The Role of Surprisal in Perceptual Chunking of Spontaneous Speech

Svetlana Vetchinnikova (University of Helsinki)
svetlana.vetchinnikova@helsinki.fi

Recent studies in cognitive neuroscience suggest that when processing continuous stimuli such as speech, humans rely on periodic neural oscillations across different frequency bands (Giraud & Poeppel, 2012). If this hypothesis holds, the rhythmicity of oscillations imposes temporal constraints on the structure of speech. Specifically, theta-band oscillations are thought to align with the duration of syllables, which tend to be relatively stable both within and across languages (Ding et al., 2017; Varnet et al., 2017). Meanwhile, delta-band oscillations appear to correspond to syntactic phrases (Ding et al., 2016; Kaufeld et al., 2020) and/or intonation units (Inbar et al., 2020). Given the primacy of cognitive constraints, it is plausible that both syntax and prosody have evolved as adaptive mechanisms, facilitating the segmentation of speech into perceptually manageable units for both speakers and listeners. However, what role does statistical information play in this process, given its recognized importance in language processing?

In an earlier study (Vetchinnikova et al. 2023), we selected 97 short extracts from spoken corpora and re-recorded them with a trained speaker to achieve uniform audio quality. We then asked 50 experiment participants to listen to the extracts and intuitively mark chunk boundaries in the accompanying transcripts through a custom-built tablet application. Next, we annotated all spaces between every two words for pause duration, prosodic and syntactic boundary strength, chunk duration, and bigram surprisal. Prosodic boundary strength was estimated automatically using continuous wavelet analysis of fundamental frequency, energy and word duration (Suni et al. 2017). To measure syntactic boundary strength, we marked the start and end of each clause with a bracket and counted the total number of brackets for each space assigning a value of 0.5 to an opening bracket and 1 to a closing bracket. Since pause duration, prosodic and syntactic boundary strength as well as chunk duration were collinear, we built a separate logistic regression model predicting chunk boundary perception for each predictor. All models included random effects for listeners and extracts.

We found that pause duration, prosodic boundary strength, syntactic boundary strength, and temporal duration significantly predicted chunk boundary perception, supporting the influence of the temporal constraint and the role of prosody and syntax in perceptual chunking. In contrast, the effect of bigram surprisal contradicted the hypothesis that perceptual chunks were multi-word units: chunk-final words tended to be less predictable while chunk-initial words tended to be more predictable. We interpreted this finding as evidence of a dissociation between perceptual chunking, or the segmentation of incoming speech into temporal units, and usage-based chunking, or the extraction of statistical regularities from input.

Given the limitations of using bigram surprisal to capture statistical information, in this paper, I use surprisal values derived from the GPT2 model that incorporate the full preceding context of each extract. Preliminary results suggest that full context surprisal does not predict chunk boundary perception. I discuss these results from the perspective of the interplay between statistical and structural information in speech processing.

References

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of Intonation Units form a ~ 1 Hz rhythm. *Scientific Reports*, *10*(1), 15846. https://doi.org/10.1038/s41598-020-72739-4

Kaufeld, G., Bosker, H. R., & Martin, A. E. (2020). Linguistic Structure and Meaning Organize Neural Oscillations into a Content-Specific Hierarchy. *The Journal of Neuroscience*, *40*(49), 9467–9475.

Suni, A., Šimko, J., Aalto, D., & Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, *45*, 123–136. https://doi.org/10.1016/j.csl.2016.11.001

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179

Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2023). Chunking up speech in real time: Linguistic predictors and cognitive constraints. *Language and Cognition*, *15*(3), 453–479. https://doi.org/10.1017/langcog.2023.8