10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Intonational cues to prosodic boundary influence perception of contrastive vowel length in Tokyo Japanese

Hironori Katsuda[1] *& Jeremy Steffman*[1]

[1]University of California, Los Angeles
katsuda1123@gmail.com, jsteffman@g.ucla.edu

Abstract

We designed two experiments to test how listeners are sensitive to intonational structure in speech perception. Specifically, we tested how phrasal position, cued by pitch in a carrier phrase, mediated Tokyo Japanese listeners' perception of contrastive vowel length. We predicted that when tonal cues signal a target as phrase-final, listeners should expect it to be lengthened due to phrase-final lengthening, effectively requiring longer vowel durations for a phonemically long vowel percept in phrase-final position. We tested this in two experiments, one with an accented target word which contrasted phonemically in the length of the word-final vowel (Experiment 1, *shi'sho* "librarian" vs. *shi'shoo* "master"; Experiment 2: *dookyo* "housemate" vs. *dookyoo* "townmate"). We placed this target word in a carrier phrase, and manipulated contextual pitch to signal it either as Intonation Phrase (IP) final, or medial. As predicted, a phrase-final target required significantly longer vowel duration to be perceived as phonemically long. The results thus highlight the importance of intonational structure as a mediating factor in listeners' processing of temporal cues in speech.

Index Terms: speech perception, intonation, contrastive vowel length, Tokyo Japanese

1. Introduction

In processing speech listeners must extract both segmental and prosodic information from the speech signal, however sometimes the acoustic features that specify these structures are the same. For example, vowel duration may signal a phonemic contrast in a language with contrastive vowel length, and may *also* vary systematically as a function of prosodic organization, e.g., undergoing phrase-final lengthening. Listeners must therefore contend with prosodically driven variation in the process of mapping acoustic information to segmental categories. Accordingly, a recent topic of interest in the literature is how listeners' perception of segmental material is mediated by prosodic structure (e.g., [1-3]).

One line of research has focused on how listeners' processing of durational cues in speech is mediated by temporally organized prosodic patterns, e.g., phrase-final and accentual lengthening [2,4-6]. This research is pursued in light of the central role that prosody plays in organizing the temporal structure of speech [7,8], and the importance of context in listeners' perception of durational cues [9,10].

Accordingly, recent research suggests that listeners incorporate prosodic context in their perception of duration cues such that perception of durational contrasts is prosodic-context dependent, offering support for the proposal that prosodic and segmental information are processed by listeners in parallel [3,11,12]. One gap in our current understanding of listeners' perception of temporal cues is the role that tonal cues to intonational structure play in signaling phrasal position. We are also unaware of previous studies which test the relevance of intonational structure for listeners' perception of contrastive vowel length. In this study we explore how listeners integrate their perception of intonational cues to prosodic structure with their perception of vowel length in the Tokyo dialect of Japanese.

Steffman [5] tested how American English-speaking listeners are sensitive to phrase-final lengthening in their perception of vowel duration as a cue to coda obstruent voicing (where vowels are longer before voiced obstruents). Listeners categorized a vowel duration continuum as either the word "coat" (voiceless) or "code" (voiced). Steffman found that when a target word was phrase-final as compared to phrase-medial, as signaled by the absence of following material in a carrier phrase, listeners required longer target vowel durations for a "code" response, that is, listeners generated an expectation of phrase-final lengthening for sounds in final position such that longer duration was required to signal voicing. This result points to the importance of phrasal boundaries as being used by listeners in their perception of prosodic structure and segmental categories. However, Steffman did not manipulate intonational cues, therefore their relevance in signaling prosodic position and mediating durational processing has not been tested (though note intonation has been shown to play a role in other domains in recent research [1,4]).

1.2. The present study

Tokyo Japanese presents a valuable test case to address this question, given that it has a well-described intonational system, and contrastive vowel length. The present study thus tests if listeners rely on purely tonal cues to compute prosodic boundaries in durational processing, extending recent research by manipulating *only* contextual pitch as a signal for intonational structure, and by testing a phonemic length contrast (as compared to [5] which tested perception of vowel length as a cue to a voicing contrast).

As in many languages, phrase-final lengthening is observed in Japanese [13,14], and is localized to approximately one mora preceding the right edge of the intonational phrase [15]. Phrase-final lengthening has also been shown to be related to the prominence system of the language: [16] show that disyllabic words with an initial pitch accent exhibited less lengthening on their final syllable compared to unaccented disyllabic words. This may be driven by the suppression of phrase-final lengthening following accented syllables, serving to preserve syntagmatic prominence contrasts (where increased duration would make an unaccented syllable more prominent to listeners). This finding is consistent with previous studies which

10.21437/SpeechProsody.2020-12

have shown an interplay between phrase-final lengthening and prominence cross-linguistically (cf. [8,17,18]).

In the present study, we ask if listeners rely on tonal cues to prosodic structure to signal a target sound as phrase-medial or phrase-final in a carrier phrase, and if these tonal cues mediate their perception of contrastive vowel length. In other words, will a phrase final sound be expected to undergo lengthening such that listeners' category boundary for the contrast shifts? In light of the accent-dependent realization of phrase-final lengthening, we created two experiments, both of which tested perception of a vowel length contrast. The minimal pair used in each experiment was different, such that in Experiment 1 the first syllable of the target word minimal pair bore a pitch accent and in Experiment 2 it did not. Our goal in testing both unaccented and accented minimal pairs was firstly to offer a basic replication of the predicted effect. Additionally, given that unaccented and accented target words undergo different modulations, exhibiting different degrees of phrase final-lengthening [16], we can test if listeners are sensitive to these accentually-differentiated prosodic patterns in perception. Specifically, we predicted that because unaccented words undergo more substantial lengthening, listeners will exhibit larger positional effects in categorizing an unaccented minimal pair as compared to an accented one.

In describing the intonational variables manipulated in our experiment, we adopt the Autosegmental-Metrical (AM) model of Japanese intonational phonology developed by Beckman and Pierrehumbert, and others [19-23]. In the model there are two tonally-defined prosodic groupings above the word level: the accentual phrase (AP) and the intonational phrase (IP). Every AP consists of a phrasal H tone (H-) on the second mora and a falling pitch movement (L%) at its right edge. It also accommodates at most one pitch accent (A in XJ-ToBI). The IP, on the other hand, serves as a domain of downstep, and is marked by an initial L boundary tone (%L).

2. Method

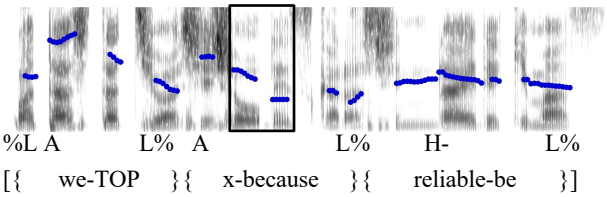
We implemented a 2AFC task, in which listeners categorized a target sound from a vowel duration continuum as phonemically long or short. In Experiment 1 listeners categorized a disyllabic minimal pair that contrasted the length of the second syllable, with an accented first syllable. The target was categorized as *shi'sho* "librarian" (司書) or *shi'shoo* "master" (師匠). In Experiment 2 listeners categorized an unaccented disyllabic minimal pair as *dookyo* "housemate" (同居) or *dookyoo* "townmate" (同郷).

2.2. Stimuli

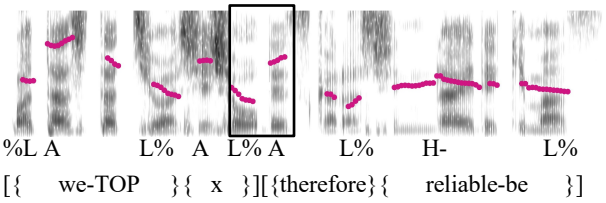
Stimuli were created by resynthesizing the speech of male speaker of the Tokyo dialect of Japanese. The speaker was first recorded at 44.1 kHz in a sound-attenuated room, using an SM10A Shure™ microphone and headset. The goal in creating the stimuli was to manipulate pitch within a carrier phrase to signal a target sound as either phrase-medial or phrase-final.

The same carrier phrase given in (1), which had a similar design to the carrier phrase from [15]'s production study, was used for both medial and final conditions. The sentence shown in (1) at right can be phrased in two different ways, which are represented in Figure 1, where brackets and XJ-ToBI labels showing phrasing. The possible presence of an IP boundary changes the position of the target sound, such that in (1a) and (1c) it is IP (and AP) medial, and in (1b) and (1d) it is IP final.

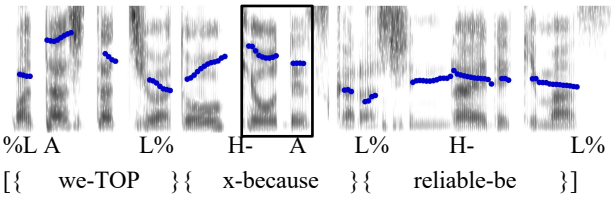
a. Experiment 1 (accented): medial condition (1 IP)



b. Experiment 1 (accented): final condition (2 IPs)



c. Experiment 2 (unaccented): medial condition (1 IP)



d. Experiment 2 (unaccented): final condition (2 IPs)

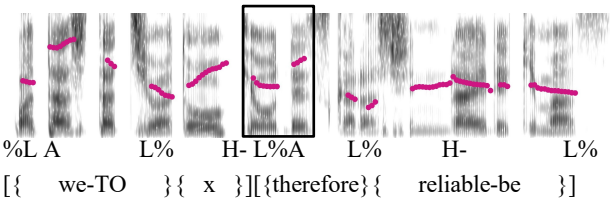


Figure 1: *Examples of stimuli from Experiment 1 and Experiment 2, in both the medial and final conditions. Spectrograms with a frequency range of 0-5000Hz are overlaid with pitch tracks (50-200Hz range). The second syllable in the target word and post-target syllable are boxed to highlight changes across conditions. XJ-ToBI labels are given below each panel, as well as glosses (the accent on "be" is reduced, and not marked). Phrases are bracketed: {…} indicates an AP boundary, […] indicates an IP boundary. In these examples the target vowel is approximately 110 ms long (step 4 on each continuum, see below).*

(1) wata'shitachi-wa *x* de'sukara shinraideki-ma'su
we-TOP *x* because/therefore reliable-be
One IP: "Because we are *x* (we are) reliable." (x = medial)
Two IPs: "We are *x*. Therefore (we are) reliable." (x = final)

In both Experiment 1 and Experiment 2, the stimuli for the medial and final conditions differ only in the pitch of the target syllable and that of the following syllable (i.e., *de'* in *de'sukara*). In the medial condition (shown in panels a. and c. in Figure 1), the target word *x* forms an AP with the following conjunction *de'sukara* "because/therefore". The accentuation

of the first syllable of *de'sukara* depends on the accentedness of the target word: if the target word is accented, the accent on the conjunction is reduced [24-26], while if the target word is unaccented, the accent on the conjunction is realized. In the final condition (shown in panels b. and d. in Figure 1), on the other hand, the target word does not phrase with the following conjunction and there is an IP boundary after the target word. Pitch range and downstep are reset after the IP boundary [21]. In this condition, the accent on the conjunction is realized regardless of the accentuation of the target word.

The final condition is characterized by lower pitch on the target syllable due to the L% associated with the right edge of the phrase boundary, and the accent realized on the first syllable of *de'sukara* as well as the absence of downstep on the accent indicating the pitch range is reset after the target syllable. In the medial condition, if the target word is accented (Experiment 1), the accent on the first syllable of *de'sukara* is reduced, resulting in a gradual fall over the AP containing the target word and *de'sukara*, while if the target word is unaccented (Experiment 2), the following accent is realized, exhibiting a sharp fall from the accented syllable to the following syllable.

To create these conditions, we resynthesized the duration and pitch of a carrier phrase and target, using the PSOLA method [27], in Praat [28]. Figure 1 provides a visual representation of the two conditions used in each experiment. In both experiments, only pitch on two syllables was manipulated: the second syllable in the target word, and the following syllable /de/. The remainder of the carrier phrase was acoustically identical across conditions within an experiment.

The starting point for pitch manipulations in Experiment 1 was an IP-medial production with a phonemically long vowel target (*shi'shoo*). Pitch values from another medial production were overlaid onto the second syllable of the target and following post-target /de/ to create the medial condition. We then overlaid the orginal target and post-target syllable with the pitch from a natural IP-final production, creating the final condition (note both conditions were thus resynthesized). These two conditions varied only in the pitch of the target-final syllable and post-target /de/ (see Figure 1). They were judged to sound like a natural phrasing of both a medial production (1a and 1c) and a final production (1b and 1d), by a ToBI-trained native speaker of Japanese. A vowel duration continuum was then resynthesized from this medial and final production. The duration of the starting vowel was approximately 100 ms in duration. It was manipulated to range from 60 to 180 ms of vowel duration. The continuum was manipulated to have 8 evenly spaced steps that included these endpoint values, with a between-step durational difference of approximately 17 ms (note that all continuum steps were created via resynthesis, the unaltered original was not used). These manipulations resulted in 16 unique stimuli (2 positional conditions × 8 continuum steps), with the only difference across conditions being the pitch on the second syllable of the target and post-target /de/ (Figures 1a and 1b).

The stimuli from Experiment 2 used the same carrier phrase as Experiment 1. The new target word, which was originally produced in another production of the same carrier phrase, was cross-spliced into the Experiment 1 carrier phrase. As with Experiment 1, the starting point for manipulation was a medial production of a long vowel target (*dookyoo)*. Pitch on the target vowel, and the immediately post-target syllable /de/ was then overlaid with the pitch values from a medial and final production of the Experiment 2 target word to create both conditions. Note the pitch values on the target word and post-target /de/ are different than that of Experiment 1, in particular in the medial condition. This is due to the unaccented status of the target word. As described above, when the target word is accented, as in Experiment 1, the accent on post-target /de/ is reduced resulting in a smooth f0 downtrend across the AP. In Experiment 2, when the target word is unaccented the accent on /de/ is realized, resulting in substantially higher post-target f0 as compared to Experiment 1

By varying only pitch we ensured that other potential explanations for shifts in categorization are ruled out, e.g. changes in adjacent segmental duration as discussed in [2]. Additionally, psychoacoustic influences of pitch height/dynamics on perceived duration are unlikely to play a role given that these effects have been shown to be restricted to isolated monosyllables [29], and that psycho-acoustic effects may also *not* occur in monosyllables when pitch has a possible prosodic interpretation [6].

2.3. Participants and procedure

We recruited 26 native speakers of the Tokyo dialect of Japanese for each experiment (no participant participated in both experiments). All participants were from the Greater Tokyo Area. Participants provided informed consent and completed a language background questionnaire prior to the beginning of the experiment. Participants were paid the equivalent of 5 USD for 20 minutes of their time. Participants completed the experimental task seated in a quiet room and using a laptop computer, with stimuli presented binaurally via a Peltor™ listen-only headset.

During each trial, participants listened to a stimulus and were presented with orthographic representations of the target words on either side of the screen. Participants were instructed to use the computer keyboard to indicate which word they heard, where an 'f' keypress indicated the word on the left side of the screen and a 'j' keypress indicated the word on the right side of the screen. The side on which each word appeared was counterbalanced across participants. Listeners categorized 12 instances of each unique stimulus for a total of 192 (12×16) trials in both Experiment 1 and Experiment 2. Stimulus presentation was totally randomized. The experiment took approximately 20 minutes to complete.

2.4. Results

Results for both experiments are assessed with a mixed-effect logistic regression model, implemented using *lme4* [30]. Models predicted listeners' response (short vowel or long vowel) as a function of continuum step, prosodic position, and the interaction of these two fixed effects. Vowel duration was treated as a continuous variable and centered at zero. Prosodic position was contrast-coded (final position mapped to 1, medial to -1). The dependent variable was coded such that a short vowel response was the reference level. The random effect structure for each model was specified with by-subject intercepts and maximal random slopes.

Results from both experiments will be discussed together in reference to the statistical models. Plotted results for both experiments are given in Figure 2, with the model summaries given in Table 1.

Table 1: *Model summaries for both experiments.*

Experiment 1	β(SE)	z	p
(Intercept)	-0.65(0.26)	-2.57	0.01
position	-0.39(0.14)	-2.84	< 0.01
step	6.08(0.34)	17.78	< 0.001
position:step	0.43(0.23)	1.86	0.06

Experiment 2	β(SE)	z	p
(Intercept)	1.25(0.23)	5.54	< 0.001
position	-0.66(0.10)	-6.42	< 0.001
step	6.08(0.37)	16.49	< 0.001
position:step	-0.41(0.20)	-2.01	< 0.05

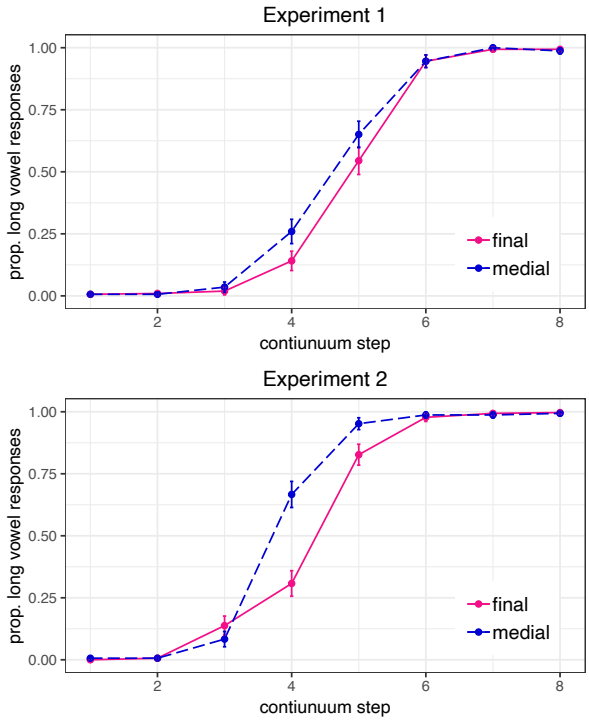


Figure 2: *Categorization responses for Experiment 1 (top) and Experiment 2 (bottom). The x-axis shows the continuum steps (step 1 = 60 ms, step 8 = 180 ms, inter-step-interval = approximately 17 ms). The y axis shows proportion of long vowel responses in each condition. Error bars give 95% confidence intervals.*

Both experiments showed an expected effect of vowel duration whereby increasing durations increased long vowel responses. Prosodic position, the predictor of interest, also showed a significant effect in both experiments, visible in both plots in Figure 2. In both experiments, final position significantly decreased listeners' long vowel responses (in Experiment 1: β = -0.39, z = -2.84; In Experiment 2: β = -0.66, z = -6.42). This finding aligns with our central prediction, showing that when the target sound in both experiments was cued as phrase final, overall longer vowels were required for a long vowel response. This general finding therefore supports the idea that listeners' computation of prosodic structure (informed by pitch patterns) plays a mediating role in their perception of contrastive vowel length.

Also of interest was the magnitude of the positional effect across experiments. As discussed above, the results of [16]'s production study indicated that the effect of final lengthening is reduced when a non-final syllable of the target word is accented. Although the stimuli in Experiment 1 and those in Experiment 2 are not directly comparable (i.e., segmental structure of the target words are different across the experiments), our results indicate unaccented target words in Experiment 2 exhibit a greater effect of position than accented words in Experiment 1. This concurs with [16]'s production results in suggesting that listeners may exploit prominence-boundary interactions perceptually such that boundary effects in unaccented words (Experiment 2) are more pronounced. Importantly, this conclusion is only speculative. Future research will benefit from addressing this point more directly, which will better our understanding of the perceptual consequences of prominence-boundary interactions.

3. General Discussion

The present study found, in two experiemnts, that pitch-based cues to intonational structure influenced listeners' perception of contrastive vowel length in Tokyo Japanese. This finding extends recent research on prosodic-segmental interactions in speech perception and showed that listeners are indeed sensitive to intonational structure, signaled by changes in pitch, in their perception of contrastive vowel length. These results are consistent with other recent findings which posit that listeners process segmental and prosodic information in parallel [1,3]. Such a parallel processing architecture is motivated on the basis of findings like these: i.e., prosodic-structural influences in listeners perception of segmental contrasts. More broadly, these findings are compatible with models of speech perception which allow for contextual factors to mediate expectations about cue values e.g., [31,32], with the additional claim that prosodic categories can be used for this purpose.

The present study can complement previous findings in showing that pitch plays an important role in this domain. We also found that unaccented target words show a larger positional effect, which is consistent with differences found between accented and unaccented target words in speech production. This difference is intriguing in suggestion that listeners incorporate prominence-boundary interactions in their sensitivity to phrasal positional effects. Further work will benefit from extending the present findings to test (1) other languages with vowel length contrasts and (2) other possible influences of intonational structure speech perception in Japanese. Testing the timecourse of these effects with eye-tracking may also present a valuable extension, to help understand how quickly prosodic information is incorporated in listeners' perception, and inform models of this process, following e.g., [1,3].

4. References

[1] S. Kim, H. Mitterer and T. Cho, "A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing," *PLOS ONE*, *13*(8), e0202912, 2018.

[2] H. Mitterer, H., T. Cho and S. Kim, "How does prosody influence speech categorization?," *Journal of Phonetics*, *54,* pp. 68-79, 2016.

[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] A. E. Turk and J. R. Sawusch, "The domain of accentual lengthening in American English," *Journal of Phonetics*, *25*(1), pp. 25-41, 1997.
[8] 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.
[9] E. Reinisch and M. J. Sjerps, "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," *Journal of Phonetics*, *41*(2), pp. 101-116, 2013.
[10] T. Wade and L. L. Holt, "Perceptual effects of preceding nonspeech rate on temporal properties of speech categories," *Perception & Psychophysics*, *67*(6), pp. 939-950, 2005.
[11] 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*, *35*(2), pp. 210-243, 2007.
[12] A. P. Salverda, D. Daha and J. M. McQueen. "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension". *Cognition*, *90*(1), pp. 51- 89, 2003.
[13] 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.
[14] M. Ueyama, "An experimental study of vowel duration in phrase-final contexts in Japanese," *UCLA Working papers in Phonetics*, *97*, pp. 174-182, 1999.
[15] M. A. Shepherd, "The scope and effects of preboundary prosodic lengthening in Japanese," *USC Working Papers in Linguistics 4*, pp. 1-14, 2008.
[16] 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.
[17] A. Katsika, "The role of prominence in determining the scope of boundary-related lengthening in Greek," *Journal of Phonetics*, *55*, pp. 149-181, 2016.
[18] S. Nakai, S. Kunnari, A. Turk, K. Suomi and R. Ylitalo, "Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics*, *37*(1), pp. 29-45, 2009.
[19] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, *3*, pp. 225-309, 1986.
[20] J. Pierrehumbert and M. E. Beckman, *Japanese Tone Structure.* Cambridge, MA: MIT Press, 1988.
[21] J. J. Venditti, "Japanese ToBI Labelling Guidelines," *Ohio State University Working Papers in Linguistics 50*, pp. 127-162, 1995.
[22] J. J. Venditti, "The J_ToBI Model of Japanese Intonation," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed., Oxford: Oxford University Press, pp. 172-200, 2005.
[23] K. Maekawa, H. Kikuchi, Y. Igarashi and J. Vanditti, "X-JToBI: An extended J_ToBI for spontaneous speech," in *Proceedings of the 7th International Congress on Spoken Language Processing, September 16-20, Denver, USA*, pp. 1545-1548, 2002.
[24] W. Poser, "The phonetics and phonology of tone and intonation in Japanese," Ph.D. Dissertation, MIT, 1984.
[25] H. Kubozono, "The organization of Japanese prosody," Ph.D. Dissertation, University of Edinburgh, 1987.
[26] 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.
[27] E. Moulines and F. Charpentier, "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," *Speech Commun.*, *9*(5-6), pp. 453-467, 1990.
[28] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program], 2019.
[29] W. Van Dommelen, "Does Dynamic F0 Increase Perceived Duration? New Light on an Old Issue," *Journal of Phonetics*, *21*(4), pp. 367-386, 1993.
[30] D. Bates, M. Maechler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, *67*(1), pp. 1-48, 2015.
[31] B. McMurray and A. Jongman. "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations". *Psychological review*, *118*(2), p. 219, 2011.
[32] R. Smits. "Evidence for hierarchical categorization of coarticulated phonemes". *Journal of Experimental Psychology: Human Perception and Performance*, *27*, pp. 1145–1162, 2001.