Eye movements reveal real-time noisy-channel inference in spoken word processing

Rachel Ryskin (University of California, Merced), rryskin@ucmerced.edu

The goal of communication is to transmit information from a producer to a comprehender, yet comprehenders often arrive at an interpretation that differs from the literal meaning of linguistic input [1]. According to the Noisy Channel theory, comprehenders infer the speaker's intended meaning from what they perceive, combining the prior probability of intended meanings with a model of likely errors [2,3,4]. However, given its potential computational cost, whether this rational inference process unfolds in real time remains an open question. The present work uses eye-tracking to demonstrate that real-time spoken word processing is well-modeled as Bayesian belief updating with a noise model that incorporates the relative probabilities of different speech or listening errors.

Methods: *Eye-tracking.* Eye-gaze of 60 participants was monitored in a visual world paradigm (VWP) study modeled on [5]. On each trial, participants heard an instruction ("Click on the <referent>") while viewing a 4-picture display (Fig 1). Condition (No competition vs. Cohort competition vs. Rhyme competition) was manipulated within-subjects and items (see Table 1). Each participant experienced all 24 item sets in 6 conditions (total = 144 trials). The order of trials and the location of images on each trial were randomly generated for each participant. We analyze the target advantage (TA) which is computed by subtracting the average proportion of fixations to a competitor (cohort, rhyme, or distractor) from the average proportion of fixations to the target within a time window of interest. *Noisy-channel model.* The process of inferring the correct referent given the auditory input can be modeled as Bayesian belief updating [6,7,8]. Adapted to the VWP, the proportion of fixations to an ROI can be viewed as a reflection of how much probability the listener is assigning to that referent, r, given the spoken input, s, they hear, $P(r|s)$. According to Bayes' rule, $P(r|s) \propto P(s|r) \cdot P(r)$, the posterior probability is proportional to the product of the prior probability that a particular image would be the referent in the first place, $P(r)$, and the probability of the particular auditory signal being produced by the speaker given the intended referent, $P(s|r)$. Crucially, this Bayesian inference can occur incrementally as the utterance unfolds, with the posteriors from each timestep being used as the priors for the next timestep. For the current purposes, the continuous acoustic input is discretized into 4 timesteps: t0 = before start of word, t1 = beginning of word shared with cohort (e.g., "bea"), t2 = disambiguation point between target and cohort (e.g., "k"), t3 = end of word shared with rhyme (e.g., "er"). Given that all 4 referents in a given trial are equally likely to be named *a priori*, $P(r) = 0.25$ for all referents at t0. The noise likelihood, $P(s|r)$, is estimated using the pairwise similarities of audio files of referents extracted from a Transformer model trained on speech data (HuBERT [9]) — a model that has learned the statistics of human speech should capture the probabilities of noise corruptions that speakers may introduce. For instance, $P(s =$ "bea"$|r =$ speaker$)$ is the cosine similarity between the embeddings for "beaker" and "speaker" within a time window determined by the onset and offset of "bea."

Results: Replicating [5], in Cohort competition, TA initially rises more slowly compared to No competition because listeners look to both the target and the cohort equally before the disambiguating point (e.g., "k" in "beaker"; Fig. 2-left). In Rhyme competition, TA initially rises as rapidly as in No competition, but then slows toward the end of the word because listeners look to the rhyme more than an unrelated distractor after disambiguation (Fig. 2-right). Noisy-channel model TA predictions averaging over all possible stimuli are summarized in Fig. 3. They qualitatively recover the TA patterns in eye-tracking data: a delayed rise for Cohort competition and a slowdown toward the end of the auditory stimulus for Rhyme competition. Moreover, stimulus/timepoint-level model predictions explain substantial variance ($R^2 = 0.63$) in eye-gaze patterns (Fig. 4).

Conclusions: The timecourse of spoken word processing is well-captured as incremental noisy-channel inference, suggesting that noisy-channel inference is a continuous, online computation at the core of language comprehension.

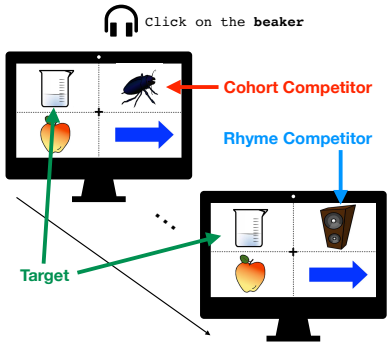

Figure 1: Example VWP displays.

Condition	Instruction *Click on the...*	ROI 1	ROI 2	ROI 3	ROI 4
No compet.	*beaker*	beaker	apple	arrow	clock
Cohort compet.	*beaker*	beaker	beetle	apple	arrow
Rhyme compet.	*beaker*	beaker	speaker	apple	arrow
Filler 1	*beetle*	beetle	beaker	arrow	clock
Filler 2	*speaker*	speaker	beaker	apple	arrow
Filler 3	*apple*	apple	beaker	arrow	clock

Table 1: Example materials from eye-tracking experiment across conditions (1 item set out of 24).

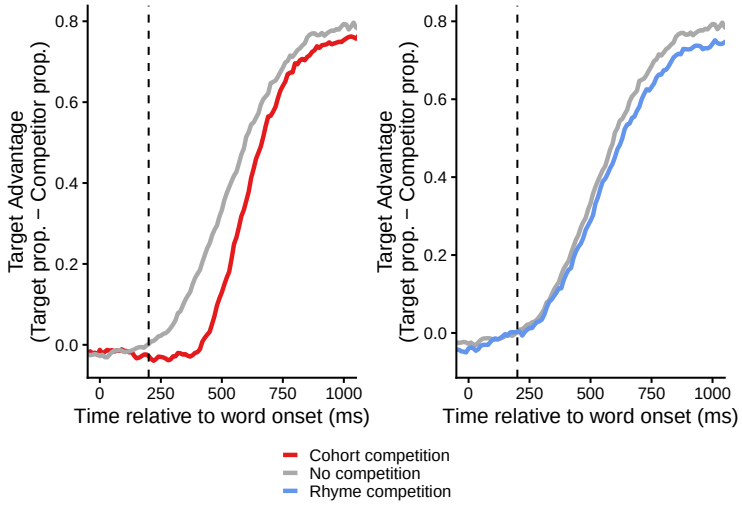

Figure 2: Timecourse of target advantage relative to word (e.g., "beaker").

References: **1.** Ferreira & Patson (2007). *Lang. & Ling. Compass*, 1. 71-83. **2.** Shannon, C. E. (1948). *Bell System Technical Journal, 27(3)*, 379–423. **3.** Levy, R. (2008). *Proceedings of EMNLP*, 234–243. **4.** Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). *PNAS*, 110(20), 8051–8056. **5.** Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). *JML, 38*, 419-439. **6.** Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). *Cognition, 108*, 804–809. **7.** Kleinschmidt, D. F., & Jaeger, T. F. (2015). *Psychological Review, 122*, 148–203. **8.** Norris, D., & McQueen, J. M. (2008). *Psychological Review, 115*, 357–395. **9.** Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). *arXiv*.

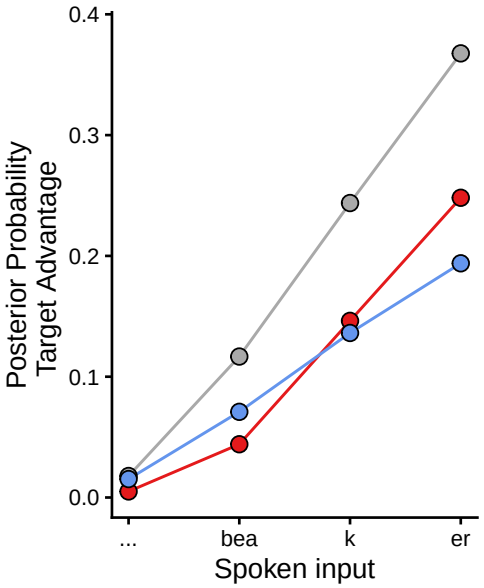

Figure 3: Posterior predictions from the incremental Noisy Channel model aggregated across all stimuli present in the eye-tracking experiment in 4 timewindows.

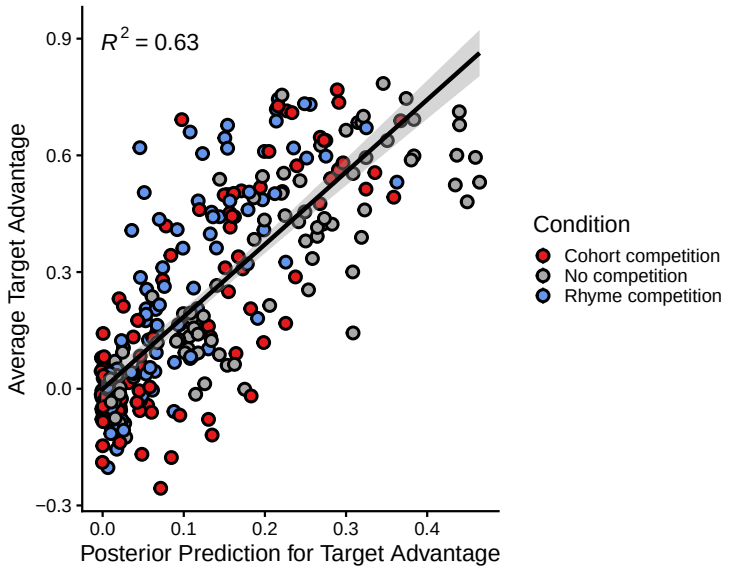

Figure 4: Model fit to eye-tracking data. Average target advantage in fixations for each item by condition in each of 4 timewindows over target advantage predicted by the model posterior.