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

Phonetic Realization of Focus in English by Taiwan Mandarin Speakers

Sherry Chien

University of California, Santa Barbara
sherrychien@ucsb.edu

Abstract

This study examines how speakers of Taiwan Mandarin, a syllable-timed tone language, realize focus in L2 English. Taiwan Mandarin marks focus by increasing the pitch range and duration of the whole focused constituent, which can be polysyllabic. English, instead, marks focus via a pitch accent on the focused word's stressed syllable, which also lengthens. An interactive question-answering experiment was conducted. Two sets of initial-stressed English words featuring identical segments in the first syllable and varying word lengths were elicited phrase-medially under three focus conditions (contrastive, narrow, unfocused). Results show that stressed syllables undergo focus-related lengthening in monosyllabic words, but remain unaffected in polysyllabic words. Instead, focused polysyllabic words present lengthening on the first post-stress syllable. Meanwhile, F_0 marks focus less robustly, being higher in focused than unfocused conditions only in polysyllabic words. Similarly to lengthening, the locus of the F_0 effect is on the syllable following the stressed one. Neither duration nor mean F_0 distinguishes between focus types. Results suggest that speakers of Taiwan Mandarin mark focus phonetically in L2 English, but the effects are not realized in the stressed syllable for polysyllabic words as is typical of English. These findings are discussed in terms of L1-L2 prosodic transfer.

Index Terms: L2 prosody, accentual lengthening, focus, Taiwan Mandarin, pitch accent

1. Introduction

Prosody is one of the essential ways to highlight what is new or important to the discourse in conversational speech, and it is employed in many languages to mark focus. The phonetic dimensions used in focus-marking and how much they are used in differentiating focus types (e.g., narrow vs. contrastive), however, are reported to be language-specific (see [1] for an overview). Moreover, the on-focus domain (i.e., the stretch of speech affected by focus realization, e.g., syllable, word) can also depend on a language's prosodic system. English, for instance, marks focus by placing a nuclear pitch accent on the stressed syllable of the focused word [2], [3], which is phonetically realized by expanded pitch range, higher intensity, and increased duration [4], [5], as well as longer and larger gestures in articulation [6], [7]. Different types of focus (broad, narrow, contrastive) are also distinguished in English by varying degrees of F_0 height [5] and accentual lengthening [7].

Taiwan Mandarin, on the other hand, marks focus by expanding the pitch range and increasing the duration of the whole focused constituent [8], [9], which can be polysyllabic, rather than associating a single pitch event (i.e., pitch accent) to the stressed syllable of the focused word like English. Much less is known about how different types of focus are prosodically distinguished in Taiwan Mandarin, although previous studies have found varying degrees of focus effect in words from different syntactic categories [9], [10].

While prosodic marking of focus may come naturally when acquiring a first language (L1), the picture is much more complex in a second language (L2) setting, especially when the L1-L2 pair are prosodically and rhythmically distinct. Here, we examine how speakers whose L1 is Taiwan Mandarin, a syllable-timed tone language, phonetically realize different types of focus in English, a stressed-accent language in which the focus prosody relies on phrase-level pitch accents. As a tone language, F_0 in Taiwan Mandarin is used phonemically on the lexical level to distinguish between word meanings. Thus, pitch is employed both on the lexical and the phrase levels. Even though tones in Taiwan Mandarin and pitch accents in English are both a form of pitch modulation, tones are associated with every syllable to convey lexical meaning, while pitch accents are associated with only the stressed syllable of the focused word to convey intonational meaning. Moreover, syllable boundaries in Mandarin typically function as stable reference points for the alignment of lexical tones [11], whereas the pitch alignment in English pitch accents depends on the type of the accent as well as various phonetic factors (e.g., the peak alignment of the nuclear H* tone is affected by speech rate and number of post-nuclear syllables) [12]. This knowledge of lexical tones in Taiwan Mandarin that operate on every syllable instead of one stressed syllable may present challenges to speakers learning the focus marking in English, where the pitch modulations of pitch accents operate on the phrase level instead of the lexical level and are associate to one syllable per word.

If the Taiwan Mandarin speakers mark focus in English using their focus prosody in L1, we should see a lengthening effect and a pitch raise on all syllables in the focused word, no matter the word length, and not just on the stressed syllable. If the Taiwan Mandarin speakers apply the focus-marking strategies in English by placing a nuclear pitch accent on the stressed syllable of the focused word, we may see the stressed syllable carrying the largest degree of focus effect (lengthening and pitch raise) when under contrastive focus, and lowest when under unfocused conditions, with the effect of narrow focus in between. However, for the L2 speakers, accentuation of the stressed syllable in the focused word may be more variable in disyllabic and trisyllabic words, as previous research [13] has found that Taiwan Mandarin speakers use weaker prosodic cues (e.g., F_0, duration, amplitude) in realizing lexical stress in English when disyllabic and multisyllabic words are embedded in higher-level prosodic contexts (e.g., narrow focus context). Moreover, if the Taiwan Mandarin speakers do realize these contrasts between focus types and between lexical stressed and unstressed syllables, but in a weaker form, then the difference may not be easily explained away by L1 prosodic transfer, but may reflect more on a different level of processing in L2 speech, as suggested by [13].

10.21437/SpeechProsody.2024-107

2. Method

2.1. Participants

To date, twelve native speakers of Taiwan Mandarin (8 males and 4 females, M_{age} = 27 years, range 23-32 years) have participated in the experiment, and the analysis has included data from eight (4 males and 4 females). All speakers were from Taiwan and were international graduate students at the same University during the conduction of the experiment. They were all experienced learners of English, having passed the TOEFL examination as a prerequisite for university admission, and reported starting to learn English as a foreign language in Taiwan at the age of 4-6 years old. The Length of Residence (LOR) in the United States for the speakers ranged from 1 year to 5 years (M_{LOR} = 2.5 years). The speakers were naïve as to the purpose of the study and had no reported speech, hearing, or vision problems. They received financial compensation for their participation.

2.2. Experimental design and procedure

Test words were two sets of initial-stressed English words featuring identical segments in the first syllable with varying word lengths (monosyllabic, disyllabic, and trisyllabic), as shown in Table 1. Segments of all stressed syllables are controlled so that the nucleus is either the monophthong /æ/or the diphthong /eɪ/, with a preceding bilabial nasal (/m/) onset and followed by an alveolar nasal (/n/) coda.

Table 1: *Test words used in the experiment.*

	monosyllabic	disyllabic	trisyllabic
Vowel /æ/	man	manor	manager
Vowel /eɪ/	mane	mainland	mania

All test words were embedded in frame sentences in phrase-medial positions. Sentence lengths were controlled by having the same syllable numbers. Stressed syllables of the test words were controlled to be five syllables away from the sentence's final boundary and at least two syllables away from the surrounding stressed syllables. An example of the frame sentences for the test word *man* in three types of focus conditions is shown in Table 2.

Table 2: *Prompt questions and expected answers to be elicited. NF = Narrow Focus, CF = Contrastive Focus, Unfoc = Unfocused*

NF	Q: Who did Gabby remember?
	A: Gabby remembered the man with the purple suit.
CF	Q: Did Gabby remember the robot with the purple suit?
	A: No, Gabby remembered the man with the purple suit.
Unfoc	Q: Did Peter remember the man with the purple suit?
	A: No, Gabby remembered the man with the purple suit.

Three types of focus conditions (narrow focus, contrastive focus, and unfocused) were targeted to be elicited for each test word using audio and visual prompts on a computer screen placed around 1.5 feet away from the participant. The audio question prompts were pre-recorded by a trained native English male speaker with the intended pitch accent patterns, while the visual prompts included relevant images of the test words to help guide participants to answer the prompt questions with different focus conditions. As the participants produced the answers on their own based on the visual prompts, the responses elicited were semi-spontaneous speech. Participants' answers were recorded with a Sennheiser shotgun microphone set at a sampling rate of 16 kHz.

2.3. Data analysis

Boundaries of the test words, syllables, and vowels were manually labeled in Praat [14]. The audio and TextGrid files were then imported to VoiceSauce [15], [16] for durational and F_0 analyses of the labeled portions, with F_0 being measured using the STRAIGHT algorithm [17] that gives an estimate of F_0 value at every millisecond throughout an interval. Duration and mean F_0 for each vowel were later normalized with respect to the same vowels within a test word, as only the stressed syllable has the same segmental structure across all test words, while the segments in other syllables vary. The vowel in the stressed syllable is referred to as V1, and the vowels in the second and third syllables in disyllabic and trisyllabic words are referred to as V2 and V3, respectively.

Two analyses were conducted for this paper. For the first analysis, each vowel (V1, V2, and V3) in all the test words was analyzed separately. Separate linear mixed effects models with normalized vowel duration and mean F_0 as response variables were fitted for each vowel in the test words using the lmerTest package [18] in R [19]. The fixed factors were Focus Condition for the models of all vowels (levels: contrastive, narrow, unfocused) and Word Length for the models of V1 (levels: monosyllabic, disyllabic, trisyllabic) and V2 (levels: disyllabic, trisyllabic). Word Length was not considered a fixed factor for V3, as V3 was only present in trisyllabic words. Random effects included intercepts by speaker and test word and slopes by the fixed factors, and Random Effects Principle Components Analysis (rePCA) was used to determine the most suitable random effect structure. Pairwise comparisons were obtained using emmeans with Holm correction ($\alpha = 0.05$) [20].

For the second analysis, polysyllabic words were analyzed. Separate linear mixed effects models with normalized vowel duration and mean F_0 as response variables were fitted for each vowel in disyllabic and trisyllabic words using the lmerTest package [18] in R [19]. The fixed factors were Focus Condition (levels: contrastive, narrow, unfocused) and Vowel Position (levels: V1 and V2 for disyllabic words; V1, V2, and V3 for trisyllabic words). Random effects again included intercepts by speaker and test word and slopes by the fixed factors, and the most suitable random effect structure was determined by rePCA. Emmeans with Holm correction [20] were used to derive pairwise comparisons (α =0.05).

Altogether, 3 focus conditions, 6 test words (manipulating word length and vowel position), 8 repetitions, and 8 speakers for a total of 1152 utterances were analyzed in this paper.

3. Results

3.1. Duration and mean F_0 of V1

For the duration of V1, statistical analyses detected an interaction effect between focus condition and word length ($F(4) = 3.9607$, $p < 0.005$). Post-hoc pairwise comparisons clarified that the contrastive focus condition had a significant lengthening effect ($p < 0.05$) and the narrow focus condition had a near-significant lengthening effect ($p = 0.06$) on the stressed vowels of monosyllabic words, compared to those in the unfocused context. This effect was insignificant for the

stressed vowels in disyllabic or trisyllabic words. No significant differences were found between contrastive and narrow focus conditions. However, as Figure 1 illustrates, a general trend can be observed in the stressed syllable of test words: the vowel duration is longest when under contrastive focus, shortest when unfocused, and intermediate in length under narrow focus.

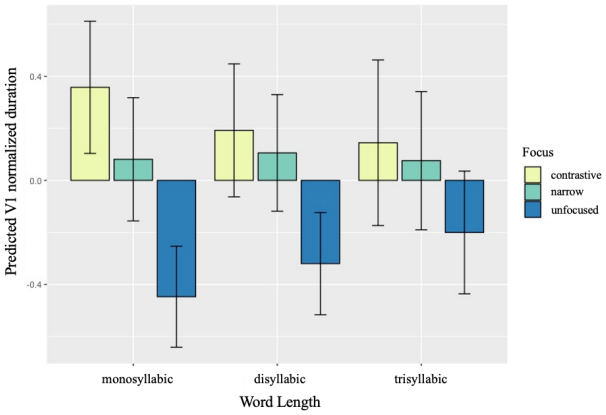

Figure 1: *Predicted duration of V1 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused) by word length (monosyllabic, disyllabic, trisyllabic)*

As for the mean F_0 of V1, although the effect of focus condition was detected in the statistical analyses ($F(2) = 9.3836$, $p < 0.005$), post-hoc pairwise comparisons show that the effect is mainly due to the difference between narrow focus and unfocused conditions ($p < 0.05$) (see Figure 2), while no significant difference was found between contrastive and unfocused conditions or between contrastive and narrow focus. Similarly to duration though, Figure 2 shows a trend of increasing mean F_0 from unfocused to narrow focus and then to contrastive focus.

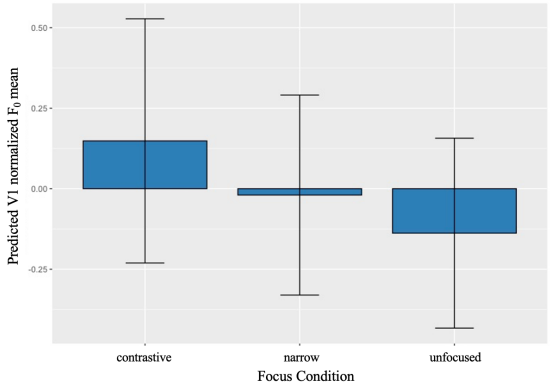

Figure 2: *Predicted F_0 mean of V1 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused)*

3.2. Duration and mean F_0 of V2

For the second vowel (V2) in disyllabic and trisyllabic words, a main effect of focus across all word lengths ($F(2) = 18.885$, $p < 0.0001$) was detected. Post-hoc pairwise comparisons showed that contrastive and narrow focus had significant lengthening effects ($p < 0.0001$) on V2 compared to unfocused conditions for all word lengths (see Figure 3). This indicated that V2 had focus-related lengthening effects in both disyllabic and trisyllabic words, but types of focus (contrastive vs. narrow) were again not distinguished.

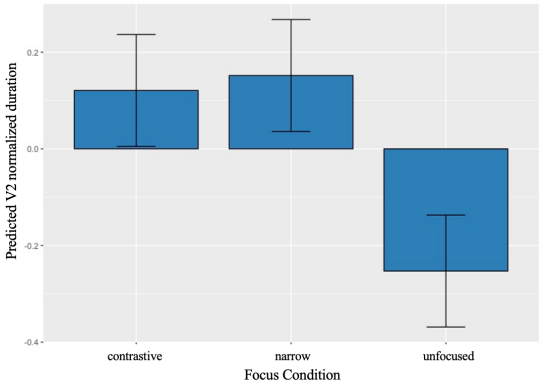

Figure 3: *Predicted duration of V2 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused)*

For the mean F_0 in V2, a main effect of focus condition across all word lengths was again detected ($F(2) = 7.5645$, $p < 0.05$). Post-hoc pairwise comparisons showed that the mean F_0 of V2 in contrastive ($p = 0.06$) and narrow focus ($p = 0.06$) were higher than in unfocused conditions, although both differences were only near-significant. Again, no difference was found between contrastive and narrow focus conditions.

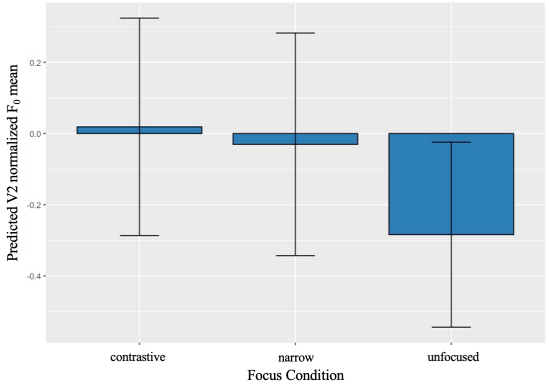

Figure 4: *Predicted F_0 mean of V2 (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused)*

3.3. Duration and mean F_0 of V3

No effect of focus was detected on the duration or mean F_0 of the third vowel (V3) in trisyllabic words.

3.4. Analysis of polysyllabic words

To better understand the profile of duration and F_0 over the whole target word, a separate analysis was conducted comparing across the syllables of polysyllabic words. Separate models were fit in disyllabic and trisyllabic words.

For vowel duration in disyllabic words, statistical analysis showed a main effect of focus condition ($F(2) = 11.138$, $p < 0.005$) across all vowel positions, but no durational differences were found between the vowel positions (V1 and V2). Post-hoc pairwise comparisons showed that both V1 and V2 were significantly longer under contrastive ($p < 0.05$) and narrow focus ($p < 0.05$) compared to unfocused conditions (see Figure

5). No significant differences were found between narrow and contrastive focus. As for mean F_0 of the vowels in disyllabic words, no effect of focus condition or vowel position was detected.

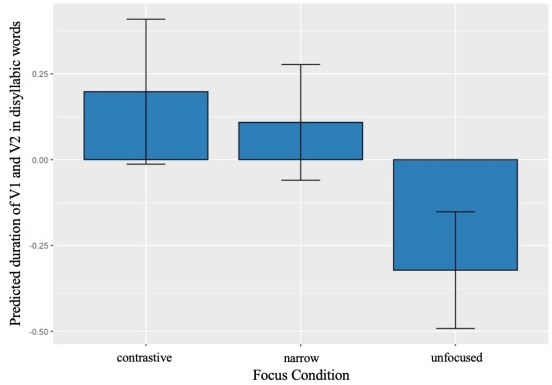

Figure 5: *Predicted duration of both V1 and V2 in disyllabic words (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused)*

For vowel duration in trisyllabic words, statistical analysis showed a main effect of focus condition ($F(2) = 10.895$, $p < 0.0001$) and a main effect of vowel position ($F(2) = 13.844$, $p < 0.005$). Post-hoc pairwise comparisons clarified that V1, V2, and V3 were all significantly longer under contrastive ($p < 0.005$) and narrow focus contexts ($p < 0.001$) than unfocused ones. Moreover, V3 had a longer duration than V2 in contrastive ($p = 0.0527$) and unfocused conditions ($p = 0.0527$), although the differences were only near-significant (see Figure 6). No significant differences were found between contrastive and narrow focus conditions. As for the F_0 mean of the vowels in trisyllabic words, no effect of focus condition or vowel position was detected.

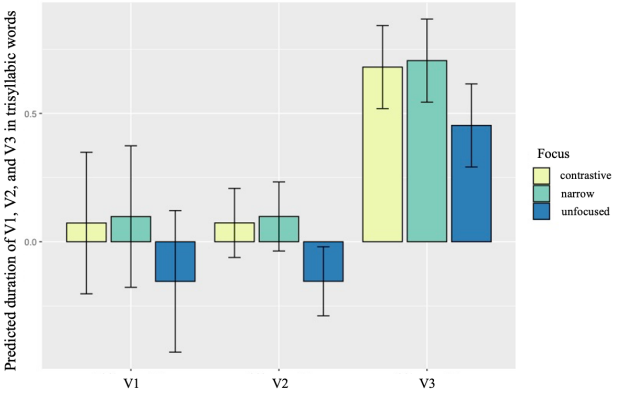

Figure 6: *Predicted duration of V1, V2, and V3 in trisyllabic words (normalized z-score, with standard error) as a function of Focus Condition (contrastive, narrow, unfocused)*

4. Discussion

To examine how Taiwan Mandarin speakers phonetically realize different types of focus in L2 English, this study investigated the duration and mean F_0 of every vowel in initial-stressed monosyllabic, disyllabic, and trisyllabic English words. When comparing only the stressed syllable (V1) in all word lengths, focus-related lengthening substantially differed between focused and unfocused conditions in monosyllabic words. Namely, vowel duration was significantly longer in contrastive focus and near-significantly longer in narrow focus compared to unfocused conditions. Even though no significant lengthening effects were found for V1 in disyllabic and trisyllabic words, a general trend of V1 being the longest under contrastive focus, shortest under unfocused conditions, and with narrow focus in between was found for both monosyllabic and polysyllabic words. Regarding mean F_0, focus-related effects were only near-significant on V2 in disyllabic and trisyllabic words.

For the analysis of polysyllabic words, significant focus-related durational effects were found for both V1 and V2 in disyllabic words, as well as for V1, V2, and V3 for trisyllabic words. Namely, all vowels were substantially lengthened under focused contexts compared to unfocused contexts. This indicated that focus-related lengthening effect was applied to all syllables in polysyllabic words, rather than just the stressed syllable. Moreover, no durational differences were found between V1 and V2 in disyllabic words, and the durational differences among V1, V2, and V3 in trisyllabic words were also minor and inconclusive. These findings indicate constant vowel duration (and possibly constant syllable duration, which will be clarified in future analysis) across the whole polysyllabic word. For mean F_0, no effect of focus condition or vowel position was found for polysyllabic words.

Overall, the results of this study show that the Taiwan Mandarin speakers did use durational differences to mark focus in their L2 English. However, the effect seemed to affect the whole focused word, instead of just the stressed syllable. This result can plausibly be attributed to L1 prosodic transfer, as the on-focus domain in Taiwan Mandarin is the entire focused constituent [8], [9], which can be polysyllabic, rather than just one syllable per word. On the other hand, mean F_0 was used much less robustly, if at all, to mark focus in English by the L2 speakers, possibly due to its strong association with lexical meaning in their L1 knowledge. This lexical component of F_0 in their L1 may have influenced their use of F_0 in the focus marking of their L2 English. To further clarify this issue, future steps will include analysis of other dimensions of F_0 (e.g., tonal alignment) via ToBI description [21] and more dynamic (e.g., curve) F_0 analysis.

Finally, even though focus in English was indeed marked (mainly by lengthening in duration) by the L2 speakers in this study, no distinction was made between the different types of focus (contrastive vs. narrow). This finding somewhat aligns with previous studies, where prosodic contrasts were reported to be much less robustly produced in the L2 English of Taiwan Mandarin speakers compared to their L1 counterparts (e.g., [13], [22]). How this under-differentiation between focus types may be connected to the speakers' L1 background is still unclear, as different focus types in Taiwan Mandarin have not been extensively studied, which is an aspect that we plan on exploring for future studies.

5. References

[1] D. Büring, "Towards a typology of focus realization," in *Information Structure: Theoretical, Typological, and Experimental Perspectives*, M. Zimmermann and C. Féry, Eds., Oxford University Press, 2009, p. 0. doi: 10.1093/acprof:oso/9780199570959.003.0008.

[2] M. E. Beckman and J. B. Pierrehumbert, “Intonational structure in Japanese and English,” *Phonol. Yearb.*, vol. 3, pp. 255–310, May 1986, doi: 10.1017/S095267570000066X.

[3] K. Silverman *et al.*, “TOBI: a standard for labeling English prosody,” *Proc. Int. Conf. Spok. Lang. Process. 2*, pp. 867–870, Jan. 1992.

[4] S. J. Eady and W. E. Cooper, “Speech intonation and focus location in matched statements and questions,” *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 402–415, Aug. 1986, doi: 10.1121/1.394091.

[5] Y. Xu and C. X. Xu, “Phonetic realization of focus in English declarative intonation,” *J. Phon.*, vol. 33, no. 2, pp. 159–197, Apr. 2005, doi: 10.1016/j.wocn.2004.11.001.

[6] T. Cho, “Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English,” in *Laboratory Phonology 8*, L. Goldstein, D. H. Whalen, and C. T. Best, Eds., De Gruyter Mouton, 2006, pp. 519–548. doi: 10.1515/9783110197211.3.519.

[7] A. Katsika, J. Jang, J. Krivokapic, L. Goldstein, and E. Saltzman, “A hierarchy of prominence: The production and perception of focus in American English,” in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic: Guarant International, 2023, pp. 1696–1700.

[8] S.-W. Chen, B. Wang, and Y. Xu, “Closely related languages, different ways of realizing focus,” in *10th Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, Sep. 2009, p. 1010. doi: 10.21437/Interspeech.2009-298.

[9] Y. Hsu and J. S. German, “Prosodic Organization and Focus Realization in Taiwan Mandarin,” in *The 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Hong Kong, Nov. 2018. Accessed: Jan. 06, 2024. [Online]. Available: https://hal.science/hal-02103873

[10] 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*, Aug. 2019, pp. 706–710.

[11] Y. Xu, “Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates,” *Phonetica*, vol. 55, no. 4, pp. 179–203, Dec. 1998, doi: 10.1159/000028432.

[12] K. Silverman and J. B. Pierrehumbert, “The timing of prenuclear high accents in English,” in *Papers in Laboratory Phonology: Volume 1: Between the Grammar and Physics of Speech*, vol. 1, J. Kingston and M. E. Beckman, Eds., in Papers in Laboratory Phonology, vol. 1. , Cambridge: Cambridge University Press, 1990, pp. 72–106. doi: 10.1017/CBO9780511627736.005.

[13] T. Visceglia, C. Tseng, C. Su, and C.-F. Huang, “Interaction of Lexical and Sentence Prosody in Taiwan L2 English,” in *SLaTE Workshop, Interspeech 2010*, Tokyo, Japan, 2010.

[14] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program].” 2024. Accessed: Jan. 06, 2024. [Online]. Available: http://www.praat.org/

[15] Y.-L. Shue, “The voice source in speech production: data, analysis and models,” phd, University of California at Los Angeles, USA, 2010.

[16] 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.

[17] H. Kawahara, A. Cheveigné, and R. D. Patterson, “An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite,” in *Proc. ICSLP’98*, Sydney, Australia, Nov. 1998. doi: 10.21437/ICSLP.1998-555.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest Package: Tests in Linear Mixed Effects Models,” *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26, Dec. 2017, doi: 10.18637/jss.v082.i13.

[19] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria., 2023. [Online]. Available: https://www.R-project.org/

[20] R. V. Lenth *et al.*, “emmeans: Estimated Marginal Means, aka Least-Squares Means.” Dec. 06, 2022. Accessed: Jan. 06, 2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[21] M. E. Beckman and G. Ayers, “Guidelines for ToBI labelling,” *OSU Res. Found.*, vol. 3, p. 30, 1997.

[22] C. Tseng, C. Su, and T. Visceglia, “Underdifferentiation of English lexical stress contrasts by L2 Taiwan speakers,” presented at the Speech and Language Technology in Education, 2013. Accessed: Jan. 07, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Underdifferentiation-of-English-lexical-stress-by-Tseng-Su/ad0109242f3799046a89624181fc22f6fed34a2d