Emergent neural principles of speech encoding in contrastive joint-embedding models

Speech perception requires transforming acoustic signals to extract relevant information, yet the computational principles guiding these transformations remain largely unknown. Recently, unsupervised deep learning models have demonstrated the ability to learn rich speech representations that correlate with neural activity. However, it remains unclear whether this correlation reflects only output similarity or extends to shared principles of sensory information extraction.

To investigate potential shared principles, our study explores contrastive learning within joint-embedding architectures (JEAs). These models learn compressed, causal representations by filtering information irrelevant for predicting future speech signals. We compared speech encoding in JEAs (specifically utilizing Wav2Vec2.0 and the CEBRA framework), with neural processing measured via EEG activity from 26 Italian participants listening to English sentences (USC-Timit dataset). A rhyming task during listening engaged participants in active feature extraction.

First, we validated the perceptual relevance of the JEA representations via neural encoding. We found that neural encoding strength within a key processing window (-0.25 s to +0.8 s lag) positively correlated with participants’ behavioral performance in the rhyming task (r = 0.434, p = 0.026), suggesting JEA features capture behaviorally relevant aspects, without any supervision.

Next, reverse-engineering the speech encoding strategies revealed both the brain and the JEA model prioritize causal information over mere signal variance. Strikingly, EEG encoded low-variance subspaces: the bottom 1% variance subspace (0.0020 ± 0.00019 bits; MEAN±SEM) shared similar information with EEG signals as the top 99% variance subspace (0.0019 ± 0.0002 bits; MEAN±SEM), indicating a focus beyond high-energy components.

Finally, evaluating the model’s objective function at inference time showed the brain encodes new predictive information identified by the model (i.e., information non-redundant with the past, identified using the model's InfoNCE loss). This neurally encoded predictive information significantly explained participants’ behavioral performance (r = 0.430, p = 0.028).

Overall, these findings underline the potential of self-supervised contrastive learning (SSL/CL) for studying fundamental cognitive mechanisms and suggest that the observed alignment between these models and brain activity may stem from shared computational principles of causal and predictive sensory information extraction.