Do speech models use phonological features?

Canaan Breiss
University of Southern California, Los Angeles, California, United States

Jon Gauthier
University of San Francisco, San Francisco, California, United States

Abstract

Distributional and phonetic considerations lead linguists to posit that speakers represent phones as members of overlapping natural classes (ex., labials [p, b, m], nasals [m, n, ŋ]), which can be represented in a feature system ([+labial], or [+nasal]). Choices of values -, +, 0 and features labial, nasal in this system make different predictions about what classes are accessible targets for generalization (Mayer 2020). Building on the hypothesis that LLMs and humans share the objective of resource-rational analysis of their environment (Leider & Griffiths 2019), we assess the canonical correlation between different proposed feature systems and representations of sounds they describe in self-supervised deep learning models trained on speech, HuBERT and Wav2Vec2. We also examine differences in representational similarity of phones implied by these alignments. We find that although differences in canonical correlations between feature systems and model representations are small, they have qualitatively distinct error patterns for novel data.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*