Prosodic Prominence in Speech Perception: The Influence of Focus Prosody on the Perception of Durational and Spectral Cues

Jeremy Steffman and Sun-Ah Jun

1. Introduction

In comprehending spoken language, listeners must determine the segmental contrasts produced by a speaker, which specify the lexical items spoken. Additionally, listeners must process the phrasal prosody of an utterance as it unfolds. Phrasal prosody will give the listener information about prosodic grouping reflecting syntactic structure, prominence, information structure, and so on. Determining segmental and prosodic categories in speech perception and processing therefore constitutes two key parts of spoken language comprehension (Cutler et al., 1997; Mitterer et al., 2019). Traditional models of perception and word recognition tend to focus on the recognition of segmental structure independent of prosody (McClelland and Elman, 1986; Norris, 1999), while phrasal prosody is usually studied in the context of later-stage processing, e.g. syntactic processing (Cutler et al., 1997; Price et al., 1991). In this sense phrasal prosody and segment could be seen as factoring into speech recognition in fairly independent ways. However, recent research suggests this is not the case (Kim et al., 2018; Mitterer et al., 2019; Steffman, 2019a,b). Instead, current evidence suggests that listeners integrate phrasal prosody in their perception of segmental contrasts, in line with how prosodic structure is encoded in segments in a phonetically detailed manner.

Most of these recent studies have focused on prosodic boundaries, testing whether listeners make reference to a boundary to determine the status of a cue that is modulated by it. For example, voice onset time (VOT) for voiceless aspirated stops is longer at the beginning of an intonational phrase (IP) as compared to IP-medial position (Keating et al., 2004). In a language where VOT is a primary cue for voicing contrasts, listeners might accordingly reconcile the VOT value of a given token with the prosodic context in which it occurs. If a segment is IP-initial, lengthened VOT would be a manifestation of prosodic structure, and therefore could be misleadingly long if not reconciled with context (see e.g. Cho et al., 2007). Adjustments for context in this light could therefore be characterized as compensatory, that is, adjusting for the expected effect of some prosodic factor on a cue's value. Kim and Cho (2013) tested this pattern for VOT in American English, and found predicted compensatory effects. Listeners categorized a word from a VOT continuum as having an initial /p/ or /b/ and modulated categorization based on whether that word was preceded by an IP boundary (and was therefore IP-initial) in a carrier phrase, or was IP-medial. Listeners required overall longer VOT for a voiceless /p/ response for an IP-initial target sound (as compared to an IP-medial target), i.e. they factored in the lengthening of VOT in initial position and adjusted their categorization of the continuum accordingly (see also Mitterer et al., 2016; Steffman, 2019a). Various other influences of prosodic boundaries in several languages have since been documented, extending to both domain-initial (as in the VOT example above), and domain-final patterns, e.g. phrase final lengthening (Kim et al., 2018; Mitterer et al., 2019; Steffman, 2019b; Steffman and Katsuda, 2020), showing listeners reference a prosodic boundary in determining how a cue maps to segmental categories.

The influence of prosodic prominence in the perception of segmental contrasts, however, has received less attention and as such the general relevance and scope of prominence effects in this domain

* Jeremy Steffman, Northwestern University, jeremy.steffman@northwestern.edu. Sun-Ah Jun, UCLA, jun@humnet.ucla.edu. We would like to thank Adam Royer for recording the materials for the stimuli, as well as Yang Wang, Danielle Bagnas and Qingxia Guo for help with data collection. We are also very grateful to the audience at WCCFL 38, and to members of the UCLA Phonetics lab for helpful feedback and commentary.

 Proceedings of the 38th West Coast Conference on Formal Linguistics, ed. Rachel Soo et al., 406-416. Somerville, MA: Cascadilla Proceedings Project.

is more of an open question. In the present study we present two tests of how prosodic prominence, and in particular whether a target word is emphasized (marked as focused) or de-emphasized in a carrier phrase, mediates how listeners perceive both spectral and durational cues to segmental contrasts. For the case of durational cues, we additionally test how other influences on the perception of duration relate to effects driven by prosodic prominence. One way in which prominence may differ from boundary effects is that a segment or word will carry intrinsic properties (pitch, duration), which convey prominence to listeners (Mo, 2011). In this light, contextual prominence might be expected to play a smaller role. The two studies presented here are accordingly a test of the current theory of prosodic and segmental interactions in perception, extending to a less studied domain. We will additionally consider several non-prosodic influences and segment-intrinsic prominence which may interact with contextual prominence effects.

2. Experiment 1

The goal of Experiment 1 was to test how the way focus is manifested in a carrier phrase influenced listeners' perception of vowels categories. The reason we expect this sort of prominence manipulation to influence vowel perception derives from the way in which vowel articulations and formant structure has been shown to vary based on prominence. One well-documented pattern in this domain is so-called "sonority expansion", where the term sonority is used in a phonetic sense, referring to the overall openness of the vocal tract (Silverman and Pierrehumbert, 1990). Phrasal prominence in these studies is usually manipulated as the presence of focus on a target word of interest, and compared to a production of that target where another word within a phrase receives focus, such that the target is un-focused and non-prominent. This sort of phrasal prominence marking influences the articulation of non-high vowels such that they show increased amplitude of jaw lowering, and backing and lowering of the tongue (Cho, 2005; Van Summers, 1987). These articulatory modulations are hypothesized to help enhance syntagmatic contrasts for prominent vowels, in relation to e.g., adjacent consonants.[1]

One acoustic consequence of these articulatory adjustments is a change in formant structure for prominent vowels, as compared to their non-prominent counterparts. For example, jaw and tongue lowering entails *raised* F1 values corresponding to a lower, more open vowel articulation (Van Summers, 1987). In this sense we could conceptualize formant structure as varying along both a segmental dimension (i.e. contrastive vowel categories) and a prosodic dimension (shifting systematically on the basis of prosodic prominence). Relatedly, changes in formant values have been shown to shape how prominent a vowel sounds to listeners in a rapid prosody transcription (RPT) task. Mo et al. (2009) found that, within a vowel category, having both higher F1 (correlated with tongue/jaw lowering) and lower F2 (correlated with tongue backing), increases perceived prominence for non-high vowels that undergo sonority expansion. This finding suggests a clear interplay between formants and the prosodic property of perceived prominence. The question addressed in Experiment 1 is essentially the reverse of that explored by Mo et al. (2009): we test how prominence, at the level of the phrase, influences the perception of a vowel contrast cued by F1 and F2. Below, the stimuli used in the experiment are discussed, after which we outline our predictions.

2.1. Materials

The materials used in Experiment 1 were created by re-synthesizing the speech of a ToBI-trained American English speaker. The speech material was recorded in a sound-attenuated booth in the UCLA Phonetics Lab, using an SM10A Shure™ microphone and headset. Recordings were digitized at 32 bits with a 44.1 kHz sampling rate.

The test case we adopt is the contrast between the American English vowels /ɛ/ and /æ/. Generally speaking, /æ/ has been described acoustically as having higher F1 and lower F2 relative to /ɛ/ (Peterson and Barney, 1952), i.e. it is a lower and less-front vowel. It should be noted that there is clearly regional variation in terms of how this contrast is manifested in F1 and F2 (Clopper et al., 2005), but the

[1] These same patterns of sonority expansion are not observed for high vowels, where expansion might jeopardize attainment of a high vowel target (Cho, 2005).

aforementioned pattern seems to be robust. It will therefore be assumed that listeners will use F1 and F2 to distinguish these vowel categories, with higher F1 and lower F2 signaling /æ/.

In Experiment 1, listeners' task was to categorize a sound drawn from a continuum as "ebb" /ɛ/ or "ab" /æ/. The continuum for the target word was created by re-synthesizing the formant values of natural speech, such that one endpoint had F1 and F2 which were matched to a naturally produced /ɛ/, and the other endpoint had F1 and F2 which were matched to a naturally produced /æ/. The continuum varied jointly in F1 and F2 between each endpoint in 8 interpolated steps (for 10 steps total including endpoints). Each target word was originally recorded in two carrier phrases. These are shown with ToBI labels in (1) and (2), where *x* represents the target sound.

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

(2) I'll SAY *x* now
L+H* L-L%

In (1), the target is prominent, being in the nuclear accented position of the phrase, which contains a standard declarative tune. In (2), the target follows narrow focus marking, realized with a sharply rising L+H* pitch accent, on the word "say". Being post-focus, the target lacks prominence (Beckman and Pierrehumbert, 1986; de Jong, 2004; Xu and Xu, 2005). Two phrasal prominence conditions were created in Experiment 1 corresponding to (1), referred to as the Nuclear Pitch Accent (NPA) condition, and (2), referred to as the Post-focus condition. These conditions were created by cross-splicing and PSOLA method synthesis. The starting point for the creation of these frames was (1) above. The NPA condition was created simply by using the frame in (1), from which the target sound was excised. To create the Post-focus condition, the vowel in the word "say" from (2), with narrow focus, was spliced into the frame, replacing the vowel in "say" from (1). The vowel in "say" in the Post-focus condition therefore has increased amplitude and duration relative to "say" in (1). Following this, the pitch on the preceding word "I'll" was re-synthesized to match the pitch values of this word in (2), i.e. a low-dipping pitch realizing the low target of the following L+H* accent. The post-target material "now" was identical across conditions, left as it was produced in (1), which was highly similar to its production in (2). In both cases it was realized as unaccented and phrase-final with a low (L-L%) boundary tone. These manipulations thus vary the pre-target pitch contour, as well as the duration, amplitude and envelope of the pre-target vowel /eɪ/ (but not other parts of the carrier phrase), as shown in Figure 1.

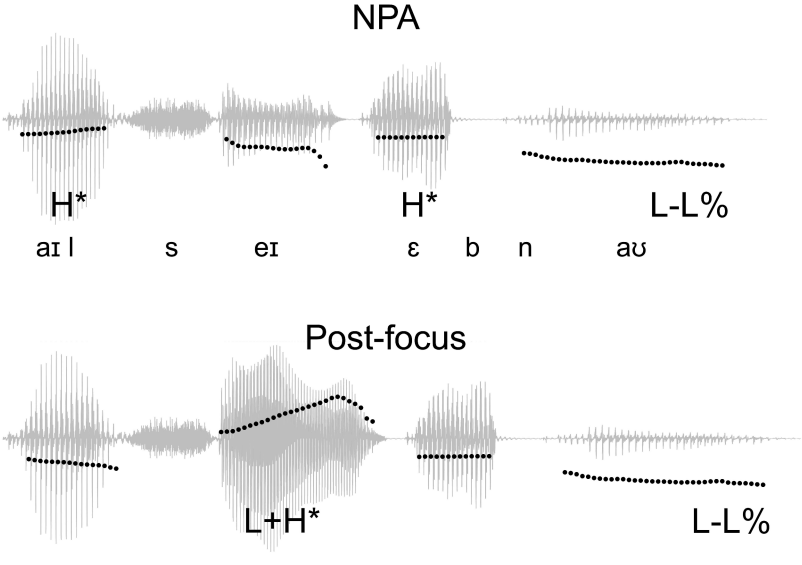

Figure 1: Waveforms of the Experiment 1 stimuli, showing the carrier phrase "I'll say ebb/ab now" overlaid with pitch tracks (lowest value = 85 Hz, highest value = 228 Hz). A segmental transcription is given in IPA aligned to the top-most waveform. The target word shown in the figure is from the /ɛ/ endpoint of the continuum.

2.2. Predictions

Given the structure of the stimuli and following the general pattern of categorization shifts driven by phrasal boundaries outlined above, we predicted that listeners should adjust their criteria for the /ɛ/ -/æ/ contrast based on prominence. In the prominent NPA condition, phrasal prominence would manifest as raised F1 and lowered F2 on the target - essentially rendering /ɛ/ acoustically more like /æ/. Accordingly, we predicted that listeners should accept more /æ/-like values as /ɛ/ in the NPA condition (in comparison to the Post-focus condition). Put differently, listeners should compensate for expected sonority expansion patterns driven by contextual prominence. The empirical prediction is therefore that we should see increased "ebb" responses in the NPA condition.

2.3. Participants and procedure

30 self-reported native English speakers with normal hearing were recruited from the UCLA student population for the experiment. Participants received course credit for their participation. The procedure was a simple 2AFC task in which participants heard a stimulus and categorized it as one of two words, "ebb" or "ab". Participants were seated in front a computer monitor, in a sound attenuated room. Stimuli were presented binaurally via PELTOR™3M™listen-only headset. Target words were represented orthographically on the monitor, each target centered in each half of the monitor. The side of the screen on which the target words appeared was counterbalanced across participants. Participants were instructed that their task was to identify the word by key press, and used the "f" and "j" keys to respond. Prior to the test trials participants completed 4 training trials. In these trials, the continuum endpoints were presented once in each prominence condition. Each unique stimulus was presented a total of 10 times, in random order, for a total of 200 trials during the experiment (20 unique stimuli $\times$ 10 repetitions). The experiment took approximately 15-20 minutes to complete.

2.4. Results and discussion

Results were assessed statistically using a mixed-effects logistic regression model. Listeners' responses, with "ebb" mapped to 1 and "ab" mapped to 0, were predicted based on continuum step (scaled and centered at zero), prominence manipulation (NPA condition mapped to 0.5, Post-focus mapped to -0.5), and the interaction of these two fixed effects. Random effects in the model consisted of random intercepts for participants and maximal random slopes. The fixed effects from the model are shown in Table 1, while the model fit for the data is shown in Figure 2.

	β	SE	z	p
intercept	0.04	0.15	0.24	0.81
prominence	0.85	0.26	3.26	< 0.01
continuum	-2.55	0.25	-10.90	< 0.001
prominence:continuum	-0.22	0.10	-2.19	< 0.05

Table 1: Model output for Experiment 1, with estimates for each fixed effect.

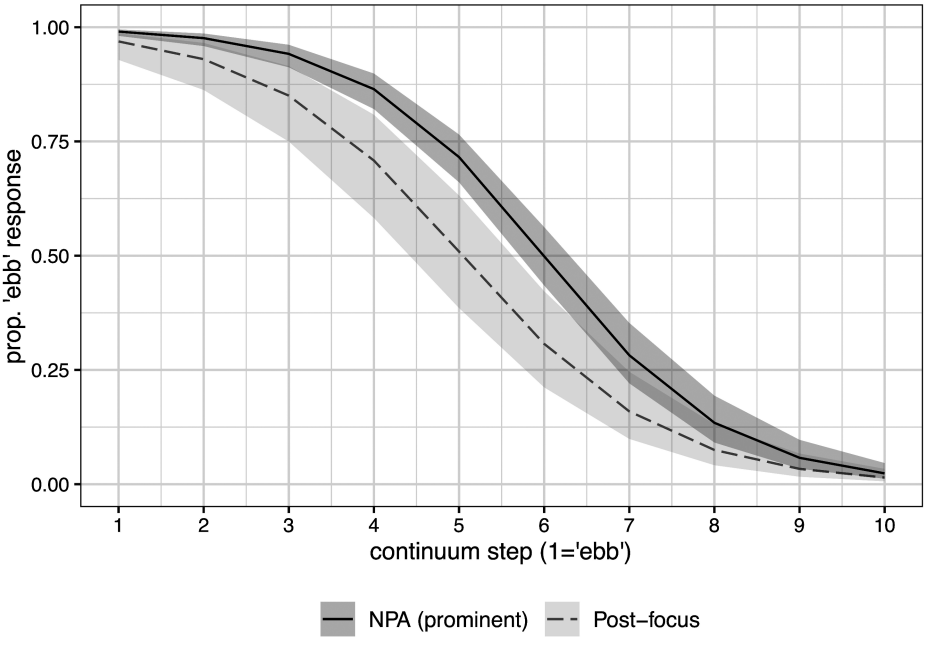

Figure 2: The model fit for categorization responses in Experiment 1. The F1/F2 continuum step is shown on the x axis. Shading around the model fit lines represents 95% CI. Categorization is split by prominence condition, indexed by line type and labeled below the plot.

As shown in Table 1 and Figure 2, the prominence manipulation had an effect on categorization such that listeners' "ebb" responses increased in the prominent NPA condition (β= 0.85, z =3.26). This aligns with our predictions above, showing that perception of the vowel contrast shifted on the basis of the prominence of the target word. This outcome suggests that, in line with the patterns we see for prosodic boundaries, phrasal prominence plays a mediating role in segmental contrast perception, based on how segmental cues (here, formants) are shaped by prominence in speech production.

Experiment 1 thus offers evidence for the relevance of phrasal prominence in the perception of spectral cues which signal vowel categories. Experiment 2 extended these findings in two principal ways. Firstly, we wanted to see if analogous effects of focus marking are observed for perception of a durational cue, also impacted by focus prosody (described below). Secondly, our goal was to increase the naturalness of our stimuli such that pitch on the target word varied between prominence conditions. Recall that in Experiment 1 the target was acoustically identical across conditions, with prominence manipulated by changing pre-target material. This was done to ensure differences in the target itself did not impact categorization, and to test prominence that was purely contextual. With the findings from Experiment 1 in hand we can accordingly test the impact of pitch on the target varying across conditions. This will additionally entail consideration of psychoacoustic effects of pitch on perceived duration, and the impact of adjacent segment durations, both relevant to the perception of durational cues, which we outline below.

3. Experiment 2

The durational cue we adopted as a test case in Experiment 2 is vowel duration as a cue to coda obstruent voicing in American English. Vowels are robustly longer before voiced coda obstruents, and this is a strong cue to voicing for listeners (Raphael, 1972). Accordingly, in Experiment 2 listeners categorized a continuum varying only in vowel duration, as the English word "coat" (word-final /t/) or "code" (word-final /d/). How should we expect focus prosody, conveyed by the same contextual changes in pitch and duration as those used in Experiment 1, to impact perception of this durational cue? One well-documented pattern in this regard is *post-focus compression.* Generally speaking, focused words are temporally expanded, while words that are post-focus within the same phrase are temporally compressed and de-accented (de Jong, 2004; Xu and Xu, 2005). We can conceptualize this as a durational analog to the sonority expansion effects on formants outlined above, i.e. focus marking within a phrase shapes how durational cues are realized. We can accordingly predict that listeners will adjust their categorization based on this pattern. Predictions are outlined below in Section 3.2.

3.1. Materials

The target word vowel duration continuum was created by PSOLA resynthesis. The vowel duration continuum ranged from 60 ms (corresponding to "coat") to 120 ms (corresponding to "code"). The starting point for the target was an accented production of the word "code". Audible voicing after stop closure was removed, to render the stop itself ambiguous, and vowel duration was subsequently manipulated. After the continuum was made, pitch was manipulated to vary across conditions. The pitch contour over the target vowel was modeled off a natural production of both a nuclear pitch accented (NPA) and post-focus target, produced with the prosodic structure of (1) and (2) shown in Section 2.1. Therefore, generally speaking, pitch was relatively high in the NPA condition (marking accentedness), and low on the de-accented Post-focus word. This manipulation was carried out by overlaying the pitch from an accented target, and post-focus target production (produced in the same frames).[2] Pitch on the target in the NPA condition was 134 Hz at vowel onset and 128 Hz at vowel offset (mean 131 Hz). Pitch on the target in the Post-focus condition was 112 Hz at vowel onset and 102 Hz at vowel offset (mean 107 Hz). The NPA and Post-focus frames were produced by the same speaker and analogous to those used in Experiment 1, though not identical. They were created in the same fashion as in Experiment 1.

3.2. Predictions

The aforementioned target pitch differences across conditions, not present in Experiment 1, raise the possibility that pitch itself will influence listeners' perception of vowel duration. The psychoacoustic literature documents a variety of influences of pitch height and dynamics, whereby changing pitch influences how long an acoustic event or speech sound is perceived to be by listeners. In this sense pitch and duration are described as being integrated, or interactive dimensions in perception (Prince, 2011). One well-documented effect is that of pitch height: higher pitch is perceived as longer by listeners, even when actual duration does not vary. This effect obtains both in explicit tasks where listeners rate or compare stimulus durations (Brigner, 1988; Yu, 2010) as well as in perceptual categorization tasks like those used here, e.g. changing perception of duration as a cue to coda voicing (Steffman and Jun, 2019b). Recall that pitch is higher on the prominent target in the NPA condition, as compared to the Post-focus condition: this difference in pitch would be predicted to lead to a longer perceived vowel in the NPA condition, and accordingly might be predicted to decrease voiceless "coat" responses.

Another general auditory influence on the perception of durational cues that merits consideration is that of the duration of adjacent segments. Durational contrast (e.g., Miller and Dexter, 1988) is the effect whereby a segment sounds relatively short or long based on the duration of a preceding or following acoustic event. This too is relevant for our stimuli given that pre-target duration varied across conditions. This variation is visible in Figure 1: a focused "say" preceding the target in the Post-focus condition is longer than its counterpart in the NPA condition. Following durational contrast effects, this difference

[2] Pitch was resynthesized in both conditions to ensure there was not a difference in naturalness based on resynthesis.

would predict that the target should be perceived as *longer* in the NPA condition, being preceded by a shorter segment.[3] A longer perceived vowel duration should lead to decreased "coat" responses, the same as predicted by pitch height differences. We can therefore say that both general auditory (non-prosodic) effects predict *decreased "coat" responses in the NPA condition.*

What, on the other hand, does the aforementioned pattern of post-focus compression (i.e., prosodic effect) predict? Following the logic of the compensatory adjustment observed in Experiment 1, compression of a post-focus target would lead to reduced vowel duration in general (as compared to an accented target). Listeners should accordingly require *shorter* vowel durations to perceive voicing for a Post-focus target, if they take this prosodic shortening into account. Overall, the Post-focus condition would accordingly generate more voiced percepts (i.e., shorter vowels can therein be perceived as voiced). This would lead to *decreased "coat" responses in the Post-focus condition*, the opposite of the effect predicted by the general auditory factors outlined above. Therefore, in testing durational cues, we must consider two possible outcomes driven by different influences: on one hand, general auditory effects related to pitch height and contextual duration, and on the other, adjustments driven by the way that phrasal prosody shapes temporal patterns in speech.

Some recent research suggests a third possibility: that the effect of prominence in this domain will vary based on vowel duration itself, i.e. the actual values from the continuum. For example, Steffman and Jun (2019b) tested if pitch height as a cue to prominence in isolated words would generate an analogous shift as the contextual effect tested here. The expected prosodic influence appeared only when target vowels were short, and was strongest when vowels were under 100 ms (see also Steffman and Jun (2019a)). At longer vowel durations, the effect of pitch was reversed, lining up with the predicted psycho-acoustic effect whereby high pitch increases perceived duration. The restriction of the prosodic effect to shorter vowels on the continuum was interpreted as originating from the fact that in natural speech, unaccented vowels (including post-focus vowels) are quite short, roughly under 100 ms in duration (Greenberg et al., 2003). Likewise, vowels that are perceived by listeners as lacking prominence (as indexed by an RPT task) also fall in this shorter range, where longer vowels tend to be perceived as more prominent (Mo, 2011), and more generally, duration serves as a clear cue to prominence for listeners. Accordingly, longer vowels on the continuum may not be readily perceived as non-prominent, even in a post-focus context in the present experiment, as a function of their inherent durational prominence. Following Steffman and Jun (2019b) we might therefore expect listeners to show stronger effects related to post-focus compression at the shorter steps of the continuum (and perhaps only at shorter steps). Observing if this contingency between the contextual prominence effect and vowel duration on the continuum obtains will therefore show if the influence of prosodic effects is restricted by the mapping of a target sound duration to context-typical realization, i.e. the fit of a target with a given prosodic context. If so, this would add some nuance to the findings of Experiment 1 in showing that the prevalence of contextual prominence effects in perception is contingent on intrinsic properties of the target itself.

3.3. Participants and procedure

39 participants were recruited from the same population as Experiment 1. The procedure was identical to Experiment 1.

3.4. Results and discussion

Results were assessed statistically in the same fashion as Experiment 1, this time predicting responses with "coat" mapped to 1 and "code" mapped to 0. Fixed and random effects in the model were specified as in Experiment 1. The fixed effects from the model are shown in Table 2, while the model fit for the data is shown in Figure 3.

[3] This effect is in fact also relevant for Experiment 1, given that /ɛ/ is overall shorter than /æ/ (House, 1961). Note that in this case a shorter preceding "say" in the NPA condition should make the target sound relatively longer, and therefore like /æ/, the opposite of what was found, suggesting that listeners are using primarily spectral information for the contrast in Experiment 1.

	β	SE	z	p
intercept	-0.32	0.06	-5.46	< 0.001
prominence	-0.07	0.18	-0.36	0.72
continuum	-0.73	0.08	-9.08	<0.001
prominence:continuum	-0.56	0.07	-8.06	<0.001
contrasts for vowel duration (ms)	β	SE	z-ratio	p
60	0.77	0.19	4.07	< 0.001
70	0.49	0.18	2.71	<0.01
80	0.21	0.18	1.18	0.24
90	-0.07	0.19	-0.36	0.73
100	-0.34	0.20	-1.78	0.09
110	-0.62	0.21	-2.92	< 0.01
120	-0.90	0.23	-3.86	< 0.001

Table 2: Model output for Experiment 2, with estimates for each fixed effect (above) and comparison of contrasts testing the effect of prominence condition at each continuum step (below).

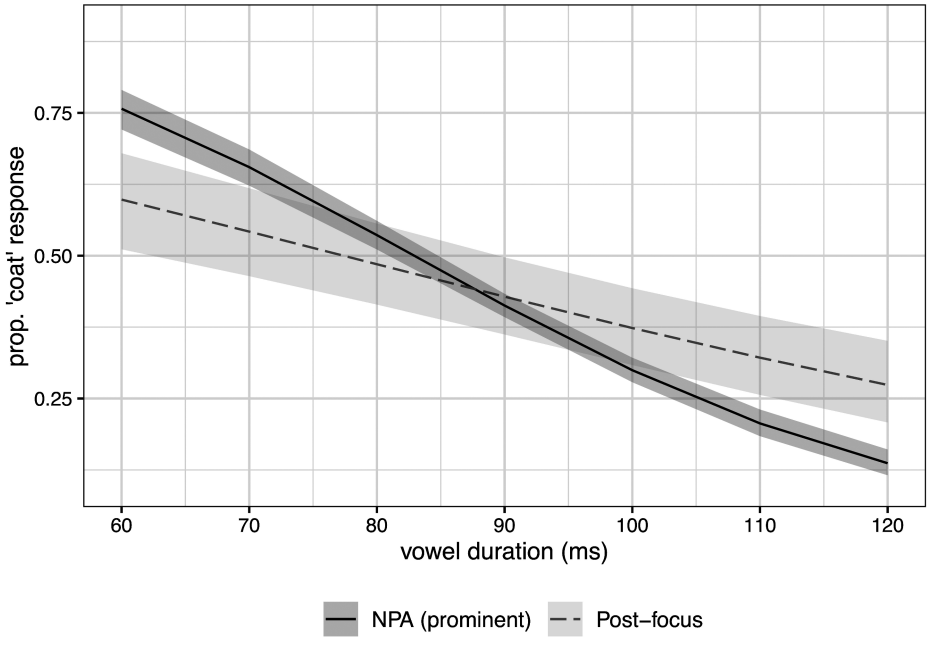

Figure 3: The model fit for categorization responses in Experiment 2. Vowel duration in milliseconds is shown on the x axis. Shading around the model fit lines represents 95% CI. Categorization is split by prominence condition, indexed by line type and labeled below the plot.

As shown in Figure 3, the directionality of the effect of prominence flips along the continuum, also visible in the robust interaction in the model (β=-0.56, z = -8.06). To look more closely at the changing directionality of the prominence effect, we compared contrasts from the model using the *emmeans* package in *R* (Lenth et al., 2018), testing the effect of prominence at each vowel duration step. This post-hoc comparison of contrasts shows clearly that at shorter vowel durations (60 and 70 ms) the Post-focus condition shows significantly decreased "coat" responses (shown in the bottom portion of Table 2). This effect aligns with the post-focus compression effects outlined above, showing that listeners incorporated prominence-related temporal patterning in their perception of these shorter continuum steps, in spite of possible competing influence of preceding duration and target pitch. For these shorter steps, we therefore see clear support for the role of phrasal prominence, even in this fairly conservative test case.

At the longer steps of the continuum (110 and 120 ms) we additionally see an effect of prominence condition, which is notably reversed: the NPA condition shows significantly decreased "coat" responses. The directionality of this effect aligns with the predicted auditory influences sketched above. That is, both higher pitch and a shorter preceding "say" in the NPA condition should lead to increased perceived target vowel duration and should therefore decrease voiceless "coat" responses.

The overall pattern that emerges in Experiment 2 therefore adds some nuance to results of Experiment 1. Firstly, we do in fact see both predicted effects, with the directionality of the effect contingent on stimulus duration. At shorter vowel durations only, results are in line with the predicted post-focus compression effects outlined above. The fact that this is restricted to shorter vowels is taken to relate to the fact that non-prominent vowels are typically quite short (under 100 ms) and vowels need to be short to be perceived as non-prominent by listeners (Greenberg et al., 2003; Mo, 2011). This result thus indicate that the contextual effects explored here are dependent on intrinsic properties of a target sound, such that they only occur when a target matches with a prosodic context. This outcome is perhaps not surprising; as mentioned above it is known that listeners' perception of prominence is also dependent on acoustic properties of a given sound, as shown in e.g., Mo (2011). However, the relation of intrinsic to contextual information that we see here suggests the importance of considering intrinsic prominence when exploring prosodic context effects in perception. The pattern that we see at longer vowel durations supports the idea that psychoacoustic processing might operate as a sort of default, something that occurs when target and context are not cohesive based on intrinsic prominence cues (here at longer vowel durations on the continuum). In this sense, prominence effects in segmental perception may be fundamentally different from boundary effects (cf. Mitterer et al., 2019; Steffman, 2019b). Further research will accordingly benefit from further exploring the relationship between intrinsic and contextual prominence in this regard, building on the results of Experiment 2.

4. General Discussion

Two experiments reported here tested how listeners incorporate phrase-level prominence in their perception of segmental cues, both spectral and durational. Both experiments find support for prominence effects, in line with how focus marking in a phrase modulates both formants and segment durations. Together, these experiments offer support for the idea that listeners interpret segmental cues in relation to their prominence in a phrase, aligning with similar claims made for phrasal boundaries (Cho et al., 2007; Mitterer et al., 2019). We also observed an interaction between contextual prominence and vowel duration in Experiment 2, which we take to reflect a moderating influence of intrinsic prominence.

This restriction of the observed prosodically-driven effect in Experiment 2 raises various questions that will benefit from further research. One open question raised by these results is how both effects seen in Experiment 2 relate to one another. Under the assumption that general auditory effects are unavoidable in processing, i.e., they occur irrespective of context and task-related factors (Green et al., 1997; Miller and Dexter, 1988), we should expect them to occur uniformly across continuum steps, early in processing. The literature suggests prosodic effects may operate at a later stage in processing, e.g. being integrated in the process of lexical competition (Cho et al., 2007; Mitterer et al., 2019). Support for both of these claims in the context of the present study could be gained by using eyetracking to assess the time-course of each effect. We hypothesize that shorter vowel durations in Experiment 2 might take longer to process in the case where listeners relate them to phrasal prominence. Timecourse evidence would accordingly be useful in explaining further the asymmetry we see across the continuum. Additionally, the question of the target sound's duration as a contributing factor to perceived prominence could be addressed with a task in which listeners provide prominence judgments for the stimuli, with e.g., a prominence rating task as in Bishop (2012).

One other line of research that would build on the present results is cross-linguistic. We claim the effects seen here originate from how focus prosody in American English shapes segmental realization, i.e., they are language-specific. Cross-linguistic differences in how focus prosody influences segmental realization should therefore be expected to show language-specific perceptual effects, and would accordingly tell us something important about how language experience shapes perception. The ways in which language prosody factors into perception in this regard remains largely unexplored (though see Mitterer et al., 2016). In the temporal domain, languages vary in the extent to which they show

post-focus compression: in Mandarin Chinese it does not occur, while post-focus *expansion* occurs in Taiwanese (Xu et al., 2012). In the spectral domain, languages further vary in the extent to which formant structure changes based on prominence (Delattre, 1969). Testing how the effects observed here extend to other languages will accordingly broaden the scope of the present study and more generally better our understanding of how listeners integrate prosodic prominence in segmental perception.

References

Beckman, Mary E and Pierrehumbert, Janet B. "Intonational structure in Japanese and English." *Phonology* 3: 255–309.

Bishop, Jason. "Information structural expectations in the perception of prosodic prominence." *Prosody and meaning* 25: 239.

Brigner, Willard L. "Perceived Duration as a Function of Pitch." *Perceptual and Motor Skills* 67.1 (1988): 301–302.

Cho, Taehong. "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ, i/ in English." *The Journal of the Acoustical Society of America* 117.6 (2005): 3867–3878.

Cho, Taehong, McQueen, James M., and Cox, Ethan A. "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English." *Journal of Phonetics* 35.2 (2007): 210–243.

Clopper, Cynthia G., Pisoni, David B., and de Jong, Kenneth. "Acoustic characteristics of the vowel systems of six regional varieties of American English." *The Journal of the Acoustical Society of America* 118.3 (2005): 1661–1676.

Cutler, Anne, Dahan, Delphine, and Van Donselaar, Wilma. "Prosody in the comprehension of spoken language: A literature review." *Language and speech* 40.2 (1997): 141–201.

de Jong, Kenneth. "Stress, lexical focus, and segmental focus in English: atterns of variation in vowel duration." *Journal of Phonetics* 32.4 (2004): 493–516.

Delattre, Pierre. "An acoustic and articulatory study of vowel reduction in four languages." *IRAL: International Review of Applied Linguistics in Language Teaching* 7.4 (1969): 295.

Green, Kerry P, Tomiak, Gail R, and Kuhl, Patricia K. "The encoding of rate and talker information during phonetic perception." *Perception & Psychophysics* 59.5 (1997): 675–692.

Greenberg, Steven, Carvey, Hannah, Hitchcock, Leah, and Chang, Shuangyu. "Temporal properties of spontaneous speech—a syllable-centric perspective." *Journal of Phonetics* 31: 465–485.

House, Arthur S. "On vowel duration in English." *The Journal of the Acoustical Society of America* 33.9 (1961): 1174–1178.

Keating, Patricia, Cho, Taehong, Fougeron, Cécile, and Hsu, Chai-Shune. "Domain-initial articulatory strengthening in four languages." *Phonetic interpretation: Papers in laboratory phonology VI* : 143–161.

Kim, Sahyang and Cho, Taehong. "Prosodic boundary information modulates phonetic categorization." *The Journal of the Acoustical Society of America* 134.1 (2013): EL19–EL25.

Kim, Sahyang, Mitterer, Holger, and Cho, Taehong. "A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing." *PloS one* 13.8 (2018).

Lenth, Russell, Singmann, Henrik, Love, Jonathon, Buerkner, Paul, and Herve, Maxime. "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2018.

McClelland, James L and Elman, Jeffrey L. "The TRACE model of speech perception." *Cognitive psychology* 18.1 (1986): 1–86.

Miller, Joanne L and Dexter, Emily R. "Effects of speaking rate and lexical status on phonetic perception." *Journal of Experimental Psychology: Human Perception and Performance* 14.3 (1988): 369.

Mitterer, Holger, Cho, Taehong, and Kim, Sahyang. "How does prosody influence speech categorization?" *Journal of Phonetics* 54: 68–79.

Mitterer, Holger, Kim, Sahyang, and Cho, Taehong. "The glottal stop between segmental and suprasegmental processing: The case of Maltese." *Journal of Memory and Language* 108: 104034.

Mo, Yoonsook. *Prosody production and perception with conversational speech*. Doctoral Dissertation, University of Illinois at Urbana-Champaign, 2011.

Mo, Yoonsook, Cole, Jennifer, and Hasegawa-Johnson, Mark. "Prosodic effects on vowel production: Evidence from formant structure." *INTERSPEECH*. 2009, 2535–2538.

Norris, Dennis. "The Merge model: Speech perception is bottom-up." *The Journal of the Acoustical Society of America* 106.4 (1999): 2295–2295.

Peterson, Gordon E. and Barney, Harold L. "Control Methods Used in a Study of the Vowels." *The Journal of the Acoustical Society of America* 24.2 (1952): 175–184.

Price, Patti J, Ostendorf, Mari, Shattuck-Hufnagel, Stefanie, and Fong, Cynthia. "The use of prosody in syntactic disambiguation." *the Journal of the Acoustical Society of America* 90.6 (1991): 2956–2970.

Prince, Jon B. "The integration of stimulus dimensions in the perception of music." *Quarterly Journal of Experimental Psychology (2006)* 64.11 (2011): 2125–2152.

Raphael, Lawrence J. "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 (1972): 1296–1303.

Silverman, K and Pierrehumbert, J. "The timing of prenuclear high accents in English." *Papers in Laboratory Phonology*. eds. M. E. Beckman and John Kingston, Papers in Laboratory Phonology. 1990. 72–106.

Steffman, Jeremy. "Intonational structure mediates speech rate normalization in the perception of segmental categories." *Journal of Phonetics* 74: 114–129.

———. "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 (2019b): EL560–EL566.

Steffman, Jeremy and Jun, S. A. "Effects of prosodic structure versus durational context on the perception of segmental categories: The case of focus realization." *Proceedings of the 19th International Congress of Phonetic Sciences*. Melboure, Australia, 2019a.

Steffman, Jeremy and Jun, Sun-Ah. "Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception." *The Journal of the Acoustical Society of America* 146.3 (2019b): EL251–EL257.

Steffman, Jeremy and Katsuda, Hironori. "Intonational Structure Influences Perception of Contrastive Vowel Length: The Case of Phrase-Final Lengthening in Tokyo Japanese." *Language and Speech* : 0023830920971842.

Van Summers, W. "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses." *The Journal of the Acoustical Society of America* 82.3 (1987): 847–863.

Xu, Yi, Chen, Szu-wei, and Wang, Bei. "Prosodic focus with and without post-focus compression: A typological divide within the same language family?" *The Linguistic Review* 29.1 (2012): 131–147.

Xu, Yi and Xu, Ching X. "Phonetic realization of focus in English declarative intonation." *Journal of Phonetics* 33.2 (2005): 159–197.

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

Proceedings of the 38th West Coast Conference on Formal Linguistics

edited by Rachel Soo, Una Y. Chow, and Sander Nederveen

Cascadilla Proceedings Project Somerville, MA 2021

ISBN 978-1-57473-479-9 hardback

Ordering information

Orders for the printed edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Steffman, Jeremy and Sun-Ah Jun. 2021. Prosodic Prominence in Speech Perception: The Influence of Focus Prosody on the Perception of Durational and Spectral Cues. In *Proceedings of the 38th West Coast Conference on Formal Linguistics*, ed. Rachel Soo et al., 406-416. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3585.