doing phonetics by computer". *Glot International*, no. 5, 2002.

[9] S. Wood, "mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation, " R Package Version 1.8-42, 2023. [Online]. Available: https://cran.r-project.org/web/packages/mgcv/index.html

[10] J. van Rij, M. Wieling, R. Baayen, and H. van Rijn, "Itsadug: Interpreting time series and autocorrelated data using GAMMs," R Package Version 2.4, 2023. [Online]. Available: https://search.r-project.org/CRAN/refmans/itsadug/html/itsadug.html.

[11] M. Lin and Z. Li, "Focus and boundary in Chinese intonation," in Proceedings of ICPhS, vol. 17, Prague, Czech, Aug. 2011, pp. 1246 – 1249.

[12] Y. Chen, "Durational adjustment under corrective focus in Standard Chinese," Journal of Phonetics, vol. 34, no. 2, pp. 176 – 201, Apr. 2006.

[13] J. Yuan and M. Liberman, "F0 declination in English and Mandarin broadcast news speech," Speech Communication, vol. 65, pp. 67 – 74, 2014.