**Compensation, classification and inferring sentence structure from acoustic duration**

Sten Knutsen, Karin Stromswold & Dave Kleinschmidt (Rutgers University)
sten.knutsen@rutgers.edu

**Introduction.** Each acoustic feature of the speech stream is affected by many underlying factors: phonetic identity, neighboring segments, talker, prosody, and even (indirectly) syntactic structure. This means each cue is a potentially rich source of information, but extracting information about a particular factor (e.g., syntactic structure) requires *compensating* for the effects of other factors. For instance, Stromswold et al. [1] found that listeners were able to guess whether sentences truncated to be syntactically ambiguous (Fig. 1) were active or passive with 83% accuracy. The only reliable acoustic difference between them was found in the duration of the verb stem vowel, but the distributions of durations for active/passive segments are highly overlapping (e.g., Fig. 2a) due to the influence of other factors (e.g., talker identity and phoneme). Here, we develop a fully Bayesian model that can compensate for sources of variation like talker and phoneme — even when the influence of those factors is not known and must be inferred — while simultaneously inferring (classifying) syntactic structure. This sets it apart from previous approaches to compensation which treat compensation and classification as two separate computational stages [2].

**Data.** We analyzed voice recordings of 8 native English speakers from [3]. Each spoke 28 temporarily ambiguous active/passive sentence pairs (Fig. 1) differing only in the choice of verb stem and agent/patient. All sentence pairs were syntactically ambiguous up until the verbal inflection. There were a total of 439 spoken sentences (9 removed due to speaker error).

**Model and procedure.** While our *ideal compensator* model knows the identity of the phoneme and talker that produced each token's duration, it still must infer *how* to compensate for variation due to talker and phoneme. The model was implemented in Stan software using the 8 speakers' data. Using Stan samples, we modeled the accumulation of evidence over segments as the cumulative posterior probability, calculated for each sentence as the mean cumulative sum of log-likelihood ratios over segments (Fig. 2d). To model listener behavior in a gating task, we thresholded our cumulative posteriors at the end of the verb stem to obtain an accuracy score.

**Results.** Focusing first on the verb stem vowel segment, a visual inspection of Fig. 2a reveals the extent to which the model reduced variance in active/passive distributions. The effect of compensation is also reflected in the classified posteriors of Fig. 2b-c. Our compensated model results for syntactic voice are visualized in Fig. 2d where each blue line shows the mean trajectory of evidence accumulation for one of the 439 sentences.The mean cumulative probability of the true structure over all sentences and samples in the non-/compensated models were 0.56 and 0.63 respectively; modeled overall accuracy scores were 62.9% and 72.7%, respectively.

**Conclusion.** Our results suggest it is possible to create a fully Bayesian model of speech perception that compensates and classifies linguistic stimuli simultaneously and dynamically. Although overall accuracy resembles accuracy scores from previously obtained behavioral studies, future work will include gathering behavioral data to assess detailed predictions of this model for listener behavior. We also plan on scaling up the model to situations where *multiple* underlying factors are unknown and must be inferred (e.g., the vowel is not known but must be inferred based on duration and formant frequencies).

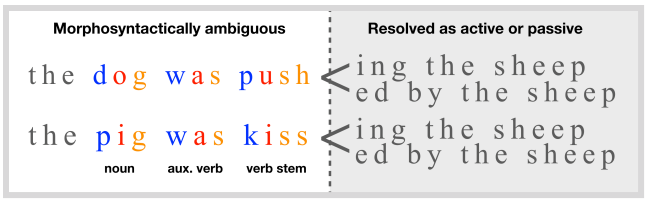



Figure 1 (left). Sentence pair examples. For each truncated sentence, we coded 9 segments leading up to the verbal inflection. These segments aligned across all sentences.

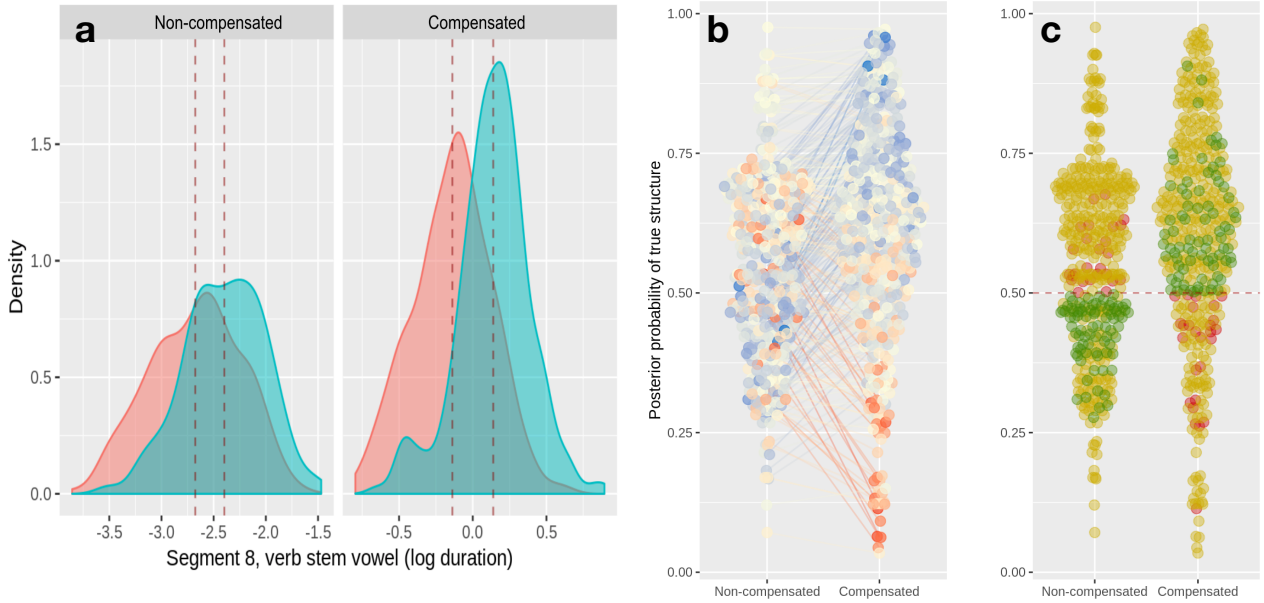
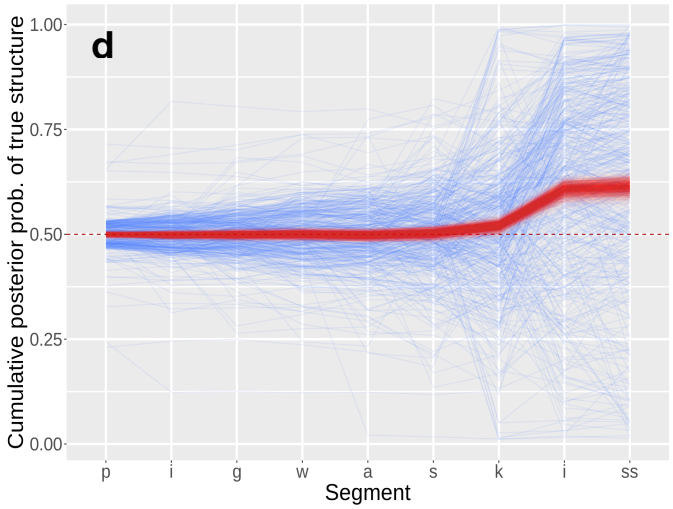



Figure 2a. Non-/compensated log duration active/passive distributions from Stan samples for verb stem vowel.
Fig. 2b. Non-/compensated verb stem vowel posteriors increase/decrease in posterior probability. Of 439 total tokens, 290 increased with compensation. Fig. 2c. With compensation, 81 posteriors flip from false to true prediction (over p=0.5 threshold/dashed line), 26 flip from true to false and 332 show no change. Fig. 2d. Each red line shows the mean cumulative probability by sample over all sentences.




[1] Stromswold, K., Kharkwal, G., & Eisenband Sorkin, J. (in review). Tracking the Elusive Passive: The Processing of Spoken Passives
[2] McMurray, B., & Jongman, A. (2011). What Information Is Necessary for Speech Categorization? Harnessing Variability in the Speech Signal by Integrating Cues Computed Relative to Expectations. *Psychological Review* 118 (2): 219–46.
[3] Stromswold, K., Lai, M., Rehrig, G., & de Lacy, P. (2016). Passive sentences can be predicted by adults. *The 29th Annual CUNY Conference on Human Sentence Processing*. Gainesville, FL.