Distal rhythmic patterns influence listeners' processing of phonetic cues

Jeremy Steffman (University of California, Los Angeles)
jsteffman@ucla.edu

A body of literature shows that distal rhythmic/prosodic patterns play an important role in word segmentation [1-3], and online processing of speech [4]. Listeners are argued to project a prosodic structure in perceiving a speech signal based on the rhythmic properties of a distal preceding context, grouping ambiguous strings of sounds into words based on patterns of alternating duration and pitch e.g. [1]. For example, listeners parse an ambiguous string as "crisis # turnip" instead of "cry # sister # nip", if preceding alternations in duration/pitch suggest the first two syllables form a unit (i.e., a foot) [2]. These findings are couched within the *perceptual grouping hypothesis*, which predicts that perceptual grouping of alternating patterns will affect listeners' expectations about incoming information in the speech signal (motivated by findings in non-speech auditory perception e.g. [5,6]).

In light of this research, the present study investigates how distal rhythmic context may influence rate-dependent speech perception, testing perception of vowel duration as a cue to coda stop voicing (a robust durational/rate-dependent cue in English [7,8]). Predictions based on rhythmic structure are contrasted with predictions based on proximal durational contrast effects e.g. [9,10] (outlined below), building on recent speech perception research which has investigated the importance of distal versus proximal cues in rate-dependent speech perception [7,11,12]. This study can thus be seen as a novel investigation of distal context effects, extending research that shows the importance of rhythmic structure in lexical processing to test its relevance for the processing of phonetic cues.

Participants (n=30) categorized a continuum (2AFC task) that varied only in vowel duration (90-150ms; 15ms steps) as "coat" or "code" (PSOLA resynthesis [13]). This target was preceded by an alternating sequence of syllables (of simple CV shape: /tɑ/) that formed a series of durational trochees (long-short), or iambs (short-long). Preceding syllables were resynthesized so that a short syllable had a vowel duration of 75ms, while a long syllable had a duration of 150ms. Three trochaic or iambic feet preceded a final foot in which the target was grouped with either a long syllable, forming a potential trochee (in the TROCHEE condition), or a short syllable, forming a potential iamb (in the IAMB condition; see Fig. 1). Following the perceptual grouping hypothesis, if listeners group the target as the second syllable in a foot based on preceding durational alternations, expectations about the duration of the target vowel might change based on whether it was the implied second syllable in trochee (where it would be shorter), or an iamb (where it would be longer). This predicts categorization would shift such that "code" responses *decrease* in the *IAMB condition* where longer vowel durations are expected (relative to the TROCHEE condition), reflecting the influence of distal rhythmic patterns. Crucially, proximal durational contrast effects predict the *opposite* shift in categorization. Given that a relatively *longer* syllable precedes the target in the TROCHEE condition (see Fig. 1), categorization would be expected to shift to *longer* required vowel durations for a "code" response (*decreasing* "code" responses in the *TROCHEE condition*). As these two influences make opposite predictions about the directionality of the effect, this design directly tests whether proximal duration, or distal rhythmic structure will influence categorization.

Results were assessed by mixed-effects logistic regression (random effects: by-subject intercepts with maximal random slopes [14]). As shown in Fig. 2, the rhythmic properties of the preceding target significantly influenced categorization, whereby the IAMB condition shows significantly decreased "code" responses ($\beta(SE)$ = 0.17(0.06), $z = 2.70$, $p < 0.01$). This suggests that listener *expectations* about duration based on distal rhythmic structure influenced the perception of vowel duration as a cue to voicing, in defiance of proximal durational contrast effects. These results thus provide novel insight into how distal rhythmic/prosodic context affects listeners' processing of durational phonetic cues, extending the research that documents their importance in lexical processing [1-4], and building on the demonstrated importance of distal context in rate-dependent speech perception [11,12]. More broadly, these results indicate that rhythmic structure is relevant both in the fine-grained perception of phonetic detail as well as word segmentation and lexical processing, consistent with the proposal that effects of temporal structure in the speech signal operate at multiple processing levels [11,15]. Results are further discussed in terms of their extension to unified segmentation/segmental perception experiments e.g. [7], their expansion to more naturalistic stimuli, and their implications for recent issues in the speech perception literature regarding proximal and distal speech rate effects e.g. [11,16].

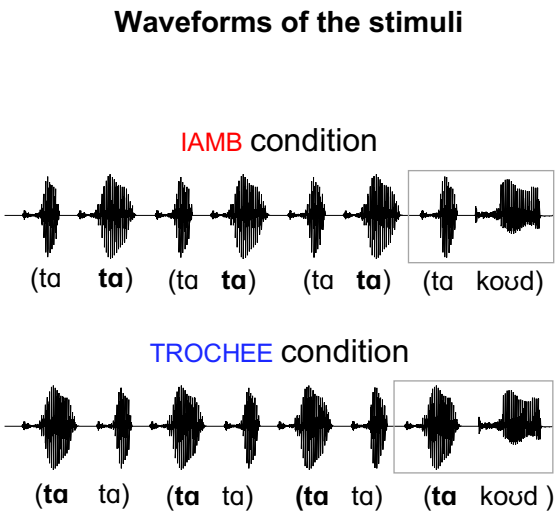

Figure 1: Waveforms are transcribed and bracketed according to listeners' hypothesized grouping. The longer precursor syllable is bolded. The target has 150ms vowel duration. The gray box highlights proximal context.

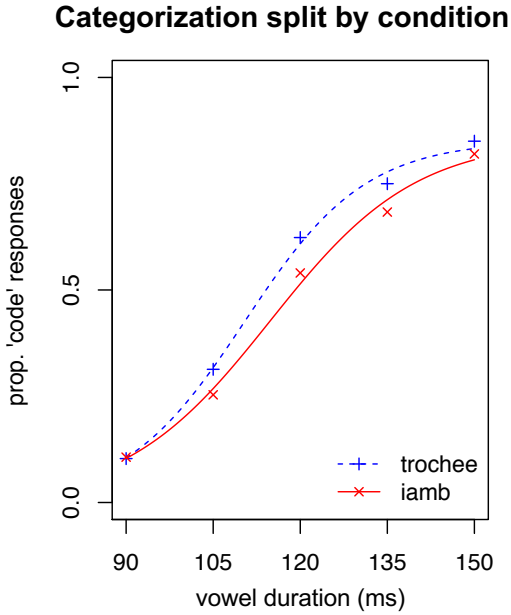

Figure 2: *x* axis shows vowel duration steps from the continuum. Points show the proportion of "code" responses (on the *y* axis) at each continuum step, in each condition. Lines are Psychometric curves, fit to show a smoothed categorization trend.

References: **[1]** 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. **[2]** 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. **[3]** 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. **[4]** 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. **[5]** Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception & Psychophysics*, *53*(6), 585–600. **[6]** Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psych. Review*, *83*(5), 323–355. **[7]** 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. **[8]** 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. **[9]** 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. **[10]** Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560. **[11]** Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, *79*(1), 333–343. **[12]** Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Lang., Cog. and Neuroscience*, *33*(8), 955–967. **[13]** Moulines, E., & Charpentier, F. (1990). Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. *Speech Commun.*, *9*(5-6), 453–467. **[14]** 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). **[15]** Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception, & Psychophysics*, *78*(4), 1203–1217. **[16]** 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.