Intonational structure influences perception of contrastive vowel length: the case of phrase-final lengthening in Tokyo Japanese

Journal:	*Language and Speech*
Manuscript ID	LAS-19-0146.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Steffman, Jeremy; UCLA, Linguistics Katsuda, Hironori; UCLA, Linguistics
Keywords:	speech perception, prosody, intonation, Japanese, vowel length
Abstract:	Recent research has proposed that listeners use prosodic information to guide their processing of phonemic contrasts. Given that prosodic organization of the speech signal systematically modulates durational patterns (e.g. accentual lengthening, phrase-final lengthening), listeners' perception of durational contrasts has been argued to be influenced by prosodic factors. For example, given that sounds are generally lengthened preceding a prosodic boundary, listeners may adjust their perception of durational cues accordingly, effectively compensating for prosodically-driven temporal patterns. In the present study we present two experiments designed to test the importance of pitch-based cues to prosodic structure for listeners' perception of contrastive vowel length in Tokyo Japanese along these lines. We tested if, when a target sound is cued as being phrase-final, listeners compensatorily adjust categorization of vowel duration, in accordance with phrase-final lengthening. Both experiments were a 2AFC task in which listeners categorized a vowel duration continuum as a phonemically short or long vowel. We manipulated only pitch surrounding the target sound in a carrier phrase to cue it as intonational phrase (IP) final, or accentual phrase (AP) medial. In Experiment 1 we tested perception of an accented target word, and in Experiment 2 we tested perception of an unaccented target word. In both experiments, we found that contextual changes in pitch influenced listeners' perception of contrastive vowel length, in accordance with their function as signaling intonational structure. Results

	therefore suggest that listeners use tonal information to compute prosodic structure and bring this to bear on their perception of durational contrasts in speech.

Intonational structure influences perception of contrastive vowel length: the case of phrase-final lengthening in Tokyo Japanese

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

*Corresponding author: jeremysteffman@gmail.com Tel: +17203019297

[1]UCLA Department of Linguistics
3125 Campbell Hall
Los Angeles, CA 90095-1543

1. Introduction

When listeners process speech, they need to extract (among other things) a segmental message and a prosodic message from the speech signal (e.g., Cho, McQueen & Cox, 2007; Kim, Mitterer & Cho, 2018). That is, listeners need to compute what segmental contrasts and lexical items are intended by the speaker, while also computing prosody, which will tell listeners how items are grouped in phrases, and what the prominence relations are among them. Sometimes the acoustic information which specifies these segmental and prosodic structures may be the same. We can consider this in light of a body of literature that shows higher-level prosodic structure is encoded in a detailed way by various properties of individual speech segments and articulations (e.g., Byrd, Narayanan & Saltzman, 2000; Cho 2015, 2016; Cho & Keating, 2009; De Jong, 1995; Fougeron & Keating, 1997; Keating, 2006; Keating, Fougeron, Hsu & Cho; 2003). One well-attested pattern in this literature is the role of prosodic boundaries in fine-tuning timing and amplitude patterns for articulations.

For example, segments are realized as "stronger" initial to prosodic domains (e.g. Keating et al., 2003; Cho, 2016), produced with increased articulatory contact, duration of closure, etc. (e.g., Keating et al., 2003). This introduces systematic patterns in the acoustic properties of speech sounds as a function of prosodic organization. Consider an example that illustrates the dual-function of acoustic cues domain-initially. Voice Onset Time (VOT), which serves to differentiate stop voicing categories (e.g., Abramson & Lisker, 1970) is also systematically modulated by prosodic structure. At the beginning of phrasal domains VOT is lengthened in some languages as a manifestation of domain-initial strengthening (Keating et al., 2003; Cho & Keating, 2009). If we consider prosodic position as a source of contextual variability for VOT along these lines, its relevance in the perception of e.g., voicing contrasts becomes evident: if VOT varies as a function

of prosody, listeners would benefit from reconciling a given VOT value with the prosodic context in which it occurs, making reference to whether VOT is lengthened as a function of initial strengthening, or is simply long to cue a voicing contrast at the lexical level. Put differently, listeners would benefit from using "prosodic information in determining whether segmental information is driven lexically or post-lexically (prosodic-structurally)" (Mitterer, Kim & Cho, 2019 p 14). Recent research, discussed below, suggests this may indeed be the case.

In similar fashion to documented "initial strengthening" at the left edges of domains, one well-described pattern in the literature that occurs at right edges is final lengthening, also called pre-boundary lengthening (e.g., Cho 2015, 2016; Turk & Shattuck-Hufnagel, 2007, Wightman, Shattuck-Hufnagel, Ostendorf & Price, 1999). Generally speaking, this refers to the temporal expansion of linguistic units preceding a prosodic boundary (Cho, 2015, 2016). Though domain-initial prosodic effects in speech perception have received recent attention in the literature (Kim & Cho, 2013; Mitterer, Cho & Kim, 2016; Mitterer, Kim & Cho, 2019), perception of segmental contrasts in phrase final-position remains relatively less studied.

Domain-final prosodic effects and their relationship to other aspects of linguistic structure can be considered more broadly as part of the prosodically-driven temporal organization of the speech signal. Turk and White (1999) for example conceptualize prosody as "hierarchical structure [that] influences the domain and distribution of durational effects" (p 171; see also Turk & Sawusch, 1997; Turk & Shattuck-Hufnagel, 2007). As discussed for the example of domain-initial VOT, prosodic organization along these lines might be thought to play an important role in speech perception and spoken language processing. Prosodic patterning in the temporal domain can be taken to systematically modulate the duration of segmental structure, and acoustic cues over time,

and may accordingly exert a mediating influence in listeners' perception of various durational contrasts in speech.

In the present study we test the influence of pre-boundary lengthening on the perception of contrastive vowel length in Tokyo Japanese. We also manipulate only f0 as a contextual cue to intonational structure, which allows for control over other possible durational context effects, discussed below. The present study therefore presents a new testing ground for the relevance of intonational structure, and its temporal encoding, in listeners' processing of durational contrasts in speech.

1.1. Previous studies

Several previous studies have presented evidence for the role prosodic boundaries play in listeners' perception of segmental contrasts, and more generally in other domains of speech processing such as word segmentation (see e.g., Cho et al., 2007). Kim & Cho (2013) tested the aforementioned pattern of initial strengthening and perception of VOT in American English. They carried out an experiment in which listeners categorized a VOT continuum as /p/ or /b/. The target sound was placed in a carrier phrases "let's hear pa/ba again", and the presence/absence of a preceding intonational phrase boundary before the target word was manipulated (among other things). The phrase boundary was cued by a low boundary tone, and phrase-final lengthening. The authors predicted that listeners may use the boundary to adjust their categorization of VOT, in line with domain-initial strengthening. Specifically, if listeners expect phrase-initial lengthening of VOT, they should effectively require longer VOT for a voiceless aspirated /p/ response. The authors found this effect, though the picture is complicated by another possible explanation. As discussed by Mitterer et al. (2016), lengthening preceding a target sound could modulate perception of VOT

independent of prosody. Because a preceding boundary was encoded by pre-boundary lengthening, perception of VOT (a durational cue) may be expected to shift on the basis of local speech rate normalization, i.e., perceptual adjustments for changes in durational context (e.g., Miller & Liberman, 1979; Newman & Sawusch, 1996; Summerfield, 1981). This presents a possible explanation because preceding lengthening may cause subsequent VOT to be perceived as relatively short, as a function of durational contrast (e.g., Diehl & Walsh, 1979). This effect would therefore decrease listeners' /p/ responses in the post-boundary condition, on the basis of adjustments for preceding speech rate. This alternative explanation does not implicate prosodic structure, and thus offers an interesting illustration of the complexity in testing for prosodic effects in perception, especially in the case of durational cues (see Mitterer et al., 2016; Steffman, 2019a for further discussion).

One promising way to extend this work which has recently been explored in the literature is testing perception of non-durational contrasts (Mitterer et al., 2019) or manipulating contextual cues that are not in the temporal domain (Kim et al., 2018). Mitterer et al. (2019) tested domain-initial effects on vowel realization in Maltese. Glottalization in Maltese is phonemic such that there exist minimal pairs such as /ʔɑ:m/ "he woke up" and /ɑ:m/ "he swam". At the same time, glottalization in vowel-initial words is a manifestation of initial strengthening, occurring at the beginning of higher-level phrasal domains. Thus, in similar fashion to VOT, glottalization cues a phonemic contrast, while simultaneously varying along a prosodic dimension. Unlike VOT however, glottalization is not an inherently temporal cue. Accordingly, in one experiment the authors created a glottalization continuum ranging from /ʔɑ:m/ to /ɑ:m/, while holding duration constant. The authors were curious how perception of this contrast would vary based on whether the target was cued as being phrase-initial, or not. Listeners categorized the target in a carrier

phrase, with the presence/absence of preceding pre-boundary lengthening manipulated, to signal the target as phrase-initial. The crucial prediction was that listeners should be more likely to interpret glottalization as phonemic (as in /ʔɑ:m/), when the target was *not* preceded by boundary cues, i.e., when domain-initial glottalization would not occur. On the other hand, a target cued as phrase-initial may be glottalized to encode prosodic structure (in lieu of a lexical contrast), and would therefore be more likely to be perceived as /ɑ:m/. The authors find this pattern, importantly with a cue that is not temporal such that speech rate normalization does not offer an alternative explanation. We can take this finding to highlight the role of prosody in the perception of domain-initial words on the basis of the way they are influenced by initial strengthening, here encoded with glottalization.

Another relevant study, Kim et al. (2018), tested if listeners reference prosodic structure, cued by pitch alone, in a phonological inferencing task. They tested Korean post-obstruent tensing (POT), whereby lax stops and affricates become tense following another obstruent. To use an example from the paper: /puri/ "beak" will become tensified [p*uri], when following an obstruent, as in the sequence /porasɛk # puri/ "purple beak", making it confusable with /p*uri/ "root". Importantly, the domain of this process is within the Korean accentual phrase (AP). Kim et al. used a visual world eye-tracking experiment in which listeners heard a color term and target word, as in the example above, and looked to colored orthographic representations as they listened to speech. The relevant test case is one in which listeners hear a tense obstruent: the question is if they will infer that POT has applied and look to an underlying lax form, e.g., /puri/, or an underlying tense form, e.g., /p*uri/. Crucially, the authors predicted that this process should be modulated by prosody: because the domain of POT is within the AP, listeners would necessarily need to reference phrasing to determine whether POT may have applied. For example, AP-internal

(porasɛk p*uri), where parentheses indicate an AP boundary, may be “beak” or “root”, however an intervening AP boundary disambiguates the meaning: (porasɛk) (p*uri), can only be “root”. Kim et al. found that when listeners heard a phonetically tense target [p*uri], they looked more to an underlyingly lax word /puri/ when that word was in a context that licensed POT (e.g., / porasɛk # puri/), as compared to when it was not (when a non-obstruent preceded the target sound). In these cases, the target and preceding color term were additionally phrased together in a single AP. This evidenced the predicted phonological inferencing effect. To test how a perceived AP boundary might modulate this effect, Kim et al. manipulated f0 alone to signal an AP boundary between the color term and following target word. Due to the role of AP boundaries in restricting application of POT, the authors predicted the effect observed for an AP-internal color-target sequence should *not* be observed when an AP boundary intervenes, showing listeners’ sensitivity to the phrasal domain of the phonological process. To cue an AP boundary, Kim et al. used rising f0, while leaving the temporal properties of their stimulus unaltered. As expected, the phonological inferencing effect disappeared in this case, presenting evidence that listeners used tonal cues to compute the phrasing for the target word such that it modulated phonological inferencing. This study can therefore be taken to show the relevance of tonal cues to prosodic domains, in this case, for the purpose of inferring whether a phonological process has applied. As Kim et al. note, an effect observed with prosodic structure signaled only by pitch, offers a strong argument for the role of language-specific prosodic structural effects in listeners’ processing of speech, as compared to a speech rate normalization account which is language-general.

These two previous studies therefore suggest that, independent of normalization for durational changes, listeners reference prosodic structure in speech processing. While Kim et al. (2018) focused on more abstract phonological inferencing effects, Mitterer et al. (2019) showed these

effects can extend to listeners' perception of phonetic detail in speech. Kim et al. also showed that tonal cues may play a central role in listeners' perception of prosodic structure. An emergent view that has come about based on these findings is that listeners process segmental and prosodic structures in parallel, as described by Mitterer et al.: "listeners compute the prosodic structure (prosodic processing), possibly in parallel with segmental processing" (see also Cho et al., 2007; Kim et al., 2018, Nakai & Turk 2011). The role of phrasal prosodic structure in this view is to guide interpretation of phonetic detail (Kim & Cho, 2013; Mitterer et al., 2019), and to modulate other domains of speech processing such as phonological inferencing (as in Kim et al., 2018) and word segmentation and recognition (e.g., Cho et al., 2007; Christophe, Peperkamp, Pallier, Block & Mehler, 2004; Salverda, Dahan, & McQueen, 2003). In this vein, we can consider the set of findings presented above to evidence the involvement of prosodically guided segmental processing for domain-initial patterns (Kim & Cho, 2013; Mitterer et al., 2019), as well as highlighting the importance of pitch as a cue to prosodic domains, as shown in Kim et al., (2018).

Several past studies have also looked at right-edge positional effects on listeners' perception of durational cues. Nooteboom & Doodeman (1980), manipulated the syntactic structure of various carrier phrases in which a Dutch target word was placed. Dutch listeners categorized it as having a phonemically short or long vowel. The prediction was that if listeners are sensitive to phrases as the domain for final lengthening (here represented by Nooteboom & Doodeman in terms of syntactic structures), they should adjust their perception of contrastive vowel length. In particular, if a given target sound is perceived as being phrase, or utterance-final, listeners should expect it to have undergone lengthening. In similar fashion to VOT in American English (as in Kim & Cho, 2013) or glottalization in Maltese (as in Mitterer et al., 2019), we can therefore consider duration as serving to cue a phonemic contrast in Dutch, while also providing listeners with information

about prosodic organization. If this prediction were borne out, we should expect to see listeners compensatorily adjust their categorization of vowel length, i.e., a vowel in phrase-final position will need to be longer than a vowel in phrase medial position to be perceived as phonemically long. Nooteboom & Doodeman find this result, which presents suggestive evidence that Dutch listeners may use prosody to guide their perception in this way. However, given that Nootebaum & Doodeman's carrier phrases consisted of different words, and were produced in different utterances, their stimuli present a high degree of variability in context: i.e., the words surrounding a target, and adjacent segmental durations, vary across condition. In light of the possible effects of durational context on listeners' perception of durational cues (as discussed in Mitterer et al., 2016; Steffman, 2019a), these results should perhaps be interpreted cautiously. More recently, Steffman (2019b) tested how American English-speaking listeners would perceive vowel duration as a cue to coda obstruent voicing, using a "coat" to "code" continuum, where longer vowels occur before voiced obstruents (e.g., Chen, 1970) and are used as a cue to voicing (e.g., Raphael, 1972). Steffman manipulated phrasal position simply as the presence/absence of following material in a carrier phrase. The stimuli were designed such that speech rate normalization effects should predict the opposite of a prosodically guided interpretation of duration, by making added post-target material lengthened such that it would predict contrast effects when present (see also Miller & Liberman, 1979). Steffman found the expected prosodically-guided effect: listeners required longer vowel duration for a voiced "code" percept when the target sound was phrase-final, suggesting a compensatory adjustment for phrase-final lengthening. These findings together suggest both Dutch and American English-speaking listeners exhibit a sensitivity to right-edge durational patterns in their perception of durational cues.

In the present study we present two experiments which seek to extend the research outlined above. We test how Tokyo Japanese speaking listeners' interpretation of a target sound as (intonational) phrase-final or phrase-medial influences their perception of contrastive vowel length, in the same vein as Nootebaum & Doodeman (1980), and Steffman (2019b). This will, broadly, help better our understanding of right-edge temporal effects on listeners' processing of durational cues cross-linguistically, and accordingly help inform recent proposals related to the parallel processing of prosodic and segmental structures. Our goal in testing Japanese is to additionally implement purely tonal cues to prosodic structure, as informed by models of Japanese intonational phonology, discussed below. To this end we manipulated only f0 in a carrier phrase, which avoids possible pitfalls of changing the durational context in which a target sound appears, discussed above. This also presents a departure from Nootebaum & Doodeman (1980), and Steffman (2019b), who either did not control for f0 across conditions, or did not manipulate it all. This will thus allow us to assess the role that pitch-based cues play for Japanese listeners in their interpretation of prosodic structure, building on Kim et al.'s (2018) finding for AP-final tonal patterns in phonological inferencing in Korean.

1.2. The present study

Tokyo Japanese presents a valuable test case, given that it has a well-described intonational system, and contrastive vowel length. The present study thus tests if listeners rely on intonation to compute prosodic boundaries, which may mediate their processing of durational cues.

In describing the intonational contexts manipulated in our experiment, we adopt the Autosegmental-Metrical (AM) model of Japanese intonational phonology developed by Beckman and Pierrehumbert (1986), Pierrehumbert and Beckman (1988), Venditti (1995; 2005) and

Maekawa, Kikuchi, Igarashi, and Venditti (2002). In the recent versions of the model, there are two tonally-defined prosodic groupings above the word level: the accentual phrase (AP) and the intonational phrase (IP). The AP is defined as having a phrasal H tone (H-) around the second mora and a subsequent gradual fall to a low tonal target (L%) at its right edge. It is also regarded as the domain of pitch accent realization: an AP can accommodate at most one pitch accent (shown as "A" in XJ-ToBI, as proposed by Maekawa et al., 2002), which is realized as a sharp f0 fall starting near the end of the accented mora. The IP, on the other hand, consists of one or more APs, and is marked by an initial L boundary tone (%L). Each IP has its own pitch range, so the effect of downstep, by which the f0 height of a pitch accent is lowered when following another pitch accent, is reset at an IP boundary.

As in many languages, phrase-final lengthening has been documented in Japanese (e.g., Takeda, Sagisaka, & Kuwabara, 1989). On the basis of the prosodic hierarchy proposed by Beckman and Pierrehumbert (1986) and Pierrehumbert and Beckman (1988) (which includes an intermediate phrase), Ueyama (1999) conducted a production experiment, comparing vowel duration at the right edge of four different prosodic levels: AP, intermediate phrase (ip), IP, and sentence. Ueyama's results showed that the four prosodic levels fell into two groups such that a vowel was significantly longer in IP-final position and sentence-final position as compared to AP-final position and ip-final position. This indicates that the duration of a domain-final vowel varies with the strength of the prosodic boundary. Seo, Kim, Kubozono, and Cho (2019) additionally find that phrase-final lengthening in Japanese appears to be mediated by the syllable structure (cf. Shepard, 2008) by showing that the effect of phrase-final lengthening on the rime of the CVN syllable (e.g., *takan "sensitive"*) was comparable to that on the final vowel of the CV syllable (e.g., *taka* "hawk").

The present study addresses the perceptual relevance of phrase-final lengthening in two experiments, testing if listeners' perception of contrastive vowel length shifts based on whether a target sound is expected to undergo phrase final lengthening (when cued as phrase final). In both experiments, listeners categorized a phrase-final vowel from a vowel duration continuum as phonemically long or short. Target words in both experiments contrasted only in the length of the final vowel, and were both disyllabic, however accentedness varied across experiments. In Experiment 1, the target word had an accent on the first syllable, while in Experiment 2 it did not. Testing both an accented and unaccented word pair allows for basic replication of the predicted effect, and allows us to help generalize the effect by testing different tonal environments across experiments, which vary based on the pitch accent status of the target word (described below). Findings from the present study will accordingly better our understanding of how prosody, and particularly tonal cues, shape listeners' perception of domain-final temporal structure in speech, extending the lines of research outlined above.

2. Experiment 1

In Experiment 1 we tested how Japanese listeners were influenced by changes in contextual f0 in their perception of contrastive vowel length, in this case for an accented target word minimal pair. We implemented a 2AFC task, in which listeners categorized a sound from a vowel duration continuum as phonemically long or short. f0 was manipulated in a carrier phrase to signal a target as IP-medial or IP-final.

2.1. Methods

Listeners categorized a disyllabic minimal pair with accent on the first syllable, that contrasted in terms of the phonemic length of the second syllable. The target was categorized as *shi'sho* "librarian" (司書) or *shi'shoo*[1] "master" (師匠). These words are both fairly low frequency based on word counts in Corpus of Spontaneous Japanese (CSJ) (Maekawa, Koiso, Furui & Isahara, 2000; Makawa, 2003) (log-transformed frequency = 0.3 for *shi'sho* , 1.28 for *shi'shoo*). The carrier phrase used for both medial and final conditions is shown in (1) below, with glosses. It had a similar design to the carrier phrase from Shepherd's (2008) production study.

(1) wata'shitachi-wa *x* (target) de'sukara shinraideki-ma'su

we-TOP *x* (target) because/therefore reliable-be

(1) can be phrased in two different ways, which are represented in (2) and (3) below, where […]$_{AP}$ represents an AP boundary and […]$_{IP}$ represents an IP boundary. Importantly, changing the phrasing along these lines alters the implied position of the target sound, such that in (2) it is IP-medial (and AP-medial), and in (3) it is IP-final. Both phrasings are possible, by virtue of the fact that Japanese sentences can end with a noun omitting a copula.[2] A translation is given as well.

(2) *One* IP: "Because we are x (we are) reliable." (x = medial)

[[wata'shitachi-wa]$_{AP}$ [*x* de'sukara]$_{AP}$ [shinraideki-ma'su]$_{AP}$]$_{IP}$

[1] Here and throughout the paper we use an apostrophe to indicate that the preceding mora is accented.
[2] The absence of the copula *desu* may decrease the perceived formality of the stimuli, which could be perceived as slightly less natural by some listeners. However, this occurs often in spontaneous speech and was judged to sound natural by a native Japanese speaker (author HK).

(3) *Two* IPs: "We are x. Therefore (we are) reliable." (x = final)

[[wata'shitachi-wa]$_{AP}$ [*x*]$_{AP}$] $_{IP}$ [[de'sukara]$_{AP}$ [shinraideki-ma'su]$_{AP}$]$_{IP}$

Visual representations of this manipulation for the stimuli used in Experiment 1 are shown in Figure 1. In creating the different possible phrasings, f0 was manipulated to differ only on the second syllable of the target word, and the following syllable (/de/ in *de'sukara* "because/therefore"). Manipulating f0 in this way was judged to be desirable because it entailed a minimal difference across conditions, but also produced a clear perceived change in phrasing as judged by a ToBI-trained native Japanese speaker (author HK). In the medial condition (Figure 1, top panel), the target word *x* forms a single AP with the following conjunction *de'sukara*. Since an AP has at most one pitch accent, the accent on the first syllable of *de'sukara* is deleted (Poser, 1984) or at least phonetically reduced (Kubozono, 1987; Maekawa, 1994). This results in a gradual f0 fall over the AP containing the target word and the following conjunction. In the final condition (Figure 1, bottom panel), on the other hand, the target word does not phrase with the *de'sukara*, and is followed by an IP boundary. This condition is signaled by lower f0 on the target syllable due to the L% associated with the right edge of the phrase, and the realization of the pitch accent on the first syllable of *de'sukara* as well as the absence of downstep on that syllable, which is signaled by the f0 height of the first syllable of *de'sukara* being as high as that of the accented syllable of the target word.

[FIGURE 1 HERE]

[FIGURE 2 HERE]

2.2. Materials

Stimuli were created by resynthesizing the speech of ToBI-trained male speaker of the Tokyo dialect of Japanese (author HK). The speaker was first recorded at 44.1 kHz in a sound-attenuated booth, using an SM10A Shure™ microphone and headset. Stimulus manipulation was carried out in Praat (Boersma & Weenik, 2019), using the PSOLA method (Moulines & Charpentier, 1990).

The starting points for the manipulation was a production in which the target was phrased IP-medially as in the top panel of Figure 1, and one in which it was phrased IP-finally, as in the bottom panel. As outlined above, only the f0 on two syllables varied across conditions: the second syllable in the target word, and the following syllable /de/. The remainder of the carrier phrase was acoustically identical across conditions, including the duration of acoustic silence following the target word (likely attributable mostly to the stop closure for following /d/), as shown in Figure 1. The duration of this interval between the end of the target word and the beginning of the following syllable /de/ was approximately 50ms in duration. As it is argued in Venditti (2005), a pause is not obligatory for Japanese listeners to perceive a disjuncture equivalent to an IP boundary. In spontaneous speech, there are many cases in which a large degree of disjuncture is solely cued by a boundary tone without an intervening pause. Likewise, a pause can be present without cueing large disjuncture. Thus, this relatively short silent interval is compatible with both phrasings.

The starting point for f0 manipulations in Experiment 1 was a naturally produced IP-medial phrasing, as in (2), with a phonemically long vowel target (*shi'shoo)*. f0 from another medial production of the second syllable in the target word, and the post-target syllable /de/ was resynthesized onto these two syllables in the carrier phrase, to create the medial condition, shown in the boxed region of the top panel of Figure 1. To create the final condition, we overlaid the

second target syllable and post-target syllable /de/ with f0 values from a natural IP-final production (Figure 1, bottom panel). In this way, f0 was resynthesized on the crucially differing syllables in *both* conditions. These manipulations were judged to sound like a natural medial and final phrasing by a ToBI-trained native speaker of Japanese (author HK). Subsequently, A vowel duration continuum was resynthesized from both medial and final conditions. The duration of the starting vowel was approximately 100ms. It was manipulated to range from 60 to 180ms of vowel duration. This manipulation was accomplished by linear compression and expansion of the target vowel material such that the entire vocalic portion was compressed or expanded. The continuum had 8 evenly spaced steps that included these endpoint values, with a between-step durational difference of approximately 17ms (note that all continuum steps were created via resynthesis, the unaltered original was not used). Spectrograms of continuum endpoints are shown in more detail in Figure 2. These manipulations resulted in 16 unique stimuli (2 positional conditions × 8 continuum steps), with the only difference across conditions being the f0 on the second syllable of the target and post-target /de/ (shown in Figure 1). By varying only f0 we ensured that changes in adjacent segmental duration are ruled out as a possible explanation, as discussed in by Mitterer et al. (2016). Additionally, psychoacoustic influences of pitch on perceived duration, whereby high f0 and more dynamic pitch contours can lead to increased perceived duration, are unlikely to play a role given that these effects have been shown to occur only in isolated monosyllables (Van Dommelen, 1993), and given that psychoacoustic effects also do not seem to occur when pitch has a possible prosodic interpretation (Steffman & Jun, 2019).

2.3. Participants and procedure

26 participants (13 males and 13 females; mean age 28) were recruited for Experiment 1. 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 presented with audio stimuli binaurally via a Peltor™ 3M™ listen-only head set, while seated in a quiet room, in front of a laptop computer, in Tokyo. Participants were presented with orthographic representations of the target words on the laptop during audio presentation, 司書 representing *shi'sho* "librarian, and 師匠 representing *shi'shoo* "master". These words were each displayed centered on either half of the computer screen. The side of the screen on which each word appeared was counterbalanced across participants, i.e., for 13 participants 司書 was on the left side of the screen and for 13 participants 師匠 was on the left side of the screen. Participants indicated their response via key press: an 'f' key press indicated the choice on the left side of the screen, a 'j' key press indicated a choice on the right side of the screen. Prior to the beginning of the test trials, participants completed 8 practice trials, in which they heard each continuum endpoint, in each positional condition twice, in random order. During experimental trials, participants categorized 12 instances of each unique stimulus, in random order, for a total of 192 trials (8 continuum steps × 2 position conditions × 12 repetitions). All responses (excluding practice trials) were analyzed.

2.4. Results and discussion

Results were assessed statistically by a linear mixed-effects model with logistic linking function, implemented using the *lme4* package in R (Bates et al., 2015). The model was fit to predict listeners'

response (*shi'sho* or *shi'shoo*) as a function of continuum step, position condition, and the interaction of these two fixed effects. The dependent variable was coded such that a long vowel (*shi'shoo*) response was mapped to 1, with a short vowel response mapped to 0. Therefore, a positive coefficient would represent an increase in long vowel responses. Position was contrast-coded, with FINAL mapped to 1, and MEDIAL mapped to -1. Given this coding scheme, we can restate our predictions in model terms. Firstly, we predict that changes in vowel duration along the continuum should significantly impact listeners' responses. Longer vowel durations should increase the log-odds of a long vowel response, something that we would expect from any continuum of this sort (showing a positive beta based on how the variables were coded). Our crucial prediction is that if listeners require longer vowel duration to perceive a phonemically long vowel in final position, they should show decreased *shi'shoo* responses in the final condition. In other words, the final position should significantly *decrease* the log odds of a *shi'shoo* response (showing a negative beta based on how the variables were coded). We did not predict a significant interaction between the two fixed effects.

Following e.g., Barr et al. (2013), we specified the model random effect structure as by-participant intercepts with maximal random slopes, and no correlation between random effects. A model that included correlations did not converge, at which point the correlation parameter was removed. The converging model therefore contained de-correlated random slopes for both fixed effects and the interaction between them. Further simplification of random slopes led to decreased model fit, assessed by inspecting AIC and by likelihood ratio tests (Matushek et al. 2017), suggesting the fully specified random effect structure is justified. The model output is shown in Table 1, with results plotted in Figure 3.

Table 1: Fixed effects from the model in Experiment 1. Values are rounded. Asterisks indicate p-values, where * = p < 0.05, ** = p < 0.01, and *** p < 0.001.

	β	SE	z
(Intercept)	-0.65	0.26	-2.57*
position	-0.39	0.14	-2.84**
step	6.08	0.34	17.78***
position:step	0.43	0.23	1.86

[FIGURE 3 HERE]

Increasing vowel duration along the continuum (the "step" factor in the table above), significantly increased long vowel responses (β = 6.08, z = 17. 78). This is expected as outlined above. Position, the predictor of interest also showed a significant effect. As can be seen visually in Figure 3, the position of the target word impacted listeners' categorization: an IP-final target showed significantly decreased *shi'shoo* responses (β = -0.39, z = -2.84). The interaction between the two fixed effects was not significant.

The effect of position found in the model supports our predictions: listeners required longer vowel durations for a long vowel (*shi'shoo*) percept when the target word was cued as phrase-final. The results of Experiment 1 can therefore be taken to suggest that Japanese listeners rely on phrasal boundaries in their perception of contrastive vowel length. Because only contextual f0 varied across conditions, we can conclude this computation of this prosodic context is crucially informed by f0. This provides a new piece of evidence in the same vein as Steffman (2019b), and additionally shows listeners use pitch-based cues alone to construe the prosodic boundary location relative to a target sound (as in Kim et al., 2018). Following these results for an accented target word, we can ask if an analogous pattern can be observed for a target that is unaccented. In doing

so, we can see if the effect replicates, and generalizes to a different context, given that an accented target word engenders a different realization of contextual f0, outlined below.

3. Experiment 2

3.1. Methods

As in Experiment 1, participants categorized a sound from a vowel duration continuum as phonemically long or short. In Experiment 2, listeners categorized an unaccented disyllabic minimal pair as *dookyo* "housemate" (同居) or *dookyoo* "townmate" (同郷), which were comparable in log-transformed frequency based on word counts in the CSJ (1.62 for *dookyo* and 0.95 for *dookyoo*). Note these are both fairly low frequency as the target words in Experiment 1. The same carrier phrase as Experiment 1 was used for both medial and final conditions.

3.2. Materials

As in Experiment 1, the stimuli for the medial and final conditions differ only in the f0 of the target syllable and that of the following syllable (i.e., /de/ in *de'sukara* "because/therefore"). Visual representations of both medial and final conditions for Experiment 2 are shown in Figure 4. In the medial condition (Figure 4, top panel), the target word *x* forms an AP with the following conjunction *de'sukara*. Unlike Experiment 1, however, the accent on the conjunction is realized, since the target word is unaccented. The medial condition is thus characterized by the higher f0 on the target syllable due to a lack of the L% boundary tone, as well as the absence of pitch reset on the following syllable. In the final condition (Figure 4, bottom panel), in which the target word does not phrase with the following conjunction, the f0 on the target syllable is lowered due to the presence of the L% associated with the right edge of the AP. Furthermore, the f0 on the following

syllable is slightly higher in the final condition than in the medial condition, reflecting a pitch reset, which cues the presence of an IP boundary after the target word. In this way, the tonal context of the target is shaped by its accentual status, and we can therefore test of how the effect observed in Experiment 1 generalizes to unaccented words. Representative examples of the conditions in Experiment 2 are shown in Figure 4. Figure 5 shows spectrograms of continuum endpoints in more detail.

[FIGURE 4 HERE]

[FIGURE 5 HERE]

The carrier phrase used for the Experiment 2 stimuli was identical to that used for Experiment 1, with one exception: the f0 on the post-target syllable /de/. Because of the different realization of accent on this word described above, the f0 on this syllable was resynthesized to match with natural productions following both an IP-final and IP-medial unaccented target word produced by the same speaker in the same carrier phase. All other parts of the carrier phrase were identical to the carrier phrase in Experiment 1. The starting point for the manipulation of the target was a phonemically long vowel target *dookyoo*, produced in IP-medial position, as in Experiment 1. This vowel was approximately 100ms in duration. f0 from another medial production was resynthesized onto the second syllable of the target word, as in Experiment 1 (to ensure both conditions were equally resynthesized). This medial target word was cross-spliced into the medial frame used in Experiment 1 (with different f0 on post-target /de/ as outlined above), shown in the top panel of Figure 3. To create the final condition target, the second syllable of the naturally produced (not resynthesized) medial target was overlaid with the f0 from an IP-final production, and cross-

spliced (with the first syllable) into the IP-final carrier sentence, shown in the bottom panel of Figure 3. Following this, a vowel duration continuum with the same endpoint durations and inter-step intervals as in Experiment 1 was resynthesized from both conditions, resulting in a total of 16 unique stimuli (2 positional conditions × 8 continuum steps). By virtue of changing f0 on post-target /de/, the new carrier phrase was now judged to sound natural for both phrasings of an unaccented target word, while remaining fairly comparable to the carrier phrase used in Experiment 1.

3.3. Participants and procedure

26 different participants (15 males and 11 females; mean age 26) were recruited for Experiment 2. Unlike Experiment 1, for logistical reasons, participants in Experiment 2 were tested in two different locations. 14 were tested in Tokyo, Japan, and 12 were tested in Los Angeles, California. Participants tested in Los Angeles has been living in the US for an average of one year. Given that all participants were native speakers of Tokyo Japanese, we did not expect testing location to impact their performance. However, to be sure, we ran a preliminary model with the same structure as that used in Experiment 1 which additionally included location (Los Angeles/Tokyo) as a predictor, as well as its interaction with position (medial/final). The model revealed, as expected, location did not impact responses overall, and that crucially, it did not interact with position (suggesting an analogous effect of position regardless of the country in which participants were tested). The model we report below does not include location as a predictor. The procedure was identical to that in Experiment 1.

3.4. Results and discussion

The model specifications used to assess the Experiment 2 results was the same as that in Experiment 1, with a long vowel (*dookyoo*) response mapped to 1. The same full random effect structure, with no correlation between random effects, was specified in the model. As in Experiment 1, this fully specified random effect structure was observed to provide the best model fit in comparison to simplified variants, and so was retained. The model output is given in Table 2 below. The results are plotted in Figure 6.

Table 2: Fixed effects from the model in Experiment 2. Values are rounded.

	β	SE	z
(Intercept)	1.25	0.23	5.54***
position	-0.66	0.10	-6.42***
step	6.08	0.37	16.49***
position:step	-0.41	0.20	-2.01*

[FIGURE 6 HERE]

As would be expected, increasing vowel duration increased listeners' *dookyoo* responses (β = 6.08, z = 16.49). As shown in Figure 6, position also had a significant effect: an IP-final target showed significantly decreased *dookyoo* responses (β = -0.66, z = -6.42). Unlike Experiment 1, a significant interaction (p = 0.045) between these two main effects was observed. This stems from the fact that the medial condition shows a sharper increase in *dookyoo* as a function of increasing vowel duration, as compared to the final condition, i.e., the effect of continuum step is (slightly) larger in the medial condition (cf. Lunden, 2013). The main effect of position observed in the model concurs with our predictions in showing that listeners required longer vowel durations for

a *dookyoo* response in the final condition, effectively requiring longer vowel durations to perceive a vowel as phonemically long when phrase-final. This can therefore be taken to replicate the effect seen in Experiment 1, showing that accent-dependent pitch patterns modulate listeners' perception of final lengthening for both accented, and unaccented words. The effect in both experiments aligns with the predictions laid out above.

One observation across Experiments 1 and 2 is that the magnitude of the effect is larger in Experiment 2 (β(SE) = -0.66(0.10)) as compared to Experiment 1 (β(SE) = -0.39(0.14)), suggesting a possible sensitivity to accent in the effect of boundary, where the boundary effect is larger for an unaccented target word (though clearly robust in both cases). The present experiments only allow us to speculate in this regard, because in addition to varying in accentedness across experiments, target words also varied in moraic and segmental make-up (e.g., short /i/ precedes the target in Experiment 1, long /oo/ precedes it in Experiment 2). Variation in acoustic context (including f0) across experiments therefore makes isolating the effect of accent on the magnitude of the positional effect impossible. Nonetheless, with the goal of motivating future research, we can consider how an accent-driven asymmetry might be explained by recent findings which show asymmetrical final lengthening in initial-accented versus unaccented words.

Seo et al. (2019) found that disyllabic words with an initial accent (e.g., *ta'ka,* a proper name) exhibited less lengthening on their final syllable compared to disyllabic words without an accent (e.g., *taka* "hawk"). Reduced final lengthening for words with a preceding accent was described as a suppression of pre-boundary lengthening by the authors and was hypothesized to serve the function of maintaining a syntagmatic contrast between the final rhyme and preceding accented syllable, consistent with previous studies which have shown an interplay between phrase-final lengthening and prominence in other languages (cf. Turk and Shattuck-Hufnagel, 2007 for

English, Katsika, 2016 for Greek, and Nakai, Kunnari, Turk, Suomi, and Ylitalo, 2009 for Finnish). One hypothesized perceptual consequence of this pattern in Japanese is that listeners may show increased sensitivity to a phrasal boundary's influence for unaccented words, which undergo more substantial lengthening (Experiment 2). In comparison, accented words may show a relatively limited effect of phrasal position, reflecting relatively limited phrase-final lengthening (Experiment 1). If we interpret the results along these lines, they could be taken to reflect an interplay between the prominence marking and boundary marking systems of the language.[3]

Given the other differences between target words across experiments, we cannot conclude anything concrete in this regard. Future research will accordingly benefit from addressing the possible involvement of pitch accent as a mediating factor for the observed boundary effect directly. Finding evidence for prominence-mediated boundary effects in perception would enrich our understanding of the amount of detail encoded in prosodic representations (e.g., reduced temporal expansion based on accentedness), and how different facets of prosodic organization interact in this domain.

4. General discussion

Taken together, Experiment 1 and 2 provide evidence that listeners compute prosodic boundary information in processing contrastive vowel length and use this information to guide their interpretation of duration. Specifically, we found that a phrase-final target required longer vowel

[3]Another reason to be tentative in this regard is the observation that overall more long vowel responses were given in Experiment 2 as compared to Experiment 1 (identifiable from the intercept in each model, the first row of the model summary). This represents the overall tendency for a short or long vowel response in an Experiment (different from the effect of position). The pattern in the intercepts is different from that predicted by Seo et al. (2019). That is, we might expect to see fewer long vowel responses *overall* in unaccented words (as in Experiment 2) as they are further lengthened in final position, leading to an expectation of a vowel needing to be longer to be categorized as phonemically long when the word is unaccented. These observed differences in intercept could also arise from the differences in segmental and acoustic context across experiments, as mentioned above.

duration to be categorized as phonemically long. We took this effect to reflect the influence of intonational structure in listeners' perception of segmental contrasts such that phrase-final sounds are expected to be lengthened, and longer vowel duration is therefore required for a long vowel percept.

Generally speaking, we can take this result to provide additional evidence for the relevance of prosodic factors in listeners' perception of segmental contrasts cross-linguistically, and to highlight the continued importance of extending recent work on this topic (e.g. Kim et al., 2018; Mitterer et al., 2019).

The effect of phrasal position, though reliable in both Experiments, is fairly small, particularly in Experiment 1. Shifts in categorization are such that listeners' responses are impacted only for several steps on the continuum, and only to the extent that they increase slightly, showing a minimal rightwards shift in the categorization function. Though the effect is larger in Experiment 2, it remains fairly restricted and localized at the most ambiguous continuum steps. It can also be noted generally that prosodic effects from previous studies discussed above seem to be fairly subtle, and to engender small adjustments in categorization and online processing. In light of this, the restricted nature of the effects seen here could relate to the *way in which* listeners make use of prosodic information in speech processing, discussed below. The small effect size might also originate from the fact that only one cue to boundary (f0) was manipulated, and other cues, such as a pause, or post-boundary initial strengthening, are lacking. Seeing if additional boundary cues create a larger effect would be a useful further direction, especially given that we see a robust shift in categorization in this conservative test case, where temporal context is totally controlled. As an example, the presence of a post-target pause could be crossed with tonal manipulations in future studies. Though this would introduce variation in temporal context surrounding the target,

it would allow us to test if a stronger boundary percept, indexed by larger shifts in categorization, obtains when boundary cues combine (cf. Nakai and Turk 2011, Mitterer et al. 2016). More generally, looking cross-linguistically to test how different boundary tones, or boundary tones in combination with other cues such as glottalization (in American English; e.g., Redi & Shattuck-Hufnagel, 2001) will help better our understanding of what informs perception of a prosodic boundary for the purposes of guiding segmental interpretation.

We can consider these results in light of two sets of previous findings outlined above. First, in regards to previous work on phrase-final effects (Nootebaum & Doodeman, 1980; Steffman, 2019b), the present experiments offer additional evidence for the relevance of right-edge temporal patterns in perception cross-linguistically. They also offer some controls which were not present in these previous studies: by holding contextual duration constant across conditions, any possible confounds related to speech rate normalization are removed, as discussed above. With position cued only by pitch we can also complement these previous findings in highlighting the importance of intonational structure and its relation to phrasal boundaries for the purpose of informing listeners' perception of prosody. More generally, we can take the present results to suggest tonal cues play an important role, in line with the status of tonal events encoded in models of the intonational phonology of Japanese. These findings also suggest future work that is informed by models of intonational phonology may help shed light on the sorts of structures which listeners compute, and the cues that specify them.

We can also consider these results in terms of recent proposals for parallel processing of segmental and prosodic structures, discussed above. In this light, these results can be taken as another piece of evidence for proposed "prosodic analysis" in which a prosodic representation guides listeners' interpretation of segmental contrasts. As discussed above, both Mitterer et al.

(2019) and Kim et al. (2018) provide evidence for this sort of role for prosodic boundaries, both in perception of phonetic detail (e.g., phrase-initial glottalization) and in phonological inferencing. Time-course evidence from these studies, not discussed above, showed these effects occur relatively late in processing, which may be taken to suggest they involve later-stage modulation of lexical competition, after segmental material activates lexical hypotheses (Cho et al., 2007). The involvement of prosodic structure in a later stage of processing might explain its relatively limited influence on categorization observed in the present experiments. That is, prosodic structure may exert a role only in the more ambiguous cases (i.e. when both possible word forms are fairly equally activated, see Newman et al. 1997 for discussion of lexical activation/competition in a forced choice perception task), and then only in a non-deterministic fashion (Cho et al. 2007). In this regard, prosodic boundary information may be integrated in perception in a secondary fashion as compared to e.g. vowel internal durational cues. The present results, being purely offline, cannot speak to this timing prediction. Accordingly, one promising extension would be to test the time-course of the observed effects with eye-tracking (using a similar paradigm as e.g., Reinisch & Sjerps, 2013; Kingston et al., 2016). Seeing if these prosodically-guided effects occur with a delayed time-course, and comparing these effects to that of vowel duration along the continuum, a segment-internal cue that is used rapidly (Reinisch & Sjerps, 2013), might offer a window into the processes that underly our observed effect. A later time-course would be a strong argument in favor of this proposed later-stage prosodic analysis. Extending the present findings in this way may thus offer useful converging evidence for this idea. These results nevertheless may tell us something informative about the role of prosodic analysis in speech perception, mainly the importance of intonational cues in guiding the perception of phonetic detail. As discussed above, Kim et al. (2018) showed intonational structure, cued only by pitch, seems to play an important

role in phonological inferencing. The present results would suggest that computation of intonational structure similarly influences processing of phonetic detail as well. Together, recent findings (Kim & Cho, 2013; Kim et al., 2018; Mitterer et al., 2016, 2019; Steffman, 2019a, 2019b) would suggest prosodic boundaries, of various types, and cued by various means, merit further research as a mediating factor at multiple levels of speech processing.

In addition to exploring these effects with online measures, other extensions of our findings will benefit from testing our speculation related to the role of accent, outlined above. Finding evidence for a mediating effect of prominence in this sort of task would point to an interplay between these two facets of prosodic organization in perception. More generally, finding other test cases which allow for researchers to exploit intonational patterns will also enrich the present findings, and could be used as a test for the relevance of various intonational properties, or claimed functions of intonational tunes in a given language, for language perceivers. Extending the present findings along these lines will better our understanding of the role of intonation in language comprehension, and will help inform a theory of the processes which underpin it.

Acknowledgements

We are grateful to all of our participants for their time, and to Sun-Ah Jun and three anonymous reviewers for helpful feedback and commentary. Further thanks to Shigeto Kawahara, Mami Gosyo, and Naoki Ishikawa for recruitment assistance. A previous version of this work was presented at the 10th International Conference on Speech Prosody. This research was funded by the UCLA Ladefoged Scholarship, awarded to author HK.

References

Abramson, A., and Lisker, L. (1970). Discrimination along the voicing continuum: Cross-language tests, in Proceedings of the 6th International Congress of Phonetic Science, Prague, 1967 (Academic, Prague), pp. 569-573.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.05, retrieved from http://www.praat.org/

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In M. Broe, & J. Pierrehumbert (Eds.), Acquisition and the lexicon: Papers in laboratory phonology V (pp. 70–88). Cambridge, UK: Cambridge University Press.

Chen, M. (1970). Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica*, *22*(3), 129–159. https://doi.org/10.1159/000259312

Cho, T. (2015). Language Effects on Timing at the Segmental and Suprasegmental Levels. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 505–529). https://doi.org/10.1002/9781118584156.ch22

Cho, T. (2016). Prosodic Boundary Strengthening in the Phonetics–Prosody Interface. *Language and Linguistics Compass*, *10*(3), 120–141.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, *35*(2), 210–243.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. Journal of Memory and Language, 51, 523–547.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, *85*(5), 2154–2164.

Fougeron, C. and Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. The Journal of the Acoustical Society of America, 101(6), 3728-3740.

Keating, P., Fougeron, C., Hsu, C., & Cho, T. (2003). Domain initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press.

Keating, P. (2006). Phonetic Encoding of Prosodic Structure. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 167–186). New York and Hove: Macquarie Monographs in Cognitive Science, Psychology Press.

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19–EL25.

Kim, S., Mitterer, H., & Cho, T. (2018). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *PLOS ONE*, *13*(8), e0202912. https://doi.org/10.1371/journal.pone.0202912

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(12), 1969–1988. https://doi.org/10.1037/xhp0000269

Lunden, A. (2013). Reanalyzing final consonant extrametricality. *The Journal of Comparative Germanic Linguistics*, *16*(1), 1-31.

Maekawa, K. (1994). Is there 'dephrasing' of the accentual phrase in Japanese? *Working Papers in Linguistics: Papers from the Linguistics Laboratory 44*, 146-165.

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. In *Proceedings ISCA and IEEE workshop on spontaneous speech processing and recognition* (pp. 7-12).

Maekawa, K., Kikuchi, H., Igarashi, Y. & Venditti, J. (2002). X-JToBI: An extended J_ToBI for spontaneous speech. In *Proceedings of the 7th International Congress on Spoken Language Processing* (pp. 1545-1548).

Maekawa, K., Koiso H., Furui, S. & Isahara, H. (2000). Spontaneous speech corpus of Japanese. Proceedings of the Second International Conference of Language Resources and Evaluation, (pp. 947-952).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457–465. https://doi.org/10.3758/Bf03213823

Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, *108*, 104034. https://doi.org/10.1016/j.jml.2019.104034

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Commun.*, *9*(5–6), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Nakai, S., & Turk, A. E. (2011). Separability of prosodic phrase boundary and phonemic information. *The Journal of the Acoustical Society of America*, *129*(2), 966–976. https://doi.org/10.1121/1.3514419

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(23), 873–889.

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287. https://doi.org/10.1121/1.383737

Raphael, L. J. (1972). Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303. https://doi.org/10.1121/1.1912974

Redi, L., and Shattuck-Hufnagel, S. (2001). "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics* 29, 407–429.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116. https://doi.org/10.1016/j.wocn.2013.01.002

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90, 51–89. https://doi.org/10.1016/S0010-0277(03)00139-2

Seo, J., Kim, S., Kubozono, H., & Cho, T. (2019). 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), 1817. https://doi.org/10.1121/1.5122191

Shepherd, M. A. (2008). The scope and effects of preboundary prosodic lengthening in Japanese. *USC Working Papers in Linguistics* 4, 1-14.

Steffman, J. (2019a). Intonational structure mediates speech rate normalization in the perception of segmental categories. *Journal of Phonetics*, *74*, 114–129. https://doi.org/10.1016/j.wocn.2019.03.002

Steffman, J. (2019b). 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), EL560–EL566. https://doi.org/10.1121/1.5111772

Steffman, J., & Jun, S.-A. (2019). Listeners integrate pitch and durational cues to prosodic structure in word categorization. *Proceedings of the Linguistic Society of America*, *4*(1). https://doi.org/10.3765/plsa.v4i1.4536

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance*, *7*(5), 1074–1095.

Poser, W. (1984). *The phonetics and phonology of tone and intonation in Japanese*. PhD. Dissertation, MIT.

Takeda, K., Sagisaka, Y. & Kuwabara, H. (1989). On sentence-level factors governing segmental duration in Japanese. *The Journal of Acoustical Society of America 86*, 2081-2087.

Turk, A. E., & Sawusch, J. R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, *25*(1), 25–41. https://doi.org/10.1006/jpho.1996.0032

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

Turk, A. E., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, *27*(2), 171–206. https://doi.org/10.1006/jpho.1999.0093

Ueyama M. (1999). An experimental study of vowel duration in phrase-final contexts in Japanese. *UCLA Working papers in Phonetics* 97, 174-182.

Van Dommelen, W. (1993). Does Dynamic f0 Increase Perceived Duration? New Light on an Old Issue. *Journal of Phonetics*, *21*(4).

Venditti, J. J. (1995). Japanese ToBI Labelling Guidelines. *Ohio State University Working Papers in Linguistics* 50, 127-162.

Venditti, J. J. (2005). The J_ToBI Model of Japanese Intonation. In S.-A. Jun, *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 172-200.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717. https://doi.org/10.1121/1.402450

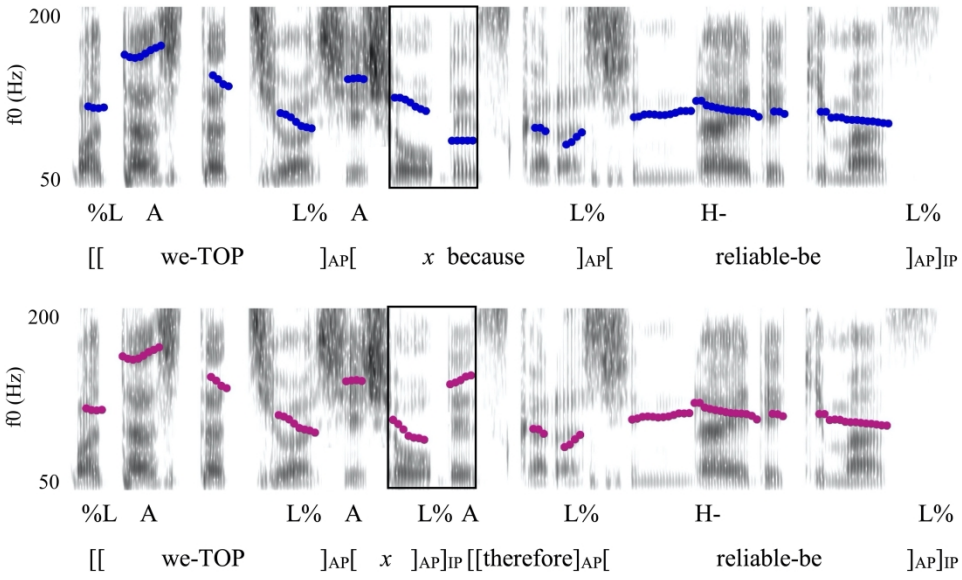

Figure 1: Example stimuli in both medial (top panel) and final (bottom panel) conditions in Experiment 1. Step 4 from the continuum is shown, with a vowel duration of approximately 110 ms. Spectrograms (0-5kHz range) overlaid with pitch tracks (50-200 Hz range) are shown. The black boxed region highlights the second syllable of the target word and the post-target syllable /de/, note this is the only difference across conditions. X-JToBI labels are given for tonal events, and below, glosses are bracketed according to their phrasing, where […]AP indicates an AP boundary and […]IP indicates an IP boundary.

171x98mm (600 x 600 DPI)

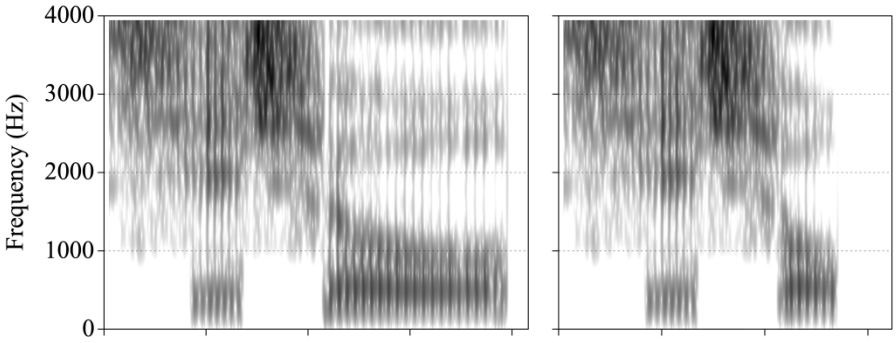

Figure 2: Spectrograms showing the continuum endpoints for the target word in Experiment 1 (the longest step at left, shortest at right). Time, marked by ticks on the x axis, is indicated in 100 ms intervals.

139x63mm (1000 x 1000 DPI)

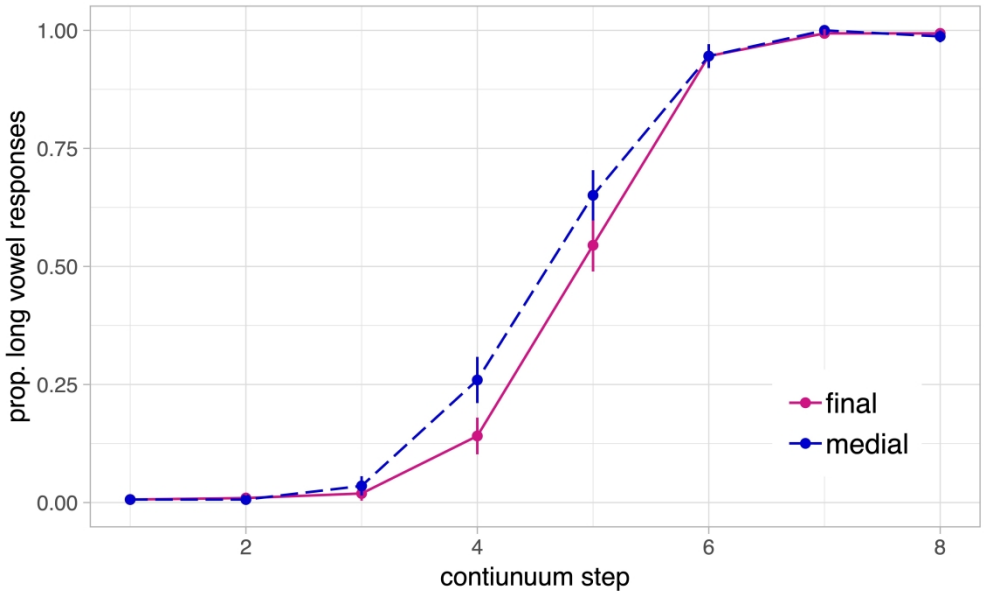

Figure 3: Categorization responses for Experiment 1, split by condition. The x axis shows numbered continuum steps (step 1 = 60 ms, step 8 = 180 ms, step intervals are approximately 17 ms). The y axis shows the proportion of long vowel responses. Error bars around each point represent 95% CI.

127x82mm (1000 x 1000 DPI)

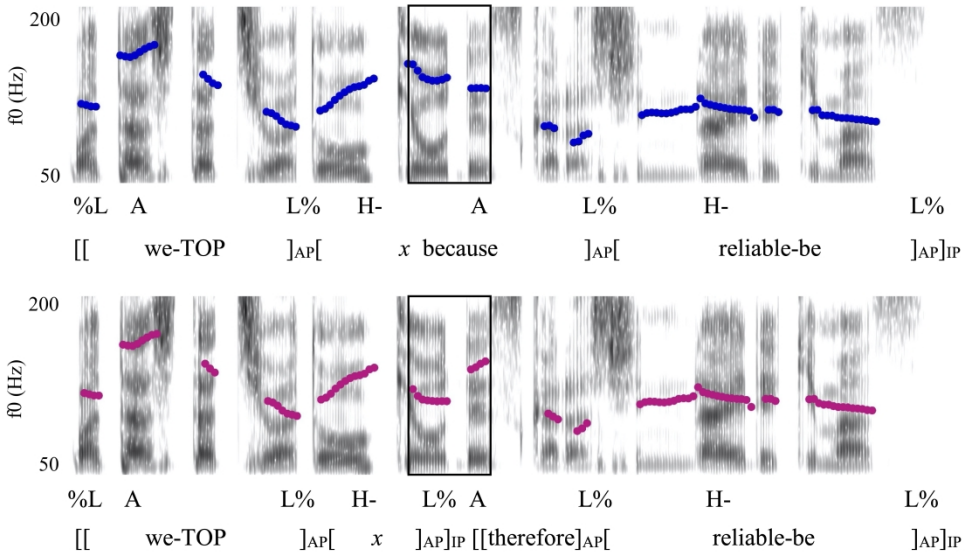

Figure 4: Example stimuli in both medial (top panel) and final (bottom panel) conditions in Experiment 2. Step 4 from the continuum is shown, with a vowel duration of approximately 110 ms. Spectrograms (0-5kHz range) overlaid with pitch tracks (50-200 Hz range) are shown. X-JToBI labels and bracketed glosses are shown, as in Figure 1. The boxed region highlights the second syllable of the target word, and post-target /de/.

2448x1362mm (72 x 72 DPI)

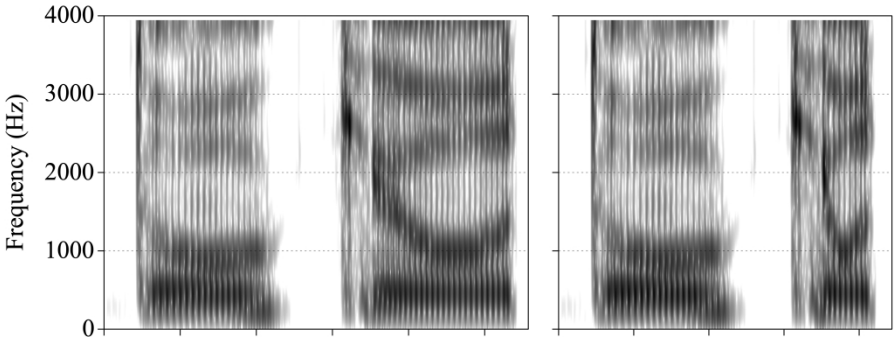

Figure 5: Spectrograms showing the continuum endpoints for the target word in Experiment 2 (the longest step at left, shortest at right). Time, marked by ticks on the x axis, is indicated in 100 ms intervals.

139x63mm (1000 x 1000 DPI)

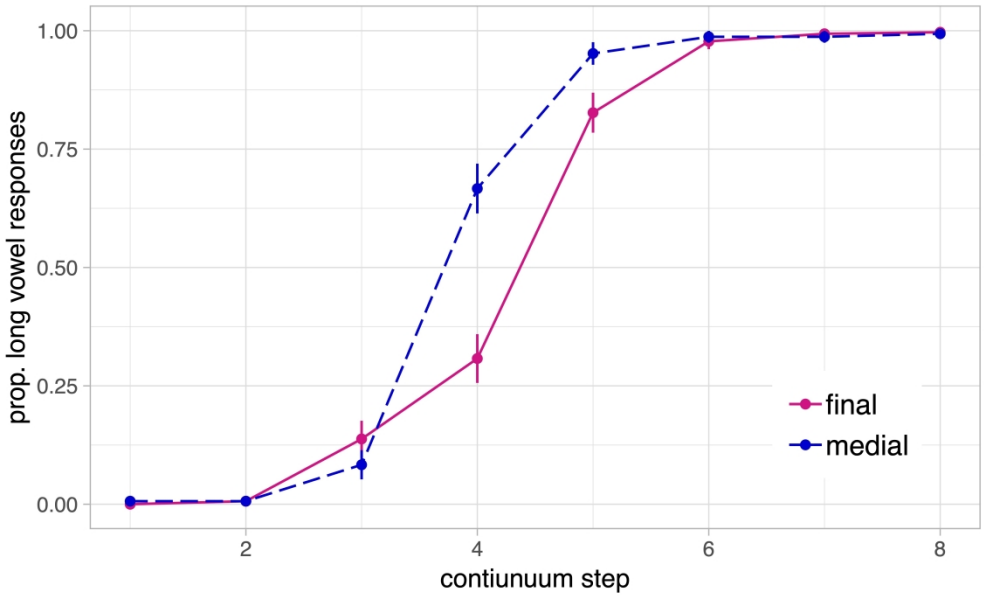

Figure 6: Categorization responses for Experiment 2, split by condition. The x axis shows numbered continuum steps (step 1 = 60 ms, step 8 = 180 ms, step intervals are approximately 17 ms). The y axis shows the proportion of long vowel responses. Error bars represent 95% CI for each point.

127x82mm (1000 x 1000 DPI)