2019. *Proc Ling Soc Amer* 4. 49:1-15. https://doi.org/10.3765/plsa.v4i1.4536

Listeners integrate pitch and durational cues to prosodic structure in word categorization

Jeremy Steffman & Sun-Ah Jun*

Abstract. In this study we investigate how listeners perceive vowel duration as a cue to voicing based on changes in pitch height, using a 2AFC task in which they categorized a target word from a vowel duration continuum as "coat" or "code". We consider this issue in light of (1) psychoacoustic perceptual interactions between pitch and duration and (2) compensatory effects for prosodically driven patterning of pitch and duration in the accentual/prominence-marking system of English. In two experiments we found that listeners' interpretation of pitch as a psychoacoustic, or prosodic event is dependent on continuum step size and range. In Experiment 1 listeners exemplified the expected psychoacoustic pattern in categorization. In Experiment 2, we altered the duration continuum in an attempt to highlight pitch as a language-specific prosodic property and found that listeners do indeed compensate for prosodically driven patterning of pitch and duration. The results thus highlight flexibility in listeners' interpretation of these acoustic dimensions. We argue that, in the right circumstances, prosodic patterns influence listeners' interpretation of pitch and expectations about vowel duration in the perception of isolated words. Results are discussed in terms of more general implications for listeners' perception of prosodic and segmental cues, and possibilities for cross-linguistic extension.

Keywords. speech perception; prosody; psychoacoustics; phonetic categorization; prominence; duration; pitch

1. Introduction. It is well established that the phonetic properties of speech segments, both acoustic and articulatory, are systematically modulated by prosodic factors (e.g. Cho 2015, 2016, Georgeton et al. 2016, Keating et al. 2003, Onaka 2003). This can be conceptualized, in a general sense, as the phonetic encoding of prosodic structure (e.g. Keating 2006). However, the extent to which listeners are sensitive to prosodically driven variation in perception remains an open question (cf. Kim & Cho 2013, Mitterer et al. 2016).

Previous studies investigating perceptual compensation for prosodic patterns (Kim & Cho 2013, Mitterer et al. 2016) have tested boundary phenomena, i.e. initial strengthening. In the present study we address this question in a new empirical domain by investigating listeners' perception of acoustic correlates of prominence marking in English. We further extend this line of research by testing how listeners' perception of isolated words may be mediated by prosodically driven variability in pitch and duration, whereas previous studies have used carrier phrases to provide prosodic context.

Specifically, we ask whether listeners' perception of duration and pitch is influenced both by psychoacoustic factors, and by the patterning of pitch and duration as correlates of accentedness (i.e. prominence marking) in the prosodic system of English. We test if listeners incorporate their

*Many thanks are due to Adam Royer for recording speech for the stimuli and Yang Wang for help with data collection. We are further grateful to attendees at the UCLA Phonetics seminar, and audience members at the 93rd annual meeting of the Linguistic Society of America for feedback on this project. Authors: Jeremy Steffman, UCLA (jsteffman@ucla.edu) & Sun-Ah Jun, UCLA (jun@humnet.ucla.edu).

experience with durational variation due to accent (signaled by pitch) in their perception of durational cues, in addition to domain-general auditory processes.

1.1. THE PRESENT STUDY. In the present study listeners categorized a "coat" ~ "code" continuum varying only in vowel duration. In English, among other languages, vowels before voiced obstruents are longer than those before voiceless obstruents (e.g. Chen 1970, Peterson & Lehiste 1960, van Santen 1992), and this is a robust cue to voicing for listeners (e.g. Raphael 1972). Pitch height on the vowel in the target word was manipulated to have one of two levels, HIGH and LOW (the creation of which is described in section 2.1). Using this continuum, we tested how listeners' perception of durational cues is influenced by changes in pitch height. Two different predictions for this manipulation are considered below in turn: (1) psychoacoustic predictions informed by documented perceptual interactions between duration and pitch and (2) linguistic/prosodic predictions informed by the patterning of duration and pitch as correlates of accentedness in English prosody.

Various psychoacoustic interactions between pitch and duration have been documented in the literature (e.g. Gruenenfelder & Pisoni 1980, Lehiste 1976, Shigeno 1986), suggesting that, to some extent, pitch and duration are *interactive* or *integrated* dimensions (e.g. Ellis & Jones 2009, Prince 2011)[1]. Importantly, the extent to which listeners integrate these cues appears to be flexible, varying on the basis of stimulus and task factors (Prince 2011). In the present study we consider just one interactive aspect: the influence of pitch height on the perception of duration. Higher pitch increases perceived duration, both for non-speech and speech stimuli (e.g. Brigner 1988, Gussenhoven & Zhou 2013, Šimko et al. 2016, Yu 2010, Yu et al. 2014). This is argued to have a domain-general auditory basis in light of the fact it is observed with non-speech and patterns similarly for speakers of different languages (Brigner 1988, Šimko et al. 2016). Previous studies find evidence of this in listeners' numerical ratings of duration, as well as explicit comparison of two stimuli.[2] In the present study, if listeners perceive increased pitch as increased vowel duration, they would be predicted to shift categorization of the target sound such that a vowel with HIGH pitch is perceived as longer, and thus more likely to be categorized as "code". In other words, listeners' perceptual integration of duration of pitch may influence their perception of vowel duration as a cue to voicing, *increasing* "code" responses when pitch is HIGH.

These psychoacoustic predictions can be contrasted with predicted compensatory effects, guided by listeners' interpretation of pitch as a correlate of prosodic structure. We first consider some structural properties of English prosody related to accentedness. Most accented syllables in English are marked with high (H*) pitch accents (Dainora 2006), while unaccented syllables tend not to have tonal targets (e.g. Beckman & Pierrehumbert 1986, Pierrehumbert 1980). It is also well established that, in general, accented syllables and vowels undergo systematic lengthening (e.g. Turk & Sawusch 1997, Turk & Shattuck Hufnagel 2007) and unaccented and unstressed vowels undergo systematic shortening. Further, focused words have expanded pitch range (Xu & Xu 2005) and are lengthened (e.g. Cooper et al. 1985, De Jong 2004, Eady et al. 1986), while post-focus words are shortened and reduced in pitch (De Jong 2004, Xu & Xu 2005).

[1] Pitch and duration are not considered strictly integral dimensions in the sense discussed in e.g. Garner 1974, as compared to, for example, pitch and loudness, which are argued to be processed holistically by listeners (Grau & Nelson 1988, Nelson 1993).

[2] These previous studies use explicit judgments of duration, in comparison to the present study in which listeners categorize a continuum that cues a phonemic contrast. Segmental categorization can be seen as an *implicit* test for perceived duration. Since implicit and explicit judgements of this sort do not necessarily align (Reinisch 2016), it is possible that previous results using explicit tasks will not be obtained in the implicit task in the present study.

These different structural properties of the prosodic/intonational system of English engender a very general acoustic consequence: accented syllables have increased duration and pitch relative to unaccented syllables (e.g. Greenberg et al. 2003, Kochanski et al. 2005). In this broad sense, increased pitch and duration can be considered acoustic correlates of accentedness in English. Aligning with this view, both increased pitch and duration have been shown to be contributing factors in listeners' perception of prominence in speech (Bishop et al. *resubmitted*, Ladd et al.1994, Ladd & Moreton 1997, Mo 2011).

We can consider this general acoustic correlation in light of prosodically driven compensatory effects (e.g. Kim & Cho 2013). Given that increased pitch correlates with accentedness and contributes to listeners' perception of prominence, we predict that if listeners interpret pitch along these lines, the HIGH pitch condition may give a percept of prominence, or accentedness. Further, because of accentual lengthening, listeners may *expect* longer vowel durations when pitch is HIGH. In other words, increased pitch as a correlate of prominence marking in English might mediate listeners' expectations about vowel duration such that they expect longer vowels to cooccur with increased pitch. Following this logic, listeners may compensatorily adjust categorization of the vowel duration continuum such that they require longer vowel durations for a "code" response in the HIGH pitch condition, *decreasing* "code" responses when pitch is HIGH. Such an effect would reflect listeners' interpretation of pitch as a correlate of accentedness and perceptual compensation for prosodically driven patterning of pitch and duration.[3] The directionality of this effect is, crucially, the opposite of that predicted based on psychoacoustics, as outlined above. The present study therefore tests whether psychoacoustic or prosodic factors will influence perception of vowel duration as a cue to voicing.

2. Experiment 1. To test these predictions, we implemented a 2AFC task in which listeners categorized a stimulus from a vowel duration continuum as one of two English words, "coat" or "code". These two words were chosen to be fairly matched for lexical frequency from the $\text{SUBTLEX}_{\text{US}}$ corpus (Brysbaert & New 2009) to minimize frequency biases in categorization.[4]

2.1. MATERIALS. The stimuli were created from the resynthesized speech of a ToBI-trained male English speaker. The speaker was first recorded at 44.1 kHz (32 bit) using SM10A Shure™ microphone and headset in a sound-attenuated room in the UCLA Phonetics Lab. Manipulation was carried out using PSOLA resynthesis (Moulines & Charpentier 1990) in Praat (Boersma & Weenik 2019). The utterances that served as a starting point for the creation of the stimuli are represented below with ToBI transcription (e.g. Beckman & Ayers-Elam 1997).

(1) I'll say code now
H* H* L-L%

(2) I'll *say* code now
L+H* L-L%

(1) is produced with neutral focus while (2) is produced with narrow focus on "say". Therefore, the word "code" in (1), being nuclear pitch-accented, has higher pitch and intensity than the post-focus production of "code" in (2). The token from which all stimuli were created was the word "code" from sentence (1), with pitch on this token manipulated. This token was excised and audible voicing after closure was removed to make the coda stop ambiguous. The intensity

[3] One additional issue here is the nature of the task itself. Given that perception of prominence is relative (e.g. Jagdfeld & Baumann 2011, Terken & Hermes 2000, Turnbull et al. 2017), listeners' attribution of prominence to HIGH pitch words would be in relation to LOW pitch words, which they are hearing together, in a series of randomized trials in the present study.

[4] The $\log_{10}$ word frequency for "coat" is 3.33; $\log_{10}$ word frequency for "code" is 3.43.

of the token was then manipulated to be the average between the productions in (1) and (2). Controlling intensity across conditions in this way is essential given that loudness and duration interact perceptually (e.g. Turk & Sawusch 1996) and would confound listeners' perception of duration as a function of pitch.

The f0 values from the nuclear pitch-accented target word "code" in (1) are referred to the HIGH pitch condition in the present study (onset = 135Hz; offset = 129Hz). The pitch values from the same target word in (2) are referred to as the LOW pitch condition (onset = 112Hz, offset = 103Hz). Two vowel length continua were resynthesized from these HIGH and LOW pitch words. The continua ranged from 60 ms to 150 ms, in 15 ms step intervals. These manipulations created 14 unique stimuli (seven continuum steps in each condition). An example stimulus representing each pitch condition is shown below in Figure 1.

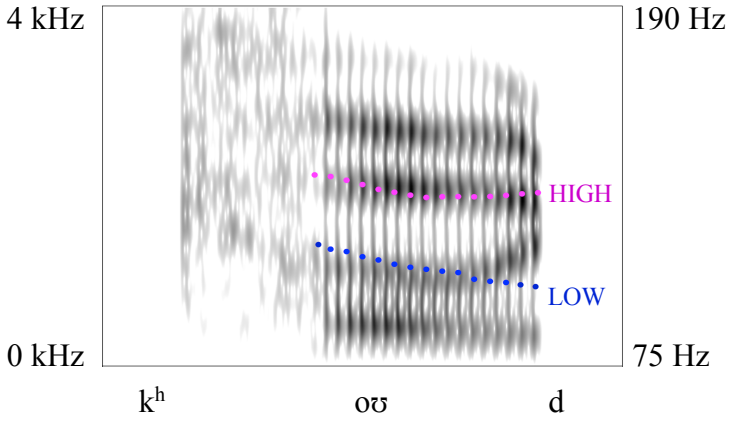

Figure 1: An example stimulus spectrogram overlaid with a pitch track for each condition. Conditions are labeled to the right. The *y* axis at right shows the f0 range in Hz, the *y* axis at left shows the frequency range for the spectrogram. A transcription is given below the spectrogram.

By using pitch values from the prosodic contexts outlined above, we ensure that they are fairly natural low and high pitch for the speaker's range, and that they are an instantiation of the intended prosodic context under investigation, giving representative pitch for an accented (HIGH pitch) and unaccented (LOW pitch) syllable.[5]

2.2. PARTICIPANTS. 30 participants were recruited for Experiment 1 (15 participants identified as female and 15 as male). Participants were self-reported native English-speaking adults with normal hearing. All participants were students at UCLA and received course credit for participation. All provided informed consent to participate. No participant responses were excluded from analysis.

[5] A related consideration is the role of pitch as a cue to voicing. Lowered pitch is one of the acoustic features associated with voiced obstruents (e.g. Lisker 1986, Ohde 1984). This is a salient cue for listeners (e.g. Kohler 1985, Winn et al. 2013), so one may wonder if the LOW pitch condition might be interpreted by listeners as a cue to voicing, predicting *increased* "code" responses in the LOW pitch condition. Two possible arguments against this are as follows. Firstly, voicing is variably realized in word final stops (e.g. Guy 1980, Ratner& Luberoff 1984), and pitch does not reliably fall as a correlate of voicing in word-final position (Gruenenfelder & Pisoni 1980). In light of this, pitch may not be as salient a cue to obstruent voicing word-finally, and previous studies have largely investigated voicing in initial position where f0 modulations are more consistent (Ohde 1984). Secondly, local f0 modulations near the coda consonant have been shown to be a more reliable cue to voicing, as compared to f0 across a vowel (e.g. Gruenenfelder & Pisoni 1980, Kohler 1985, Kohler & Van Dommelen 1986). This suggests our manipulations, which alter pitch across the vowel may not be interpreted as a cue to voicing by listeners.

2.3. PROCEDURE. Testing was carried out in a sound-attenuated room in the UCLA Phonetics Lab, with participants seated in front of a desktop computer. Stimuli were presented binaurally via a Peltor™ 3M™ listen-only headset, adjusted to a comfortable listening level. Before testing began, participants were told they would listen to a native English speaker saying one of two English words, "coat" or "code", and that their task was to select which word they had heard. During the trials, participants heard a stimulus and were presented visually with "coat" and "code", one on each side of the screen. Participants indicated their choice via a key press on the computer keyboard, where an 'f' keypress indicated the left side choice, and a 'j' keypress indicated a right side choice. The side of the screen on which "coat" and "code" appeared was counterbalanced. The inter-trial-interval was 250 ms. Before testing, participants performed eight training trials to familiarize themselves with the procedure. In these trials, participants heard the endpoints of the continuum for both pitch conditions. These training stimuli were randomized by pitch, such that participants heard two instances of each pitch condition for each endpoint (for a total of four randomized-by-pitch trials for each endpoint block). It was random which endpoint block came first. In the subsequent test trials the stimuli were totally randomized (by pitch and vowel duration). Participants categorized a total of 16 instances of each of unique stimulus, for a total of 224 (16*14) test trials. They were prompted to take a short self-paced break halfway through. The experimental procedure took approximately 10-15 minutes.

2.4. RESULTS AND DISCUSSION. Results were assessed using a linear mixed-effect model with a logistic linking function. Fixed effects in the model were vowel duration (treated as continuous and centered at zero), two levels of pitch (LOW and HIGH), and their interaction. Pitch was contrast-coded (HIGH was mapped to -1 and LOW was mapped to 1). The random effect structure of the model consisted of by-subject random intercepts, with maximally specified random slopes (e.g. Barr et al. 2013). Data visualization was carried out in RStudio (RStudio Team 2018), analysis used lme4 (Bates et al. 2015) and emmeans (Lenth et al. 2018). The output of the model and comparison of contrasts with emmeans are contained in the appendix. Results from Experiment 1 are visualized in Figure 2 below.

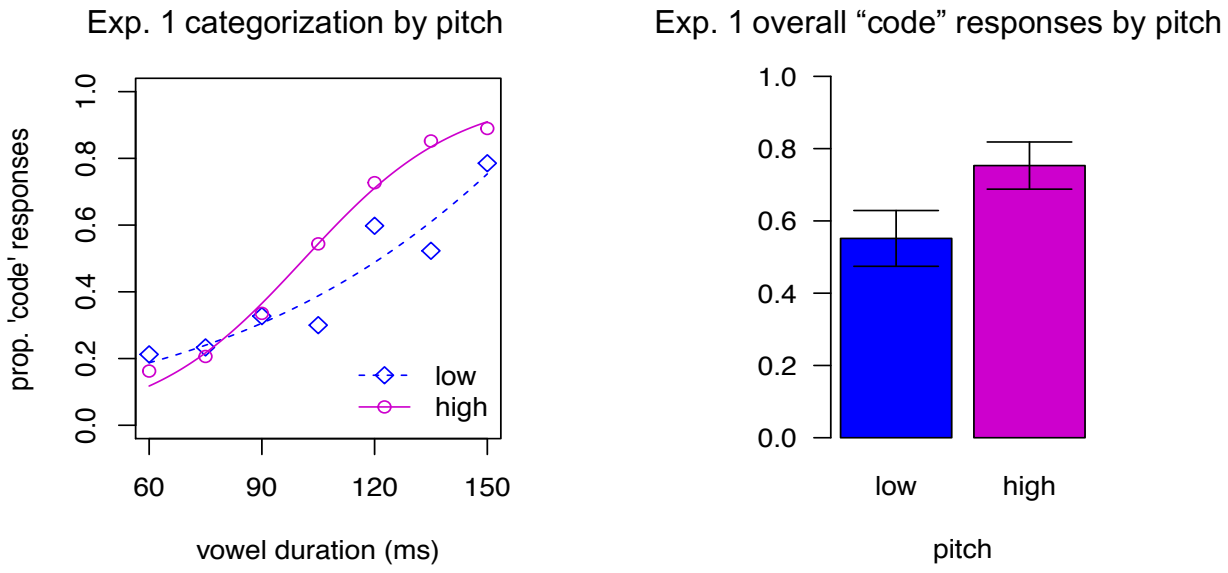

Figure 2: Categorization along the continuum split by pitch condition (at left) and the overall proportion of "code" responses in each pitch condition (at right). In the categorization plot, the

x axis shows vowel duration values. Points show the raw proportion of "code" responses, lines are psychometric curves which are fit to show a smoothed categorization trend. Error bars in the bar plots at right show one SE from the model.

Firstly, as would be expected from any such vowel duration continuum, increasing vowel duration significantly increased "code" responses (β = 1.34, z = 11.85, p < 0.001). Pitch, the predictor of interest, also showed a significant effect (β = -0.35, z = -2.73, p < 0.01), whereby overall LOW pitch significantly decreased "code" responses. As shown in Figure 2, listeners more readily categorized the target sound as "code" when it bore HIGH pitch. As outlined above, this effect is expected if HIGH pitch increased perceived vowel duration (as a cue to voicing). In this sense, the main effect of pitch observed in Experiment 1 is consistent with the psychoacoustic integration predictions outlined above and concurs with the results from the previously mentioned explicit rating studies.[6] A robust interaction between duration and pitch was also observed in the model (β = 0.33, z = 6.87, p < 0.001). Post-hoc testing with emmeans shows that pitch has no significant effect at the three lowest steps of the continuum, and at higher steps the effect increases in magnitude as vowel duration increases (see Table 3 in the Appendix).

These results overall suggest that listeners' interpretation of a duration as a cue to obstruent voicing can be directly influenced by pitch, such that increased pitch increases perceived duration, resulting in increased "code" responses. The presence of the interaction in the model highlights that this effect is contingent on vowel duration itself and is only observed at longer vowel durations (greater than 90 ms on the continuum).

It can also be noted that previous studies which found this effect (with explicit listener ratings of duration) all used stimuli which are substantially longer than our own continuum, both in terms of minimum and maximum values. The lowest minimum in these previous studies is 100 ms (as compared to our 60 ms minimum). Table 1 in the appendix offers a full summary of the durational ranges used in different stimuli in these previous studies. The fact that previous studies which consistently found this effect employed longer durations than our own, coupled with the finding that in Experiment 1 pitch only exerted an influence at longer vowel durations is suggestive of the possibility that the influence of pitch may vary based on vowel duration.

In light of this potential issue, we return to the question of how pitch and duration pattern as correlates of accentuation in English. Following the logic that compensatory processes related to prosodic structure are learned from patterns in the language (in the sense discussed in e.g. Holt et al. 2001, Wade & Holt 2005), we need to consider the duration of vowels observed in spoken corpora of English with the goal of seeing how these durational values compare to the stimuli used in Experiment 1.

Previous studies which have systematically investigated this topic show that vowels which are analyzed as unstressed (Greenberg et al. 2003, SWITCHBOARD corpus) and perceived by naïve listeners as lacking prominence (Mo 2011, Buckeye corpus) are both under 100 ms in duration on average and can be much shorter. Notably, Greenberg et al. also measure the longest *stressed* vowel (in terms of vowel quality) to be 200 ms long on average. All other vowel qualities are on average shorter than 200 ms when stressed. However, previous studies with explicit listener ratings of duration all used durational maxima that are above 200 ms, which puts their stimuli outside of the typical range of unaccented *or* accented vowels in natural speech in English. When listeners are presented with these longer durations, they may not incorporate their

[6] This result can also be taken as a clear indication that listeners are not interpreting pitch as a cue to voicing in these stimuli, as discussed in footnote 5.

expectations about accentual prominence at all. Our own stimuli in Experiment 1 have a maximum duration (i.e., 150ms) that aligns fairly well with the average durations of accented vowels. However, the longer steps of our continuum are well above the range for unaccented vowels, which raises the possibility that listeners would not interpret pitch differences in the stimuli as correlating with accentedness, because longer continuum steps are too long to be unaccented. Following this logic, we predict that listeners' sensitivity to LOW pitch as a correlate of un-accentedness may be enhanced when duration is relatively short such that vowels are more plausibly interpretable as lacking prominence. In other words, listeners' interpretation of acoustic cues as conveying prosodic information may relate to their expectations about how those cues (here duration and pitch) typically pattern as a function of prosody. When durations are longer than typical accented vowels (as in previous studies), or when durations are longer than typical unaccented vowels (as in Experiment 1) listeners simply may not generate expectations about accentual lengthening and pitch. Using shorter vowel durations, i.e. those that span a range of more plausibly accented *and* unaccented vowels (and reducing the presence of durations that are too long to be unaccented), may highlight pitch as a property related to prominence-marking for listeners.

We therefore predict that *if* listeners compensate perceptually for pitch as a correlate of prominence marking, shorter vowels (as compared to those used in Experiment 1) will make duration more salient as correlate of un-accentedness and will therefore have a greater tendency to exhibit this effect. In a second experiment we address this point by investigating how changing aspects of the stimuli might influence listeners' interpretation of pitch along these lines.

3. Experiment 2. In Experiment 2, listeners categorized a modified continuum with the same HIGH and LOW pitch conditions. Modifications are outlined below.

3.1. MATERIALS. Two changes were made to the continuum used in Experiment 1. Firstly, the maximum value was reduced from 150 ms to 120 ms, meaning the new range was 60-120 ms. Secondly, the step size was reduced from 15 ms to 10 ms. This new continuum therefore had seven steps total, as in Experiment 1.

These changes were made with the intent of encouraging a linguistic/prosodic interpretation of pitch. By reducing the range of the continuum, listeners are exposed to less extreme variability in duration, rendering durational differences less pronounced. Further, in the absence of longer vowel durations, shorter continuum steps are probably more salient to listeners, where we predict an effect of pitch as a correlate of accentuation should be greatest. Importantly, accented vowel durations also fall roughly within this continuum range (Greenberg et al. 2003), though they can be longer. Accordingly, we predict that, in terms of duration, listeners will have exposure to continuum steps that can plausibly be interpreted as accented or unaccented, though unlike in Experiment 1, durations that may more plausibly be interpreted as *only* being accented are less present.

The reduced step size further makes changes in duration less perceptible (Healy & Repp 1982, Repp 1984). Importantly, a 10 ms step size is quite small for a vowel duration continuum and approaches the JND for continuum steps 100 ms and longer (Klatt 1976, Klatt & Cooper 1975). In this sense, vowel duration would become a less reliable cue to voicing for listeners, including, potentially, perceived duration as a function of pitch. Listeners may therefore be pushed to interpret pitch as prosodic property, compensating for differences in pitch height as originally predicted. If listeners do indeed adjust categorization along these lines, we predict that LOW pitch should significantly *increase* listeners' "code" responses, as outlined in section 1.1. Such a result would be a reversal of the effect observed in Experiment 1.

3.2. PARTICIPANTS AND PROCEDURE. 30 (different) participants were recruited for Experiment 2 (22 participants identified as female and 8 as male). No participant responses were excluded from analysis. The procedure was identical to Experiment 1.

3.3. RESULTS AND DISCUSSION. The statistical assessment and model fitting procedure was the same as in Experiment 1. Results from Experiment 2 are visualized in Figure 3 below.

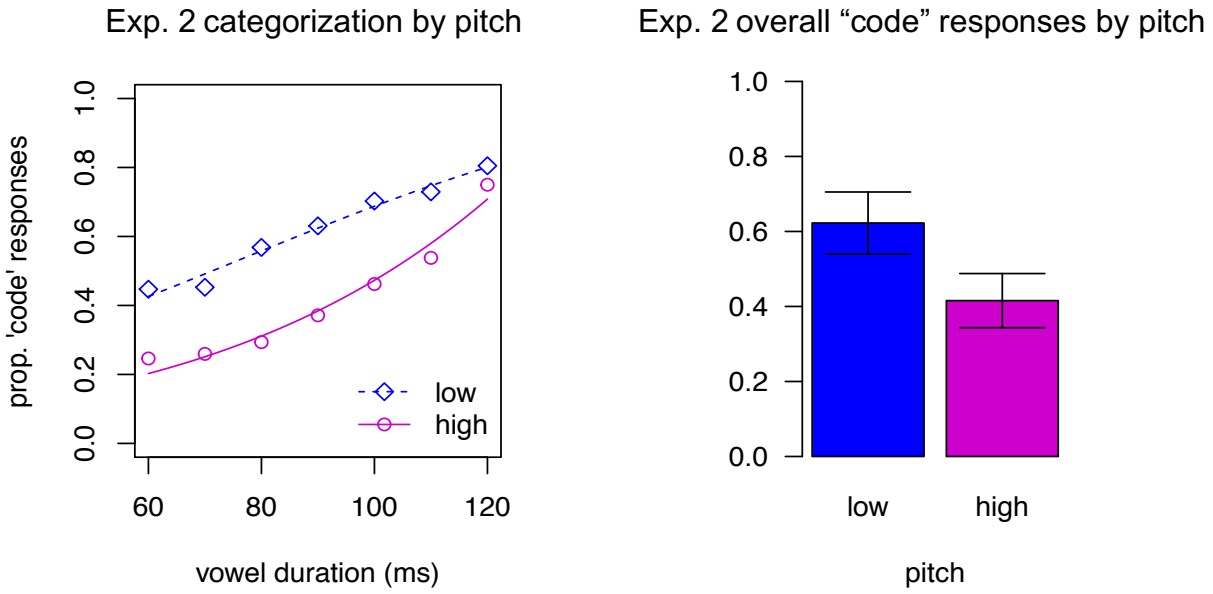

Figure 3: Categorization along the continuum split by pitch condition for Experiment 2 (at left) and the proportion of "code" responses in each pitch condition (at right).

As in Experiment 1, increasing vowel duration significantly increased "code" responses (β = 0.76, z = 10. 64, p < 0.001). It can be noted that the effect is smaller than that in Experiment 1, suggesting that, as expected, vowel duration has become a less reliable cue to voicing. Visually, we can see in Figure 3 that the categorization functions are quite shallow, and the endpoints of the continuum are not anchored, indicating that stimuli are overall fairly ambiguous to listeners. Crucially, in this context of ambiguity, pitch also showed a significant main effect (β = 0.51, z = 3.28, p < 0.01), whereby LOW pitch significantly *increased* "code" responses. The effect of pitch is therefore the opposite of that in Experiment 1 (cf. Figure 2). A significant interaction was also observed in the model (β = -0.07, z = -2.54, p < 0.05), showing that the magnitude of the effect of pitch is largest at the lowest endpoints of the continuum, and decreases systematically as vowel duration increases, though there is a significant effect of pitch at each continuum step (see Table 3 in the appendix). The presence of this interaction suggests that listeners are more sensitive to pitch differences at shorter vowel durations, aligning with the hypothesis that shorter durations may be more plausible as unaccented, or non-prominent vowels. In other words, the shorter end of the continuum may provide listeners with a range of stimuli where LOW pitch more reliably correlates with duration as a function of prosody, and therefore enhances listeners' interpretation of pitch as prosodic at these continuum steps.

4. General discussion. The present study gives us a nuanced view of how prosodic structure mediates listeners' perception of durational cues, and interfaces with domain-general perceptual processes. The results of Experiment 1 suggest that listeners' perception of vowel duration as a

cue to obstruent voicing can be influenced by pitch height, reflecting psychoacoustic perceptual integration of pitch and duration. This aligns with previous literature which used explicit judgements to test listeners' perception of duration as a function of pitch height (e.g. Brigner 1988, Yu et al. 2014) and supports the view that these dimensions can be integrated perceptually by listeners (e.g. Prince 2011). Importantly, in Experiment 1 the effect of pitch was observed to be contingent on vowel duration itself, showing no effect at shorter continuum steps.

In Experiment 2, which highlighted shorter durations and reduced the perceptibility of durational differences, the effect of pitch was reversed entirely. As outlined above, the pattern seen in Experiment 2 suggests that listeners are interpreting pitch as a correlate of prosodic structure. This interpretation crucially mediates their expectations about vowel duration and engenders robust compensatory shifts in categorization. We propose that the results in Experiment 2 can be interpreted as reflecting *prosodic integration*, that is, the integration of expectations about duration with listeners' interpretation of pitch as a linguistic/prosodic property.

Taken together, these results suggest that compensation for prosodically driven variation in pitch and duration can indeed be observed under the right circumstances, crucially, when patterns in the stimuli map more closely onto prosodic patterns in the language. These results therefore align with the view that perceptual integration of pitch and duration is flexible and varies based on stimulus factors (e.g. Ellis & Jones 2009, Prince 2011). They further support the claim that listeners are sensitive to prosodically conditioned variation in phonetic detail (e.g. Kim & Cho), and that perceptual compensation of this sort may be based on learned patterns from language input (following e.g. Holt et al. 2001). This is consistent with the idea that "structured experience shapes perception" (Holt et al. 2001, p 772; see also Holt & Lotto 2006, 2010). The present results can complement this view in showing that acoustic patterns driven by English prosody can apparently structure perceptual experience in a way that mediates listeners' perception of speech segments, overriding psychoacoustic effects under the right circumstances.

The present results extend our understanding of how prosody influences speech perception in two ways. First, extending from the previous finding that cues to a prosodic boundary influences listeners' perception of duration (e.g. Kim & Cho 2013, Mitterer et al. 2016, Steffman 2018), the present study shows that patterns associated with prominence marking can also influence how listeners interpret durational cues. These results thus highlight that multiple facets of prosodic organization influence the way listeners interpret segmental cues in the speech signal. Secondly, the present results suggest that these prosodically-driven compensatory effects for accentual prominence can occur even in isolated words, whereas previous research on prosody and segmental perception placed words within carrier phrases that varied prosodic context.

The fact that listeners are adjusting categorization of isolated words on the basis of prosodic factors suggests that they may be able to interpret and perceptually access prosodic information for words in isolation. This general notion aligns with the view that listeners retain phonetically rich representations of sounds in memory as couched in exemplar theories of speech perception (e.g. Johnson 2006, Pierrehumbert 2001). The present results may support this view, in the sense that prosodic-structural factors introduce patterned acoustic variability (in duration and pitch), which is encoded and retained by listeners, and influences categorization of words, even those that are dissociated from an explicit prosodic context. A variety of previous studies have suggested that listeners retain phonetically rich representations of prosodic information in perception (e.g. Braun et al. 2006, D'Imperio et al. 2014, Kimball et al. 2015, Schweitzer et al. 2015), and the present study may offer evidence of this in the form of a categorization task. Additionally, it has been argued that compensatory (or normalization) effects in categorization may arise as a

natural consequence of exemplar storage (e.g. Johnson 1997). This offers a useful lens for considering the compensatory effects seen in Experiment 2, which may occur when acoustic properties in the stimuli map more closely on to stored exemplars. Further exploration of this idea may prove useful as a way of investigating the perceptual mechanisms underlying these effects.

Additionally, the present results make several concrete predictions for cross-linguistic extension, which may prove a valuable test for the way that learned language patterns mediate listeners' perception of durational cues. For example, vowels with lexical low tones tend to be longer than vowels with lexical high tones in both Thai (Gandour 1977) and Beijing Mandarin (Ho 1976). This general correlation between duration and pitch is therefore the opposite of that introduced by English prosody. This would predict that compensatory perceptual adjustments for speakers of these languages would show the opposite directionality as that observed in Experiment 2, in for example, a task where listeners categorize phonemic vowel length contrasts in Thai (e.g. Abramson & Ren 1990). The effect for speakers of these languages is therefore predicted to be uniform, in that both psychoacoustic and experience-based prosodic factors predict the same directionality. Testing if this uniformity is observed with comparable stimuli to the experiments reported here would provide a useful exploration for the role of language experience in this domain.

Further extending these results along these lines will therefore better our understanding of how listeners' interpretation of prosodic aspects in the speech signal mediates their perception of speech segments, and how language experience with prosody constrains listeners' interpretation of segmental contrasts. Testing how these prosodically driven effects interface with domain-general perceptual processes will further help us explore how the perceptual system integrates acoustic dimensions on the basis of both psychoacoustic and language-specific factors.

Appendix

Type of stimuli	Stimuli duration range (ms)	Study
Non-speech	110-659	Bringer 1988
	150-450	Šimko et al. 2016
Speech (vowels)	100-300	Gussenhoven & Zhou 2013
	150-250	Yu 2010
	230-320	Yu et al. 2014

Table 1: Summary of duration ranges used in previous studies, which found increased pitch increases perceived duration. The type of stimuli (speech versus non-speech) is given at left

EXP. 1 : LONG CONTINUUM			EXP. 2 : SHORT CONTINUUM		
	β(SE)	z-value		β(SE)	z-value
Intercept	-0.12(0.08)	-1.46	Intercept	0.08(0.06)	1.18
pitch	-0.35(0.13)	-2.73**	pitch	0.51(0.15)	3.28**
vdur	1.34(0.11)	11.85***	vdur	0.76(0.07)	10.64***

Table 2: Summary of the regression analyses for the Experiment 1(at left), and Experiment 2 (at right). Values are rounded. Approximate p-values are indicated by asterisks, where * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$.

EXP. 1: LONG CONTINUUM			EXP. 2: SHORT CONTINUUM		
duration (ms)	Estimate(SE)	z-ratio	duration (ms)	Estimate(SE)	z-ratio
60 ms	-0.28(0.29)	-0.98	60 ms	-1.26(0.33)	-3.83***
75 ms	0.04(0.7)	0.17	70 ms	-1.19(0.32)	-3.68***
90 ms	0.37(0.26)	1.44	80 ms	-1.11(0.32)	-3.49***
105 ms	0.69(0.25)	2.73**	90 ms	-1.04(0.32)	-3.28**
120 ms	1.03(0.26)	3.91***	100 ms	-0.96(0.32)	-3.03**
135 ms	1.25(0.27)	4.89***	110 ms	-0.89(0.32)	-2.76**
150 ms	1.68(0.30)	5.63***	120 ms	-0.82(0.33)	-2.49*

Table 3: Summary of the comparison of contrasts with emmeans (Lenth et al. 2018) for the Experiment 1 (at left), and Experiment 2 (at right).

References

Abramson, Arthur S. & Nianqi Ren. 1990. Distinctive vowel length: Duration vs . spectrum in Thai. *Haskins Laboratories status report on speech research*. 256-268.

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Beckman, Mary E. & Janet B. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology* 3(01). 255–309.

Beckman, M. E. & Gayle Ayers Elam. 1997. Guidelines for ToBI labeling, version 3.0. Unpublished ms. Ohio State University.

Bishop, Jason, Grace Kuo & Boram Kim. resubmitted. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription. *Journal of Phonetics (Special issue on integrating phonetics and phonology in the study of linguistic prominence)*.

Boersma, Paul. & David Weenik. 2019. Praat: Doing phonetics by computer [Computer program].Version 6.0.37, retrieved from http://www.praat.org/.

Braun, Bettina, Greg Kochanski, Esther Grabe & Burton S. Rosner. 2006. Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America* 119(6). 4006. https://doi.org/10.1121/1.2195267.

Brigner, Willard L. 1988. Perceived duration as a function of pitch. *Perceptual and Motor Skills* 67(1). 301–302. https://doi.org/10.2466/pms.1988.67.1.301.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4). 977–990. https://doi.org/10.3758/BRM.41.4.977.

Chen, Matthew. 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, Taehong. 2015. Language effects on timing at the segmental and suprasegmental levels. In Melissa A. Redford (ed.), *The handbook of speech production*. 505–529. Hoboken, NJ: John Wiley & Sons.

Cho, Taehong. 2016. Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass* 10(3). 120–141.

Cooper, William E., Stephen J. Eady & Pamela R. Mueller. 1985. Acoustical aspects of contrastive stress in question-answer contexts. *The Journal of the Acoustical Society of America* 77(6). 2142–2156.
Dainora, Audra. 2006. Modeling intonation in English: A Probabilistic Approach to Phonological Competence. In Louis Goldstein, Douglas H. Whalen & Catherine T. Best (eds.), *Laboratory Phonology 8*. Walter de Gruyter.
de Jong, Kenneth. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics* 32(4). 493–516. https://doi.org/10.1016/j.wocn.2004.05.002.
D'Imperio, Mariapaola, Rossana Cavone & Caterina Petrone. 2014. Phonetic and phonological imitation of intonation in two varieties of Italian. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.01226.
Eady, Stephen J., William E. Cooper, Gayle V. Klouda, Pamela R. Mueller & Dan W. Lotts. 1986. Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech* 29. 233–251. https://doi.org/10.1177/002383098602900304.
Ellis, Robert J. & Mari R. Jones. 2009. The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology. Human Perception and Performance* 35(1). 264–280. https://doi.org/10.1037/a0013482.
Gandour, Jack T. 1977. On the interaction between tone and vowel length: Evidence from Thai dialects. *Phonetica* 34. 54–65.
Garner, Wendell R. 1974. *The processing of information and structure*. Oxford, England: Lawrence Erlbaum.
Georgeton, Laurianne, Tanja Kocjančič Antolík & Cécile Fougeron. 2016. Effect of domain initial strengthening on vowel height and backness contrasts in French: Acoustic and ultrasound data. *Journal of Speech, Language, and Hearing Research* 59(6). S1575–S1586. https://doi.org/10.1044/2016_JSLHR-S-15-0044.
Grau, James W. & Deborah K. Nelson. 1988. The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General* 117(4). 347–370. https://doi.org/10.1037/0096-3445.117.4.347.
Greenberg, Steven, Hannah Carvey, Leah Hitchcock & Shuangyu Chang. 2003. Temporal properties of spontaneous speech-a syllable-centric perspective. *Journal of Phonetics* 31(3). 465–485. https://doi.org/10.1016/j.wocn.2003.09.005.
Gruenenfelder, Thomas M. & David B. Pisoni. 1980. Fundamental frequency as a cue to postvocalic consonantal voicing: Some data from speech perception and production. *Perception & Psychophysics* 28(6). 514–520.
Gussenhoven, Carlos & Wencui Zhou. 2013. Revisiting pitch slope and height effects on perceived duration. *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association.* 1365–1369.
Guy, Gregory. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed.), *Locating language in time and space.* 1–36. New York: Academic Press.
Healy, Alice F. & Bruno H. Repp. 1982. Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance* 8(1). 68–80. https://doi.org/10.1037/0096-1523.8.1.68.

Ho, Aichen T. 1976. The acoustic variation of Mandarin tones. *Phonetica* 33(5). 353–367. https://doi.org/10.1159/000259792.
Holt, Lori L. & Andrew J. Lotto. 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America* 119(5). 3059–3071. https://doi.org/10.1121/1.2188377.
Holt, Lori L. & Andrew J. Lotto. 2010. Speech perception as categorization. *Attention, Perception, & Psychophysics* 72(5). 1218–1227. https://doi.org/10.3758/APP.72.5.1218.
Holt, Lori L., Andrew J. Lotto & Keith R. Kluender. 2001. Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *The Journal of the Acoustical Society of America* 109(2). 764–774. https://doi.org/10.1121/1.1339825.
Jagdfeld, Nils & Stefan Baumann. 2011. Order effects on the perception of relative prominence. *17th International Congress of Phonetic Sciences*. 958–961.
Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & J. W. Mullennix (eds.), *Talker variability in speech processing*. 145–165. San Diego: Academic Press.
Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4). 485–499. https://doi.org/10.1016/j.wocn.2005.08.004.
Keating, Patricia. 2006. Phonetic encoding of prosodic structure. In Jonathan Harrington & Marija Tabain (eds.), *Speech production: Models, phonetic processes, and techniques*, 167–186. New York and Hove: Macquarie Monographs in Cognitive Science, Psychology Press.
Keating, Patricia, Cécile Fougeron, Chai-Shune Hsu & Taehong Cho. 2003. Domain initial articulatory strengthening in four languages. In John Local, Richard Ogden & Rosalind Temple (eds.), *Phonetic interpretation: Papers in laboratory phonology VI*. Cambridge University Press.
Kimball, Amelia, Jennifer Cole, Gary Dell & Stefanie Shattuck-Hufnagel. 2015. Categorical vs. episodic memory for pitch accents in American English. *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–4.
Kim, Sahyang & Taehong Cho. 2013. Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America* 134(1). EL19–EL25.
Klatt, Dennis H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59(5). 1208–1221. https://doi.org/10.1121/1.380986.
Klatt, Dennis H. & William E. Cooper. 1975. Perception of segment duration in sentence contexts. *Structure and Process in Speech Perception*, 69–89. Springer, Berlin, Heidelberg.
Kochanski, Greg, Esther Grabe, Jim Coleman & Bernard Rosner. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America* 118(2). 1038–1054. https://doi.org/10.1121/1.1923349.
Kohler, Klaus J. 1985. F0 in the perception of lenis and fortis plosives. *The Journal of the Acoustical Society of America* 78(1). 21–32. https://doi.org/10.1121/1.392562.
Kohler, Klaus J. & Wim Van Dommelen. 1986. Prosodic effects on lenis /fortis perception : Preplosive F0 and LPC synthesis. *Phonetica* 43. 70–75.
Ladd, D Robert & Rachel Morton. 1997. The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics* 25(3). 313–342. https://doi.org/10.1006/jpho.1997.0046.

Ladd, D. Robert, Jo Verhoeven & Karen Jacobs. 1994. Influence of adjacent pitch accents on each other's perceived prominence: Two contradictory effects. *Journal of Phonetics* 22(1). 87–99.
Lehiste, Ilse. 1976. Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics* 4(2). 113–117.
Lenth, Russell, Henrik Singmann, Jonathon Love, Paul Buerkner & Maxime Herve. 2018. *Emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans.
Lisker, Leigh. 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1). 3–11. https://doi.org/10.1177/002383098602900102.
Mitterer, Holger, Taehong Cho & Sahyang Kim. 2016. How does prosody influence speech categorization? *Journal of Phonetics* 54. 68–79. https://doi.org/10.1016/j.wocn.2015.09.002.
Moulines, Eric & Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication.* 9(5-6). 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z.
Mo, Yoonsook. 2011. *Prosody production and perception with conversational speech*. University of Illinois at Urbana-Champaign Dissertation.
Nelson, Deborah K. 1993. Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance* 19(5). 1105–1113. https://doi.org/10.1037/0096-1523.19.5.1105.
Ohde, Ralph N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America* 75(1). 224–230.
Onaka, Akiko 2003. Domain-initial strengthening in Japanese: An acoustic and articulatory study. *Proceedings of the 15th International Congress of Phonetic Science (ICPhS)*. 2091–2094.
Peterson, Gordon E. & Ilse Lehiste. 1960. Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America* 32(6). 693–703. https://doi.org/10.1121/1.1908183.
Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*. Massachusetts Institute of Technology Dissertation.
Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency lenition and contrast. In Joan L. Bybee & Paul Hopper (eds.), *Typological studies in language*, vol. 45. 137–158. Amsterdam: John Benjamins.
Prince, Jon B. 2011. The integration of stimulus dimensions in the perception of music. *Quarterly Journal of Experimental Psychology* 64(11). 2125–2152. https://doi.org/10.1080/17470218.2011.573080.
Ratner, Nan B. & Arlene Luberoff. 1984. Cues to post-vocalic voicing in mother–child speech. *Journal of Phonetics* 12(3). 285–289.
Reinisch, Eva. 2016. Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception, & Psychophysics* 78(4). 1203–1217. https://doi.org/10.3758/s13414-016-1067-x.
Repp, Bruno H. 1984. Categorical perception: Issues, methods, findings. In Norman J. Lass (ed.), *Speech and Language*, vol. 10. 243–335. New York: Elsevier.
RStudio Team. 2018. RStudio: Integrated development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Schweitzer, Katrin, Michael Walsh, Sasha Calhoun, Hinrich Schütze, Bernd Möbius, Antje Schweitzer & Grzegorz Dogil. 2015. Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation. *Speech Communication* 66. 65–81. https://doi.org/10.1016/j.specom.2014.09.006.

Shigeno, Sumi. 1986. The auditory tau and kappa effects for speech and nonspeech stimuli. *Perception & Psychophysics* 40(1). 9–19. https://doi.org/10.3758/BF03207588.

Šimko, Juraj, Daniel Aalto, Pärtel Lippus, Marcin Włodarczak & Martti Vainio. 2015. Pitch, perceived duration and auditory biases: Comparison among languages. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*.

Steffman, Jeremy A. 2018. *Intonation mediates speech rate normalization in the perception of segmental categories*. UCLA MA Thesis.

Terken, Jacques & Dik Hermes. 2000. The perception of prosodic prominence. In Merle Horne (ed.), *Prosody: Theory and experiment: Studies presented to Gösta Bruce* (Text, Speech and Language Technology): 89–127. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-9413-4_5.

Turk, Alice E. & James R. Sawusch. 1996. The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America* 99(6). 3782–3790. https://doi.org/10.1121/1.414995.

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

Turnbull, Rory, Adam J. Royer, Kiwako Ito & Shari R. Speer. 2017. Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience* 32(8). 1017–1033. https://doi.org/10.1080/23273798.2017.1279341.

van Santen, Jan P. H. 1992. Contextual effects on vowel duration. *Speech Communication* 11(6). 513–546. https://doi.org/10.1016/0167-6393(92)90027-5.

Wade, Travis & Lori L. Holt. 2005. Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics* 67(6). 939–950. https://doi.org/10.3758/BF03193621.

Winn, Matthew B., Monita Chatterjee & William J. Idsardi. 2013. The roles of voice onset time and F0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research* 56(4). 1097–1107. https://doi.org/10.1044/1092-4388(2012/12-0086).

Xu, Yi & Ching X. Xu. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 33(2). 159–197. https://doi.org/10.1016/j.wocn.2004.11.001.

Yu, Alan. 2010. Tonal effects on perceived vowel duration. In Cécile Fougeron, Barbara Kuehnert, Mariapaola Imperio & Nathalie Vallee (eds.), *Laboratory Phonology 10*. 151-168. Berlin: Walter de Gruyter.

Yu, Alan, Hyunjung Lee & Jackson Lee. 2014. Variability in perceived duration: Pitch dynamics and vowel quality. *Proceedings of the 4th International Symposium on Tonal Aspects of Languages*. 41–44.