Prosodic Disambiguation of Wh-Words: In the Case of 'ma' in Tianjin Mandarin

Tianqi Geng[1,2], *Hui Feng*[2], *Cong Zhang*[3]

[1]Carnegie Mellon University
[2]Tianjin University
[3]Newcastle University

tianqig@andrew.cmu.edu, fenghui@tju.edu.cn, Cong.Zhang@newcastle.ac.uk

Abstract

In Tianjin Mandarin, the wh-word *mà* (嘛, 'what') conveys either an interrogative or a negative function. This study investigates whether string-identical utterances containing *mà* are semantically ambiguous and examines the prosodic strategies that native speakers use to disambiguate them.

Fifteen native speakers of Tianjin Mandarin participated in a perception experiment and eighteen in a production experiment. The perception experiment showed that either syntactic or contextual cues alone enabled accurate interpretation of the function of *mà*. When both cues were absent, participants' judgments were highly variable. Introducing prosodic cues in this ambiguous condition significantly improved accuracy, indicating that prosody can aid disambiguation. Acoustic analysis from the production experiment revealed distinct prosodic patterns under the two interpretations of *mà*. At the word level, without syntactic marking, negative *mà* shows a longer duration and a wider F0 range. At the sentence level, prosodic prominence was consistently assigned to interrogative *mà*, whereas for negative *mà*, such prominence occurred only when syntactically specified; in unspecified contexts, prosodic prominence shifted to the subject or verb. These findings suggest that prosody plays a crucial role in disambiguating and encoding multifunctional morphemes in Tianjin Mandarin.

Index Terms: wh-word, prosodic pattern, disambiguation, Tianjin Mandarin

1. Introduction

In Mandarin Chinese, certain wh-words, such as *shénme* (什么, 'what') and *nǎlǐ* (哪里, 'where'), can be interpreted either interrogatively ('what', 'where') or with non-interrogative readings ('something', 'somewhere'). In Tianjin Mandarin, the interrogative word *mà* (嘛, 'what') shares the interrogative meaning 'what' with Standard Mandarin *shénme*. However, apart from an interrogative word, *mà* can also function as a negative polarity item [1], [2], [3]. Ambiguity can therefore arise, as in (1):

(1) 张三 买 嘛
ZhāngSān mǎi mà
Zhang San buy what

a. 'What does Zhangsan want to buy?' (interrogative)

b. 'Zhangsan shouldn't buy anything.' (negative)

When someone wonders what Zhang San buys, (1a) would be used; while (1b) would be used when the speaker is advising or implying that Zhang San should not buy anything (perhaps in a context of disapproval, caution, or suggesting an alternative action). Previous work in various languages (i.e., English, French, Korean, Chinese) shows that when declarative and interrogative strings coincide, prosodic cues can provide the critical disambiguating signal [4], [5], [6], [7]. Building on these findings, the present study aims to (1) test whether sentences with *mà* are perceived as ambiguous between negative and interrogative interpretations through a perception experiment, and (2) identify the acoustic correlates Tianjin Mandarin speakers use to distinguish the multiple functions of *mà* through a production experiment.

2. Methods

2.1. Perception experiment

Fifteen native speakers of Tianjin Mandarin (age: 18-25, M = 21.67, SD = 1.11) completed an online identification task via jsPsych [8]. Participants listened to each stimulus and judged the function of *mà*, by selecting "wh" (interrogative), "neg" (negation), "both" (both interpretations are possible), or "neither" (neither interpretation is possible). The responses and reaction times were collected via Firebase [9].

Stimuli for the perception experiment were organized by crossing groups and disambiguation methods. Groups comprised three levels – Target (T), Question (Q), and Negation (N), while the disambiguation method comprised four levels: none (non), contextual (con), syntactic (syn), and prosodic (pro). In the non-condition, *mà* appeared without any disambiguation cues. In the con-condition, each *mà* was followed by a sentence context. In the syn-condition, *mà* was embedded in a disambiguating syntactic frame. For Negation, the pattern is V-*mà*-V (V stands for verb) [1] while for Question, it is V- *mà* -le, where the perfective marker *le* contributes to the interrogative interpretation [10]. In the pro-condition, participants heard audio recordings produced by a female native speaker of Tianjin Mandarin, guided by context, to highlight the difference in prosody when *mà* represents various functions. Serving as control groups, all Q- and N-group stimuli were presented with one of the con, syn, or pro disambiguation methods. In contrast, T-group items were presented only in the non-condition. To mitigate the influence of lexical tone, all four tones in Tianjin Mandarin were considered during stimulus creation. This yielded a total of 28 unique stimuli (2 groups*3 disambiguation methods*4 tones + 1 group*1 disambiguation method*4 tones).

2.2. Production experiment

The speech materials used in this experiment were drawn from the INTO-CASS corpus [11] created as part of a larger project. Eighteen native speakers (age: 18-25, M = 21.78, SD = 1.96) participated in the production study, where they read the

materials presented on the screen. The speech materials comprised two groups: a target set and a control set. Both sets include two grammatical functions: negation and interrogation. To mitigate the influence of lexical tone, all four tones in Tianjin Mandarin were considered as well, yielding 16 items (2*2*4) per participant.

In the target set, participants produced the target utterances after reading contextual cues designed to elicit either a negative or an interrogative interpretation; the contextual prompts were provided in written form but were not read aloud as part of the utterance. The control set used the same syntactic structures as in the perception experiment.

Recordings were made in a soundproof booth at 44.1 kHz/16 bit. Speech was annotated in Praat [12] at sentence, word, and syllable tiers. Acoustic measures (F0, intensity, duration) were extracted using VoiceSauce [13], with F0 estimated via STRAIGHT [14]. Each vowel pitch was time-normalized to 11 equidistant points. F0 values were converted to semitones and normalized using a log-z transformation [15] to control individual speaker differences; intensity and duration were z-score normalized, and the duration of *mà* was further scaled as a proportion of the total sentence duration, which is coded as word/sen below.

3. Results and discussion

3.1. Perception experiment result: evidence of ambiguity

3.1.1. Response distribution

Figure 1 presents the percentage of participant responses, with participant responses in different colors and disambiguation methods in different columns. For Negation items, contextual and syntactic cues yielded accurate negation responses (>90%), while prosodic cues also biased interpretation toward negation but produced more "both" responses (16.7%). Question items showed a parallel pattern that syntactic cues reached ceiling (100% interrogative), contextual cues produced strong interrogative responses (83%), and prosodic cues remained effective but again increased the proportion of "both" and "none". The Target group, which lacked any cue, displayed the widest spread of responses: interrogative readings dominated (70.8%), but "both" reached 16.7% and negation 12.5%. This dispersion confirms that *mà* is inherently ambiguous when no disambiguating cues are present. Across groups, contextual and syntactic cues strongly stabilized interpretation, while prosody remained helpful but less categorical, reflected in a reliably higher rate of "both." Besides, this preference for interrogative interpretations in underspecified contexts is not unique to Mandarin. Jones et al. [16] also reported a bias toward wh-questions in Korean question disambiguation, attributing it to the influence of the lexical meaning of the question word. A similar case is observed in Tianjin Mandarin in a different context [17, p. 146].

The effect of GROUP and METHODS on participants' response choice was tested through generalized linear mixed-effects models (GLMM) with a binomial link. Separate models were fitted for each response choice (wh, neg, both, none) to account for the multinomial nature of the outcome. SUBJECT was included as a random intercept to control for repeated measures. Odds ratios (ORs) and 95% confidence intervals (CIs) were derived from the fixed effects.

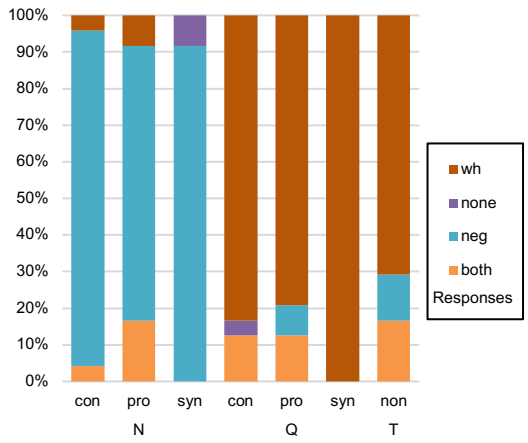
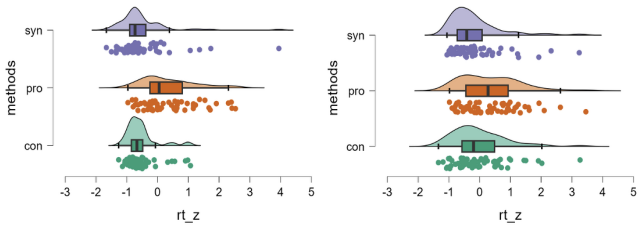


Figure 1: *Percentage distribution of response choices by group and disambiguation cues. Groups: N – negation; Q – question; T - target. Cues: syn – syntactic; pro – prosodic; con – contextual.*

From the GLMM result, it can be observed that, compared to the Negation group, in Question group and Target group, participants were more likely to choose 'wh', with odds ratios of 53.4 and 20.9 respectively ($p < 0.001$). Methods such as pro and syn increased the probability of choosing 'wh', though some effects were not statistically significant. In Question group and Target group, participants were significantly less likely to choose 'neg' compared to the Negation group (OR = 0.031, 0.0036, $p < 0.001$). The model for the "both" option exhibited instability, likely due to the sparse data and separation issues arising from the limited distribution of participants' choices between "both" and "none."

3.1.2. Response time

A linear mixed model was conducted to examine the effects of GROUP and METHODS on response time, accounting for random effects of SUBJECT. The results revealed significant effects of GROUP , $F(1,36.05)=8.86$, $p=.005$, and disambiguation METHOD, $F(2,20.08)=11.71$, $p<.001$.



(a) Negation Group *(b) Question Group*

Figure 2: *Cloud-Rain plot of normalized reaction time by disambiguation cues and group (Cues: syn – syntactic; pro – prosodic; con – contextual; rt_z – normalized reaction time).*

Tukey's HSD tests showed that contextual cues were responded to significantly faster than prosodic cues ($p<.001$), and syntactic cues were likewise faster than prosodic cues ($p<.001$). Contextual and syntactic cues did not differ ($p=.881$). The prosodic condition consistently produced the slowest and most variable responses as shown in Figure 2, suggesting

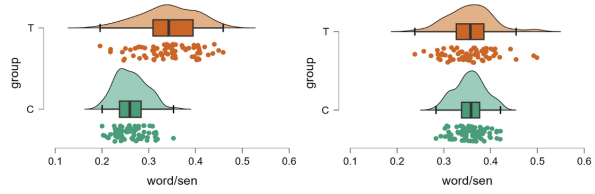
heavier processing demands or greater individual variability in prosodic sensitivity.

3.2. Prosodic disambiguation strategies

3.2.1. Acoustic Features of mà

Figure 3 shows that within the negation condition, the Target group, which contains no disambiguation cues, produced substantially longer durations than the Control group, which contains syntactic cues. The two groups showed minimal differences in the question condition. Overall, the Control group displayed more consistent durations across functions, while the Target group showed greater variability, suggesting heavier reliance on durational cues.

Preliminary linear mixed-effects models with GROUP and FUNCTION as fixed effects and SUBJECT as random intercepts failed to converge due to singular fits. Consequently, we reduced the analysis to a two-way ANOVA. This applies to this analysis and all following analyses.



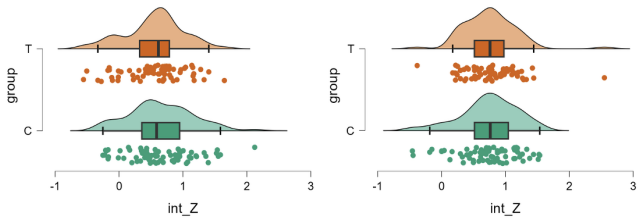
(a) Negation Group *(b) Question Group*

Figure 3: *Cloud-Rain plot of normalized duration of mà (T - target set; C – control set; word/sen – duration index).*

A two-way ANOVA revealed significant main effects of GROUP and FUNCTION, as well as a significant interaction. The GROUP effect was significant, $F(1, 283) = 56.76$, $p < .001$, $\eta^2_p = .167$, indicating that GROUP accounted for about 16.7% of the variance. The effect of FUNCTION was even larger, $F(1, 283) = 106.81$, $p < .001$, $\eta^2_p = .274$. The interaction was also significant, $F(1, 283) = 56.17$, $p < .001$, $\eta^2_p = .166$, indicating that the two groups were affected differently by functional context.

These results suggest that duration is a robust cue for distinguishing negation from question, and that the Target group relies on this cue more heavily. Tukey's HSD tests confirmed that *mà* was significantly longer in the Target group than in the Control group (mean difference = 0.04, $p < .001$) and significantly longer in negation than in question (mean difference = 0.055, $p < .001$).

Figure 4 displays the distribution of normalized mean intensity for *mà* across functions and groups. Under negation, the Target group produced higher intensity than the Control group, but the groups showed little difference under the question condition. When grouping by function, *mà* in the question condition showed higher intensity overall.



(a) Negation Group *(b) Question Group*

Figure 4: *Cloud-Rain plot of normalized mean intensity of mà (T - target set; C – control set; int_Z – normalized intensity).*

A two-way ANOVA showed no significant main effect of GROUP, $F(1, 283) = 0.74$, $p = .390$, $\eta^2_p = .003$, but a significant main effect of FUNCTION, $F(1, 283) = 7.55$, $p = .006$, $\eta^2_p = .026$. The interaction was nonsignificant. Tukey's HSD test showed that mean intensity was significantly lower in negation than in question ($p = .006$). These results suggest that intensity is sensitive to functional contrast. Interrogative contexts elicited prosodic strengthening through higher intensity.

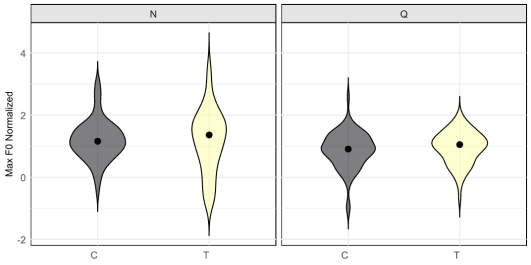


Figure 5: *Violin plot of normalized max F0 of mà (T - target set; C – control set; N – negation; Q – question).*

Figure 5 illustrates the distribution of normalized maximum F0. The Target group showed greater variability under negation, but the two groups are similar in the question condition. A two-way ANOVA revealed a significant main effect of FUNCTION, $F(1, 283) = 10.332$, $p = .001$, $\eta^2_p = .035$, but no effect of GROUP and no interaction. Tukey's HSD showed that maximum F0 was significantly higher in negation than in question (mean difference = 0.273, $p = .001$).

Figure 6 presents the distributions of normalized pitch range. Both groups displayed wider pitch ranges under negation than under question, with little difference between groups. A two-way ANOVA found a significant effect of FUNCTION, $F(1, 283) = 22.466$, $p < .001$, $\eta^2_p = .074$, but no GROUP effect and no interaction. Tukey's HSD test confirmed that the F0 range was significantly greater in negation than in question (mean difference = 0.440, $p < .001$). Pitch range therefore serves as a clear marker of functional contrast, but it does not differentiate the two speaker groups.

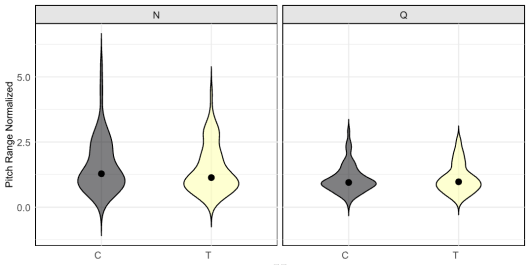


Figure 6: *Violin plot of normalized pitch range of mà(T - target set; C – control set; N – negation; Q – question).*

Non-significant effects were not analyzed in detail. Both mean F0 and minimum F0 showed no reliable effects of group or function, and no interaction. These measures, therefore, do not appear to contribute meaningfully to distinguishing negation from questioning and are not further discussed.

3.2.2. Sentence-level prosodic patterns

Figure 7 shows sentence-level F0 contours across conditions. Under the question condition, the Control and Target groups displayed similar overall contours, apart from differences in the final syllable attributable to segmental effects. The Target group produced a slightly higher overall pitch.

Under negation, the Target group differed more noticeably from the Control group, showing a rising tone on the first syllable, elevated F0 on the second, and a lowered F0 on the third, suggesting heterogeneous interpretive strategies. Within the Control group, negation and question differed mainly in the F0 of *mà*, which was higher in negation. Within the Target group, contrasts were concentrated in the first two syllables, consistent with earlier findings.

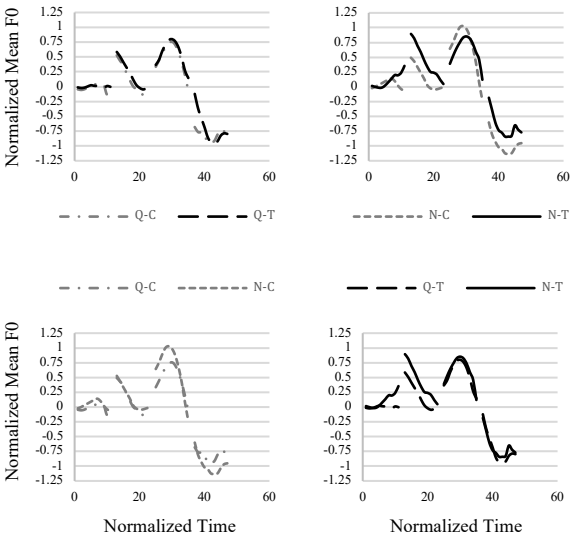


Figure 7: *Sentence-level F0 contours (T - target group; C – control group; N – negation; Q – question).*

Prosodic variation across syllables was examined under Negation and Question in both groups. Figure 8 displays normalized mean F0, relative duration, and normalized intensity for each syllable.

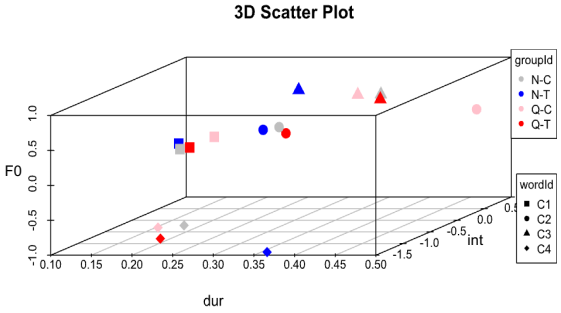


Figure 8: *3D Scatter plot of syllables (T - target group; C – control group; N – negation; Q – question; C1 – first character [subject]; C2 – second character [verb]; C3 – third character [mà]; C4 – last character [verb/particle]).*

Within-condition comparisons, while influenced by inherent segmental differences, still reveal clear positional patterns after normalization. In Negation, the Control group assigned the highest intensity to the verb and the longest duration, as well as the highest F0 to *mà*. The Target group showed a similar intensity peak on the verb but shifted the longest duration to the final syllable, reflecting a different strategy for encoding negation.

Under the interrogative condition, the Control group showed increased duration and intensity of the verb and the highest F0 on *mà*. The Target group again placed the intensity peak on the verb but shifted both the longest duration and the highest F0 to *mà*, strengthening its interrogative role.

Across groups, interrogatives elicited a global pitch rise, particularly robust in the Control group. Target speakers produced a similar but reduced elevation, possibly reflecting a more selective use of cues when syntactic marking is absent.

4. Conclusions

This study examined how native speakers of Tianjin Mandarin interpret and produce utterances containing the multifunctional morpheme *mà*, which can signal either interrogation or negation. The findings from the perception experiment confirm the existence of semantic ambiguity when syntactic and contextual cues are absent, with subjects showing varied interpretations and slower reaction times. On the other hand, findings from the production experiments reveal that speakers utilize prosodic strategies, particularly variation in duration and pitch, to encode functional distinctions. These strategies are not uniform but reflect group-specific adaptations based on the availability of syntactic cues. In utterances where negation is syntactically indicated, prosodic prominence tends to fall on *mà* itself. In contrast, when negation must be inferred prosodically, speakers shift focus to other elements in the sentence, such as the subject or verb.

Overall, the findings of this paper show that prosodic cues can partially compensate for the absence of syntactic disambiguation. Meanwhile, prosodic variation plays a critical role in the interpretation of multifunctional morphemes like *mà* in Tianjin Mandarin. The findings may have practical implications for natural language processing tasks involving tonal languages, particularly in improving disambiguation in speech synthesis and recognition systems.

5. Acknowledgements

This work was supported by the Independent Topic Selecttion Project of Cultural Experts and “Four Batches” of Talents awarded to Aijun Li, and also Key Laboratory of Linguistics, Chinese Academy of Social Sciences (2024SYZH001).

6. References

[1] S. Qi, *The study of Tianjin Grammar*. Shanghai Jiao Tong University Press, 2020.

[2] S. Lv, *Eight hundred words in modern Chinese*. The Commercial Press, 1980.

[3] W.-T. Dylan Tsai, “On the Topography of Chinese Modals,” in *Beyond Functional Sequence: The Cartography of Syntactic Structures, Volume 10*, U. Shlonsky, Ed., Oxford University Press, 2015, p. 0. doi: 10.1093/acprof:oso/9780190210588.003.0015.

[4] M. Vion and A. Colas, “Pitch Cues for the Recognition of Yes-No Questions in French,” *J. Psycholinguist. Res.*, vol. 35, no. 5, pp. 427–45, Sep. 2006, doi: 10.1007/s10936-006-9023-x.

[5] C. Gussenhoven, “Intonation and interpretation: phonetics and phonology,” in *Speech Prosody 2002*, ISCA, Apr. 2002, pp. 47–57. doi: 10.21437/SpeechProsody.2002-7.

[6] Y. Yang, S. Gryllia, and L. L.-S. Cheng, “*Wh*-question or *wh*-declarative? Prosody makes the difference,” *Speech Commun.*, vol. 118, pp. 21–32, Apr. 2020, doi: 10.1016/j.specom.2020.02.002.

[7] S. M. Jones, Y. Kim, and C. Zhang, “The syntax-prosody interface in LFG: Revisiting Korean question focus,” in *Proceedings of the LFG Conference*, University of Konstanz, 2024, pp. 186–206.

[8] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt, “jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments,” *J. Open Source Softw.*, vol. 8, no. 85, p. 5351, May 2023, doi: 10.21105/joss.05351.

[9] Google LLC, “Firebase documentation,” Google Firebase. [Online]. Available: https://firebase.google.com/docs

[10] L. L. S. Cheng, “On the Typology of Wh-Questions,” *Mass. Inst. Technol.*, 1991.

[11] A. Li and Z. Xiong, “Into-Cass: A Corpus for the Study of Intonation and Prosody in Chinese Dialects and Ethnic Languages,” in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Nov. 2021, pp. 53–58. doi: 10.1109/O-COCOSDA202152914.2021.9660534.

[12] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]*. (Jan. 23, 2022). [Online]. Available: https://www.praat.org

[13] Y.-L. Shue, P. Keating, and C. Vicenik, “VOICESAUCE: A program for voice analysis.,” *J. Acoust. Soc. Am.*, vol. 126, no. 4_Supplement, pp. 2221–2221, Oct. 2009, doi: 10.1121/1.3248865.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds1,” *Speech Commun.*, vol. 27, no. 3, pp. 187–207, Apr. 1999, doi: 10.1016/S0167-6393(98)00085-5.

[15] A. Li, “Phonetic Correlates of Neutral Tone in Different Information Structures,” *Contemp. Linguist.*, vol. 19, no. 3, pp. 348–378, 2017.

[16] S. Jones, Y. Kim, and C. Zhang, “Perceiving and modelling the scope of question focus in Korean. To appear in Language and Speech,” *Lang. Speech*, in press.

[17] C. Zhang, “Tianjin Mandarin tones and tunes,” University of Oxford, 2018. doi: 10.5287/ORA-JNNQAVEN0.