INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Tonal coarticulation as a cue for the upcoming prosodic boundary

Jianjing Kuang[1], *May Pik Yu Chan*[1], *Nari Rhee*[1]

[1]University of Pennsylvania, USA

kuangj@sas.upenn.edu, pikyu@sas.upenn.edu, nrhee@sas.upenn.edu

Abstract

It has been established that pitch reset or the lack of tonal coarticulation is a salient cue for the beginning of a large prosodic domain, however, it is yet unclear whether tonal coarticulation can be an informative cue for the end of a prosodic domain. We examined this question with two continuous speech corpora of Mandarin, and both expert and crowd-sourced perceptual annotations were used. The FPCA model of the holistic tonal contours shows that the carry-over effect of the preceding tone is significantly affected by the strength of the following boundaries. Stronger carry-over effects are associated with the end of larger prosodic boundaries. Moreover, machine learning classification shows that the fine-grained tonal coarticulation patterns are salient cues for predicting larger prosodic boundaries. This result is further validated by crowd-sourced boundary perceptual ratings from human listeners. This study has important implications for the understanding of prosodic phrasing.

Index Terms: tonal coarticulation, boundary perception, prosody

1. Introduction

In connected speech, the phonetic realization of segmental and suprasegmental features is influenced by the neighboring sounds. This mechanism is known as coarticulation. For languages with lexical tones, the phonetic realization of the pitch contours of the current tone contains the information from the preceding tone (i.e. carry-over effect) and the following tone (i.e. anticipatory effect). Although coarticulation is bidirectional, for tonal articulation, carry-over effect is generally much stronger than anticipatory effect [1, 2, 3, 4]. Crucially, tone coarticulation is perceptually salient, and is important for tone identification in connected speech [1, 5].

Coarticulation is sensitive to prosodic contexts, such as prosodic boundaries. It has been well documented that neighboring segments are more strongly coarticulated within the same prosodic domain, but minimally coarticulated when the two adjacent segments cross major prosodic boundaries [6, 7, 8]. This effect also applies to suprasegmental features. As for pitch, post-boundary pitch reset has been found to be a salient cue for major prosodic boundaries cross-linguistically [9, 10, 11]. Pitch reset effect for lexical tones is more complex, and involves two aspects: 1) the pitch declination trend (i.e. the gradual trend of the reduction and lowering of overall pitch range) ends, and the post-boundary/domain initial tones are produced with much a higher and larger pitch range [12, 13, 14]; 2) large domain-initial tones have little carry-over coarticulation effects from preceding tones [15, 14, 16]. Importantly, the degree of carry-over effect across prosodic break is conditioned by the strength of prosodic boundaries [17]. Therefore, previous studies have established that the lack of carry-over tone coarticulation is a salient cue to indicate the beginning of a large prosodic domain.

However, it is still unclear whether tonal coarticulation is a cue for the upcoming prosodic boundary. That is, whether the shape and timing of the tonal contours can indicate the end of the prosodic domain. Pitch cues are known to be sensitive to domain-final positions. Phrase-final pitch reduction and lowering are fairly common across languages [11]. For tonal languages, these pitch scaling effects are shown as the reduction of the overall pitch range and lower pitch height. Moreover, because of the final lengthening effect, pitch contours are stretched in a longer time window. However, it is still unclear whether the shape and timing of pitch contours of the tones also vary according to the strength of the following boundary, in addition to the duration and pitch scaling effects. Some evidence from vowel coarticulation suggest that coarticulation effects can be independent from final lengthening effects [6]. Sun and Shih's paper also suggests that boundary-conditioned tonal coarticulation can be independent from the final lengthening effect [18].

Taken together, this paper will test whether the carry-over coarticulation effect from the preceding tone vary as a function of the relative strength of the following boundary in production, and whether the detailed tonal coarticulatory cues are useful in perceiving the strength of the upcoming prosodic boundaries. We will address these questions with two continuous speech corpora, focusing on tonal coarticulation of the two pre-boundary syllables.

2. Methods

2.1. Corpora

The first corpus used in this study is COSPRO [14, 19]. Specifically, COSPRO_01 and COSPRO_02 were used here, as they are phonetically balanced such that the effects of tonal categories can be tested. Both corpora consist of read sentences designed to include all possible syllables in Mandarin (about 1300), the most frequently used 2- to 4-syllable lexical words, all possible segmental combinations and concatenations, and all possible tonal combinations. The two corpora have the same design and only differ in the number of sentences and speakers. COSPRO_01 contains speech data from six Taiwan Mandarin speakers (3M, 3F), and each speaker read 599 short discourses (between 1 and 180 characters in length); COSPRO_02 includes 100 sentences from COSPRO_01, but contains recordings from 90 speakers (40F, 50M). Therefore, the two corpora share the same nature and thus were combined in the analysis. All the recordings were made in soundproof chambers and were digitized at 16 kHz with 16-bit resolution.

The second corpus we used is a spoken corpus of the Chi-

10.21437/Interspeech.2023-1033

nese Tree Bank [20, 13]. The texts were news articles from Chinese Tree Bank 9.0 [21], which contained 132,076 sentences, of which, 12.46% are included in the spoken corpus. The speech corpus was both segmented and syntactically parsed. All speakers were native Mandarin speakers who had achieved Class 2 Level 1 or better on the national standard Mandarin proficiency test. The recordings were made at Shanghai JiaoTong University in a sound-treated booth. The recordings had a sampling frequency of 44.1kHz and a sample depth of 16-bit. A subset of the corpus (22 sentences) was chosen for crowd-source annotation of boundaries, which included one male and one female speaker reading two different passages. These speakers were chosen because their recordings had fewer disfluencies, had consistent pacing throughout their readings, and were expressive. Only sentences of around 30 seconds or less were included.

2.2. Perceptual annotations of prosodic boundaries

For COSPRO, prosodic boundaries were annotated by trained phoneticians. The break index system in COSPRO corpora follows Tseng's Mandarin ToBI model [14]. B2 refers to "below the prosodic word level," including both syllable and prosodic word boundaries. B3 is defined as prosodic phrase boundary. Two larger prosodic boundaries annotated in the corpus are B4 and B5. B4 is defined as the "breath group boundary", and B5 is the "prosodic group boundary", both associated with final lengthening and weakening.

For the Chinese Tree Bank corpus, the perceptual judgments of the prosodic boundaries were obtained through a crowd-sourcing annotation task. We recruited 29 native Mandarin speakers (18-35 years old; 18 female) from university student communities. They were asked to listen to the 22 chosen sentences over an online experience conducted in Qualtrics. A boundary detection task similar to that in [22] and the Rapid Prosodic Transcription tasks in [23, 24] was used. The selected sentences were presented one at a time, with the audio being played automatically once the participant entered the trial, the corresponding sentence transcript was displayed simultaneously. Participants were asked to select where they believed there were boundaries within the sentence. Participants were not able to select the beginning or the end of the sentence, and they could replay the audio as many times as they wanted. Three sample questions were presented at the beginning to ensure that participants understood how to use the experimental interface, involving the same sentence read with different prosodic focus and structures. No further definition of 'boundaries' was given, allowing participants to interpret the instructions in their own way. There were a total of 458 potential boundaries for participants to rate. In the upcoming analysis, we quantified boundary strength as the rate of boundary perception (number of boundary responses divided by the total number of participants).

2.3. Pitch Contour Measurements

For the COSPRO corpus, F0 was measured with the STRAIGHT algorithm [25] in VoiceSauce [26]. F0 values are z-score normalized to individual speakers' mean F0. F0 contours are time-normalized to 15 equal intervals. Therefore, duration information is factored out from the contour analysis. Functional Principal Component Analysis (*fda* package in *R* [27]) was used to parameterize F0 contours. This method is able to capture the detailed timing and shape variation of the F0 contours and represent the information in a low-dimensional space. Before running functional data analysis, B-splines was first applied to transform a contour into a smooth continuous function of time [28]. In order to find the optimal values for two parameters (k; lamda), generalized cross-validation (GCV), proposed by [27] was employed. Then Principal Component Analysis was applied to the smoothed functional data to generate the mean curve and principal component curves, as well as weights for the principle component curves. As the regular PCA, the rank of the PCs reflects the decreasing percentage of variance in the input data that the PCs explain.

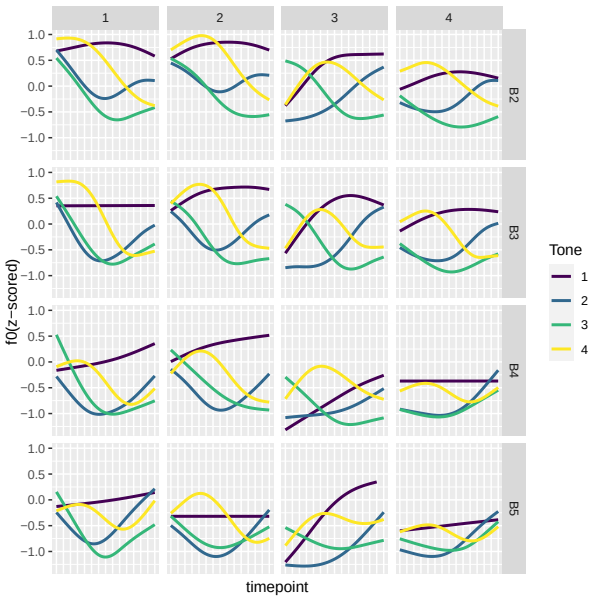

Figure 1: *Tonal coarticulation effects conditioned by following boundary (rows) and preceding tones (columns).*

For the Chinese Tree Bank, as the corpus was segmented at the word and not the syllable level, only disyllabic words entered the acoustic analysis. For each disyllabic word, pitch extraction was done using a Praat script at 11 equidistant points from the mid-point of the word to the end of the word (inclusive). This allowed us to infer a time-normalized tonal contour of the second syllable of each disyllabic word. F0 measurements were z-scored by speaker.

3. Results

3.1. Boundary-conditioned carry-over coarticulation

Starting with the COSPRO analysis, tonal contours as a function of the following break and preceding tone is plotted in Figure 1. The first four FPCs are illustrated in Figure 2: PC1 represents the overall pitch height, PC2 represents the slope of rising or falling, PC3 is the extent of dipping or convex shape, and PC4 is related to more complex shape with double peaks/valleys. The first four FPCs account for 97.2% of the variance.

Linear mixed-effect models were used to examine the effects of prosodic breaks and preceding tones on the shape of the current tones. We fit FPC1 to FPC4 each in a separate model with the main effects of the following break, the target tone and the preceding tone, as well as their 2-way and 3-way interactions. Across all four FPCs' models, there are significant main effects of tone (FPC1: $F = 112.55, Pr(> F) < 0.001$; FPC2: $F = 249.48, Pr(> F) < 0.001$; FPC3: $F = 267.84, Pr(> F) < 0.001$; FPC4: $F = 13.86, Pr(> F) < 0.001$), the fol-

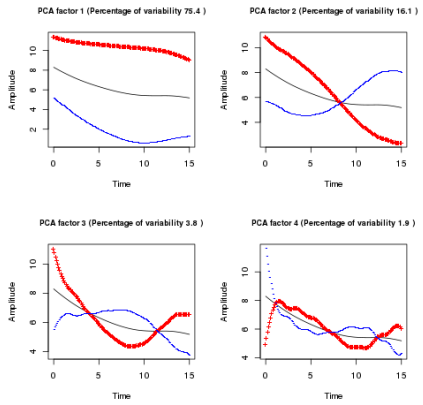

Figure 2: *The first four Functional Principal Components for Mandarin tones.*

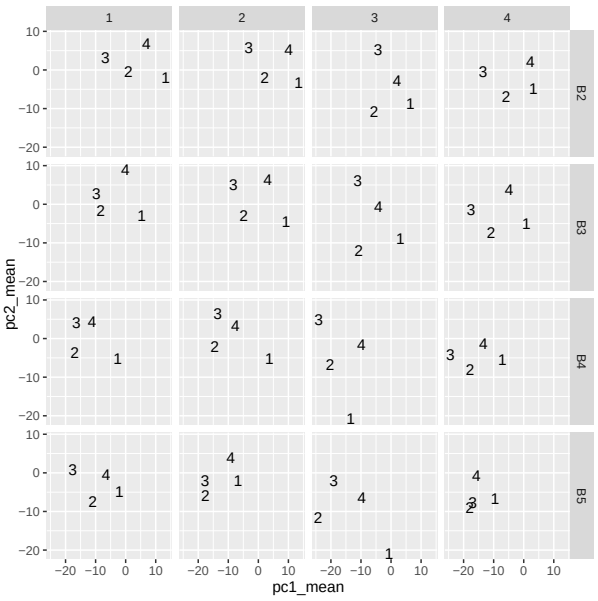

Figure 3: *FPC1 and FPC2 conditioned by following boundary (rows) and preceding tones (columns).*

lowing break (FPC1: $F = 226.83, Pr(>F) < 0.001$; FPC2: $F = 38.27, Pr(>F) < 0.001$; FPC3: $F = 4.23, Pr(>F) = 0.005$; FPC4: $F = 5.23, Pr(>F) = 0.001$), preceding tones (FPC1: $F = 33.23, Pr(>F) < 0.001$; FPC2: $F = 68.34, Pr(>F) < 0.001$; FPC3: $F = 41.78, Pr(>F) < 0.001$; FPC4: $F = 5.71, Pr(>F) < 0.001$) and interaction effects between the following break and tone (FPC1: $F = 3.22, Pr(>F) < 0.001$; FPC2: $F = 9.91, Pr(>F) < 0.001$; FPC3: $F = 2.38, Pr(>F) = 0.011$; FPC4: $F = 4.31, Pr(>F) < 0.001$). More crucially, there are also significant interactions between tone, the preceding tone and the following boundary across all four models (FPC1: $F = 2.03, Pr(>F) = 0.001$; FPC2: $F = 2.12, Pr(>F) < 0.001$; FPC3: $F = 1.59, Pr(>F) = 0.027$; FPC4: $F = 1.57, Pr(>F) = 0.031$), indicating that the effects of preceding tones are modulated by the strength of the upcoming/following boundary. PC1 is about the overall pitch height. As expected, the overall pitch height of the tones is lower when they are followed by a larger break (e.g., B4 or B5). Also unsurprisingly, tones are lower (especially the onset portion) when preceded by low offset tones (Tone4 and Tone3). Positive FPC2 values indicate falling, negative FPC2 values indicate rising, and close to zero line means level. Therefore, as illustrated in Figures 3 and 1, the shape of the contours is conditioned by the strength of breaks. For example, Tone 3 has a falling contour before smaller boundaries, but primarily has a rising contour before B4 and B5. As mentioned, there is also a significant effect of preceding tones, and significant interaction effects between the tone, following break and preceding tone. For example, Tone 2 rises more when preceded by Tone 3 and Tone 4, the low offset tones; by contrast, Tone 4 falls more when preceded by Tone 1 and Tone 2, the high offset tones. For FPC3, positive values indicate a more dipping shape, and negative values indicate a more convex shape. Like FPC2, FPC3 is also significantly conditioned by tone, the following break and the preceding tone. As shown in Figure 4, Tone 2 and Tone 3 have a deeper dipping contour when preceded by Tone 1 and Tone 2; Tone 1 and Tone 4 have a convex shape when preceded by Tone 3 and Tone 4. The extent of dipping and convex is reduced when the tones are followed by larger boundaries. As shown in Figure 4, even though FPC4 only accounts for a small variance, it is particularly sensitive to the strength of the following break.

Overall, these results demonstrate that the carry-over coarticulation is highly fine-grained and sensitive to the strength of the following break. In general, contours are more complex in a reduced pitch range due to greater coarticulation. Then, we can ask whether this information is useful for predicting/perceiving the upcoming boundaries.

3.2. Predicting upcoming break with tonal coarticulatory cues

The Random Forest (RF) algorithm [29] was used to examine the relative contribution of the timing and shape of tonal contours in classifying different levels of following breaks (i.e., B2-B5). The interactions between pitch contour shapes (PFCs) and preceding tonal categories were included in the model to evaluate the tonal coarticulation effects. Ten repeats of 10-fold cross-validation (CV) were used to estimate the performance of the model. This procedure automatically chooses tuning parameters associated with optimal model performance. Our final candidate model was selected based on the best accuracy for determining what percentage of probability is accepted in classifying a case. To visualize the relative importance generated by the model, we utilized the model-agnostic approach to provide an interpretation of our optimal RF model. The relative contribution of each input variable was estimated by calculating the increase in the model's prediction error after permuting the variable. Friedman's H-statistics was employed to measure the weights of the interaction effects. Overall, the model achieves 92% accuracy, suggesting that the cues included are informative. As shown in Figure 5, the relative weight of the preceding tone is greater for larger boundaries, suggesting that tonal coarticulation from the preceding tone is an important cue to signify large upcoming boundaries such as B4 and B5.

To cross-validate these results with crowd-sourcing annotated data, we modeled crowd-sourced boundary perception data from the Chinese Tree Bank to examine how F0 contours and their word timepoints interact with tonal information. We use a generalized additive model (GAM) to see how perceived

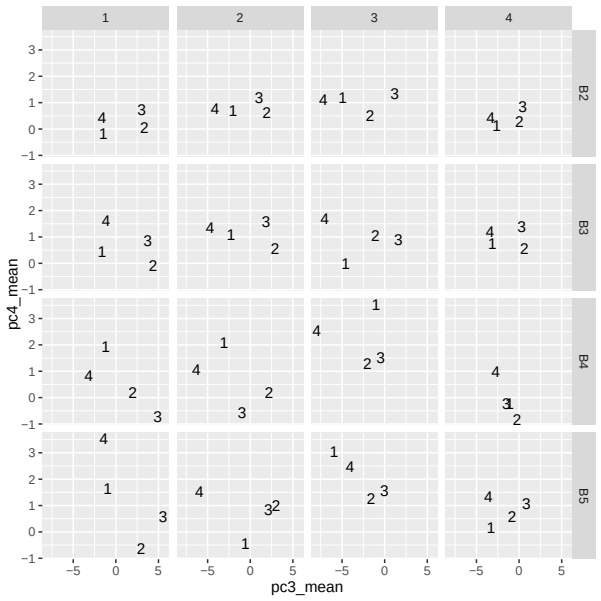

Figure 4: *FPC3 and FPC4 conditioned by following boundary (rows) and preceding tones (columns).*

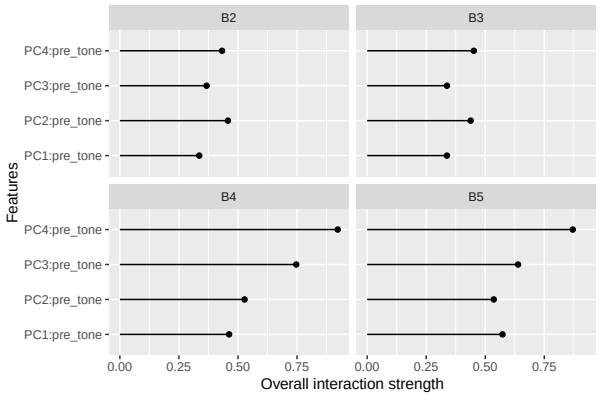

Figure 5: *Weight plot from the Random Forest model.*

boundary strength varies as a function of the by speaker z-scored F0 contour, the preceding tone and the target tone. We start by creating a categorical predictor `PrecTrgt` which acts as a parametric term that represents the interaction between the offset of the preceding tone (High: T1, T2 vs Low: T3, T4) and the onset of the target tone (High: T1, T4 vs Low: T2, T3), resulting in 4 levels. We then included in our model a smooth function with word timepoint and F0 interactions for each level of `PrecTrgt`. In this model, we use a tensor smooth for word timepoint (ranges from 0-10) and F0 interactions as they vary on their own scales (`te(WordTimepoint, F0, by = PrecTrgt)`). The model formula is therefore `gam(FollowingBoundaryPercent ~ PrecTrgt + te(WordTimepoint, F0, by = PrecTrgt)`. Smoothing was done via the restricted maximum likelihood method. Results show significant main effects of PrecTrgt for High-Low ($beta = -0.05, p < 0.001$) and Low-High ($beta = 0.04, p = 0.003$) combinations relative to High-High. Results for the smooth terms are summarized in Table 1.

Table 1: *Summary of the smooth terms for the GAM Model.*

Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
te(time,f0):HH	8.90	11.30	4.48	< 0.0001
te(time,f0):HL	9.37	11.64	6.98	< 0.0001
te(time,f0):LH	11.11	13.89	6.04	< 0.0001
te(time,f0):LL	4.21	4.76	4.29	0.0008

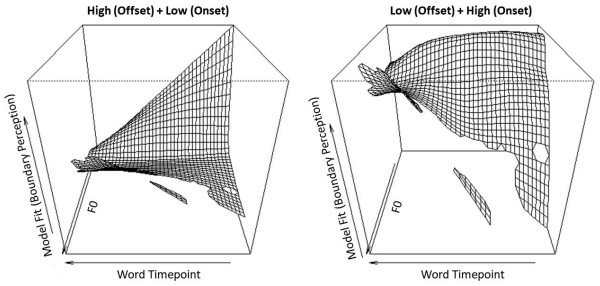

Figure 6: *Crowd-sourced boundary perception ratings as a function of F0 and word timepoint, by conflicting offset of the preceding tone and onset of the target tone (i.e. HL and LH.)*

Figure 6 illustrates the model fit for two of the smoothed terms, `te(time,f0):HL` and `te(time,f0):LH`, which are the conflicting combinations of preceding tonal offset and target tonal onset. The x-axis shows the word timepoint, the y-axis shows the z-scored F0, and the z-axis shows the fitted perceived boundary strength. As can be seen in the figure, the relationship between the F0 contours and boundary perception are complex. At different degrees of perceived boundary strength, the timing and the shape of the tonal contours modulate in not only F0 height but also the complexity of the contour.

4. Discussions and Conclusions

This paper investigated whether tonal coarticulation from the preceding tone is a salient cue in predicting/perceiving the strength of the following boundary. The Functional Principal Component Analysis on tonal contours in COSPRO corpus demonstrated that the tonal coarticulation patterns of the preceding tone vary as a function of the strength of the following prosodic boundaries. As expected, tones followed by larger boundaries are produced with a reduced pitch range, and overall lower pitch height (c.f. FPC1 effect). A crucial and novel finding of this study is that the timing and shape of the coarticulated tonal contours (FPC2-FPC4) are also subject to change, and an overall stronger carry-over effect from the preceding tone is found for tones followed by larger prosodic boundaries (B4 and B5 in the corpus). This result is consistent with the notion that weak prosodic positions are more vulnerable to coarticulation. Moreover, both machine learning classification and crowd-sourced human boundary perception ratings suggest that tone coarticulation is a salient cue for predicting/perceiving the strength of the upcoming boundary, and the weight of tonal coarticulation is greater for the larger boundaries. Together with the pitch reset/coarticulation reset effect associated with domain-initial tones, the extent of tonal coarticulation is very informative for boundary strength. These results significantly advance our knowledge of the interaction between tonal coarticulation and prosodic phrasing, and have important implications for speech planning.

5. References

[1] Y. Xu, "Production and perception of coarticulated tones," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2240–2253, 1994.

[2] ——, "Asymmetry in contextual tonal variation in Mandarin," *Advances in the study of Chinese language processing*, vol. 1, pp. 383–396, 1994.

[3] J. Gandour, S. Potisuk, and S. Dechongkit, "Tonal coarticulation in Thai," *Journal of Phonetics*, vol. 22, no. 4, pp. 477–492, 1994.

[4] G. Kochanski, C. Shih, and H. Jing, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–645, 2003.

[5] S.-h. Peng, "Production and perception of Taiwanese tones in different tonal and prosodic contexts," *Journal of Phonetics*, vol. 25, no. 3, pp. 371–400, 1997.

[6] T. Cho, "Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English," *Journal of Phonetics*, vol. 32, no. 2, pp. 141–176, 2004.

[7] H.-h. Pan, "The effects of prosodic boundaries on nasality in Taiwan Min," *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3755–3769, 2007.

[8] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *Journal of Phonetics*, vol. 26, no. 2, pp. 173–199, 1998.

[9] P. Keating, T. Cho, C. Fougeron, and C.-S. Hsu, "Domain-initial articulatory strengthening in four languages," *Phonetic interpretation: Papers in laboratory phonology VI*, pp. 143–161, 2004.

[10] T. Cho, J. M. McQueen, and E. A. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," *Journal of Phonetics*, vol. 35, no. 2, pp. 210–243, 2007.

[11] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[12] H.-h. Pan and Y.-h. Tai, "Boundaries and tonal articulation in Taiwanese Min," in *Proceedings of Speech Prosody*, vol. 51, 2006.

[13] J. Kuang, M. P. Y. Chan, and N. Rhee, "The effects of syntactic and acoustic cues on the perception of prosodic boundaries," *Speech Prosody*, 2022.

[14] C.-y. Tseng, C.-H. Chang, and Z.-y. Su, "Investigating f0 reset and range in relation to fluent speech prosody hierarchy," *Technical Acoustics*, vol. 24, pp. 279–284, 2005.

[15] S.-h. Peng, M. K. Chan, C.-y. Tseng, T. Huang, O. J. Lee, and M. E. Beckman, "Towards a pan-Mandarin system for prosodic transcription," *Prosodic typology: The phonology of intonation and phrasing*, pp. 230–270, 2005.

[16] Q. Li and Y. Chen, "An acoustic study of contextual tonal variation in tianjin Mandarin," *Journal of Phonetics*, vol. 54, pp. 123–150, 2016.

[17] W. Lai and J. Kuang, "Prosodic grouping in Chinese trisyllabic structures by multiple cues–tone coarticulation, tone sandhi and consonant lenition," *Tonal Aspects of Languages 2016*, pp. 157–161, 2016.

[18] Y. Sun and C. Shih, "Boundary-conditioned anticipatory tonal coarticulation in standard Mandarin," *Journal of Phonetics*, vol. 84, p. 101018, 2021.

[19] C.-y. Tseng, Y.-C. Cheng, and C. Chang, "Sinica cospro and toolkit—corpora and platform of Mandarin Chinese fluent speech," in *Oriental COCOSDA*, 2005, pp. 23–28.

[20] J. Kuang, M. P. Y. Chan, N. Rhee, M. Liberman, and H. Ding, "The mapping between syntactic and prosodic phrasing in English and Mandarin," *Proc. Interspeech 2022*, pp. 3443–3447, 2022.

[21] N. Xue, X. Zhang, Z. Jiang, M. Palmer, F. Xia, F.-D. Chiou, and M. Chang, *Chinese Treebank 9.0*. Philadelphia: Linguistic Data Consortium, 2016.

[22] A. Buxó-Lugo and D. G. Watson, "Evidence for the influence of syntax on prosodic parsing," *Journal of Memory and Language*, vol. 90, pp. 1–13, 2016.

[23] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[24] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," *Computer Speech & Language*, vol. 45, pp. 300–325, 2017.

[25] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3933–3936.

[26] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *Proceedings of the ICPhS XVII*, 2011, pp. 1846–1849.

[27] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd*. New York, NY: Springer Verlag, 2005.

[28] C. de Boor, *A Practical Guide to Splines, Revised Edition*. New York: Springer, 2001.

[29] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/