10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Durational and Pitch Marking of Rhetorical *Wh*-questions in Mandarin

Roger Yu-Hsiang Lo[1], *Angelika Kiss*[2]

[1]Department of Linguistics, University of British Columbia, Canada
[2]Department of Linguistics, University of Toronto, Canada

roger.lo@ubc.ca, angelika.kiss@mail.utoronto.ca

Abstract

In a production experiment, we explore the prosody of different types of questions in Mandarin. In each trial, we manipulated the context to elicit a set of string-identical *wh*-questions with the following three readings: information-seeking questions (ISQs), positive rhetorical questions (RQ+s): *Who wants to drink coffee? - (Of course) John*, and negative rhetorical questions (RQ−s): *Who wants to drink coffee? - Nobody*.

The results suggest that ISQs tend to have a shorter utterance duration and a longer, higher-pitched sentence-final particle (SFP), whereas RQ−s are more likely to have a longer utterance duration and a shorter, lower-pitched SFP, with RQ+s lying in between these two extremes. These trends are found to mirror the extent of speaker commitment and addressee engagement across the three question types.

Overall, our results suggest that the prosodic difference helps to differentiate between question types, though the mapping between acoustic properties and question types is language-specific.

Index Terms: rhetorical questions, *wh*-questions, discourse pragmatics, Mandarin, speech production, sentence-final particles, Functional Principal Component Analysis

1. Introduction

Aside from their canonical use as information-seeking questions (ISQs), interrogatives also have non-canonical uses, such as rhetorical questions (RQs). Semantically, ISQs express lack of information. RQs, on the other hand, presuppose that the answer to the question is already known by both the speaker and the addressee. RQs can be further distinguished on the basis of whether the suggested answer denotes the empty set or some salient element from the set denoted by the question word. For instance, *Who wants to drink coffee?*, as an RQ, could suggest an answer that 'Nobody wants to drink coffee', or that someone specific, who is known to everyone in the discourse, wants to drink cofee [1, 2]. We label the former type of rhetorical questions as a negative rhetorical question (RQ−) and the latter as a positive rhetorical question (RQ+) [3].

The difference between the three question types can also be characterized in terms of the speaker's level of commitment and the address's level of engagement [4, 5, 6]. In ISQs, the speaker raises an issue, and the addressee is tasked to provide an answer to resolve the issue. ISQs therefore require no commitment from the speaker but full engagement on the part of the addressee. In RQ−s, the speaker is fully committed to a proposition corresponding to the empty-set reading of the *wh*-word (e.g., 'Nobody wants to drink coffee.'). As such, there is no issue for the addressee to resolve; the addressee is therefore not engaged. RQ+s are similar to RQ−s in that the speaker is fully committed (as RQ+s presuppose that everyone knows the answer), but differ from RQ−s in the extent of addressee engagement. While an RQ− indicates that the answer is the empty set — and the addressee can interpret the utterance regardless of whether she knew the answer previously or not — RQ+s are only understood if the addressee does the effort of finding the right answer. However, RQ+s require a smaller effort compared to ISQs, since the answer is already known. This property sets RQ+s apart from RQ−s at the level of addressee engagement. Table 1 summarizes the above characterization.

Table 1: *Speaker commitment and addressee engagement in the three question types.*

Question type	Speaker	Addressee
ISQ	Not committed	Fully engaged
RQ+	Fully committed	Somewhat engaged
RQ−	Fully committed	Not engaged

Prosody represents yet another dimension where ISQs and RQs can differ. Indeed, an emerging literature has tapped into the prosodic distinction between ISQs and RQs. In particular, the distribution of boundary tones and nuclear pitch accent types, duration of utterances, and voice quality are found to contribute to the difference between ISQs and RQ−s in non-tonal languages like English, Icelandic, and German, although these cues are weighted differently across the languages [7, 8, 9, 10, 11]. For tonal languages, ISQ, RQ−, and RQ+ readings of Cantonese *wh*-interrogatives are found to diverge in the fundamental frequency (F0) contour and the duration of sentence-final particles (SFPs) [3].

In the current study, we explore the prosodic marking of ISQs, RQ+s, and RQ−s in Mandarin *wh*-interrogatives. Mandarin was chosen because there is a relatively large body of literature on Mandarin statement and question prosody (e.g., [12, 13, 14, 15, 16]), but much less is known about the prosody of RQs in the language. Here we further restrict our discussion to phonetic properties, such as duration and fundamental frequency. These two acoustic dimensions are by no means the only correlates of prosody, but these metrics are relatively easy to measure and have been targeted in almost all studies. Given the similarity between Cantonese and Mandarin and the finding that Cantonese marks the contrast between ISQs and RQs prosodically, we expect the three question types to have different prosodies in Mandarin. However, the mapping between acoustic dimensions and question types might be language-specific.

2. Experiment

2.1. Participants

The data analyzed in this study come from 25 adults (9 males, 16 females; mean age = 20.6 years; *SD* = 1.6 years), who were

10.21437/SpeechProsody.2020-88

undergraduate students at the University of Toronto. All participants are native speakers of (Mainland) Mandarin, who had resided in Canada for less than four years at the time of participation and reported no speech or hearing impairments.

2.2. Stimuli

Target sentences consist of eight *wh*-questions that are ambiguous between information-seeking and rhetorical readings, have the same syntactic structure, and contain the same number of syllables (i.e., seven syllables).[1] The *wh*-word in all target sentences is *shei2* ‘who’ and always occurs in the same position within the sentence. In addition, all questions end with the SFP *a5*. To mitigate the effect of tonal co-articulation, the lexical tone of the syllable immediately preceding the SFP is controlled for — as Mandarin has four lexical tones, one fourth of the target sentences has Tone 1 before the SFP, one fourth has Tone 2, and so on.

For each of these eight target *wh*-questions, we created three short contexts, each favoring one of the three readings respectively (see Table 2 for illustration). The context descriptions are as concise and informative as necessary, with the target questions being placed at the end of each context. In addition, each target question is prefaced by another short sentence that is semantically congruent with the target question. The purpose of these additional sentences is to strengthen the intended reading. Additional eight filler sentences, which are polar question versions of the eight target sentences (e.g., *Does anyone want to drink coffee?* for *Who wants to drink coffee?*), were constructed. Each of these eight filler polar questions is also paired with two contexts, which bias the question toward either an ISQ or an RQ− reading.[2]

Table 2: *The three contexts for the target sentence*
You3 shei2 xiang3 he1 ka1fei1 a5?
have who want drink coffee SFP
‘Who wants to drink coffee?’

ISQ: You are having a family gathering with many of your relatives. After the meal, you’d like to serve tea and coffee, but you don’t know how much to prepare. You prepare the coffee first, and you want to find out how many cups are needed. So you ask:
“I don’t know how many people want coffee. **Who wants to drink coffee?**”
RQ+: Your relative came to visit you and brought you five different kinds of coffee. Unfortunately, neither you nor your flatmate, John, drinks coffee, so you guys decide to give all the coffee away. The first person that comes in mind to you is John, your American neighbor, because you both see him with his coffee on the balcony every day. So when John wonders who you should give all the coffee to, you point to John, who is drinking coffee on his balcony right now, and say:
“Just look at that balcony. **Who wants to drink coffee?**”
RQ−: Mary is throwing a party. It’s 1am, and everybody is about to leave, but she is wondering if she should still make some coffee. You think nobody would want that, as everyone is preparing to leave. So when Mary is telling you that she’ll start the coffee machine, you say:
“Everyone is ready to leave. **Who wants to drink coffee?**”

In the experiment, after three practice trials, 40 context-question doublets (8 *wh*-targets × 3 contexts + 8 polar fillers × 2 contexts) were presented in a pseudo-random order, with the restriction that two string-identical targets or fillers did not follow each other. The entire experiment took about 20 minutes.

2.3. Procedure

The experiment was programmed in PsychoPy v3.2.3 [17] and conducted in a sound-attenuated booth. In each trial, the context was presented both auditorily (as recorded by a male native speaker of Mandarin) through a headphone and visually in simplified Chinese characters on a monitor simultaneously. The corresponding *wh*-target or polar filler, along with their facilitating sentence, then appeared on the monitor. Participants were tasked to read both sentences exactly as written and as naturally as possible. After reading the sentences, participants pressed the space bar to proceed to the next trial.

2.4. Analysis and results

Target *wh*-questions were extracted from the recordings and checked manually. A target *wh*-question was excluded from the subsequent analysis if (i) there was disfluency in the sentence, (ii) the SFP was omitted or devoiced, or (iii) the participant made mistakes (e.g., by adding or ignoring a word). In total, these restrictions excluded 39 tokens, leaving 561 *wh*-targets (ISQ: 193, RQ−: 180, RQ+: 188) that entered the analyses.

The analyses and results for duration and F0 are presented below respectively.

2.4.1. Duration

Syllables in all target sentences were annotated with Praat [18]. Total utterance duration and the duration of the *wh*-word *shei2* and the SFP *a5* were extracted automatically. The duration of the *wh*-word and the SFP was relativized to the total utterance duration, the distributions of which are shown in Figure 1.

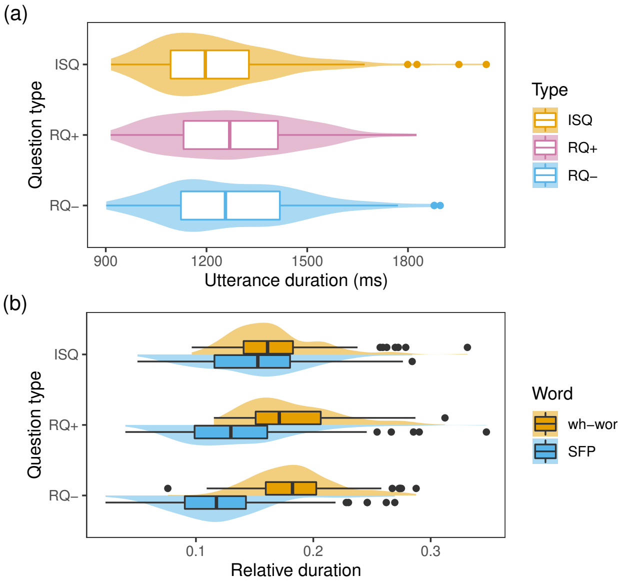

Figure 1: **(a)** *Distributions of total utterance durations across the three question types.* **(b)** *Distributions of the relative durations of the* wh*-word* shei2 *and the SFP* a5 *across the three question types.*

[1] These eight target *wh*-questions were selected from 12 candidate *wh*-questions, based on the results of a rating task that asked 31 additional participants to rate the accessibility of different readings associated with the 12 candidate sentences when embedded in contexts.

[2] We did not have contexts for the RQ+ reading, as an RQ+ interpretation of a polar question is very hard to elicit in Mandarin.

Durational data were analyzed using generalized additive models for location, scale, and shape (GAMLSS) [19], with the Box-Cox t distribution as the conditional distribution of duration. For the utterance duration model, the fixed effect only has QUESTION TYPE (ISQ, RQ+, RQ−; successive difference coded: RQ+ vs. ISQ, RQ− vs. RQ+). For the word duration model, the fixed effects included QUESTION TYPE (ISQ, RQ+, RQ−; successive difference coded: RQ+ vs. ISQ, RQ− vs. RQ+) and WORD (*wh*-word, SFP; dummy coded; reference level: SFP), as well as their two-way interaction. The random effects for both models involved only random intercepts for participants and items, as models with more complex random structures failed to converge.[3]

For the utterance duration model, there is an effect for RQ+ vs. ISQ ($\beta = 46.8$, $SE = 10.8$, $p < 0.001$), but the effect for RQ− vs. RQ+ is not significant ($\beta = 3.7$, $SE = 11.2$, $p = 0.74$). This indicates that the average utterance duration of ISQs is shorter than that of RQ+s, but that the average durations of RQ+s and RQ−s are similar.

For the word duration model, there is an effect for WORD of *wh*-word vs. SFP ($\beta = 0.042$, $SE = 0.0022$, $p < 0.001$), indicating that the duration of the *wh*-word *shei2* is longer than that of the SFP *a5* in ISQ contexts. The effects for QUESTION TYPE of RQ+ vs. ISQ ($\beta = -0.018$, $SE = 0.0036$, $p < 0.001$) and of RQ− vs. RQ+ ($\beta = -0.014$, $SE = 0.0032$, $p < 0.001$) are also significant, suggesting that the duration of SFPs is shorter in RQ+ than in ISQ, and that the SFPs are in turn shorter in RQ−, compared to that in RQ+. In addition, both interaction terms are significant (*wh*-word×(RQ+ vs. ISQ): $\beta = 0.033$, $SE = 0.0053$, $p < 0.001$; *wh*-word×(RQ− vs. RQ+): $\beta = 0.017$, $SE = 0.0052$, $p = 0.001$), showing that the duration of the *wh*-word under both RQ− and RQ+ is longer than the duration of the SFP in ISQ.

2.4.2. *Fundamental frequency (F0)*

The F0 of target sentences was estimated with REAPER (`https://github.com/google/REAPER`), a pitch tracker that uses glottal closure instants to calculate F0. We used REAPER's default setting, but lowered the F0 floor from 40 Hz to 20 Hz to improve detection accuracy of low F0 [20].[4] However, as the tracking results were likely to be inaccurate (e.g., with huge F0 jumps) in the regions of creaky voice, we further fitted a robust smoothing spline to each F0 contour spanning the entire utterance, as illustrated in Figure 2.[5] We then sampled F0 values from the fitted spline at an interval of 25 ms between the beginning and the end of the utterance. For some tokens ($n = 17$), the fitted spline produced negative values, due to a sudden drop of F0 typical of creaky voice toward the end of the utterance. For these tokens, we manually checked and corrected F0 using the Pitch objects produced by Praat's cross-correlation periodicity detection algorithm. We then refitted robust smooth splines on the F0 values of these tokens and sampled fitted F0 values from these splines.

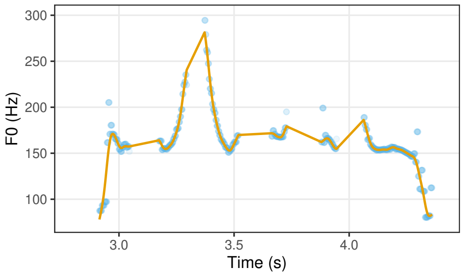

Figure 2: *Raw F0 data (blue points) superimposed with the smoothed estimates (orange line) produced with robust smoothing splines.*

We analyzed these utterance F0 contours with Functional Principal Component Analysis (FPCA) [22], as it offers a way to succinctly quantify global variation in F0. In essence, FPCA treats each F0 contour as a composition of a grand mean curve and a small number of Principal Component (PC) curves (each of which encodes a different deformation of the mean curve) weighted by corresponding PC scores (which can be studied using conventional statistical tests). Here we focus on the first two PC curves (i.e., PC1 and PC2), whose effects on the grand mean F0 curve and corresponding scores for all utterances are shown in Figure 3 and 4 respectively. Focusing on the *wh*-word and the SFP, PC1 coordinates both the F0 height/contour on the *wh*-word and the SFP, while PC2 mainly alters the pitch contour of the SFP. Specifically, PC1 captures the tendency that when the *wh*-word has a lower F0, the SFP has a higher F0, and vice versa. PC2, on the other hand, raises or depresses the F0 contour on the SFP.

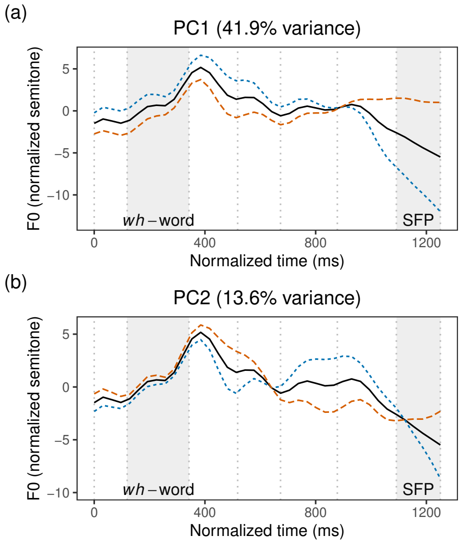

Figure 3: *Each panel shows in solid the grand mean curve $\mu(t)$ and in dashed lines the curves obtained by adding (orange) or subtracting (blue) from $\mu(t)$* **(a)** $\sigma(\text{PC1 scores}) \cdot \text{PC1}(t) = 85.3 \cdot \text{PC1}(t)$ *and* **(b)** $\sigma(\text{PC2 scores}) \cdot \text{PC2}(t) = 48.6 \cdot \text{PC2}(t)$ *respectively, where σ denotes the standard deviation.*

Now that the shape of F0 contours has been parameterized by the first two PC scores, we ask if the question type has an

[3] We also fitted two linear mixed effects models with the same effect structure to the data. However, Q-Q plots suggested a violation of the Gaussian assumption for data points with peripheral values. The results of the linear mixed effects models were otherwise parallel to those of the GAMLSS models.

[4] The default setting of REAPER sampled F0 at intervals of 1 ms, which resulted in many sampling points having the same F0 values. This caused problems for the fitting procedure of the subsequent robust smooth spline, so we added random Gaussian noise ($\mu = 0, \sigma = 0.1$) to break ties.

[5] For this purpose, we used the `qsreg` function from the R package `fields`, version 10.0 [21], with the smoothing parameter λ being $2{\times}10^{-13}$.

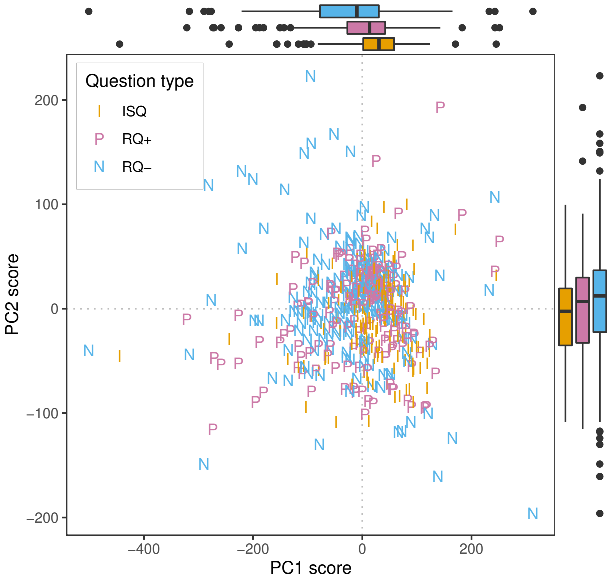

Figure 4: *A scatter plot and box plots of PC1 and PC2 scores of the 561 F0 contours labeled according to the question type each contour belongs to.*

effect in the distributions of these two PC scores. For this purpose, we fitted two GAMLSS model, with the *t* distribution as the conditional distribution for both PC1 and PC2 scores. The fixed effect in both models was QUESTION TYPE (ISQ, RQ+, RQ−; successive difference coded: RQ+ vs. ISQ, RQ− vs. RQ+), and the random effects included random intercepts for participants and items.[6]

For PC1 scores, there is an effect for QUESTION TYPE of RQ+ vs. ISQ ($\beta = -22.3$, $SE = 4.8$, $p < 0.001$), suggesting that, in comparison with ISQ, RQ+ is more likely to have a higher F0 on the *wh*-word but a lower F0 on the SFP. The effect of RQ− vs. RQ+ is also significant ($\beta = -21.6$, $SE = 5.2$, $p < 0.001$), which indicates that RQ− tends to have an even higher F0 on the *wh*-word and an even lower F0 on the SFP than RQ+. For PC2 scores, only QUESTION TYPE of RQ− vs. RQ+ is significant ($\beta = 9.3$, $SE = 3.4$, $p = 0.006$), but not for RQ+ vs. ISQ ($\beta = 3.7$, $SE = 3.1$, $p = 0.23$), indicating that PC2 serves to counteract some of the F0 lowering effect from PC1.

The finding that both RQ− and RQ+ have a lower F0 on the SFP is corroborated by the proportion of utterances where the SFP was realized with a creaky voice across the three question types. We listened to each token and determined if the SFP was perceived to have a creaky quality. The proportion of tokens that contained a creaky SFP, calculated on the basis of participants, is shown in Figure 5, separately for each question type. A generalized linear mixed effects model, with the presence of a creaky SFP as the binary dependent variable and QUESTION TYPE as the fixed effect (ISQ, RQ+, RQ−; successive difference coded: RQ+ vs. ISQ, RQ− vs. RQ+) plus random intercepts for participants and items, points to RQ− having significantly more creaky SFPs than RQ− ($\beta = 0.9$, $SE = 0.24$, $p < 0.001$), which in turn has more creaky SFPs than ISQ ($\beta = 1.3$, $SE = 0.26$, $p < 0.001$).

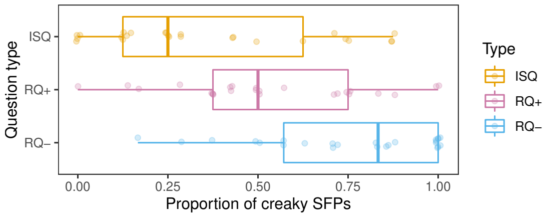

Figure 5: *Proportion of creaky SFPs across the three question types.*

3. Discussion

The results of the experiment are summarized in Table 3:

Table 3: *Summary of experiment results.*

Question type	Utterance duration	SFP duration	F0 on *wh*-word	F0 on SFP
ISQ	shorter	longest	lowest	highest
RQ+	longer	In between ISQ and RQ−		
RQ−	longer	shortest	highest	lowest

First, our results indicate that different question types are prosodically distinct in Mandarin, in terms of duration and F0. Second, our results suggest a link between addressee engagement and prosodic marking. With respect to both duration and F0, ISQs and RQ−s represent the two ends of a scale in each case, with RQ+s being in between. Third, the longer utterance duration of RQ−s and RQ+s, compared to ISQs, seems to reflect the difference in speaker commitment [6]. However, given the exploratory nature of the current study, this potential tie between the subtle semantic characteristics of questions and duration/F0 has to be explicitly examined before a relation can be concluded.

Our results both echo and deviate from the observations reported in previous studies. For instance, our results are consistent with the findings in German that RQ−s tend to be longer than ISQs and that RQ−s end with a falling contour more often than do ISQs [7, 10]. In comparison with Cantonese, even though both languages make a three-way prosodic distinction on the SFP in terms of duration and F0 for the three question types, the patterning of duration and F0 involved in making the distinction is different across the two languages.

4. Conclusion and outlook

In this exploratory study, we show that ISQs, RQ+s, and RQ−s in Mandarin are distinguished in production by duration and F0, which might signal the degree of speaker commitment and addressee engagement associated with the interrogative. In future studies, we plan to further investigate the interaction between SFPs, question types, and prosody.

5. Acknowledgements

This research was supported by the Arts Graduate Student Research Awards from the University of British Columbia. We thank Zhanao Fu for his assistance with translation, Yadong Liu for recording the contexts, and Maxime Tulling for implementing the context rating experiment.

[6] Again, we fitted two linear mixed effects models with the same effect structure. Q-Q plots indicated that data points with more extreme values incurred a violation of the Gaussian assumption. The results, however, showed the same patterns as those of the GAMLSS models.

6. References

[1] H. Rohde, "Rhetorical questions as redundant interrogatives," *San Diego Linguistics Papers*, vol. 2, pp. 134–168, 2006.

[2] I. Caponigro and J. Sprouse, "Rhetorical questions as questions," in *Proceedings of Sinn und Bedeutung 11*, E. Puig-Waldmüller, Ed., Barcelona, Spain, 2007, pp. 121–133.

[3] R. Y.-H. Lo, A. Kiss, and M. A. Tulling, "The prosodic properties of the Cantonese sentence-final particles *aa1* and *aa3* in rhetorical wh-questions," in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escodero, M. Tabain, and P. Warren, Eds., Melbourne, Australia, 2019, pp. 502–506.

[4] C. Petrone and O. Niebuhr, "On the intonation of German intonation questions: The role of the prenuclear region," *Language and Speech*, vol. 57, no. 1, pp. 108–146, 2014.

[5] P. Prieto and J. Borràs-Comes, "Question intonation contours as dynamic epistemic operators," *Natural Language & Linguistic Theory*, vol. 36, no. 2, pp. 563–586, 2018.

[6] J. Heim, "Commitment and engagement: The role of intonation in deriving speech acts," Ph.D. dissertation, University of British Columbia, Vancouver, BC, 2019.

[7] D. Wochner, J. Schlegel, N. Dehé, and B. Braun, "The prosodic marking of rhetorical questions in German," in *INTERSPEECH 2015*, Dresden, Germany, 2015, pp. 987–991.

[8] N. Dehé, B. Braun, and D. Wochner, "The prosody of rhetorical vs. information-seeking questions in Icelandic," in *Proceedings of the 9th International Conference on Speech Prosody 2018*, K. Klessa, J. Bachan, A. Wagner, M. Karpiński, and D. Śledziński, Eds., Poznań, Poland, 2018, pp. 403–407.

[9] J. Neitsch, B. Braun, and N. Dehé, "The role of prosody for the interpretation of rhetorical questions in German," in *Proceedings of the 9th International Conference on Speech Prosody 2018*, K. Klessa, J. Bachan, A. Wagner, M. Karpiński, and D. Śledziński, Eds., Poznań, Poland, 2018, pp. 192–196.

[10] 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, no. 4, pp. 779–807, 2019.

[11] N. Dehé and B. Braun, "The prosody of rhetorical questions in English," *English Language and Linguistics*, pp. 1–29, 2019.

[12] J. Yuan, C. Shih, and G. P. Kochanski, "Comparison of declarative and interrogative intonation in Chinese," in *Proceedings of the International Conference on Speech Prosody 2002*, Aix-en-Provence, France, 2002, pp. 711–714.

[13] X. Zeng, P. Martin, and G. Boulakia, "Tones and intonation in declarative and interrogative sentences in Mandarin," in *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, China, 2004, pp. 225–228.

[14] F. Liu, D. Surendran, and Y. Xu, "Classification of statement and question intonations in Mandarin," in *Proceedings of the 3rd International Conference on Speech Prosody 2006*, R. Hoffmann and H. Mixdorff, Eds., Dresden, Germany, 2006, pp. 232–235.

[15] J. Yuan, "Mechanisms of question intonation in Mandarin," in *Proceedings of the 5th International Symposium on Chinese Spoken Language Processing 2006*, Q. Huo, B. Ma, E.-S. Chng, and H. Li, Eds., Singapore, 2006, pp. 19–30.

[16] B. R. Xu and P. Mok, "Intonation perception of low-pass filtered speech in Mandarin and Cantonese," in *Proceedings of the 3rd International Symposium on Tonal Aspects of Languages*, Nanjing, China, 2012, pp. O2:1–6.

[17] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, pp. 195–203, 2019.

[18] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2019, computer program. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[19] C. Coupé, "Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape," *Frontiers in Psychology*, vol. 9, pp. 513:1–21, 2018.

[20] K. Dorreen, "Fundamental frequency distributions of bilingual speakers in forensic speaker comparison," Master's thesis, University of Canterbury, Christchurch, 2017.

[21] D. Nychka, R. Furrer, J. Paige, and S. Sain, "fields: Tools for spatial data," University Corporation for Atmospheric Research, Boulder, CO, 2017, R package version 10.0. [Online]. Available: https://github.com/NCAR/Fields

[22] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *Journal of Phonetics*, vol. 49, pp. 16–40, 2015.