Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech

Aditya Vaidya, Shailee Jain,
Alexander Huth

huthlab

THE UNIVERSITY OF TEXAS AT AUSTIN
PRAESIDIUM
DISCIPLINA
CIVITATIS

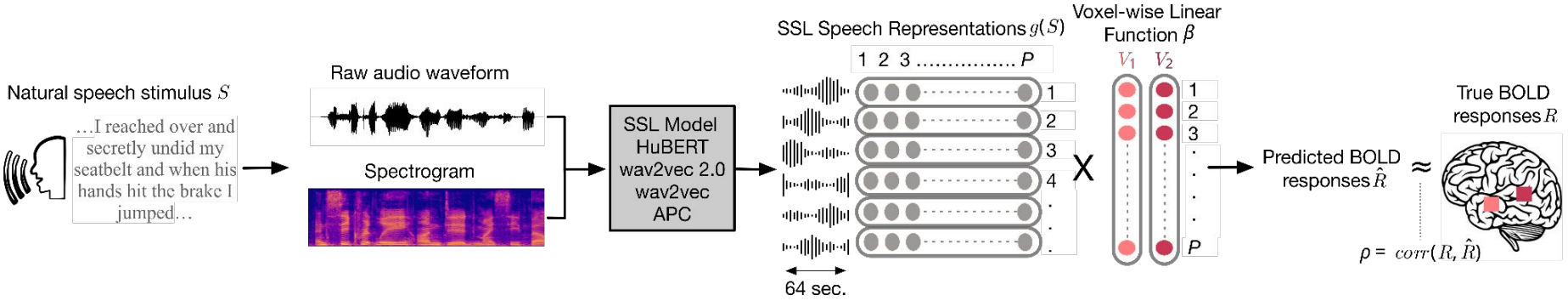
Introduction

- Self-supervised learning has produced powerful representations that capture linguistic structure without labeled data, and are even effective in modeling the brain.

- But, the best models of the auditory system are still either hand-engineered or supervised.

- We bridge the gap between recent speech representation methods and computational models of the human auditory system.

Encoding models



Predict fMRI response (*R*) from a stimulus *(S)* using features from a layer of a self-supervised speech model.

Feature spaces

Self-supervised speech models:

HuBERT

wav2vec 2.0

wav2vec

APC

Baselines:

Spectrotemporally-modulated spectrograms

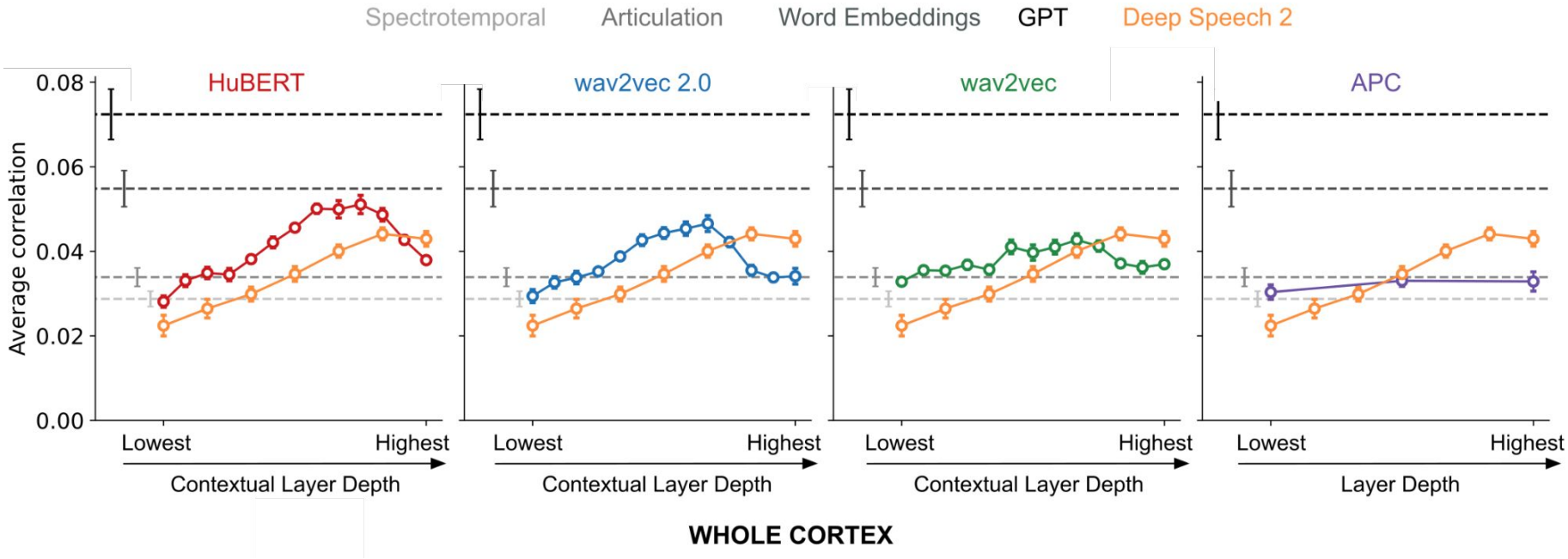
Articulatory features

Word embeddings

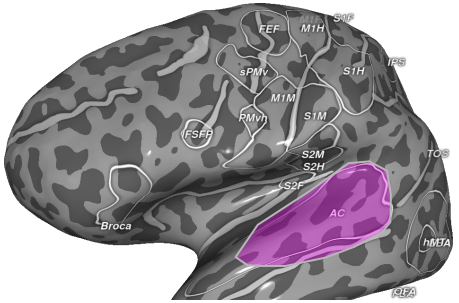
LM: GPT

ASR (supervised): Deep Speech 2

Model encoding performance



Model encoding performance



Spectrotemporal Articulation Word Embeddings GPT Deep Speech 2

HuBERT wav2vec 2.0 wav2vec APC

Average correlation

0.30
0.25
0.20
0.15
0.10
0.05
0.00

Lowest Highest
Contextual Layer Depth

Lowest Highest
Contextual Layer Depth

Lowest Highest
Contextual Layer Depth

Lowest Highest
Layer Depth

AUDITORY CORTEX

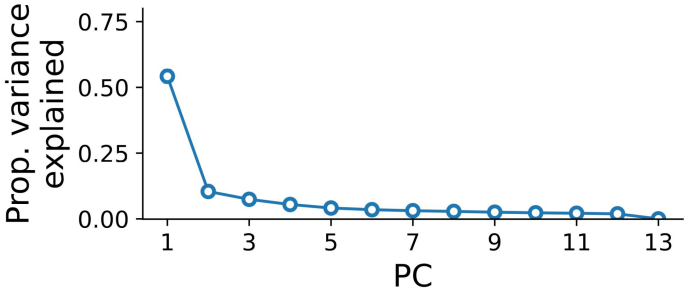
Layer selectivity

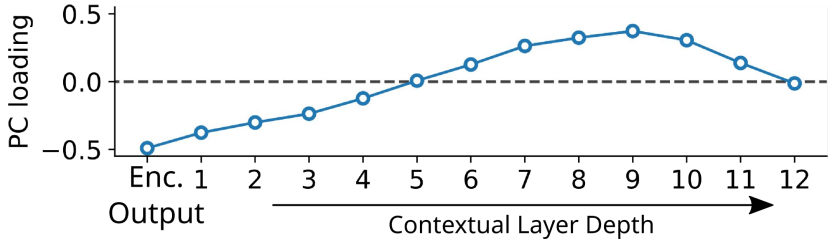
Are different brain areas better predicted by specific layers?

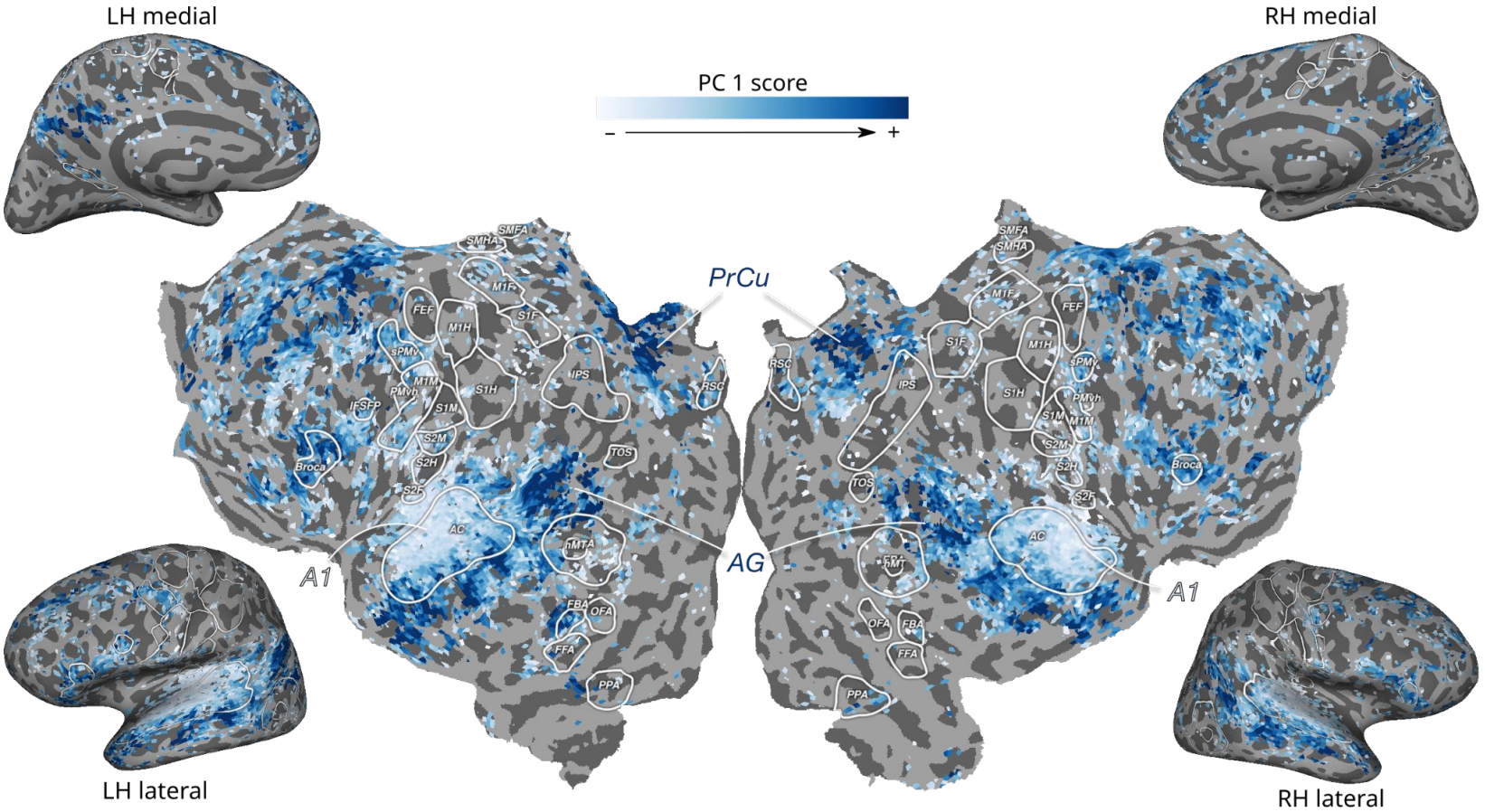
Layer selectivity

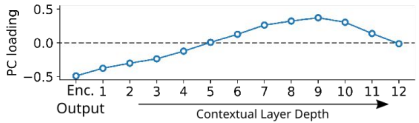
Are different brain areas better predicted by specific layers?

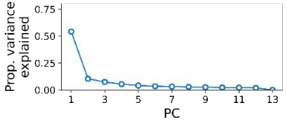
PCA on the performance of each layer & voxel reveals a primary dimension of variance:



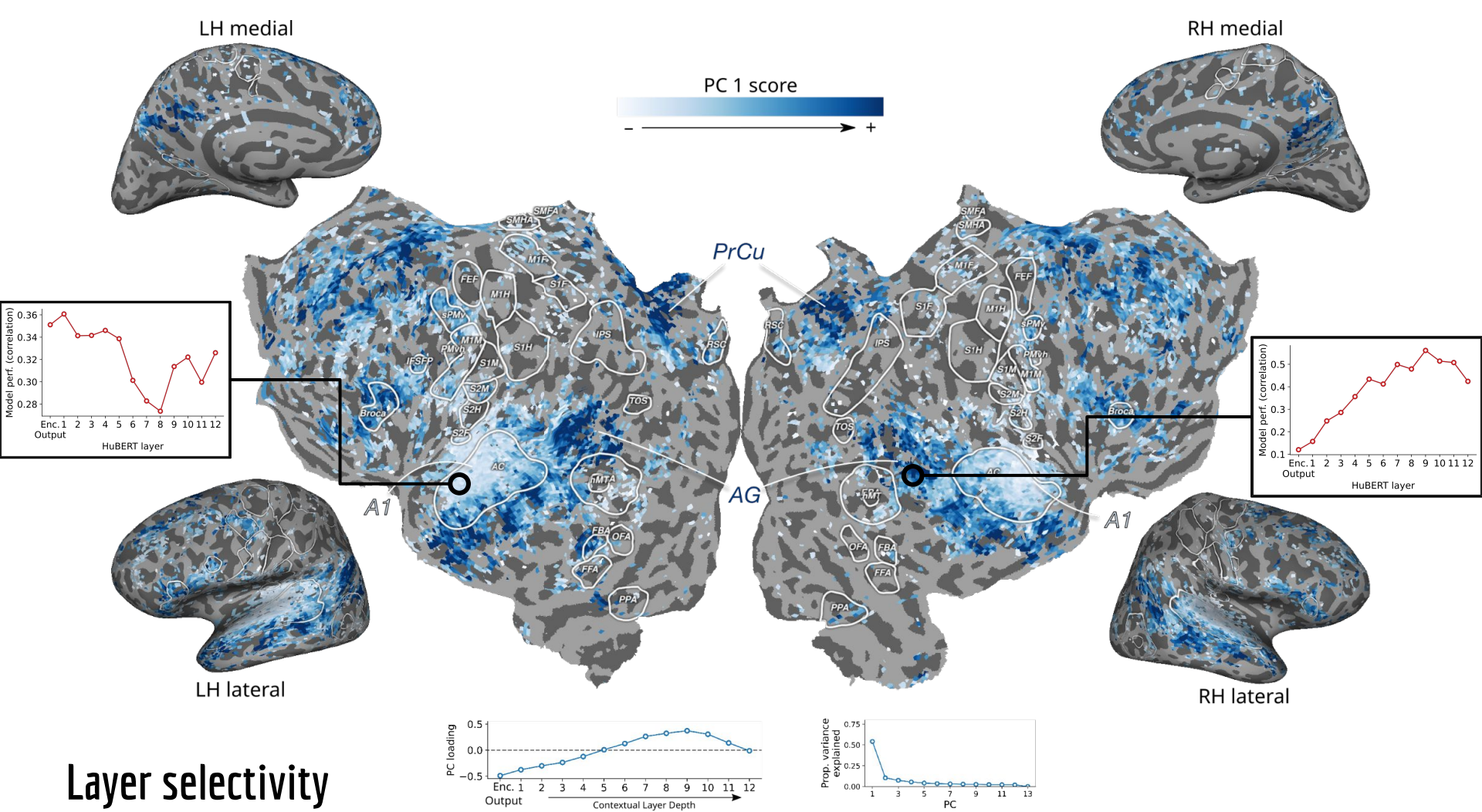








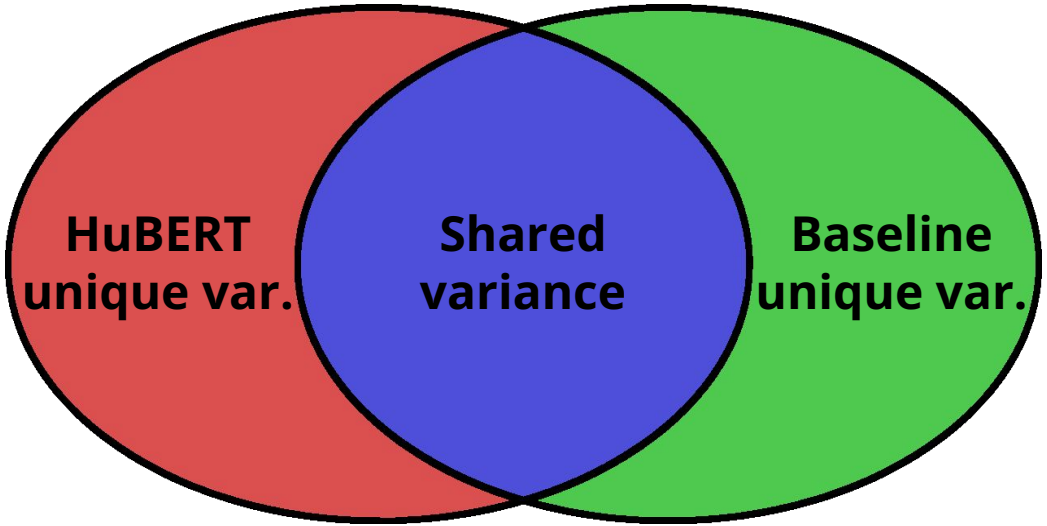
Layer selectivity



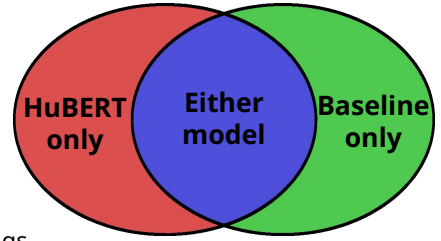
Layer selectivity

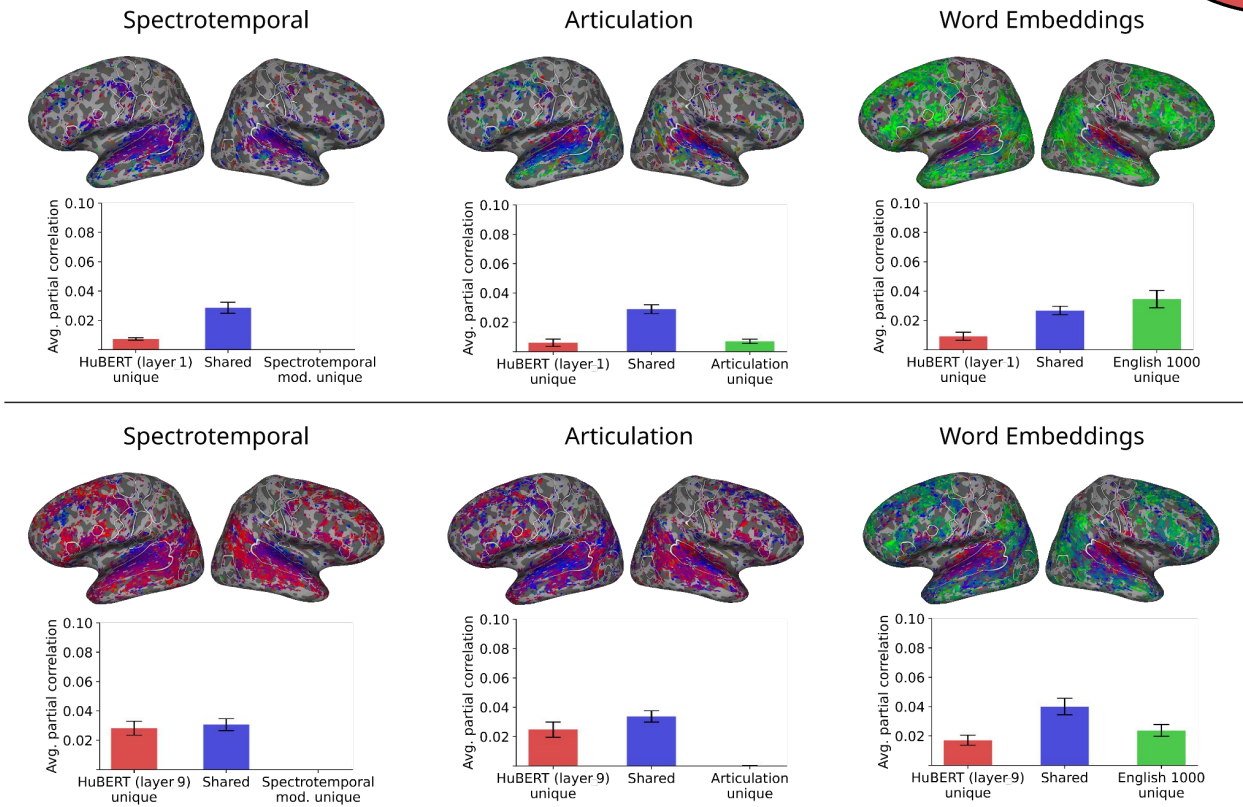
Partitioning explained variance

Do the self-supervised models explain the variance as known linguistic features?



Partitioning explained variance





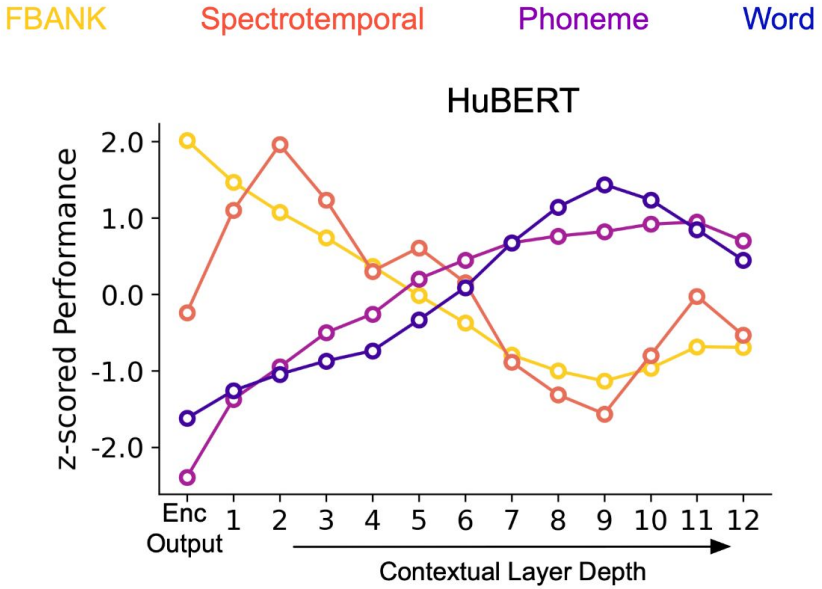
Probing

How do linguistic representations change through the layers of the model?

FBANK Spectrotemporal Phoneme Word

Probing

How do linguistic representations change through the layers of the model?



Conclusions

- Self-supervised speech models are the best models of auditory areas. Supervised tasks are not necessary.

- Lower layers best modeled low-level areas, and upper-middle layers were most predictive of phonetic & semantic areas.

- Layer representations follow the accepted hierarchy of speech processing.

Thank you!

Questions? Let me know!

@_avaidya

avaidya@utexas.edu

huthlab