**Manuscript Outline**

**Intro**

We set out to establish a computational framework that can dynamically extract multiple levels of information from ongoing speech. This is a nontrivial task (the “now or never bottleneck”), and existing models for speech understanding focus on either the low-level recognition from ongoing speech to lexicon, or the high-level linguistic operations with little temporal constraint.

Speech perception is unlikely a purely bottom-up process. Humans apply their internal knowledge during perception: high-level knowledge such as the context of the communication, the knowledge about the speaker, and personal preferences, can lead to different understandings of the same sentence. The application of internal knowledge, however, does not necessarily entail constantly trying to predict the identity of next input item, as is implied in the design of many neurophysiological/modeling studies. The goal for speech perception is to optimize the estimation of information conveyed in a piece of signal (and conduct appropriate behavior accordingly), and the optimization of prediction is a possible computational mechanism in the realization of such a goal. Our question, then, becomes: what computational paradigm might be used by humans that applies (top-down) internal knowledge in extracting (bottom-up) high-level information from speech signal in real time? How can we identify different computational processes from neurophysiological response? And with a computational model that does not use next-input prediction as the goal for optimization, can we gain insight about the role of top-down prediction during speech perception?

Based on the above principles of integrating internal knowledge to optimize information extraction, we model speech perception as the inversion of a generative model from abstract context to continuous acoustic representation, where top-down and bottom-up messages are multiplexed between temporally organized hierarchies (Figure 1). Such perceptual inference process can be characterized by two information theoretic (IT) metrics that reflect the model’s internal computation: 1) KL divergence characterizes the information carried in an (instantaneous) update between levels, and 2) entropy characterizes the information maintained between two updates within the level.

**Results**

By simulating the model with simple sentences, **we first check the model’s “understanding” of high-level information of sentences**: does it assign reasonable values to semantic roles and determines a plausible context? How does it resolve the situation when the speech input is semantically ambiguous (Figure 2)? Simulations in Figure 2 show that the model’s estimation of semantic and context states is dependent on high-level (context) preference, which influences the magnitude of belief update (divergence) and information maintained (entropy). **We then seek how this model may guide our understanding of neural response during speech perception**. We exploit the above-mentioned context dependency and contrast the model response in the form of IT metrics between two pairs of sentences used in an MEG experiment (Figure 3). We show that what’s commonly described as “ambiguity”, “surprisal”, and “disambiguation” in neurophysiological experiments may imply different combinations of computational motifs within and across hierarchies, and model-derived IT metrics may help unravel neural processings underlying signature sensor-space responses. **Last, we explore the role of top-down prediction** using the two IT metrics (Figure 4 and 5). We show that the influence of informative vs. noninformative top-down prediction is most pronounced at the levels of syllable and lemma recognition, where both strategies are possible and the model demonstrates a tradeoff between accuracy and processing effort under unfavorable scenario, consistent with behavioral findings.

**Figure 1:** construction of the model under the following main ideas:

1. The goal of speech perception is to extract high-level information (semantic roles and context)
2. Applying hierarchical internal knowledge to understanding ongoing speech can be realized by organized top-down and bottom-up message passing

**Figure 2**: high-level state estimation results are consistent with behavior observations, in that:

1. semantic ambiguity of the word “ace” is reflected in the state estimation of both levels
2. the model can resolve earlier semantic ambiguity by applying contextual knowledge upon receiving a disambiguating input “the tennis”
3. if the model doesn’t receive any disambiguating input, it relies on its prior knowledge in interpreting the ambiguous speech

**Figure 3**: semantic uncertainty and surprisal have been used to characterize neural processing of speech, and neural signatures are often defined by contrasting sentences with different levels of uncertainty/surprisal. However, the computational functions of these two classes of signatures are often entangled and not clearly discussed. We contrast model-derived metrics between sentences to show that the observed neural signatures may be the combined difference in the amount of effort in updating state estimations across hierarchies (KL divergence) and in maintaining states within hierarchies (entropy), which likely imply different types of neural activities thus summed in a nonlinear fashion (and with different temporal dynamics and electrophysiological signatures)

**Figure 4**: the model utilizes its generative model by alternating top-down and bottom-up message passing. However, the top-down information does not necessarily call for informative prediction for the next input item. Under favorable conditions (e.g. perfect sensory system and noiseless input), accurate perception can be achieved without informative top-down prediction, but may require extra effort in terms of magnitude of update signal (divergence), processing speed and short-term memory at low levels (entropy), compared to perception with informative prediction.

**Figure 5**: the model commits reasonable mistakes when the periphery impaired. When precisions in the continuous level are lowered, the model demonstrates a tradeoff between processing effort and accuracy: explicit prediction saves effort in terms of entropy, but can result in wrong perception favoring the model’s preference.

Conclusions from Figure 4 and 5: informative and non-informative top-down predictions are both possible during speech perception; relying on top-down prediction reduces processing effort, but can also lead to mistake due to overconfidence.

**Discussion**

Based on the model construction and the simulation result, we discuss the following topics:

**Language comprehension as semantic role assignment.** Our model fulfills comprehension “on the fly” from continuous speech signal, which has not been implemented in existing models of language (speech) processing.

**Understanding neural information transfer through divergence and entropy.** The using of “word surprisal” for the functional interpretation of neurophysiological signal is rooted in Levy’s theoretical work (Levy 2008) but may have limitations in characterizing the information transfer during speech processing. We propose to combine entropy and divergence to better understand the functional roles of neurophysiological responses.

**Insights for biophysical implementation by integrating neural oscillations.** Some possible way to relate entropy and divergence to neurophysiological observations.

**Limitations of the model and further development.** We point out two major limitations of the model and suggest possible (but nontrivial) methods to lift these limitations.