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

Tonal patterns of the Mandarin Third Tone Sandhi produced by Japanese-speaking L2 learners

Tong Shu, Zhiqiang Zhu, Peggy Mok

The Chinese University of Hong Kong
tongshu@link.cuhk.edu.hk, zhiqiangzhu@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

Abstract

While extensive research has been conducted on the L2 perception and production of Mandarin lexical tones, the higher prosodic patterns, such as tone sandhi, remain less explored. This study examined the L2 production of the Mandarin Third Tone Sandhi (T3 Sandhi) by Japanese speakers at two Mandarin proficiency levels (intermediate and advanced). The participants read disyllabic stimuli with all possible tonal combinations of the T3 Sandhi. Different from the common approach which mainly relied on native speakers' categorization of L2 learners' tone production, we adopted a data-driven approach using hierarchical clustering to identify the distinct tonal patterns for each T3 Sandhi combination within each group.

The results revealed a complex interplay of various factors influencing L2 production of the Mandarin T3 Sandhi, such as L1 Japanese pitch accent patterns, phonetic motivation of different T3 Sandhi, and L2 Mandarin tone inventory. The suspected influence from L1 Japanese pitch accent patterns is noted in intermediate-level learners, but advanced learners can overcome such influence. In both L2 learner groups, we found over-generalization of T3 Sandhi. In general, our study showed the transfer of L1 phonological processing to L2 tone sandhi production at an earlier stage of L2 acquisition.

Index Terms: Mandarin Third Tone Sandhi, L2 tone acquisition, Japanese pitch accent

1. Introduction

Mandarin has four lexical tones: T1: high-level (55); T2: mid-rising (35); T3: low-dipping (214); T4: high-falling (51). Figure 1 illustrates the two context-conditioned tone sandhi processes involving Mandarin T3 [1]. (1) The Full T3 Sandhi: when T3 is followed by another T3, the first T3 changes to a T2-like rising contour, i.e., 214 → 35 /_T3. (2) The Half T3 Sandhi: when T3 is followed by T1, T2, or T4, it changes to a low-falling pitch contour, as if its pitch contour is "halved" compared to its citation form, i.e., 214 → 21 /_T1/T2/T4.

Zhang & Lai (2010) have argued that the Half T3 Sandhi was more phonetically motivated than the Full T3 Sandhi because it is natural to simplify a complex pitch contour in connected speech, while the reason for T3 changing into a T2-like rising contour in the Full T3 Sandhi is less clear, and therefore, less phonetically motivated. Importantly, this difference in phonetic motivation may affect productivity. Their study has revealed that for native speakers, the pitch contour of T3 in the Half T3 Sandhi context did not significantly differ between real words and wug words, while in the Full T3 Sandhi context, the pitch contour of T3 varied, with the T3 in wug words resembled more its citation form, showing a later turning point and lower pitch height. This suggests that the Half T3 Sandhi, being more phonetically motivated, is applied more consistently and is more productive than the Full T3 Sandhi.

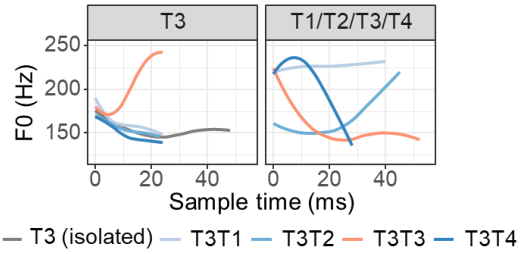

Figure 1: *Illustration of the pitch contour of T3 (left panel) when followed by different lexical tones (right panel).*

The difference in phonetic motivation has two implications for L2 acquisition of the Mandarin T3 Sandhi. First, the Half T3 Sandhi, being more phonetically transparent, may be easier to acquire than the Full T3 Sandhi. Second, we may anticipate a similar productivity difference between the two T3 Sandhi processes in L2 learners as observed in native speakers. However, previous studies on L2 learners have shown mixed results for both directions [2-5]. This could be due to factors such as small sample size, L1 influence, L2 pedagogy, and so on, all warranting further investigation.

In addition, the limited range of L1 backgrounds tested in previous studies (English in [2-4], Cantonese in [4], Korean in [5]) is insufficient to evaluate the influence of L1 on the acquisition of L2 prosodic processes at a higher level. Previous studies have found that Cantonese speakers were better at fine-grained pitch manipulation than English speakers were, which may be attributed to their L1 experience with lexical tones [4]. In our study, we further explore the L2 production of the Mandarin T3 Sandhi by Japanese speakers. We selected Japanese speakers because, like Mandarin, Japanese encodes pitch at the lexical level, albeit in the form of lexical pitch accent rather than tone [6]. More importantly, both Japanese lexical pitch accent and Mandarin T3 Sandhi operate within the domain of word, involving computational processes to determine the surface pitch form. This parallel allows us to investigate the potential transferability of L1 phonological pitch processing to a different L2 pitch-related phonological pattern.

In Japanese, the pitch contour of a word is determined by the presence and the location of the accented mora. This accented mora carries a high tone, which extends to any preceding morae (except the first mora which is assigned a low tone when it is not accented), while morae that follow are assigned low tones [6]. Note that the initial lowering of the first mora does not apply when the first two morae form a heavy syllable, resulting in a HH instead of a LH sequence [7-8]. The

10.21437/SpeechProsody.2024-23

pitch accent is realized as an abrupt pitch fall immediately after the accented mora. Typically, a word has at most one accented mora, and this pattern carries over to compound words [6]. As a result, the pitch contour of a Japanese word is generally flat, with a single peak being the norm. While predicting precisely how Japanese speakers will produce the Mandarin T3 Sandhi is difficult, it is reasonable to expect some influence from their L1 pitch accent pattern, such as a tendency towards a flatter overall pitch contour for the entire word.

Additionally, the L2 production of the Mandarin T3 Sandhi must be considered within the broader context of the acquisition of individual lexical tones. A prevalent difficulty for L2 learners of Mandarin is distinguishing between T2 and T3 in perception and production [9-11]. Consequently, due to the different lexical tone inventories of L2 learners compared with native speakers, i.e., a less distinct boundary between T2 and T3, we may expect variations in the conditions that trigger the T3 Sandhi. For instance, L2 learners who struggle to differentiate between T2 and T3 may not only apply the Full T3 Sandhi in a T3T3 context but also extend it to a T3T2 context.

Finally, we also explored a new approach to evaluating L2 tone production. A common practice in previous studies is to rely on native speakers' auditory judgment to categorize each syllable produced by the L2 learners into one of the four lexical tone categories [2-3]. However, it is very common for L2 learners to produce tones that are too ambiguous to be classified as any of the four categories. Also, relying on the categorical perception of native speakers may introduce biases stemming from their preconceived notions about L1 tones.

To address the above issues, our study examined the L2 production of the Mandarin T3 Sandhi by Japanese speakers with different proficiency levels, focusing on the following questions: (1) How does L1 phonological processing influence L2 Mandarin T3 Sandhi production? (2) Is the Half T3 Sandhi easier and more productive than the Full T3 Sandhi? Different from native judgment, we used hierarchical clustering to determine the number of distinct tonal patterns present in each T3 Sandhi combination produced by the Japanese L2 learners.

2. Method

2.1. Participants

We examined data from three groups of speakers: intermediate (IJ), advanced (AJ) Japanese learners of Mandarin, and a baseline group of native Mandarin (NM) speakers. The IJ group participants were recruited from a Japanese university where they were enrolled in a Mandarin course (*N* = 8; 6f, 2m; *Mean age* = 19.25 yr, *SD* = 1.03; *Mean learning duration* = 0.97 yr, *SD* = 0.69). The AJ group participants were mainly living in and recruited in Beijing (*N* = 8; 8f; *Mean age* = 38.75 yr, *SD* = 9.25; *Mean learning duration* = 14 yr, *SD* = 7.52). The NM participants were university students born and raised in Beijing (*N* = 6; 6f; *Mean age* = 21 yr, *SD* = 2.28).

2.2. Materials

We selected 60 disyllabic T3 Sandhi words, including all possible tonal combinations of T3 Sandhi. The number of stimuli in each combination is: T3T1 (12), T3T2 (12), T3T3 (24), T3T4 (12). Half of the stimuli in each condition were real words and the other half were wug words. The stimuli were adapted from Zhang and Lai (2010) and Yang (2015). The wug words consisted of individually existing syllables in Mandarin, but their combinations do not exist. We carefully selected the real word stimuli and the syllables used in the wug word stimuli to ensure that they were of high frequency and introduced early in L2 Mandarin teaching, matching our participant's proficiency.

2.3. Procedure

We presented the participants with a randomized word list on PowerPoint Slides and asked them to read the words aloud. The stimuli were presented as simplified Chinese characters. Notably, *pinyin* annotations were provided only for the second syllable, which is essential for the participants to determine the applicable T3 Sandhi rule. *Pinyin* for the first syllable was intentionally omitted to obscure the purpose of the experiment and prevent any potential bias in the participants' responses. After reading the entire list once, participants were asked to repeat the process twice more so that each word was read aloud three times. We collected 60 words × 3 repetitions × 22 participants = 3960 tokens in total.

2.4. Applying hierarchical clustering

First, we used ProsodyPro [12] to extract the F0 values of 10 equal-distant timepoints for the rime of each syllable. Therefore, there were 20 measure points for each disyllabic token. We manually added the missing pulses in ProsodyPro. Then, we converted the raw F0 values of each token into z-scores for each speaker using their grand means.

Then, to identify the number of distinct tonal patterns for each T3 Sandhi combination within each group, we applied hierarchical clustering to the normalized F0 values at the 20 time points for all tokens per group and tonal combination. Initially, dendrograms were used to suggest the potential minimum number of clusters. However, this method often failed to yield meaningful groupings. This limitation arose from the principle of hierarchical clustering, which organizes data into a specified number of clusters by mathematically minimizing within-cluster distances and maximizing between-cluster distances. This can inadvertently group tokens with similar within-cluster distances but distinct overall pitch contours, especially when the specified number of clusters is small.

To mitigate this problem, increasing the number of clusters can improve the granularity, making it more likely that tokens within each cluster share similar pitch trajectories. Therefore, we began with the smallest number of clusters, two, and visually inspected the pitch contours within each cluster. If we observed multiple distinct pitch trajectories grouped together under the current number of clusters, we increased the number of specified clusters. This process of incremental adjustment continued until the pitch contours within each cluster were mostly similar. For the IJ group, which displayed higher variability in patterns for each T3 Sandhi combination, up to 25 clusters were sometimes necessary to achieve satisfactory within-cluster similarity in terms of pitch contour.

However, a large number of distinct clusters is undesirable, as it can lead to redundancy. After being able to clearly identify the patterns within each cluster, we reduced redundancy by merging clusters that were linguistically similar—those with comparable pitch trajectories that might have been separated due to differences in the overall pitch height or the pitch gap between syllables. Clusters that have less than 10 observations and hard to merge with any other clusters were grouped into "Uncategorized".

This two-step approach, first increasing the number of clusters to ensure homogeneity in each cluster and then merging clusters based on linguistic similarity, was essential. It provides a nuanced solution to the inherent limitation of hierarchical clustering, which can sometimes group data with distinct pitch trajectories based on the mathematical maximization of between-cluster distances and minimization of within-cluster distances. The number of distinct clusters identified in each group for T3T1, T3T2, T3T3, and T3T4 were: IJ: 4, 6, 7, 3; AJ: 5, 5, 6, 2; NM: 1, 2, 3, 1, respectively (more details of the whole process can be accessed via the following link: https://drive.google.com/drive/folders/1KhAhjKTld5YQMo3vBOBtsc70JpnQrfv-?usp=sharing).

3. Results

3.1. Tonal patterns of T3 Sandhi in different groups

After determining the distinct tonal patterns for each tonal combination within each group, we visualized them using generalized additive models (GAM) by using the "mgcv" package [13] in R. Considering native speakers' patterns for each combination as a benchmark, tonal patterns by the two L2 learner groups that resembled those of native speaker patterns were labeled "correct", while all others were classified as incorrect. For the incorrect patterns observed in the AJ and IJ groups, we thought of potential phonological explanations for each and named them accordingly. Patterns that did not clearly align with any identified process were categorized as "Others". The major patterns identified in each group are summarized in Figure 2. To conserve space, we only displayed the patterns found in real words, as the patterns in real words and wug words were similar. Additionally, we focused on showcasing patterns associated with potential phonological processes, while the "Others" category was omitted to prevent cluttering the figure (visit the supplementary link for the full figure).

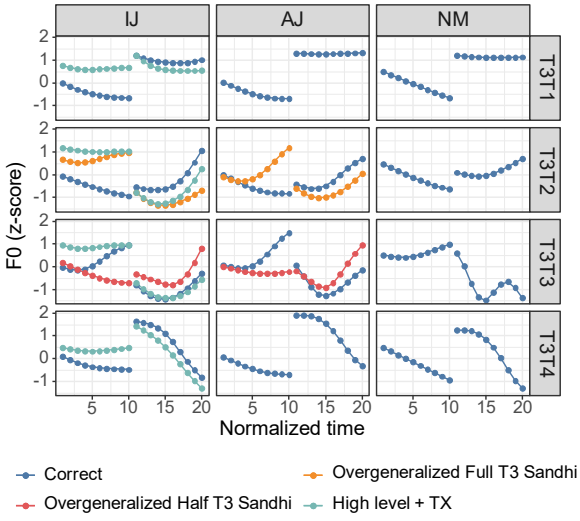

Figure 2: *T3 Sandhi patterns by different groups. Normalized time 10 represents the syllable boundary.*

The NM group's T3 Sandhi production aligns with the well-documented patterns: low-falling after T1/T2/T4 (Half T3 Sandhi) and mid-rising after T3 (Full T3 Sandhi). Note that all groups occasionally produce the final T3 in the correct T3T3 combinations as a low-falling contour, which is also an acceptable variant. However, only the instances with a low-dipping final T3 were displayed in Figure 2 to avoid clutter (visit the supplementary link for the full figure).

Both the IJ and the AJ groups could produce T3T1, T3T2, T3T3, and T3T4 combinations with native-like pitch patterns. A prevalent error unique to the IJ group, while absent in the AJ group, was to produce the initial T3 as a high-level tone across all tonal combinations. Despite this deviation on the initial T3, the tone of the second syllable was correctly produced. Thus, this error pattern was labeled "High level + TX."

For the T3T2 and T3T3 combinations, both L2 learner groups exhibited similar error patterns. In the T3T2 combinations, learners produced a rising initial T3 followed by a low-dipping final T3. This pattern closely mirrored the Full T3 Sandhi, which should be applied in T3T3 combinations and was therefore termed "Overgeneralized Full T3 Sandhi". Likewise, for T3T3 combinations where the Full T3 Sandhi should apply, an error pattern closely resembling the Half T3 Sandhi was observed. Learners realized the initial T3 as a low-falling pitch contour, akin to the initial T3 in the T3T1, T3T2, and T3T4 combinations, while the final T3 retained its citation form. Thus, we named this error as "Overgeneralized Half T3 Sandhi".

3.2. Effects of L2 proficiency, lexical status, and tonal combinations on the probability of correct production

Figure 3 further shows the proportion of each tonal pattern in the IJ and AJ groups in real and wug words. We fitted a mixed-effect logistic regression model using the "lme4" package [14] in R to examine the effects of Group (IJ, AJ), Lexical status (Real, Wug), and Combination (T3T1, T3T2, T3T3, T3T4) on the log odds of correct production. The reference level is Group = IJ, Lexical status = Real, Combination = T3T1.

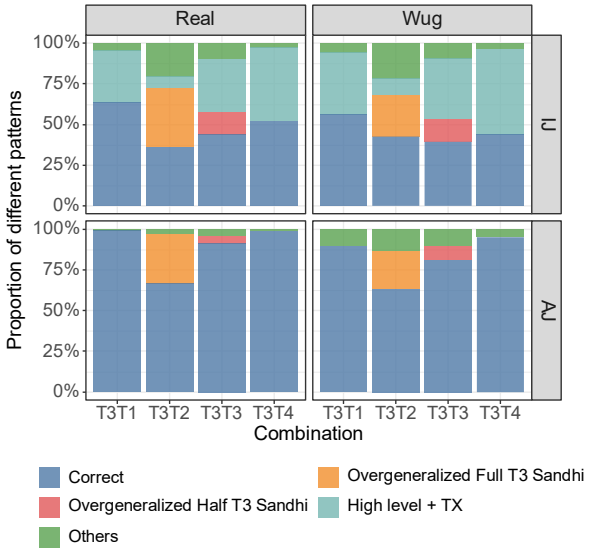

Figure 3: *Proportion of different tonal patterns for each T3 Sandhi combination in the IJ and AJ groups.*

The results showed that wug words were associated with a decrease in the log odds of correct production compared to real words, though not statistically significant (β = -0.38. *SE* = 0.28, p = 0.17), indicating that lexical status did not influence the probability of correct production in the IJ group. When examining tonal combinations, T3T2, T3T3, T3T4 were all significantly associated with a decrease in the log odds of correct production, compared to T3T1, with the largest decrease

observed for T3T2 (β = -1.46, SE = 0.29, $p < 0.001$; β = -1.04, SE = 0.24, $p < 0.001$; β = -0.58, SE = 0.28, p = 0.03). This suggested that the four tonal combinations posed different difficulties for the IJ group, with T3T1 being the easiest.

The AJ group had significantly increased log odds of correct production compared to the IJ group, holding other variables constant (β = 5.28, SE = 1.29, p = < 0.001), suggesting that the AJ group generally outperformed the IJ group. Also, the interaction term GroupAJ:LexicalstatusWug was significant (β = -2.5, SE = 1.07, p = 0.02), suggesting that while the AJ group generally demonstrated a higher accuracy, this advantage was attenuated for wug words. In addition, the significant interaction term GroupAJ:CombinationT3T2 (β = -3.20, SE = 1.03, p = 0.002) suggested that the AJ group's increased odds of correct production were not as pronounced for the T3T2 combination as for the other tonal combinations.

Post-hoc pairwise comparisons with Benjamini-Hochberg (BH) adjustment further revealed that both groups found T3T2 and T3T3 to be the most challenging combinations. For the AJ group, T3T2 was even more difficult than T3T3.

3.3. Distribution of each error pattern in real vs. wug words

Then, to examine the distribution of error patterns in real vs. wug words for each tonal combination, several chi-square tests were conducted for both the IJ and AJ groups across different tonal combinations. For the IJ group, the analysis revealed no significant differences in the distribution of each error pattern between real vs. wug words for any of the combinations tested.

In contrast, the AJ group displayed significant variation in the distribution of error patterns between real and wug words for the T3T1 (χ^2(1, N = 288) = 12.97, p = 0.002), T3T2 (χ^2(2, N = 275) = 11.53, p = 0.003), and T3T3(χ^2(2, N = 568) = 12.80, p < 0.001). Post-hoc analysis revealed that the "Others" error pattern occurred significantly more frequently in wug words than in real words for these combinations. In both groups, although there were observable tendencies for the "Overgeneralized Full T3 Sandhi" error to occur more in real words and the "Overgeneralized Half T3 Sandhi" error to occur more in wug words, these differences did not reach statistical significance. To sum up, in the AJ group, error patterns not conforming to clear phonological processes occurred more frequently in wug words.

4. Discussion

This study examined the tonal patterns of the T3 Sandhi in L2 Mandarin produced by two groups of Japanese speakers with intermediate and advanced proficiency. First, concerning the influence of L1 phonological processing, we found that the IJ group frequently produced the initial T3 as a T1-like high-level pitch contour, irrespective of the tonal combination, while the second syllable was often produced correctly. This "High level + TX" pattern in the IJ group may reveal the influence of L1 Japanese pitch accent patterns in a sense that the second syllable may be treated as the "accented" syllable so that its tonal specifications are realized, while the initial syllable was unaccented and assigned high tones throughout, similar to Japanese pitch accent patterns. Moreover, the initial lowering rule does not seem to apply here. Otherwise, we would see a LH pattern in the first syllable rather than a HH pattern. This is probably because 80% of the initial syllables used in our stimuli correspond to heavy syllables (e.g., CVi, CVN, etc., see [15]) in Japanese, where this rule does not apply [7-8]. This "High level + TX" pattern was not found in English speakers [2-3], suggesting that it is potentially unique to Japanese speakers. However, this pattern was not reported in previous studies examining Japanese speakers' production of other disyllabic Mandarin tone combinations [9]. We hypothesize that our study's presentation method of stimuli may have inadvertently opened the window for the L1 pitch accent rule to intervene. Stimuli were presented to the participants in Chinese characters without pinyin annotations for the first syllable but with annotations for the second syllable to ensure an accurate environment for tone sandhi rules to apply. However, this may visually "accentuate" the second syllable, arousing a Japanese pitch-accent-like production. This "High level + TX" error is not found in the AJ group, suggesting that advanced learners could overcome the influence of L1 phonological processing.

Another focus of this study was whether the Half T3 Sandhi is easier and more productive than the Full T3 Sandhi due to stronger phonetic motivation. Our findings showed that both learner groups demonstrated higher accuracy for T3T1 and T3T4 combinations, which require the Half T3 Sandhi, compared to the Full T3 Sandhi required in T3T3 combinations. However, the accuracy in T3T2 combinations did not align with this pattern, as they were as challenging as T3T3 combinations. The most common error for T3T2 was a Full T3 Sandhi-like pattern. This is probably related to the T2-T3 confusion in L2 learners. As a result of a fused lexical tone category for T2-T3, the domain for the Full T3 Sandhi was also extended, indicating an interplay between individual L2 lexical tone categories and L2 tone sandhi application. In addition, in the T3T3 combination, we also found overgeneralized application of the Half T3 Sandhi, suggesting an interaction between the two phonological processes in L2 learners. Excluding the exceptional T3T2 combination, the higher accuracy rates of T3T1 and T3T4 supported the claim that the Half T3 Sandhi is easier than the Full T3 Sandhi based on the phonetic motivation.

Regarding whether the Half T3 Sandhi is more productive than Full T3 Sandhi, we investigated the influence of lexical status on accuracy rates and the occurrence of error patterns in real vs. wug words. Our findings did not find any supporting evidence for the hypothesized higher productivity of the Half T3 Sandhi in either the IJ or the AJ group. We only observed a general influence of lexical status in the AJ group: wug words had lower accuracy than real words, and wug words showed more random errors for both types of T3 Sandhi. This project is currently in progress, and we are actively analyzing data from an expanded cohort of participants to explore this issue further.

Finally, our study incorporated hierarchical clustering as a novel machine-learning approach to analyze L2 tonal production, bypassing traditional reliance on native speaker judgments. However, the method required refinement as hierarchical clustering sometimes grouped pitch contours that were mathematically close but distinct in pitch trajectories. To address this, we initially increased cluster granularity and then merged linguistically similar clusters, a somewhat convoluted process which also involved human judgments. This indicates a need for improved clustering techniques to minimize labor and subjective decision-making. Future research could consider decomposing continuous F0 measurements into discrete features like pitch height and curvature before clustering. In our study, the accuracy of the tonal patterns produced by the L2 learners was assessed by visual comparison with native speaker's patterns. However, to what extent they also sound natural and correct needs further validation by native judges. A combined approach using machine learning and native ears may offer the most comprehensive analysis.

5. References

[1] J. Zhang and Y. Lai, "Testing the role of phonetic knowledge in Mandarin tone sandhi," *Phonology*, vol. 27, no. 1, pp. 153–201, 2010, doi: 10.1017/S0952675710000060.

[2] C. Yang, "Acquisition of Mandarin Tone 3 sandhi Interaction of phonology, phonetics, and pedagogy," in *The Acquisition of L2 Mandarin Prosody : From experimental studies to pedagogical practice*, John Benjamins, 2016.

[3] H. Zhang, "The effect of theoretical assumptions on pedagogical methods: a case study of second language Chinese tones," *Int. J. Appl. Linguist. (United Kingdom)*, vol. 27, no. 2, pp. 363–382, 2017, doi: 10.1111/ijal.12132.

[4] S. Chen, Y. J. He, R. Wayland, Y. K. Yang, B. Li, and C. W. Yuen, "Mechanisms of tone sandhi rule application by tonal and non-tonal non-native speakers," *Speech Commun.*, vol. 115, pp. 67–77, Dec. 2019, doi: 10.1016/j.specom.2019.10.008.

[5] Z. Qin, "The Second-Language Productivity of Two Mandarin Tone Sandhi Patterns," *Speech Commun.*, vol. 138, no. January, pp. 98–109, 2022, doi: 10.1016/j.specom.2022.02.009.

[6] S. Kawahara, "The phonology of Japanese accent," in *Handbook of Japanese phonetics and phonology*, M. Shibatani and T. Kageyama, Eds. De Gruyter, 2015, pp. 445–492.

[7] S. Haraguchi, *A Theory of stress and accent*. Dordrecht: Foris, 1991.

[8] T. J. Vance, *An introduction to Japanese phonology*. NY: SUNY Press, 1987.

[9] J.-Y. Tu, Y. Hsiung, M.-D. Wu, and Y.-T. Sung, "Error patterns of Mandarin disyllabic tones by Japanese learners," in *Interspeech 2014*, 2014, pp. 2558–2562.

[10] J.-Y. Tu, Y. Hsiung, J.-H. Cha, M.-D. Wu, and Y.-T. Sung, "Tone production of Mandarin disyllabic words by Korean learners," in *Proceedings of the International Conference on Speech Prosody*, 2016, pp. 375–379, doi: 10.21437/speechprosody.2016-77.

[11] E. Pelzl, "What makes second language perception of Mandarin tones hard?," *Chinese as a Second Lang. J. Chinese Lang. Teach. Assoc. USA*, vol. 54, no. 1, pp. 51–78, Sep. 2019, doi: 10.1075/csl.18009.pel.

[12] Y. Xu, "ProsodyPro - A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, 2013, pp. 7–10.

[13] S. N. Wood, *Generalized additive models: an introduction with R*. CRC press, 2017.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2014.

[15] L. Labrune, "Special segments," in *The Phonology of Japanese*, Oxford University Press, 2012, pp. 132–141.