Speaking rate normalization with and without segregation of simultaneous context sentences

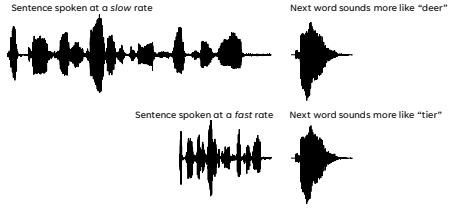
Dawson Stephens, Caleb J. King, Anya E. Shorey, & Christian E. Stilp

Department of Psychological and Brain Sciences, University of Louisville

Introduction

Speech perception, like all perception, occurs in context. Acoustic properties of earlier sounds influence perception of later sounds. This results in **acoustic context effects**, where acoustic differences between sounds are perceptually magnified.

Here we are studying speaking rate normalization, aka **temporal contrast effects (TCEs):**



These effects have a strong bottom-up component tied to differences in duration / speaking rate. Can they be altered by top-down attention?

- Bosker, Sjerps, & Reinisch (2020) claim no. Attending to one of two simultaneous talkers did not alter the sizes of TCEs.
- However, these talkers were easy to segregate because they were different people and presented dichotically (one to each ear).

Here, we eliminated talker variability and varied spatial cues for a stricter test of whether attention can alter TCEs.

Method

Participants

20 (Expt. 1) and 22 (Expt. 2) undergraduate native English speakers with self-reported normal hearing

Stimuli

Context sentences: TIMIT sentences recorded by the last author. Speaking rates were decreased by 33% (duration x 1.5) or increased by 100% (duration / 2) via PSOLA in Praat

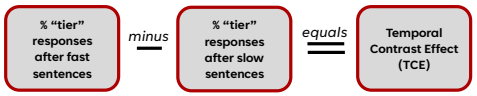
Target words: 10-step series morphing from "deer" to "tier" (Winn, 2020) spoken by the last author

Procedure

Practice: neutral-rate context sentence plus "deer"-"tier" endpoints; 80% correct required to advance

Test: 4 blocks of slow/fast context sentence(s) before "deer"-"tier" target word. Both experiments shared these blocks:

- Sentence 1 alone
- Sentence 2 alone
- Sentences 1 & 2 at the same rate presented diotically
- Sentences 1 & 2 at the same rate presented dichotically

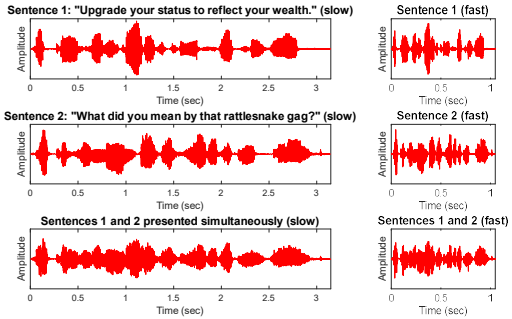


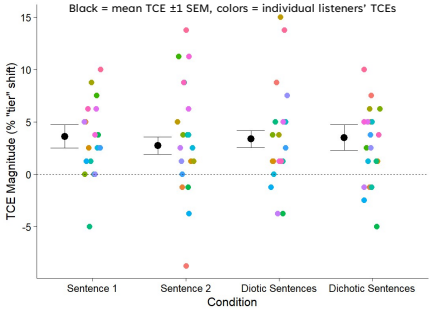
Results

Speaking rate normalization was similar when hearing one context sentence, two sentences diotically, or two sentences dichotically on each trial.

Experiment 1

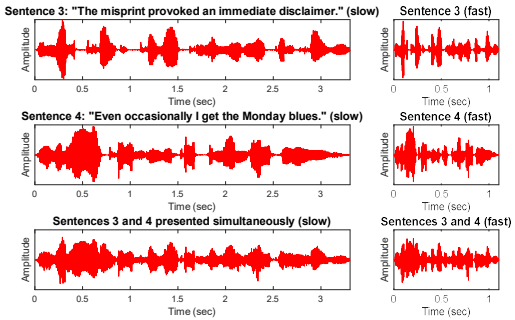
- 10-syllable sentences spoken at 3.18 (slow) or 9.53 (fast) syllables/sec
- Each condition produced a TCE (GLMER: all $Z > 2.10$, $p < .04$), but TCEs did not differ across conditions (all $Z < 1.66$, $p > .09$)

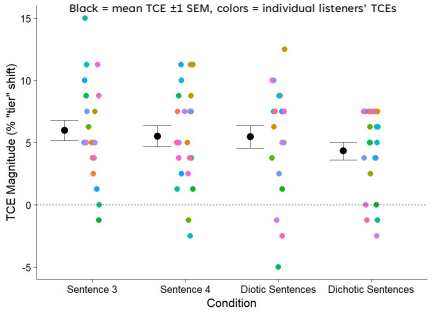




Experiment 2

- 13-syllable sentences spoken at 3.93 (slow) or 11.79 (fast) syllables/sec
- Each condition produced a TCE (GLMER: all $Z > 5.70$ $p < 2e\text{-}8$), but TCEs did not differ across conditions (all $Z < 1.51$, $p > .13$)
- Across experiments: TCEs in each condition of Expt. 2 (grand mean = 5.3%) were larger than those in Expt. 1 (grand mean = 3.3%) (GLMER: all $Z > 2.10$, $p < .04$)

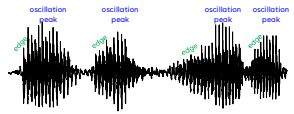




Discussion

Speaking rate normalization (TCEs) was not affected by the (in)ability to segregate simultaneous talkers. Our results are consistent with Bosker et al. (2020): attention does not play an important role in generating TCEs.

What is the mechanism underlying TCEs? The leading candidates are:



1. *Oscillatory entrainment* (Bosker & Ghitza, 2018)
   - Cortical neural entrainment to speaking rates in the theta range (3–9 Hz), but not outside it, drives rate normalization
   - But, TCEs in Expt. 2 were larger, not smaller, when fast rates exceeded the theta range; 9 Hz may not be a hard limit
2. *Acoustic edges* (Oganian & Chang, 2019; Kojima et al., 2021)
   - ECoG and MEG responses in the delta-theta range (1–10 Hz) are better predicted by modulation onsets, not their peaks
   - But, edges and entrainment only make different predictions at slower speaking rates, which were very similar across experiments here

These behavioral data cannot distinguish between these competing accounts, but clever stimulus manipulations in future experiments might be able to.

How do modulation rate and modulation depth in the context sentences contribute to TCEs?

- Combining sentences (in diotic and dichotic conditions) effectively increased the number of syllables per second
- This also decreased their modulation depth
- TCEs did not differ from those produced by single-sentence conditions
- Isolating the contributions of modulation rate and depth to TCEs will be illuminating

Limitation: Bosker et al. (2020) manipulated attention by conducting a concurrent keyword detection task. We did not; this was impractical given the difficulty of the diotic condition when the same talker spoke both sentences simultaneously (Brungart, 2001).

- Given equal TCEs across all conditions, we would likely observe the same results had we used that task in single-sentence and dichotic conditions

References

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Lang. Cogn. Neurosci.*, *33*(8), 955–967.

Bosker, H. R., Reinisch, E., & Sjerps, M. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Sci. Rep.* *10*(1), pp. 1-11.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *JASA*, *109*(3), pp. 1101-1109.

Kojima, K., Oganian, Y., Cai, C., Findlay, A., Chang, E., & Nagarajan, S. (2021). Low-frequency neural tracking of speech amplitude envelope reflects the convolution of evoked responses to acoustic edges, not oscillatory entrainment. *BioRxiv*, 2020-04.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.*, 5(11), eaay6279, 1-13.

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *JASA*, 147(2), 852-866.