TESTING THE INFLUENCE OF DISTAL RHYTHMIC STRUCTURE ON LISTENERS' PERCEPTION OF DURATIONAL CUES

Jeremy Steffman

University of California, Los Angeles
jsteffman@ucla.edu

ABSTRACT

This study investigated whether distal (non-adjacent) rhythmic patterns or proximal (adjacent) durational differences would influence listeners' perception of temporal cues, testing categorization of a "coat"~"code" vowel duration continuum. The pattern of alternating long/short syllables preceding the target was manipulated such that the target was the second syllable in either a trochee (…long-short, long-*target*) or iamb (…short-long, short-*target*). Two competing predictions are tested: (1) If the target is grouped perceptually as the second, longer, syllable in an iamb (versus trochee), listeners might expect longer vowel durations for a "code" response (decreasing "code" responses). (2) Proximal durational contrast effects predict the opposite shift in categorization, where a shorter syllable precedes the target in the iamb condition (increasing "code" responses). Results substantiate prediction (1): categorization shifts in line with *expectations* about rhythmic grouping. Results are discussed in terms of distal/proximal speech rate effects, and the importance of rhythmic patterns for word segmentation/lexical processing.

Keywords: rate-dependent perception, prosody, speech rhythm, speech rate, speech perception.

1. INTRODUCTION

Listeners must contend with highly variable acoustic cues in perceiving speech. One dimension of variability is speaking rate (e.g. [17,24]), which influences the duration of cues to segmental contrasts (e.g. [19]). In light of this, it is well established that listeners interpret durational cues relative to their context, influenced by both *proximal* [11,20] (typically defined as adjacent in terms of syllables/segments) and *distal* (i.e. non-adjacent) [15,26] speech rate. One significant issue in investigating the influence of speech rate on listeners' interpretation of durational cues is the relative importance of proximal durations (i.e. "durational contrast" as defined in [11]), and more distal changes in rate [5]. The present study addresses a related question by testing how the rhythmic properties (involving relative timing) of a more distal context influence rate-dependent perception of a segmental contrast. This question is pursued in light of the demonstrated importance of distal prosodic/rhythmic structure in word segmentation and lexical processing [12,13,21].

1.1. Distal and proximal context effects

In a series of experiments, Bosker (2017) [5] crossed the duration of pure tones preceding a target (long versus short), with their rate of repetition (fast versus slow), and observed how these manipulations influenced listeners' categorization of a subsequent Dutch vowel length contrast, reflecting their perception of duration. [5] showed that the rate of repetition of tones drives listeners' adjustment of categorization, where faster repetition increased long vowel responses (i.e. following a fast rate listeners more readily categorized a subsequent vowel as phonemically long). Perhaps surprisingly, proximal duration did not influence categorization in any of [5]'s experiments. These results highlight the importance of distal rate effects, and the apparent insignificance of proximal durational contrasts when distal context is present ([5] suggests proximal effects may be better understood as originating from cue-integration processes, following e.g. [30]). Given that distal context appears to be of central importance in rate-dependent perception [6,14] the present study investigates how the timing structure of distal patterns in speech may influence listeners' perception of durational cues.

1.2. Distal prosodic effects in speech processing

Another body of literature documents the important role that alternating sequences of pitch and duration play in word segmentation and lexical processing, e.g. [12]. For example, given an ambiguous string of sounds that can be parsed in two ways, e.g. "cry#sister#nip" versus "crisis#turnip" [13], listeners will parse the sequence such that the two first syllables form a single word "crisis", when the preceding context matches such a parse, i.e. when a sequence of strong (S) and weak (w) preceding syllables implies that the upcoming string should be grouped as one S-w unit (i.e. S-w S-w crisis # turnip).

The other parse is obtained with distal rhythmic cues imply a different grouping for the first syllable in the ambiguous string. Similar effects have also been demonstrated with eye tracking, suggesting they play an important role in online speech processing [8]. These findings are couched in the *perceptual grouping hypothesis* [12,21], which predicts that alternating patterns of pitch and duration should influence listeners' expectations about the grouping of upcoming material in the speech signal, as informed by findings in domain-general auditory perception of pitch and duration [4, 18]. Given that listeners clearly incorporate expectations about distal rhythmic grouping in word segmentation/lexical processing, the present study extends the perceptual grouping hypothesis to test how perceptual grouping may influence processing of durational cues.

1.3. The present study

The present study is a first step in extending both of these lines of research. The durational cue chosen as a test case is vowel duration as a cue to coda obstruent voicing (where vowels are longer preceding voiced obstruents, e.g. [10]). This is a robust cue for voicing in English that is influenced by changes in durational context [16,25]. This study examines how listeners' perception of vowel duration shifts on the basis of distal rhythmic information. Specifically, based on whether the target is preceded by a series of (durational) trochees (long-short) or iambs (short-long). Following the logic of the perceptual grouping hypothesis, listeners may group the target syllable as the second syllable in either a trochee, or iamb, and expectations about this grouping may mediate their perception of vowel duration as a cue to voicing. Proximal context effects are predicted to generate a different perceptual adjustment (outlined below).

The present study can thus be viewed from two angles: on one hand, it extends the literature showing the importance of distal prosodic/rhythmic patterns in word segmentation to explore how they may influence the perception of durational cues. On the other hand, it investigates the relative importance of distal and proximal speech rate effects by testing how the timing patterns of a distal context influence categorization, in competition with proximal context effects.

2. THE EXPERIMENT

The experiment was a 2AFC task. Listeners categorized a target sound from a vowel duration continuum as one of two English words: "coat" or "code". These particular words were chosen as they are relatively matched for frequency (from [9]). The crucial manipulation in the experiment is whether a series of *trochees* or *iambs* preceded the target. Based on the grouping of preceding syllables, the target formed the second syllable of either type of foot.

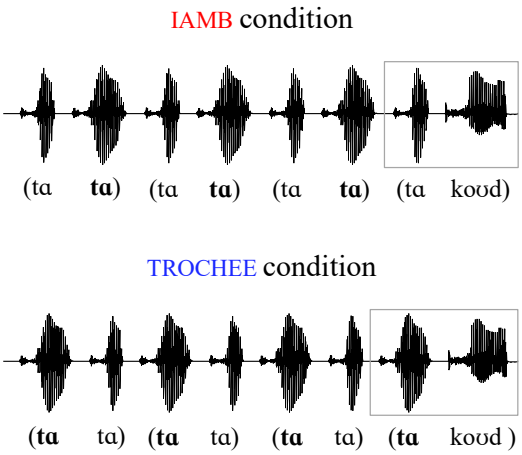
2.1. Materials

Stimuli were created by PSOLA resynthesis [22] of the natural speech of a male speaker of American English, in Praat [3]. In creating the vowel duration continuum, the word "code" was excised from the carrier phrase "I'll say code now". Audible voicing after closure was removed, to render the target stop ambiguous. The vowel duration of the original token was approximately 170 ms. The continuum was synthesized by manipulating the vocalic portion of the target word (all continuum steps were created by resynthesis). The continuum had 5 steps, each separated by 15 ms. The shortest endpoint of the continuum was set to be 90 ms (corresponding to a "coat" response), the longest endpoint of the continuum was set to be 150 ms (corresponding to a "code" response). These endpoint durations were determined based on pilot experiments.

A target sound from this continuum was placed following one of two precursors, which manipulated the implied rhythmic grouping of the target. To allow for tight control of the durational properties of the precursor, a CV syllable [$t^h\alpha$], produced by the same speaker, was resynthesized to have one of two vocalic durations, 75 ms or 150 ms. Only vowel duration was manipulated, VOT was identical in all precursor syllables. These short and long syllables were iterated in a two different patterns to create the IAMB and TROCHEE conditions. In creating the IAMB condition, a short syllable was placed preceding a long syllable to create at short-long iambic foot. This pattern was then repeated three times. In a final, fourth foot, a short syllable was followed by the target sound (which had different vowel durations based on the continuum step). The target was thus grouped with a preceding short syllable to form the second syllable of an iambic foot (see Figure 1). In creating the TROCHEE condition the relative ordering of long and short precursor syllables was switched such that three trochaic feet preceded the final foot, which consisted of a long syllable and the target sound (see Figure 1). The two conditions thus present different rhythmic structures preceding the target, and differ in the implied status of the target, as either the second syllable in an iambic, or trochaic foot. All syllables were separated by 50 ms of silence. So that duration alone distinguished the precursor syllables from the target, the average intensity and pitch (which was monotonized) of every syllable in the stimulus was manipulated to be the same (72 dB; 131 Hz). Crucially, in terms of proximal context, the syllable

preceding the target is *shorter* in the IAMB condition and *longer* in the TROCHEE condition. This is highlighted in Figure 1 which shows the two conditions below.

Figure 1: Waveforms showing all eight syllables of the stimuli. The target has 150 ms vowel duration. The longer syllable in the precursor is bolded in transcription. Parentheses represent hypothesized grouping. Proximal context is boxed.



Given these conditions, two predictions can be contrasted. Firstly, consider the proximal context only. As noted above, a longer syllable precedes the target in the TROCHEE condition, relative to the IAMB condition. Based on local speech rate normalization (i.e. durational contrast [11]) this proximal difference would predict that listeners should require *longer* vowel durations for a "code" response in the TROCHEE condition, given that preceding lengthening shifts the perception of durational cues. The effects of proximal context would therefore predict a *decrease* in "code" responses in the TROCHEE condition relative to the IAMB condition.

Next consider what might be predicted based on the perceptual grouping hypothesis. If listeners perform the expected perceptual grouping, in the TROCHEE condition the target would be grouped as the second syllable of a trochaic foot. Similarly, listeners would group the target as the second syllable in an iambic foot in the IAMB condition. Following this logic, listeners might expect *shorter* vowel durations for a "code" response in the TROCHEE condition, given that the target is perceptually grouped as being the second, *shorter*, syllable in a series of long-short feet. Likewise, in the IAMB condition, listeners would expect *longer* target durations when the target is the second syllable in an iambic foot. This would predict *increased* "code" responses in the TROCHEE condition, where shorter vowels are more readily perceived as "code". This result would implicate listeners' *expectations* about vowel duration and rhythmic grouping based on distal context, where the relative timing of the precursor modulates perception of vowel duration. As outlined above, this predicts that categorization will shift in the opposite direction as predicted by proximal context.

2.2. Participants

Thirty-two self-reported native English-speaking adults with normal hearing participated in the study. Participants were students at UCLA and received course credit for participation.

2.3. Procedure

Testing was carried out in a sound-attenuated room in the UCLA Phonetics Lab. Participants were seated in front of a desktop computer. Stimuli were presented binaurally via a Peltor™ 3M™ headset. The platform used for experiment presentation was Appsobabble [29]. During testing, participants heard a stimulus and saw "code" on one side of the screen and "coat" on the other (counterbalanced across participants). They indicated their choice by keypress where 'f' indicated the choice on the left side of the screen and 'j' indicated the choice on the right side of the screen. Participants heard 10 repetitions of each of the 10 unique stimuli, for a total of 100 trials. The ITI was 250 ms. Stimuli were completely randomized. The experimental trials were separated by a short self-paced break halfway through.

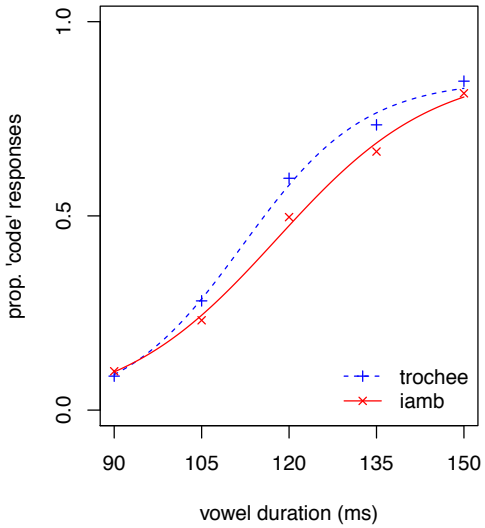
2.4. Results and discussion

The results from the experiment are discussed in reference to the statistical model used in their evaluation. Results are assessed by a mixed-effects logistic regression. The analysis was performed in RStudio [27], using lme4 [2]. The dependent variable in the model is the listeners' response ("code" mapped to 1). Fixed effects are target vowel duration (centered at 0), rhythm, which was effect-coded (IAMB mapped to -1, TROCHEE mapped to 1) and their interaction. Random effects in the model are by-subject intercepts, with maximally specified by-subject random slopes [1]. Table 1 gives the model output. Figure 2 shows listeners' categorization split by rhythm condition.

Table 1: Model output. Estimates are rounded.

	β(SE)	z-value	p-value
intercept	-0.02(0.19)	-0.13	0.90
rhythm	0.16(0.06)	2.86	0.004**
vdur	1.76(0.14)	12.58	< 0.001***
rhythm:vdur	0.07(0.05)	1.32	0.18

Figure 2: Categorization by rhythm condition. Points show the proportion of "code" responses (on the *y* axis) at each continuum step (on the *x* axis). Lines are psychometric curves representing a smoothed categorization trend.



As shown in Figure 2 and Table 1, a significant effect of rhythm was found, whereby the TROCHEE condition shows increased "code" responses relative to the IAMB condition (β(SE) = 0.16(0.06), z = 2.86, p < 0.01). This suggests that preceding rhythmic patterns did indeed modulate listeners' perception of vowel duration. Specifically, when the target was grouped perceptually as being the second syllable in an iamb, an expectation of relative lengthening shifted categorization such that longer vowel durations were required for a "code" response, in comparison to the TROCHEE condition. This effectively increases "code" responses when there is a preceding trochaic context. As emphasized above, this shift in categorization is in contrast to what would be expected on the basis of proximal context.

These results can be looked at in two ways. In terms of recent research on distal effects in rate-dependent speech perception, they can be taken to extend conclusions reached by [5], in showing that the relative timing of preceding distal material plays a role in listeners' perception of durational cues. They further indicate that proximal durational contrast effects [11] are not observed when certain distal contexts are provided. As highlighted by [5], this underscores the necessity of further research into the relative importance of proximal and distal cues, and which predominate under what circumstances. One basic empirical step in furthering these results would be testing how increasing the temporal distance of rhythmic repetitions from the target, or inverting the pattern at different points in a precursor would influence listeners' categorization. Given that [6] has shown that distal speech rate effects persist even after an interval of non-manipulated speech, one might predict that the effects of distal rhythmic patterns may persist over a certain interval with a neutral rhythmic context. Another empirical extension of the present result is to investigate how rhythmic alternations interact with changes in rate of repetition of distal material. Observing the relative importance of each and potential interactions between these factors may be an important step in taking a full account how different properties of distal context influence listeners' perception of temporal cues.

On the other hand, these results can be thought of as an extension of the research showing distal rhythmic/prosodic effects in word segmentation and lexical processing, outlined above. Given that rate-dependent speech perception is typically seen as originating from general auditory processes (e.g. [5]) these results can be taken as showing that rhythmic structure is relevant in what is thought of as low-level speech processing. Following the idea that this perceptual grouping is a general auditory process (as discussed in [21]), the present results can be taken as suggesting such an auditory grouping effect can have important consequences for listeners' uptake of linguistic information, both in the processing of durational cues (in the present study) and in word segmentation [21]. The present results therefore align with the claim that temporal structure in the speech signal is incorporated at multiple levels of processing [5,7], both in what is thought of as more early-stage processing [5] (as in the present results) and processing which is post-lexical [21]. Another possible extension is to investigate how these effects play out online, using a similar eye tracking paradigm as e.g. [26]. More broadly, using more naturalistic stimuli with different intonational and metrical properties is a further step in scaling up the present results. In this vein, testing different prosodic patterns, and looking cross-linguistically, may help inform our understanding of how linguistic information is relevant in rate-dependent speech perception, and how it interacts with domain-general auditory processing, which remains a pertinent question (e.g. [7,23,28]).

3. CONCLUSIONS

In sum, these results are a first step in showing the importance of rhythmic context in listeners' perception of vowel duration, and durational cues more generally. These rhythmic patterns apparently override proximal contrast effects and show that distal rhythmic structure should be further investigated as a factor in listeners' processing of temporal cues.

4. REFERENCES

[1] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3).

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

[3] Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved from http://www.praat.org/

[4] Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics*, *53*(6), 585–600.

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

[6] Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955–967.

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

[8] Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Exp. Psych. Human Perception and Performance*, *41*(2), 306–323.

[9] 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.

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

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

[12] Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Journal of Mem. and Lang.*, *63*(3), 274–294.

[13] Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Mem. and Lang.*, *59*(3), 294–311.

[14] Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. Frontiers in Psychology, 3, 238.

[15] Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. Perception & Psychophysics, 43(2), 137–146.

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

[17] Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, *128*(2), 839–850.

[18] Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psych. Review*, *83*(5), 323–355.

[19] Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168.

[20] Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512.

[21] Morrill, T. H., Dilley, L. C., & McAuley, J. D. (2014). Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics*, *46*, 68–85.

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

[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. *Attention, Perception, & Psychophysics*, *78*(1), 334–345.

[24] Quené, H. (2008). Multilevel modeling of between-speaker and within- speaker variation in spontaneous speech tempo. The Journal of the Acoustical Society of America, 123(2), 1104–1113.

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

[26] Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

[27] RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL http://www.rstudio.com/.

[28] 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), 060001.

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

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