EFFECTS OF PROSODIC STRUCTURE VERSUS DURATIONAL CONTEXT ON THE PERCEPTION OF SEGMENTAL CATEGORIES: THE CASE OF FOCUS REALIZATION

Jeremy Steffman & Sun-Ah Jun

University of California, Los Angeles
jsteffman@ucla.edu; jun@humnet.ucla.edu

ABSTRACT

The present study investigates the role prosodic context plays in the perception of durational contrasts. A "coat"~"code" vowel duration continuum was placed in two frames, one in which it received a nuclear accent (as a control condition), and another in which it followed a focused word. We tested two competing hypotheses: (1)Based on durational context alone, listeners might require *longer* vowel duration for a "code" response when following a focused (lengthened) word, compared with the control condition, normalizing for proximal duration. (2)Alternatively, listeners may expect *shorter* vowel durations for a post-focus target, because vowels are compressed post-focus. Two experiments, using different length continua, find that listeners in fact show both patterns of categorization. When target vowel duration is short, categorization shifts in line with expectations about prosodic structure, negating proximal duration effects. Results thus suggest that under the right circumstances, sensitivity to prosodic patterns mediates the perception of durational cues.

Keywords: prosody, speech perception, speech rate normalization, focus marking.

1. INTRODUCTION

It is well established that the phonetic realization of segments varies systematically by prosodic position [14,15]. However, the way in which listeners' perception of segmental contrasts is mediated by prosody is an open question [16,19,26]. The present study addresses this in light of recent research.

In one recent study, Kim & Cho [16] explored how listeners might be perceptually sensitive to initial strengthening [15] of VOT. Given that VOT is robustly longer in phrase-initial than phrase–medial position, the authors predicted that listeners would require *longer* VOT to categorize a sound as voiceless when Intonation Phrase (IP)-initial in a carrier phrase. A VOT continuum (/pɑ/ ~ /ba/) was placed in the carrier phrase "Let's hear *x* again". In one condition, the target was IP-initial, preceded by an IP-boundary marked by phrase-final lengthening and low boundary tone on "hear*"* ("Let's hear % *x* again"). In another condition the target was IP-medial and the carrier phrase was one unitary IP. The authors found listeners required *longer* VOT for a /p/ response when the target was IP-initial. They interpreted this as compensation for initial strengthening, reflecting sensitivity to the phonetic encoding of prosodic structure, and prosodic context.

More recently, Mitterer, Cho & Kim [19] suggested that this observed effect may be due to speech rate normalization. Listeners adjust categorization of VOT (and other temporal cues) based on speech rate, where longer segmental durations preceding a target shift categorization to higher VOT values for a voiceless stop response [29]. Because the IP boundary used in [16] was cued by phrase-final lengthening, and because rate normalization occurs based on *local*/*proximal* slowdowns [29], the shift observed by [16] may have originated from differences in preceding length alone. [19] shows that global slowdowns shift categorization similarly, and that the removal of intonational cues to boundary (via flattening of F0) does *not* alter categorization, suggesting that the shift occurs on the basis of duration alone. These results offer potential support for the speech rate normalization account, though as [19] highlights, in this particular case both accounts predict *the same effect*. It is therefore unclear to what extent listeners modulate categorization on the basis of prosodic patterns. Speech rate normalization is typically seen as domain-general auditory processing (e.g. [19,21]), and does not implicate linguistic structure (contra [16]). As these two accounts attribute the same shift in categorization to different mechanisms, adjudicating between them is of theoretical interest.

1.2. The present study

In light of this unresolved issue and recent work suggesting that prosody may indeed be relevant in rate-dependent speech perception [26,27], the present study investigated a different prosodic pattern as a test case. We selected a case where, unlike [16,19], a shift in categorization based on prosody would

predict the *opposite* of what would be expected in speech rate normalization, namely, focus realization.

In English, a focused word is expanded in duration and pitch, while words that follow are compressed in both (and unaccented phonologically), known as post-focus compression (PFC) [8,33]. As previous literature on PFC documents the temporal compression of vowels, we selected vowel duration as a cue to coda obstruent voicing [7] as a test case. This is a robust rate-dependent cue to voicing in English [13,24]. We used a "coat"-"code" continuum (chosen to be frequency-matched from [6]), varying only in vowel duration.

We placed the target in a carrier phrase in two prosodic conditions: (i) in the nuclear pitch accented position as a control (the NPA condition), and (ii) immediately after a focused word (the POST-FOCUS condition). (1) and (2) show ToBI [2] transcribed representations of this manipulation, where *x* is the target. The name of each condition is given at right.

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

(2) I'll *say* *x* now (POST-FOCUS)
L+H* L-L%

"Say" in (2), being focused, is longer than "say" in (1). Therefore proximal speech rate normalization (e.g. [19,29]) would be expected to shift categorization to require *longer* vowel durations for a "code" response following a relatively longer preceding "say" in (2) (causing *decreased* "code" responses in (2) relative to (1)).

On the other hand, according to the prosody account, we would predict the *opposite* shift in categorization. Specifically, if listeners attribute the lengthening in "say" in (2) as marking focus, they would expect a post-focus target to be shorter in duration, as compared to (1). In other words, listeners would require *shorter* vowel durations for a "code" response following a focused "say" in (2) (causing *increased* code responses in (2) relative to (1)). The present experiments thus directly test the influence of proximal durational cues on segment categorization in comparison with prosodic context effects.

2. EXPERIMENT 1

Experiment 1 was a 2AFC task, where listeners categorized a target sound as "coat" or "code".

2.1. Materials

The target was placed in two carrier phrases corresponding to (1) and (2) above, spoken by a ToBI-trained English speaker. In these sentences the target was produced as "code".

In creating the prosodic conditions for the stimuli (using PSOLA [20], in Praat [3]), the words "I'll" and "say" were excised from both (1) and (2), and the duration of the word "I'll" and the [s] in "say" was averaged across frames. The duration of the vowel in "say" was manipulated in two ways, creating the two prosodic contexts. In the POST-FOCUS condition, the duration of the vowel [eɪ] in "say" from (2) was set to be 205ms. In the NPA condition, the duration of the vowel [eɪ] in "say" from (1) was set to be 125ms. These vowel durations were only slightly different from the natural productions of the vowel in each sentence. The word "now" from (1) was appended to *both* frames, creating "I'll say __ now". The only durational difference across frames is thus in the vowel [eɪ]. Crucially, the frame created from (2) has increased pitch and amplitude on "say" (marking focus) relative to that in the frame created from (1). The POST-FOCUS prosody condition will refer to this frame with a longer vowel (205ms) and other cues to focus in "say" and a *post-focus target*. The NPA prosody condition will refer to the frame created from (1), with a shorter vowel in "say" (125ms) and the target bearing the nuclear pitch accent.

The word "code" produced in (1) was used to create the continuum. Voicing during closure was removed to render the stop ambiguous. The mean intensity and pitch of the target was then set to be the average of these values for the targets in (1) and (2). Rendering these values as averages ensures that the pitch on target will not bias categorization (cf. [28]) and leaves it relatively prosodically ambiguous such that its interpretation will be a function of context. The continuum was created by resynthesizing the vocalic portion of the target word, ranging from 60ms to 150ms in 15ms steps. Each of the 7 steps of the continuum was inserted into the two frames, to create a total of 14 unique stimuli. The silent interval preceding the onset [kʰ] in the target was set to 50ms. The silent interval following the word-final stop was set to 80ms, a relatively ambiguous value given that stop closures are longer for [t] than [d] (e.g. [10]).

2.2. Participants and procedure

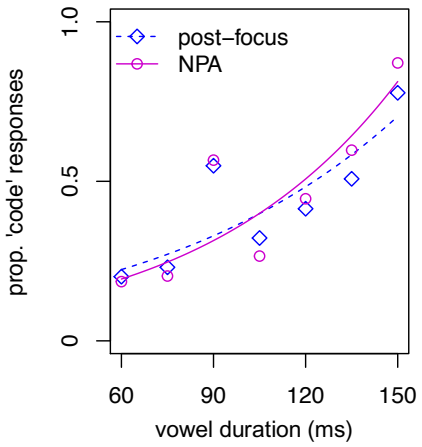
Thirty-two native American English-speaking adults participated in the study. Participants were students at UCLA and received course credit. Participants were tested in a sound attenuated room in the UCLA Phonetics Lab, seated in front of a desktop computer. Stimuli were presented binaurally on a Peltor™ 3M™ listen-only headset. Participants heard a stimulus and saw "code" on one side of the screen and "coat" on the other (counterbalanced across participants),

presented using Appsobabble [30]. They indicated their choice by keypress ('f' for the word on the left side of the screen, and 'j' for the word on the right). There were 16 repetitions of each of the 14 unique stimuli (224 trials). The ITI was 250 ms. Stimuli were totally randomized (for each participant). Participants took a short break halfway through the experimental trials.

2.3. Results and discussion

Results were assessed by mixed-effects logistic regression, in RStudio [25], using *lme4* and *emmeans* [1, 18]. The dependent variable in the model was the listeners' response ("code" mapped to 1). Fixed effects were duration (centered at 0), prosody (effect-coded: POST-FOCUS mapped to -1, NPA mapped to 1), and their interaction. Random effects were by-subject intercepts, with maximal by-subject random slopes. Figure 1 shows listeners' categorization as the proportion of "code" responses at each continuum step. The model output is shown in Table 1.

Figure 1: Categorization by prosody condition. Points are fit with psychometric curves to show a smoothed categorization trend.



As shown in Table 1, prosody did not have a significant main effect (it is also not clear why the 90ms step of the continuum shows an unusually high proportion of "code" responses). However, a significant interaction ($p < 0.001$) shows an asymmetry in the effect of prosody along the continuum. To investigate the interaction, the effect of prosody at each continuum step was tested. The POST-FOCUS condition showed significantly decreased "code" responses at the two longest continuum steps (see Table 1). This is expected based on proximal speech rate: a longer preceding vowel in the POST-FOCUS condition shifts categorization to longer required durations for a "code" response.

Table 1: Output from the Experiment 1 model (top),with post hoc comparison of contrasts at each continuum step (bottom).

	β(SE)	z value	p value
Intercept	-0.36(0.11)	-3.35	< 0.001
prosody	0.05(0.13)	0.38	0.70
v dur	0.79(0.08)	10.26	< 0.001
prosody: v dur	0.25(0.07)	3.48	< 0.001
Comparison of contrasts with emmeans			
duration (ms)	Estimate(SE)	z ratio	p value
60	0.32(0.17)	1.87	**0.06**
75	0.20(0.15)	1.32	0.18
90	0.08(0.13)	0.56	0.57
105	-0.05(0.12)	-0.38	0.70
120	-0.17(0.13)	-1.34	0.18
135	-0.29(0.14)	-2.13	**0.03**
150	-0.42(0.17)	-2.66	**0.007**

However, The effect disappears at all other steps of the continuum, and a marginally significant effect ($p = 0.06$) in the *opposite* direction (predicted by sensitivity to PFC) is shown at the shortest step. These results suggest that the effect of context is contingent on vowel duration itself, and that when duration is short enough, listeners may be sensitive to PFC. The possibility of an asymmetrical effect of context on listeners' categorization may be related to the fact the unaccented vowels in English tend to be shorter than the longer steps of the continuum used here, both in isolated sentences [9] and corpora of running speech [11]. Therefore, an effect of prosodic context may only emerge when the duration of the vowels in question more closely matches the typical duration of unaccented (here, post-focus) vowels. This idea is consistent with the hypothesis that sensitivity to prosodic patterns in perception is to some extent learned from language input. To test this idea, we shortened the range of the continuum in a second experiment (60-120ms), testing how listeners might interpret contextual information when the two longest steps are absent and the target vowel overall maps onto more typical post-focus durations. In other words, Experiment 2 tests if contextual effects of prosody will be sensitive to how closely a durational pattern matches the typical duration of sounds *in that context*.

3. EXPERIMENT 2

3.1. Materials

In Experiment 2, the continuum ranged from 60-120ms in 10ms steps. Steps were smaller and the longest continuum step was 30ms shorter than that in Experiment 1. All other aspects of the stimuli were identical to those in Experiment 1.

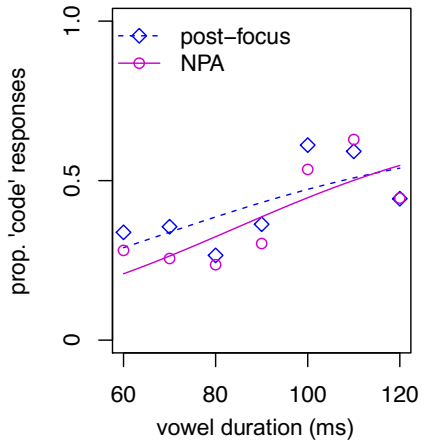
3.2. Participants and procedure

Another thirty-two native American English-speaking adults participated in the study. The procedure was the same as in Experiment 1.

3.3. Results and discussion

Model specifications were the same as for Experiment 1. Figure 2 shows categorization split by condition. Table 2 shows the model output and post-hoc comparisons.

Figure 2: Categorization by prosody condition.



Though vowel duration significantly influenced categorization as expected ($p < 0.001$), the categorization functions of the continuum are quite shallow, reflecting a high degree of ambiguity in the stimuli. As in Experiment 1, there was a significant interaction between prosody and vowel duration ($p = 0.002$). Testing the interaction showed the POST-FOCUS prosody condition significantly *increased* "code" responses at the two lowest steps of the continuum, with a marginally significant effect ($p = 0.07$) at the third lowest step. There was no significant effect of prosody at higher steps on the continuum. The directionality of the effect at all other steps, except the longest, matched this pattern showing that proximal speech rate effects were absent.

The effect of prosodic context can therefore be taken to reflect listener sensitivity to the temporal realization of PFC, contingent on the duration of the target vowel itself. Reducing continuum step size (approaching the JND for vowel duration for some continuum steps [17]), may also have encouraged listeners to rely more heavily on prosodic context and less so on contextual durations, contributing to the observed effect.

Table 2: Output from the Experiment 2 model (top), with post hoc comparison of contrasts at each continuum step (bottom).

	β(SE)	z value	p value
Intercept	-0.36(0.11)	-3.20	0.001
prosody	-0.25(0.19)	-1.28	0.20
v dur	0.39(0.08)	4.83	< 0.001
prosody: v dur	0.21(0.07)	3.124	0.002
Comparison of contrasts with emmeans			
duration (ms)	Estimate(SE)	z ratio	p value
60	0.56(0.22)	2.53	**0.01**
70	0.46(0.20)	2.20	**0.03**
80	0.35(0.20)	1.78	**0.07**
90	0.25(0.19)	1.28	0.20
100	0.14(0.19)	0.74	0.46
110	0.04(0.20)	0.20	0.84
120	-0.06(0.21)	-0.295	0.77

4. CONCLUSIONS

The experiments reported here show that, under the right circumstances, perceptual compensation for prosodic context can indeed occur, supporting the general argument made by [16, 26]. These results can thus broadly be taken to suggest that prosodic patterns can mediate listeners' perception of durational cues, though the restricted nature of the effect highlights the need for further extension of these results.

As shown in Experiment 1, speech rate normalization *also* occurs with longer vowel durations, and only with shorter durations does perceptual sensitivity to PFC appear (Experiment 2). The observed contingency between the duration of the target itself and the effect of prosodic context can be taken as suggesting that sensitivity to prosodic patterns plays a role only when durational values map to expected durations for that prosodic context, supporting the idea that sensitivity to prosodic patterns is learned (in the sense discussed in [32, 12]). These results may thus reflect an interplay between domain-general, and language-specific effects in rate dependent speech perception (following [22,23]). The mechanisms underlying proximal speech rate effects more generally remain a topic of investigation [4,31], and the present results can complement these lines of research by showing prosodic patterns merit consideration as a mediating influence in the perception of durational cues. Further exploring these questions with speakers of languages with different degrees of post-focus temporal compression [34] may provide insight into the extent learned language patterns account for the observed effect. Extension of these results will therefore help to improve our understanding of how prosodic patterns are integrated into the perception of segmental categories, and how they may interact with domain-general perceptual processes.

5. REFERENCES

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

[2] Beckman, M. E. and Hirschberg, J. (1994), The ToBI Annotation Conventions. Online MS.

[3] Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.43.

[4] Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attn., Perception, & Psychophysics*, *79*(1), 333–343.

[5] Bosker, H. R., & Reinisch, E. (2017). Foreign Languages Sound Fast: Evidence from Implicit Rate Normalization. *Frontiers in Psychology*, *8*.

[6] Brysbaert, M., & New, B. (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.

[7] Chen, M. (1970). Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica*, *22*(3), 129–159.

[8] de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *JPhon*, *32*(4), 493–516.

[9] Dimitrova, S., & Turk, A. (2012). Patterns of accentual lengthening in English four-syllable words. *JPhon*, *40*(3), 403–418.

[10] Flege, J. E., Munro, M. J., & Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin, and Spanish. *JASA*, *92*(1), 128–143.

[11] Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. J*Phon*, *31*(3), 465–485.

[12] Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(5), 1166–1183.

[13] Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attn., Perception, & Psychophysics*, *79*(3), 964–988.

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

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

[16] Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *JASA*, *134*(1), EL19–EL25.

[17] Klatt, D. H., & Cooper, W. E. (1975). Perception of Segment Duration in Sentence Contexts. In *Structure and Process in Speech Perception* (pp. 69–89). Springer, Berlin, Heidelberg.

[18] Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Retrieved from https://CRAN.R-project.org/package=emmeans

[19] Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *JPhon*, *54*, 68–79.

[20] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Commun.*, 9(5-6), 453–467.

[21] Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *JPhon*, 37(1), 46– 65.

[22] Pind, J. (1998). Auditory and linguistic factors in the perception of voice offset time as a cue for preaspiration. *JASA*, *103*(4), 2117–2127.

[23] Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attn., Perception, & Psychophysics*, *78*(1), 334–345.

[24] Raphael, L. J. (1972). Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English. *JASA*, *51*(4B), 1296–1303.

[25] RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA.

[26] Steffman, J. (2018). Phrase final lengthening modulates categorization of vowel length as a cue to obstruent voicing in English. *Proceedings of Meetings on Acoustics*, *33*(1).

[27] Steffman, J. (2018). *Intonation mediates speech rate normalization in the perception of segmental categories.* (MA Thesis). UCLA.

[28] Steffman, J., & Jun, S.-A. (2019). Listeners integrate pitch and durational cues to prosodic structure in word categorization. *Proceedings of the Linguistic Society of America*, *4*(1).

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

[30] Tehrani, H. (2015). Appsobabble [online applications platform] http://www.appsobabble.com.

[31] Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attn., Perception, & Psychophysics*, *74*(6), 1284–1301.

[32] Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950.

[33] Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *JPhon*, *33*(2), 159–197.

[34] Xu, Y., Chen, S., & Wang, B. (2012). Prosodic focus with and without post-focus compression: A typological divide within the same language family? *Tlir*, *29*(1), 131–147.