Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Investigating Mandarin Tone and Focus Prosody Production in Hong Kong Cantonese Speakers

Wenxi Fei, Yu-Yin Hsu

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
wen-xi.fei@connect.polyu.hk, yu-yin.hsu@polyu.edu.hk

Abstract

Second-language (L2) acquisition is influenced by the differences and similarities between a learner's first language (L1) and their target language. Because prior research has shown that Cantonese and Mandarin speakers employ different acoustic strategies to express focus in speech, this study examines how Hong Kong Cantonese (HKC) speakers, who are L2 Mandarin speakers, produce Mandarin focus prosody. Twenty HKC speakers completed a tone identification-production task with 72 monosyllabic Mandarin words used in the main experiment. Based on how well they performed this task, they were divided into two proficiency groups. The members of both groups then performed a speech-production task in which they answered pre-recorded wh-questions, focusing on either a numeral (ANUM), or a noun phrase (ANP). The results showed that the sampled HKC speakers did not use the typically observed HKC focus-marking strategy when producing Mandarin focus, but instead adopted a new acoustic strategy consisting of partial Cantonese focus marking strategy (lengthening) and some post-focus F0 compression, when speakers have a higher level of Mandarin proficiency. The two proficiency groups' acoustic approaches to expressing Mandarin focus differed from each other. These results represent an important contribution to our understanding of how HKC speakers perceive and produce Mandarin tone and prosody, and shed new light on L2 speech acquisition at both the suprasegmental and the sentential levels.

Index Terms: speech production, focus prosody, L2 production, acoustic analysis

1. Introduction

The acquisition of a second language (L2) can be influenced by the differences [1] and similarities [2] between the sound systems of a learner's first language (L1) and the target L2. This view has been extended to investigate L2 intonation by the L2 Intonation Learning Theory [3]; according to this theory, four dimensions are used to categorize L2 intonation learning: the systemic, realizational, semantic, and frequency dimensions. However, studies of the interactions among those dimensions remain lacking. Previous research has shown that Cantonese and Mandarin speakers' acoustic strategies for conveying prosodic focus in their native speech differ [4, 5, 6]. However, few, if any, studies have examined how Cantonese speakers produce focus prosody in Mandarin; thus, it remains unclear whether they transfer their L1 knowledge, or successfully learn to use Mandarin's focus-marking strategies. Accordingly, this study investigates the production of Mandarin focus prosody by Hong Kong Cantonese (HKC) speakers who are L2 learners of Mandarin. Its results can provide insights into the process of L2 acquisition on both the suprasegmental and sentential levels.

One of the greatest challenges faced by HKC speakers learning Mandarin is mastering Mandarin tones, particularly Tone 2 (T2) and Tone 3 (T3), which diverge considerably from their native language's phonological system. Learners of Mandarin often encounter difficulties in distinguishing between the rising (T2) and falling-rising (T3) tones because of their acoustic similarity [7, 8]. Moreover, T3 may sometimes become almost identical to T2 when two T3 syllables are adjacent – a phenomenon referred to as *tone three sandhi* (T3S). For example, the word "tiger" in Mandarin contains two T3 syllables (老lao3-虎hu3), and its first T3 syllable, 老lao3, should be pronounced like T2 in Mandarin. Cantonese speakers tend to neutralize Mandarin T3 to T2, even in single-word production [7]. Such difficulties in distinguishing between the two tones may be attributed, in part, to L1 to L2 language transfer. That is, unlike Mandarin, Cantonese has rising tones but no falling-rising tone [9], and most words that are pronounced with T3 in Mandarin have a rising tone in Cantonese [10]. Additionally, Cantonese does not employ tone sandhi in connected speech [9].

Cross-linguistic variations in prosody often present challenges to the learning and processing of language [11]. However, prosody is rarely a core topic in language curricula or language-learning research. Previous studies of Mandarin speakers have shown that they tend to enlarge the fundamental frequency (F0), duration, and intensity of focused syllables while maintaining their lexical tonal contours [5, 12]. Cantonese speakers, on the other hand, primarily use duration to express focus [4, 6, 12]. Thus, the most divergent feature of Mandarin vs. Cantonese focus prosody comprises F0 changes. Nevertheless, no research has yet investigated how Cantonese speakers produce Mandarin focus prosody.

Cantonese lacks certain prosodic features found in Mandarin: notably contour tone (T3), T3S, and F0 for focus marking. Our main hypothesis, based on the language-transfer theory [1] (cf. the version extended to L2 intonation [3]), is that missing features in one's L1 may pose difficulties in L2 learning, and could be overlooked in L2 speech production. Specifically, we expected that HKC speakers would struggle to produce T3 more than other Mandarin tones, not perform T3S as consistently as Mandarin speakers typically do, and apply focus-marking strategies from their native language, resulting in a lack of F0 rise in the focused units of their Mandarin speech. After testing this hypothesis, we will discuss the L1-L2 transfer of acoustic features and prosodic phrasing in the context of focus.

2. Methods

2.1. Participants

Twenty participants (16 female; age $M = 22.20$, $SD = 3.72$) were recruited from The Hong Kong Polytechnic University. All were native Cantonese speakers and self-assessed their Mandarin proficiency as intermediate, i.e., 4 or 5 points on a seven-point Likert scale ranging from 1= very little knowledge to 7 = native fluency. All were right-handed and had no history

10.21437/SpeechProsody.2024-77

of language-related impairments. Participants provided written consent before the experiment and received compensation of HK$100 (about US$13) upon completing it.

2.2. Design and stimuli

All items were shown to the participants in Traditional Chinese characters. Prior to the experiment, we evaluated participants' proficiency in Mandarin tones through their identification and production of monosyllabic tones. The word list consisted of 72 frequently used Mandarin monosyllabic words, i.e., 18 words with each of the four tones (4 tones × 18 words). Four additional words were added for each tone for practice.

The main part of our study, guided by a discourse-production paradigm, used a question sentence and an answer sentence in each trial [12]. All 72 words from the proficiency assessment were then used to form three-syllable phrases in the same tone, consisting of a numeral (Num), a classifier (Class), and a noun (N), such as the phrase 三san1-隻zhi1-貓mao1 (literally, "three-classifier-cat") in T1. These noun phrases (NPs) were then used to form the natural two-utterance dialogues in which the target sentences – containing such an NP in the subject position – served as the responses to a *wh*-question (Table 1). To avoid utterance–initial boundary effects, the target NPs were preceded by a three-syllable adverbial phrase and followed by a two-syllable verb phrase and a sentence-final particle. All sentences were simple active sentences consisting of nine syllables, and the character immediately before the target NPs was consistently T1, to ensure that all the target phrases had the same tone environment. None of the sentences contained a syntactic focus structure or potential focus-bearing words (i.e., words similar to English *only*, *exactly*, etc.) that could impede the use of prosody to express focus. We manipulated the target phrases' focus span using short leading questions to elicit one of two types of focus: 1) answers to a question about an unknown numeral (ANUM), and 2) answers to a question about an unknown NP (ANP). Each participant completed six practice trials followed by 144 trials (4 tones × 18 words × 2 foci).

Table 1: *An example trial in Tone 1 (Focus span is underlined).*

Focus	Leading question	Target sentence
ANUM	In the park, how many cats are full of food? 公園中幾隻貓吃飽了？	In the park, three cats are full of food. 公園中三隻貓吃飽了。
ANP	In the park, what is full of food? 公園中什麼吃飽了？	In the park, three cats are full of food. 公園中三隻貓吃飽了。

2.3. Procedure

The experiment was built using *PsychoPy* v.3.0 [13]. Participants were tested individually in a sound-attenuated speech lab and seated approximately 28 inches in front of a 24.5-inch computer screen. An AT2020 microphone was used to record their speech. The recordings were saved as .wav files at a sampling rate of 44.1 kHz with 32 bits per sample.

Participants initially filled out a language-background questionnaire and signed our consent form. Then, they were asked to use the numbers 1-4 on the keyboard to identify, within five seconds, the tone type of a character shown on the screen. The accuracy of their tone identifications was recorded. Next, they were asked to read the target word aloud. Each trial had a six-second recording duration, allowing the participants to read the word aloud at least twice.

After a three-minute break, the participants were instructed to proceed to the main experiment. The *wh*-questions were pre-recorded in the same lab by a female native Mandarin speaker from northern China. Each was about three seconds long and played in .wav audio format. In each trial, participants listened to a pre-recorded question, saw a response sentence on the screen, and were required to read the response aloud twice. Their oral productions were recorded, and the whole experiment lasted approximately 45 minutes.

2.4. Data analysis

First, we assessed the participants' proficiency in Mandarin lexical tones in terms of their tone-identification and production accuracy. The former was measured based on the above-mentioned pre-test responses, and the latter was judged by three native Mandarin speakers. There was a substantial agreement between the raters' judgments as indicated by Fleiss' kappa, $K = .752, p < .0001$. A two-thirds majority of the judges was deemed sufficient to yield a final result. Next, based on the data distribution, the participants were divided into two subgroups: a high-proficiency group having 12 participants with perception and production accuracy rates both above 80%, and a low-proficiency group (*N* = 8) whose members all had perception accuracy rates below 80% and production accuracy rates of about 80%. The grouping was because the data distribution is roughly delimited by 80%. In our subsequent analyses, we were thus able to take account of proficiency as a factor affecting acoustic features. Production accuracy rates were calculated by tone, and raw accuracy data for each tone were compared using logistic regression.

Second, to examine HKC speakers' application of T3S, we identified the surface patterns of T3 targets. Three native Mandarin raters (two PhDs and one with an MA, all in Linguistics) examined each syllable of the T3 NPs one by one on a binary scale, with 2 representing tokens similar to canonical Mandarin T2 and 3 representing tokens similar to canonical Mandarin T3. Their identification was based only on the tonal properties and disregarded the segmental information. Fleiss' kappa showed that there was a moderate agreement between the raters' judgments, $K = .581, p < .0001$. Agreement on patterns was reached by at least two of the listeners in each case. We excluded 1.28% of the trials (*N* = 37) from the analysis due to mispronunciation of the underlying tones; for example, the classifier "bundles" (捆kun3) was mispronounced as Tone 4. Then, we summarized the remaining T3 target NPs' surface patterns.

Third, to study how the sampled HKC speakers produced Mandarin focus prosody, we analyzed the intensity, duration, and F0 of each syllable in the target NPs. For this acoustic analysis, the data were segmented using *Praat* v.6.4.01 [14]. Syllable boundaries were determined using both visual (the waveform and spectrogram) and auditory information. *ProsodyPro* v.5.7.8.7 [15] was used to extract time-normalized F0, mean F0, duration, and intensity of each syllable. For time-normalized F0, ten F0 measurements were taken in each labelled interval, yielding 30 data points for each NP. All F0 data were converted into semitones using individual speakers' mean F0 values as references (i.e., $ST = 12\log_2\left(\frac{F0}{\text{mean}F0}\right)$).

Mean intensity, duration, and F0 were then analyzed by fitting the generalized linear mixed-effects models using the *lme4* package [16] in R v.4.3.0 [17]. In those models, focus types, syllables, proficiency, and tones were considered the main vari-

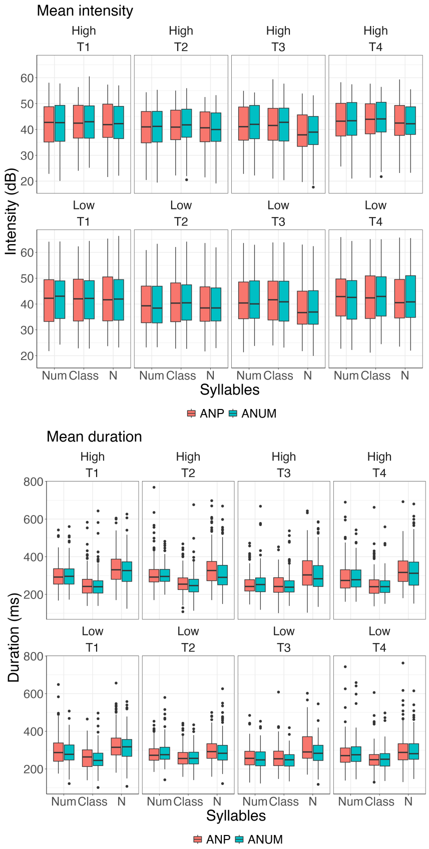

Figure 1: *Mean intensity (top) and mean duration (bottom) by focus type, proficiency, and tone. (ANP: answers to a question of wh-NP; ANUM: answers to a question of wh-numeral.*

ables, and items and participants were the random variables. Time-normalized F0 was analyzed for contour differences using smoothing spline ANOVA (SSANOVA, [18]). The shading on the resulting splines shows 95% Bayesian confidence intervals; therefore, if two shaded contours do not overlap, they are considered statistically significantly different from each other.

3. Results

3.1. Accuracy of tone identification and patterns of T3S

The participants' accuracy at identifying monosyllabic words' tone was significantly lower for T3 words (77.8%), than for other tones: i.e., T1 (94.7%; $\beta = 2.81, z = 4.36, p < .001$), T2 (93.3%; $\beta = 2.65, z = 4.17, p < .001$), and T4 (89.2%; $\beta = 1.32, z = 2.39, p = .083$). The high-proficiency group reached 95.9% overall accuracy, with T1 at 100%, T2 at 96.8%, T3 at 88.9%, and T4 at 97.9%. The overall performance of the low-proficiency group was 78.1%, with T1 at 86.8%, T2 at 88.2%, T3 at 61.1%, and T4 at 76.4%. In sum, T3 was difficult for all participants, while for the low-proficiency group, T4 was also challenging.

Our analysis of the frequency of T3S patterns was based on 683 trials. The most common patterns were “223” (393 trials, 57.5%) and “323” (276 trials, 40.4%). There were no clear differences among the patterns used for different focus types, but the high-proficiency group tended to express three adjacent T3 words with the “223” pattern (36.5%), or the “323” pattern (23%). Only 1% of the trials exhibited a “233” pattern, mostly for ANUM.

The low-proficiency group was equally likely to produce T3 NPs as either “223” (144 trials, 21%) or “323” (119 trials, 17%). Other observed patterns included “222” (1 trial) for ANP and “322” (1 trial) for ANUM in the high-proficiency group, whereas “233” (1 trial) and “333” (2 trials), both for ANUM, were produced by the low-proficiency group. These results suggest that the high-proficiency group could produce T3 sandhi in a manner similar to native Mandarin speakers, albeit in only around one-third of the relevant trials. The low-proficiency group's respective usage of the “223” and “323” patterns did not differ significantly in frequency. Another interesting finding was that focusing on number alone (ANUM) seemed to prompt some participants to produce more pattern variations than the whole NP focus condition did.

3.2. Acoustic analyses

For intensity (Fig. 1), we found no significant effects of focus type ($\chi^2(1) = 0.38, p = .54$), syllable ($\chi^2(2) = 0.44, p = .804$), or proficiency groups ($\chi^2(1) = 0.003, p = .987$). For duration (Fig. 1), we found no significant effects of focus ($\chi^2(1) = 0.18, p = .671$) or proficiency ($\chi^2(2) = 0.001, p = .972$). It was noteworthy that the duration of nouns (the phrasal end syllable) was consistently longer than that of other words within the NP, regardless of focus types, and that in certain tones, the number syllable was longer than the subsequent classifier syllable ($\chi^2(2) = 219.19, p < .001$).The only significant contrast between focus types was the duration of T2 nouns in the high-proficiency group. Specifically, that group's members produced shorter nouns in ANUM than in ANP ($\beta = -0.09, z = -3.26, p = .013$), in line with Mandarin post-focus reduction. There were no other differences in the classifier-noun syllables between the ANUM and ANP conditions ($p > .05$).

We found no significant differences in overall time-normalized F0 or mean F0 by focus type, proficiency, or syllable ($ps > .05$). These results might suggest that our sampled speakers continued using the Cantonese prosodic strategies to express focus; Nonetheless, some interesting findings emerged from our SSANOVA analysis of contrasts in F0 contours. As can be seen in Figure 2, while the overall F0 contours seem similar across the ANUM and ANP conditions, differences between them emerged in certain tones. That is, in the high-proficiency group, we observed a compression of T3 nouns' F0 curves in ANUM relative to those in ANP. A similar tendency was found among the T4 nouns, in line with Mandarin post-focus F0 reduction, although in a later syllable. Additionally, when dealing with T2, the high-proficiency group produced significantly higher F0 of the classifier in ANUM than in ANP. Similarly, for T3, the low-proficiency group produced significantly higher F0 of the classifier in ANUM than in ANP. These patterns were different from the typical focus prosody reported for both Cantonese and Mandarin.

4. Discussion and Conclusion

This study examined the acoustic patterns of HKC speakers' production of Mandarin tones, tone sandhi, and focus prosody. Our results shed light on L1-L2 phonetic and phonological transfer, as well as on some aspects of general prosodic phrasing. First, at the phonetic-phonological level, we found that our sampled HKC speakers faced challenges producing T3 and T3S

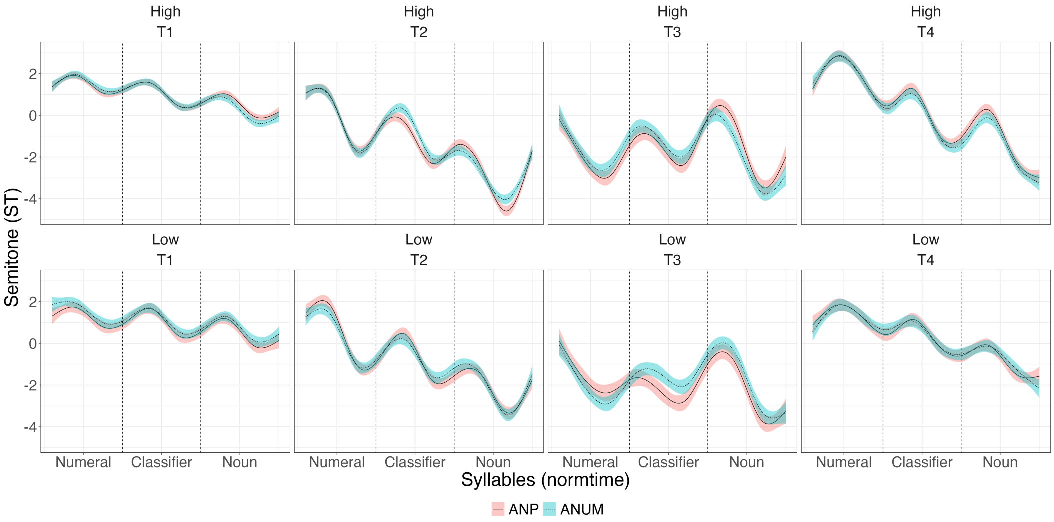

Figure 2: *Smoothing-spline analysis of variance (SSANOVA) of F0 contours by focus, proficiency, and tone.*

in Mandarin. This provides some evidence from the sentential and suprasegmental levels in support of findings from previous segmental phonetic research [7, 8], and disconfirms the hypothesis based on the basic language-transfer theory [1], that speakers tend to transfer knowledge from their L1 when producing an L2 that differs from it phonologically. However, production accuracy varied based on the L2 speakers' proficiency levels. In our high-proficiency group, although T3 was more challenging than other tones, the accuracy rate with it remained high (88.9%), and with T2, an even higher accuracy was achieved (93.3%). They were most likely (36.5%) to perform T3 sandhi with the "223" pattern, like native Mandarin speakers. The low-proficiency group, in contrast, correctly identified T3 only 61.1% of the time and produced T3 NPs as "223" or "323" with similar frequencies independent of focus types. These findings suggest that specific categories in L2 (i.e., T3 and T3S in Mandarin) are rather challenging for low-proficiency L2 learners, but can be acquired as language proficiency increases. Future studies can explore whether specific training in phonological patterns can improve the L2 learners' speech performance.

Second, our participants exhibited a distinctive focus-marking pattern for Mandarin speech that was not fully in line with their native or typical Mandarin ones. We found no effect of focus or syllable on mean intensity, which is consistent with the Cantonese focus-marking strategies reported in studies using similar materials [4, 12]. HKC focus units tend to be lengthened [6], whereas Mandarin tends to both lengthen the focus units and shorten the post-focus units [5]. Nonetheless, our HKC speakers generally exhibited no lengthening differences between types of focus in most cases. The high-proficiency group shortened nouns to the post-focus duration reduction typically found in Mandarin — but only T2 nouns, and only in the ANUM condition. Moreover, most F0 contours across the focus types were similar, especially so in T1 and T4. This, coupled with our duration results, suggests that HKC speakers primarily applied Cantonese focus marking strategies in their L2 speech. However, we also found some indications that the high-proficiency learners had acquired some Mandarin focus knowledge. That is, albeit only in T2 and T3 conditions, those participants produced post-focus F0 compression to express ANUM: a phenomenon typically observed only in Mandarin, not in Cantonese [5, 6]. Overall, T2 and T3 pose challenges to L2 learners, but the distinctiveness of contour tone and the acoustic properties of T2 and T3 seem to also provide additional cues that assist learners in recognizing these special features of Mandarin.

Some of our interesting findings may be related to HKC speakers' prosodic phrasing of Mandarin. Regardless of focus type, the participants lengthened the nouns in the target noun phrases. This resembles the focus prosody patterns of Taiwan Mandarin [19], in which both numerals and nouns are longer when being part of the focus, but the duration of classifiers is not affected. Since both Cantonese and Mandarin are languages of head-prominence in prosodic phrases [20], our findings of noun lengthening in the NPs support this view. Thus, it also contributes to the general understanding of prosodic learning in L2 by considering the interactions of semantics and prosody [3].

However, our findings also show interesting implications for the interaction of syntax and prosody. The classifier in Cantonese has been argued to play a more prominent role in marking its individuality compared with Mandarin, in which the classifier only facilitates counting [21]. That indicates a potential knowledge development in their L2 production. Moreover, our participants tended to produce a "323" pattern for T3 NPs, whereas the most common pattern used in Mandarin is "223".This prevalence of the "323" pattern could have been caused by the time limit of speech leading to a more T2-like production [8, 22] of the middle syllable. However, the fact that the high-proficiency group exhibited post-focus F0 compression only at the beginning of nouns, and not immediately after the focused numeral syllable (i.e., the classifier in ANUM), may suggest a disyllabic preference on the part of these speakers. This is not unexpected, insofar as HKC also prefers disyllabic feet, and the final foot in Cantonese tends to be non-monosyllabic [23, 24].

In conclusion, we have shown that HKC learners of Mandarin could accurately identify tones, but developed focus-marking strategies that differed from those of both their L1 and their L2. As such, this study provides important sentential evidence about L2 tone acquisition and insights into cross-lingual prosodic phrasing. Future studies can further explore the systems of L2 prosodic organization and what kind of prosodic training can be effective for L2 learners.

5. Acknowledgements

This project is supported by the General Research Fund of the Research Grants Council of Hong Kong (GRF: 15612922). We would also like to thank the anonymous reviewers for their feedback and suggestions. Special thanks to Zhihong Chen for some of the technical support.

6. References

[1] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[2] C. T. Best, "A direct realist view of cross-language speech perception," *Speech perception and linguistic experience*, vol. 171, 1995.

[3] I. Mennen, "Beyond segments: Towards a l2 intonation learning theory," in *Prosody and language in contact: L2 acquisition, attrition and languages in multilingual situations*. Springer, 2015, pp. 171–188.

[4] Y.-Y. Hsu, A. Xu, H. Ngai *et al.*, "Focus prosody in cantonese and teochew noun phrases," in *Proceedings of Speech Prosody 2018*, 2018, pp. 961–965.

[5] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[6] W. L. Wu and Y. Xu, "Prosodic focus in hong kong cantonese without post-focus compression," in *Speech prosody 2010-fifth international conference*, 2010.

[7] Y.-C. Hao, "Second language acquisition of mandarin chinese tones by tonal and non-tonal language speakers," *Journal of Phonetics*, vol. 40, no. 2, pp. 269–279, 2012.

[8] Y.-J. Lin and Y.-Y. Hsu, "Whether and how mandarin sandhied tone 3 and underlying tone 2 differ in terms of vowel quality?" in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, 2018.

[9] M. Y. Chen, *Tone sandhi: Patterns across Chinese dialects*. Cambridge University Press, 2000, vol. 92.

[10] 張凌, "香港人學習普通話的聲調偏誤之聲學分析," 中國語文通訊, vol. 100, no. 1, pp. 31–39, 2021.

[11] A. Arnhold, B. Braun, and M. Romero, "Aren't prosody and syntax marking bias in questions?" *Language and Speech*, vol. 64, no. 1, pp. 141–180, 2021.

[12] Y.-Y. Hsu and A. Xu, "Focus acoustics and prosodic organization in hong kong cantonese and taiwan mandarin," in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 706–710.

[13] 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.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2023.

[15] Y. Xu, "Prosodypro—a tool for large-scale systematic prosody analysis," in *Laboratoire Parole et Langage, France*, 2013.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Soft.*, vol. 67, no. 1, pp. 1–48, 2015.

[17] R. C. Team, "R: a language and environment for statistical computing," *R Foundation for Statistical Computing*, 2023. [Online]. Available: https://www.R-project.org/

[18] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[19] Y.-Y. Hsu and J. German, "Prosodic organization and focus realization in taiwan mandarin," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, 2018.

[20] S.-A. Jun, "Prosodic typology: By prominence type, word prosody, and macro-rhythm," in *Prosodic typology II*. Oxford University Press, 2014, pp. 520–539.

[21] L. L.-S. Cheng and R. Sybesma, "The syntactic structure of noun phrases," in *The Handbook of Chinese Linguistics*. Wiley Online Library, 2014, pp. 248–274.

[22] A. C. Yu, H. Lee, and J. Lee, "Variability in perceived duration: pitch dynamics and vowel quality," in *Fourth International Symposium on Tonal Aspects of Languages*, 2014.

[23] C. Perry, M.-K. Kan, S. Matthews, and R. K.-S. Wong, "Syntactic ambiguity resolution and the prosodic foot: Cross-language differences," *Applied psycholinguistics*, vol. 27, no. 3, pp. 301–333, 2006.

[24] W. Y. P. Wong, "Syllable fusion in hong kong cantonese connected speech," Ph.D. dissertation, The Ohio State University, 2006.