1st International Conference on Tone and Intonation (TAI)
6-9 December 2021, Sonderborg, Denmark

Prominence-boundary interactions in speech perception: Evidence from Japanese vowel length

Hironori Katsuda[1], *Jeremy Steffman*[2]

[1]UCLA
[2]Northwestern University

katsuda1123@gmail.com, jeremy.steffman@northwestern.edu

Abstract

This study examines prominence-boundary interactions as they relate to the perception of durational cues in Tokyo Japanese. We tested if the lexical pitch accent (lexical prominence) status of a word mediates the effects of a prosodic boundary in the perception of contrastive vowel length. We implemented a two-alternative forced choice perception task in which listeners categorized a vowel duration continuum as a phonemically short or long vowel, while we manipulated pitch accentuation and phrasing as contextual cues. We first replicated a recent finding (Steffman & Katsuda [1]) that listeners require longer phrase-final vowel durations (as compared to phrase-medial) to perceive vowel as phonemically long: a compensatory perceptual adjustment for final lengthening. We further find that this boundary effect is mediated by pitch accent, consistent with recent speech production results (Seo et al. [2]) which show that a pitch accent reduces the magnitude of final lengthening in a word (i.e., unaccented words undergo greater final lengthening). Our perception results indicate that listeners accordingly require even longer vowel duration for a long vowel percept when a target word is both phrase-final *and* unaccented. Overall, our results show that listeners take both prominence and prosodic boundaries into consideration when they compute vowel length: a perceptual analog to intricate prominence-boundary effects in speech production.

Index Terms: speech perception, speech prosody, contrastive vowel length, pitch accent, Tokyo Japanese.

1. Introduction

To comprehend a speaker's intended message, a listener needs to identify both the individual segments and the prosodic structure of an utterance. However, the acoustic cues that crucially distinguish segments can systematically vary based on the prosodic structure in which a segment occurs [3-6]. One clear case of this is evident if we consider the critical role of prosodic structure in organizing durational patterns in speech. A well-known and cross-linguistically common example is that of domain-final, or phrase-final, lengthening [e.g., 7]. This refers to the phenomenon whereby linguistic units (syllables, segments) at the right edge of a phrasal domain are lengthened in duration. Given that the duration of acoustic events is also critical for cuing phonological contrasts in language (e.g., voice onset time, vowel duration), we can conceptualize both segmental and prosodic structure as jointly shaping the duration of speech sounds, with prosodic factors (e.g., prosodic boundaries) potentially introducing overlap in the distributions of (durational) cues to phonological contrasts.

A specific example of this pattern is vowel duration in Tokyo Japanese (henceforth "Japanese"), a language in which vowel length is contrastive (e.g., /toko/ 'bed' vs. /tokoo/ 'travel'). Additionally, vowels are generally longer phrase-finally due to final lengthening [e.g., 8,9]. This creates overlap in the distribution of contrastive vowel length categories, in particular between phrase-final short vowels ([…o]$_{\text{phrase}}$) and phrase-medial long vowels ([…oo…]$_{\text{phrase}}$) [10].

Patterns such as this raise a key question: do listeners integrate information about phrasal prosodic structure (here, phrasing) in their perception of phonemic contrasts? In the specific case of Japanese mentioned above we could rephrase this question to ask: do Japanese listeners take phrase-final lengthening into account in their perception of phonemic vowel length? Steffman and Katsuda [1] examined how the perception of prosodic structure modulates the perception of vowel length in Japanese along these lines. Specifically, they tested whether the perception of vowel length is systematically influenced by whether the vowel is located phrase-finally or phrase-medially. They conducted a 2AFC (two-alternative forced choice) perceptual categorization task in which listeners categorized a target sound from a vowel duration continuum as phonemically long or short. Results showed that listeners required longer duration to perceive the vowel as phonemically long in the phrase-final position than in the phrase-medial position, suggesting that listeners take prosodic structural context into consideration when they compute vowel length. This result, among other recent studies [3-6], suggest listeners' fine-grained sensitivity to the effects of phrasal prosodic structure on segmental realization, which lead to a shift in perceptual responses based on phrasal prosodic context.

Adding some nuance to the picture however, phrasal prosodic effects on segment duration have been documented to interact with lexical prosodic features, including lexical prominence (e.g., lexical stress, lexical pitch accent). Japanese is a pitch accent language, in which a pitch-accented mora is phonetically realized with a fall in f0 [11]. Words can contrast based on the location of pitch accent (e.g., *hási* 'chopsticks' vs. *hasí* 'bridge', where an acute accent mark indicates a pitch accent) and the presence or absence of pitch accent (e.g., *hási* 'chopsticks' vs. *hasi* 'edge'). Seo et al.'s [2] recent production study revealed an interaction between pitch accent and final lengthening, such that unaccented disyllabic words (e.g., *taka* 'hawk') exhibit greater final lengthening than disyllabic words with an initial pitch accent (e.g., *táka,* a personal name). They argue that prominence of the accented mora is preserved by suppressing final lengthening. Analogous prominence-boundary interactions are also observed in stress accent languages (e.g., [12,13]).

10.21437/TAI.2021-41

Given these intricate interactions between (lexical) prominence, and phrasal boundaries, one unanswered question is the extent to which such interactions are relevant in perception. In other words, do listeners calibrate their perceptual responses (of the sort described in [1, 3-6] above) in relation to both phrasal prosodic and lexical prosodic (prominence) features? In the case of Japanese specifically, we can ask if Japanese listeners expect *more* phrase-final lengthening on unaccented words, as compared to pitch-accented words, and bring this to bear on perception of contrastive vowel length. If confirmed, this would constitute a perceptual analog of the interaction between lexical prominence and phrasal boundaries shown in [2].

Steffman & Katsuda [1] in fact tested both accented and unaccented minimal pairs in their speech perception experiments. They used minimal pairs in which the members of each pair can be distinguished by the length of the word-final vowel: *shísho* 'librarian' (司書) versus *shíshoo* 'master' (師匠) for accented words (their Experiment 1), and *dookyo* 'housemate' (同居) verusus *dookyoo* 'townmate' (同郷) for unaccented words (their Experiment 2). They found that the unaccented minimal pair exhibited larger positional effects than the accented minimal pair, which is consistent with [2]'s production results showing that pitch accent reduces the degree of final lengthening. However, Steffman and Katsuda's conclusion regarding this mediating effect of pitch accent was only speculative, mainly because their accented and unaccented minimal pairs were not directly comparable due to their segmental differences, and further because the two pairs were tested on different sets of participants (i.e., in a between-subjects design). It thus remains an open question if prominence boundary interactions of the sort described in [2,12,13] play a role in speech perception.

The present study addresses this question by directly comparing positional effects observed in segmentally matched accented and unaccented minimal pairs using a within-subject design. The findings thus contribute to our understanding of the relevance of prominence-boundary and lexical-phrasal prosodic interactions in speech perception, and how these factors impact perception of phonological contrasts.

2. Methods

We implemented a 2AFC task in which listeners categorized a target word with the duration of the word-final vowel drawn from a vowel duration continuum. The target word was categorized as one that contained a phonemically long or short vowel.

To directly compare accented and unaccented minimal pairs, we prepared a quadruplet which contrasted length of the word-final vowel, and the pitch accent status of the word. The four items used are *jísyu* 'voluntary' (自主), *jísyuu* 'next week' (次週), *jisyu* 'surrender onself' (自首), *jisyuu* 'self-study' (自習). These words are almost equally frequent based on word counts in the Balanced Corpus of Contemporary Written Japanese [14]: 1.46 for *jísyu*, 1.68 for *jísyuu*, 1.79 for *jisyu*, and 1.63 for *jisyuu*. These four words constituted the continuum endpoints in our experiment, with the vowel duration continuum ranging between *jísyu* ~ *jísyuu* on one hand and *jisyu* ~ *jisyuu* on the other. The target was thus categorized as *jísyu* or *jísyuu* in the accented condition, and as *jisyu* or *jisyuu* in the unaccented condition.

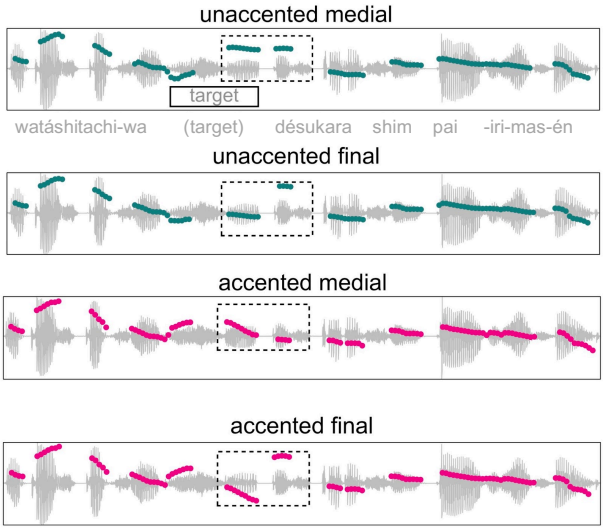

Figure 1: Waveforms and overlaid pitch tracks for the four conditions in the experiment. The dashed box encompasses the final syllable of the target and the following syllable /de/ (see topmost panel for transcription). Note that the first syllable of the target additionally varies based on pitch accent (top two panels versus bottom two panels).

In describing the prosodic structural manipulations in our experiment, we adopt the autosegmental-metrical (AM) model of Japanese intonational phonology developed by [11,15-17]. Specifically, we assume that there are two tonally-defined prosodic groupings above the word level: the accentual phrase (AP) and the intonational phrase (IP). The AP is the domain of pitch accent realization (i.e., only one pitch accent is realized in an AP) while the IP consists of one or more APs and its left edge is marked by a low boundary tone (L%) in statements.

The target was embedded in the same frame sentence, shown in (1). Crucially, the frame can be phrased with either one IP or two IPs, shown by bracketing in (1), without a substantial semantic difference (AP-phrasing is omitted as it is not relevant here). The first phrasing in (1) positions the target IP-medially, while the latter positions the target IP-finally.

(1) *wátashitachi-wa* ***x*** **(target)** *désukara shimpai-iri-mas-én*
We-TOP **x (target)** because/therefore worry-need-be-NEG
IP-medial: [Because we are x (we are) fine]IP
IP-final: [We are x.]IP [Therefore (we are) fine]IP

In sum, we have three independent variables: vowel duration (see 2.1), pitch accent (accented or unaccented), and prosodic boundary (IP-medial or IP-final).

2.1. Stimuli

Stimuli were created by manipulating the speech of a ToBI-trained male speaker of the Tokyo dialect of Japanese. The speaker was first recorded at 44.1 kHz in a sound-attenuated room, using an SM10A Shure™ microphone and headset. Stimuli were manipulated in Praat [18]. As in Steffman and Katsuda, the goal of stimulus manipulation was to manipulate fundamental frequency (f0) cues only (i.e., without changing

temporal context) to signal a target sound as either IP-medial or IP-final. As shown in Figure 1, the IP-final condition (the second and fourth from the top) was characterized by a L% on the target vowel (i.e., /u/~/uu/) and a high tone associated with the pitch accent on the first syllable of the conjunction *désukara* 'because/therefore'. On the other hand, the IP-medial condition was marked by a relatively high f0 on the target vowel, due to lack of L%. Additionally, the pitch accent on the post-target syllable *dé* is only realized after the unaccented target (top) while it is reduced/deleted after the accented target (third from the top). This is because the target word in the IP-medial condition forms an AP with the following conjunction *désukara* (i.e., IP-medial is also AP-medial), and if an AP contains more than one underlying pitch accent only the leftmost one is realized [19-21].

The starting point for creating the stimuli was a naturally produced IP-medial production with an accented target word and phonemically long target vowel (i.e., *jísyuu*). We first inserted 50 ms of silence between the target vowel and the following syllable, which is compatible with both IP-medial and IP-final conditions. We then created an eight-step vowel duration continuum of the target vowel ranging from 20 to 140ms. We then excluded the shortest and longest steps, resulting in six-step vowel duration continuum ranging from 35 to 125ms. These six tokens were used as accented IP-medial stimuli. To create IP-final stimuli, f0 of the target vowel and the following syllable was manipulated for each continuum step: the f0 of the target vowel was lowered by 30Hz while that of the following syllable was raised by 45Hz. The resulting tokens were used as accented IP-final stimuli. The unaccented IP-final stimuli were created based on the accented IP-final stimuli. Specifically, the f0 of the first syllable of the target word (i.e., *ji* in *jisyuu*) was lowered by 30Hz and that of the target vowel was flattened and raised by 5 Hz. Finally, the unaccented IP-medial stimuli were created by manipulating the unaccented IP-final stimuli: the f0 of the target vowel was raised by 35Hz and that of the following syllable is lowered by 10Hz.

2.2. Predictions

Given this experimental design and the conditions described above we can make the following predictions. We first expect that, following [1], overall longer phrase-final vowel durations should be required by listeners for a long vowel response, evident empirically as *decreased* long vowel responses in the final conditions. Additionally, if pitch accent status mediates this effect, we should see a further difference between phrase-final accented words and phrase-final unaccented words, with unaccented words showing *further* decreased long vowel responses as compared to accented words (in line with greater final lengthening in unaccented words). Note this also entails a *larger* difference between medial and final unaccented words, as compared to medial and final accented words. We did not predict any difference between medial accented words and medial unaccented words. Thus, statistically, we expect an interaction between pitch accent and boundary variables in our model.

2.3. Participants and procedure

A total of 42 native speakers of the Tokyo dialect of Japanese (17 males and 25 females; mean age 34) participated in the experiment remotely. All participants were native Japanese speakers from the greater Tokyo area. Participants provided informed consent to participate and were paid for their time.

During the experiment, participants were instructed to use their own headphones and take the experiment in a quiet room. Participants were presented with orthographic representations of the target words on either side of the screen. Participants were instructed to indicate which word they heard via key press, using 'f' for the word on the left side of the screen while 'j' for the word on the right side of the screen, on their computer. Trials were blocked based on accent condition, with block order counter-balanced across participants. Stimulus presentation was totally randomized within each block. The side of the screen on which each word appeared was also counter-balanced. Listeners categorized 12 instances of each unique stimulus for a total of 288 (12×24) trials. The experiment took approximately 30 minutes to complete.

2.4. Statistical Modeling

We analyzed the results using a Bayesian logistic mixed-effects regression model, implemented in [22]. We fit the model to predict listeners' categorization response (a short vowel response mapped to 0, a long vowel response mapped to 1), as a function of vowel duration (scaled and centered), pitch accent, and boundary. Both of these categorical predictors in the model were contrast-coded (accented mapped to -0.5, unaccented mapped to 0.5; phrase-medial mapped to -0.5, phrase-final mapped to 0.5). The model was fit with these fixed effects and all interactions between them. Random effects in the model were specified as random intercepts for speaker, with random slopes for all fixed effect and interactions. We fit the model with weakly informative priors for the intercept and fixed effects. We specified the prior for the fixed effect of vowel duration as normal(2,1.5) in log-odds (that is, a prior expectation that increasing vowel duration will increase the log-odds of a long vowel response, though with a very wide distribution). All other fixed effects and the intercept were specified as normal(0,1.5), encoding no prior expectation of an effect (and for the intercept, the prior expectation of 50% long vowel responses in the middle of the vowel duration continuum, given that the vowel duration variable was centered). We report the full model summary in Table 1, including median estimates and 95% credible intervals (CrI) for an effect. The credible interval characterizes the location of the bulk of the posterior distribution and the range of estimates for the effect. When CrI include the value 0, this suggests the model shows substantial variation in the estimated directionality of an effect and produces a non-trivial amount of estimates near zero (no effect). Thus, when 95% CrI *exclude* zero, we have reliable evidence for the presence of an effect. In reporting the results, we also include the "probability of direction" (pd) metric, computed with [23], which gives the percentage of the posterior distribution for an effect which has a given sign and corresponds more intuitively to a frequentist p-value. When the value of pd > 95% we can take this as reliable evidence for an effect, that is, a consistently estimated, and clearly non-zero, effect directionality.

3. Results

Table 1 shows a summary of fixed effects in the model. Figure 2 plots the results. We first find an expected effect of vowel duration, whereby increasing vowel duration along the continuum increases listeners' long vowel responses (pd = 100%). This can be seen clearly in Figure 2A, as listeners' long vowel responses increase from left to right along the *x* axis. We additionally find a main effect of boundary, whereby a phrase-

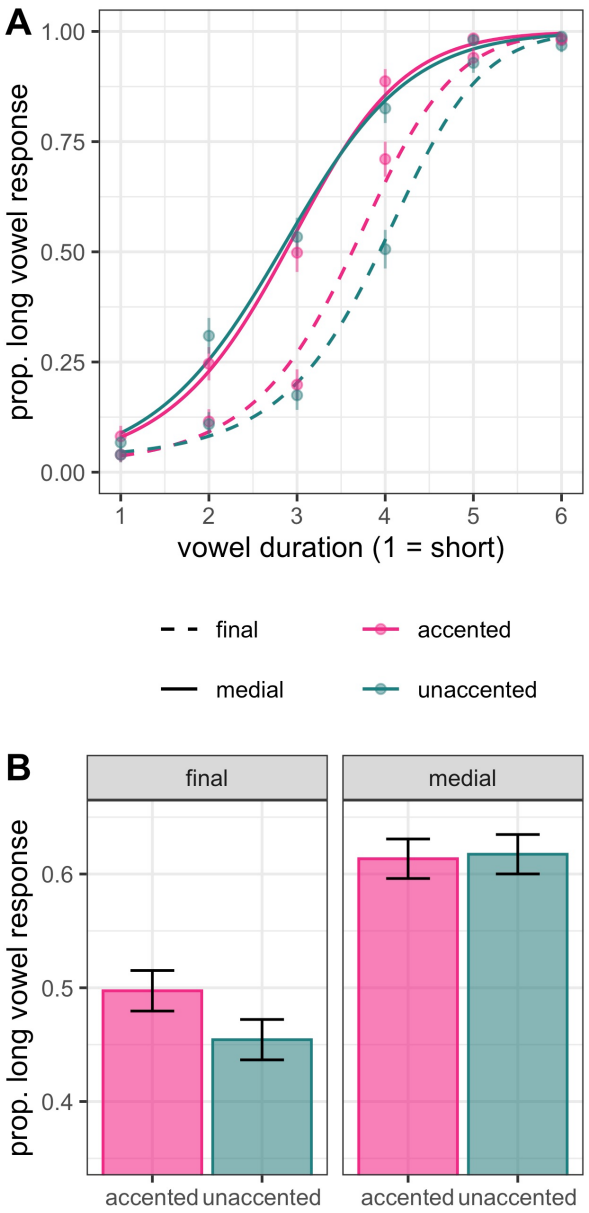

Figure 2: Categorization responses along the continuum (panel A) and pooled by continuum step (panel B) in all four conditions. Panel A shows the means for each condition as points with lines showing a logistic function fit. Error bars in both panels show 95% confidence intervals from the empirical data.

final context shows decreased long vowel responses (pd = 100%). This effect is clear in Figure 2A in the separation of the solid (phrase-medial) and dashed (phrase-final) lines, and across the two panels in Figure 2B, with phrase-final targets showing overall fewer long vowel responses. This finding replicates [1] and shows that listeners need longer acoustic vowel duration to perceive a vowel as phonemically long when phrase-final, effectively accounting for phrase-final lengthening. There was not strong evidence for a main effect of prominence (pd = 92). Importantly, we find a credible interaction between boundary and prominence (pd = 98%), which we examined further using [24] to test the effect of the pitch accent manipulation in each boundary condition. This examination confirms an effect of prominence in the final condition (β = 0.52, 95%CrI = [0.11,0.92], pd = 99%), and finds no effect in the medial condition (β = -0.05, 95%CrI = [-0.47,0.36], pd = 60%). The interaction is visible in Figure 2 as the separation between accented and unaccented conditions *only* when the target is phrase-final, with essentially no difference between the conditions when the target is phrase-medial. This interaction supports the predictions laid out above, suggesting that listeners evidence an expectation of asymmetrical lengthening effects based on pitch accent.

Table 1: *Model summary for fixed effects, ":" indicates an interaction*

Effect	Est (Err)	95%CrI	pd
(Intercept)	0.37(0.17)	[0.02,0.71]	98
vowel dur.	3.68(0.23)	[3.24,4.14]	100
prominence	-0.23(0.16)	[-0.55,0.08]	92
boundary	-1.80(0.15)	[-2.10,-1.50]	100
prom.:boundary	-0.58(0.27)	[-1.09,-0.06]	98
vowel dur.:prom.	-0.26(0.21)	[-0.69,0.13]	90
vowel dur.: boundary	0.27(0.18)	[-0.05,0.65)	95
vowel dur.: boundary: prom.	0.05(0.31)	[-0.56,0.64]	57

4. Discussion

The present study first replicated [1]'s finding that Japanese listeners require longer phrase-final vowel durations to perceive a vowel as phonemically long. By comparing the accented and unaccented conditions directly, it additionally showed that listeners' sensitivity to prosodic structure in segmental perception is more fine-grained. Listeners required even longer phrase-final vowel durations for a long vowel percept when the target word was unaccented as opposed to accented, mirroring the mediating effect of lexical prominence on final lengthening observed in production [2].

To our knowledge, this is the first study to explicitly examine perception effects based on prominence-boundary interactions, whereas previous perception studies have examined effects of prosodic structure on segmental perception based on either phrasal boundaries [1,3-5] or phrasal prominence [6]. The present study thus provides evidence for a fine-grained and interactive view of various prosodic dimensions in speech perception. The results highlight the importance of examining prominence-boundary interactions to get a more comprehensive picture of the elasticity of speech perception, and the way in which different pieces of prosodic information are integrated. Future work will benefit from testing perceptual effects related to prominence-boundary interactions in other languages, and from expanding the set of cues under consideration (e.g., pauses, voice quality changes, etc.) to test how multiple cues to prosodic structure are weighted and combined.

5. Acknowledgements

We would like to thank Sun-Ah Jun, Megha Sundara, Pat Keating, members of the UCLA Phonetics lab, and attendees at the First International Conference on Tone and Intonation for valuable feedback on this project.

6. References

[1] J. Steffman and H. Katsuda, "Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese," *Language and Speech*, *64*(4), pp. 839-858, 2021.

[2] J. Seo, S. Kim, H. Kubozono, and T. Cho, "Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter?" *The Journal of the Acoustical Society of America*, *146*(3), p. 1817, 2019.

[3] H. Mitterer, S. Kim and T. Cho, "The glottal stop between segmental and suprasegmental processing: The case of Maltese," *Journal of Memory and Language*, *108*, 104034, 2019.

[4] J. Steffman, "Intonational structure mediates speech rate normalization in the perception of segmental categories," *Journal of Phonetics*, *74*, pp. 114-129, 2019a.

[5] J. Steffman, "Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing," *The Journal of the Acoustical Society of America*, *145*(6), pp. EL560-EL566, 2019b.

[6] J. Steffman and S.-A. Jun, "Perceptual integration of pitch and duration: prosodic and psychoacoustic influences in speech perception," *The Journal of the Acoustical Society of America*, *146*(3), pp. EL251- EL257, 2019.

[7] J. Vaissiére, "Language-independent Prosodic Features," In A. Cutler and D. R. Ladd (eds.), *Prosody: Models and Measurements*, pp. 53-66, 1983. Berlin: Springer-Verlag.

[8] K. Takeda, Y. Sagisaka and H. Kuwabara, "On sentence-level factors governing segmental duration in Japanese," *Journal of the Acoustical Society of America*, *86*, pp. 2081-2087, 1989.

[9] M. Ueyama, "An experimental study of vowel duration in phrase-final contexts in Japanese," *UCLA Working papers in Phonetics*, *97*, pp. 174-182, 1999.

[10] M. A. Shepherd, "The scope and effects of preboundary prosodic lengthening in Japanese," *USC Working Papers in Linguistics 4*, pp. 1-14, 2008.

[11] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, *3*, pp. 225-309, 1986.

[12] K. Kohler, "Prosodic boundary signals in German," *Phonetica, 40*, pp. 89-134, 1983.

[13] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, *35*(4), pp. 445–472, 2007.

[14] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, and Y. Den, "Balanced corpus of contemporary written Japanese," *Language resources and evaluation*, *48*(2), pp. 345-371, 2014.

[15] J. J. Venditti, (1995). "Japanese ToBI labelling guidelines," *Ohio State University Working Papers in Linguistics* 50, pp. 127–162, 1995. https://kb.osu.edu/bitstream/handle/1811/81780/WPL_50_July_1997_127.pdf

[16] J. J. Venditti, "The J_ToBI model of Japanese intonation," In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 172–200), 2005. Oxford University Press.

[17] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti (2002). "X-JToBI: An extended J_ToBI for spontaneous speech," *Proceedings of the 7th International Congress on Spoken Language Processing,* pp. 1545–1548, 2002. http://www1.cs.columbia.edu/~jjv/pubs/icslp02_final.pdf

[18] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program], 2019.

[19] W. Poser, *The phonetics and phonology of tone and intonation in Japanese*, PhD Dissertation, Massachusetts Institute of Technology, USA, 1984.

[20] H. Kubozono, *The organization of Japanese prosody*, Kurosio, 1993.

[21] K. Maekawa, "Is there 'dephrasing' of the accentual phrase in Japanese?" *Working Papers in Linguistics: Papers from the Linguistics Laboratory* 44, pp. 146–165, 1994. http://hdl.handle.net/1811/81865

[22] P. Bürkner, "Bayesian Item Response Modeling in R with brms and Stan," *Journal of Statistical Software*, 100(5), pp. 1-54, 2021. doi:10.18637/jss.v100.i05L.

[23] D. Makowski, M. Ben-Shachar, and D. Lüdecke, "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework," *Journal of Open Source Software*, 4(40), p. 1541, 2019. doi:10.21105/joss.01541.

[24] R. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means," R package version 1.7.1-1, 2021. https://CRAN.R-project.org/package=emmeans