

Neural Architecture of Speech

June 6, 2023

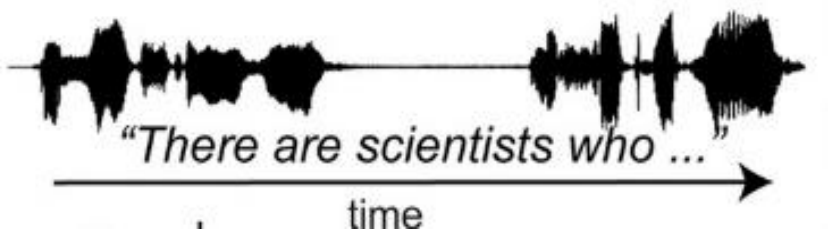
ICASSP 2023

Subba Reddy Oota[1], Khushbu Pahwa[2], Mounika Marreddy[3], Manish Gupta[3,4], Bapi Raju Surampudi[3]

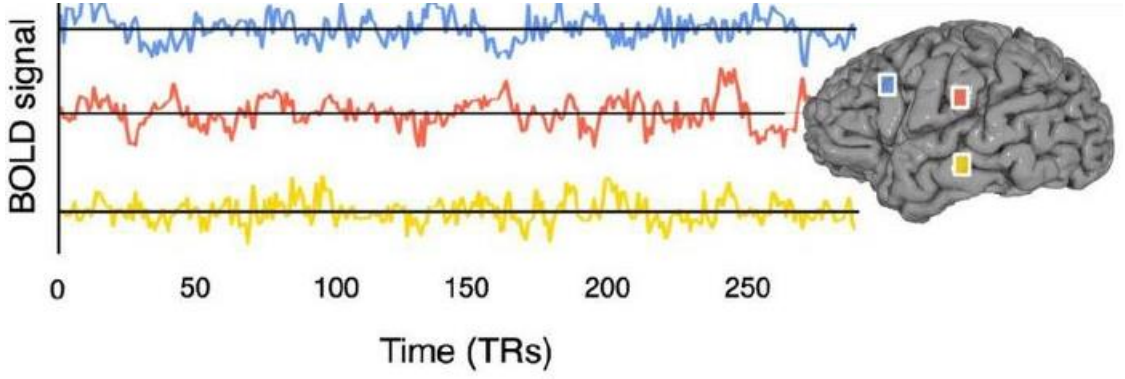
[1]Inria Bordeaux France, [2]University of California LA, [3]IIIT-Hyderabad, [4]Microsoft India

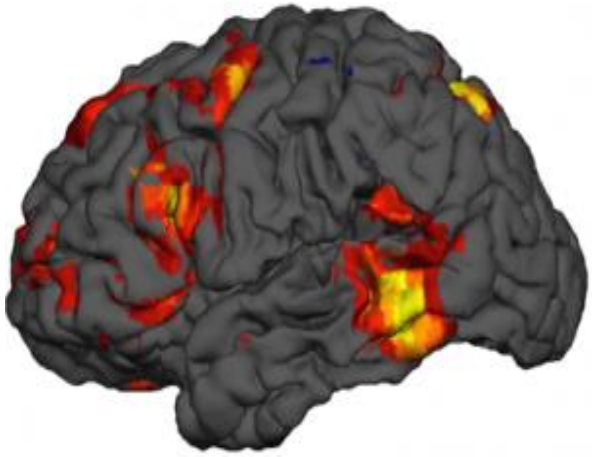
What is fMRI?

Narrative Story



A listening task in the scanner



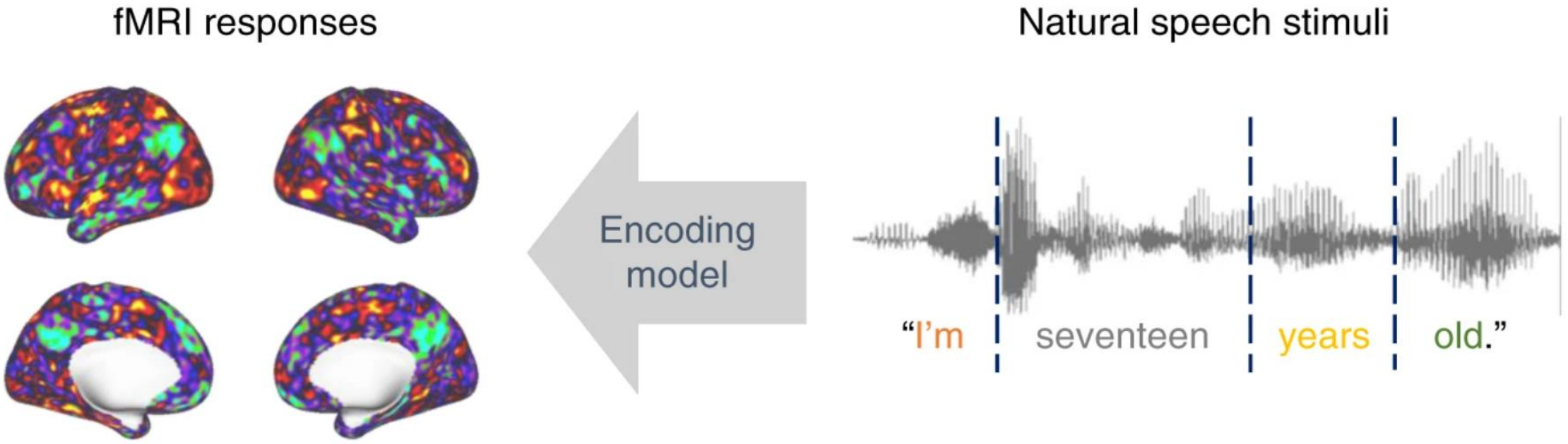


fMRI Brain Activity

https://www.biopac.com/events/fmri-psych/

Brain Encoding vs Decoding

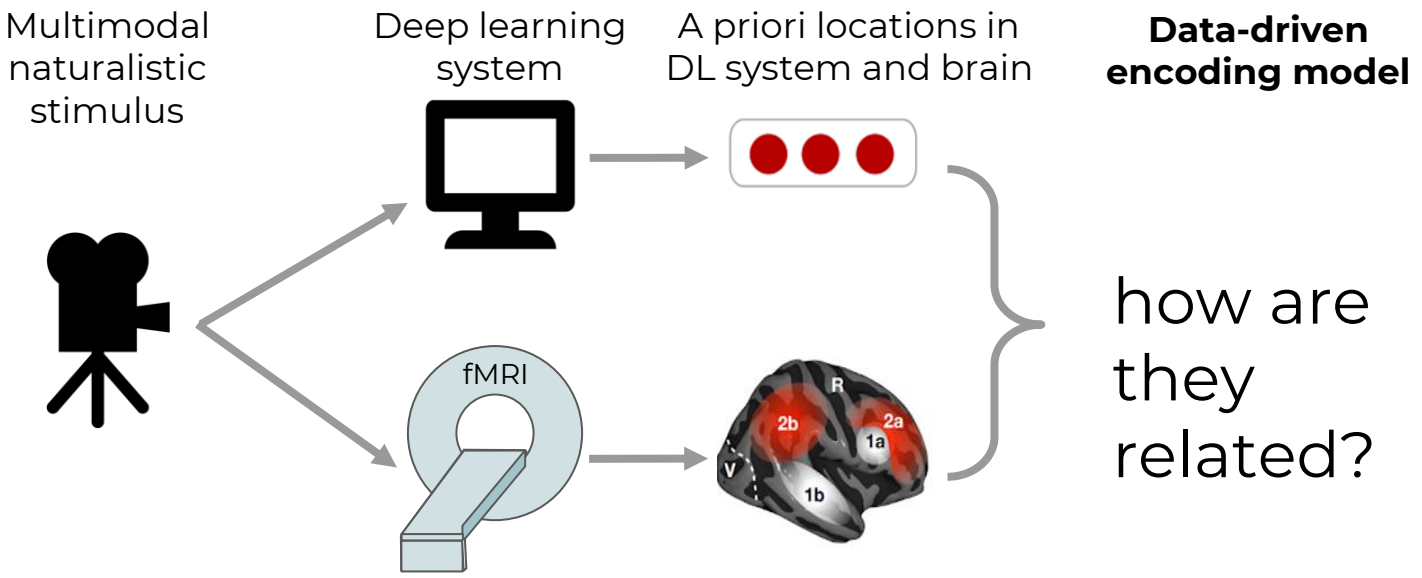
fMRI → Decoding → Stimulus Representation



fMRI ← Encoding ← Stimulus Representation

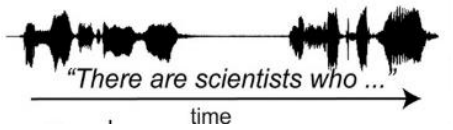
Yizhen Zhang, Kuan Han, Robert Worth & Zhongming Liu. "Connecting concepts in the brain by mapping cortical representations of semantic relations" Nature (2020).

Data-driven encoding models evaluate the relationships between brains and deep learning models



Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). Advances in Neural Information Processing Systems, 32.

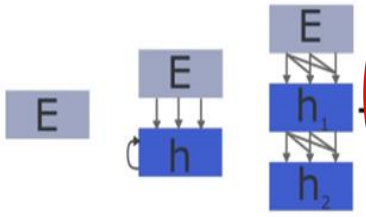
Brain Encoding?



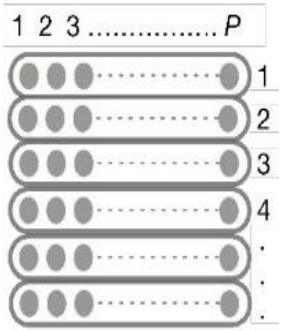
Stimulus

input

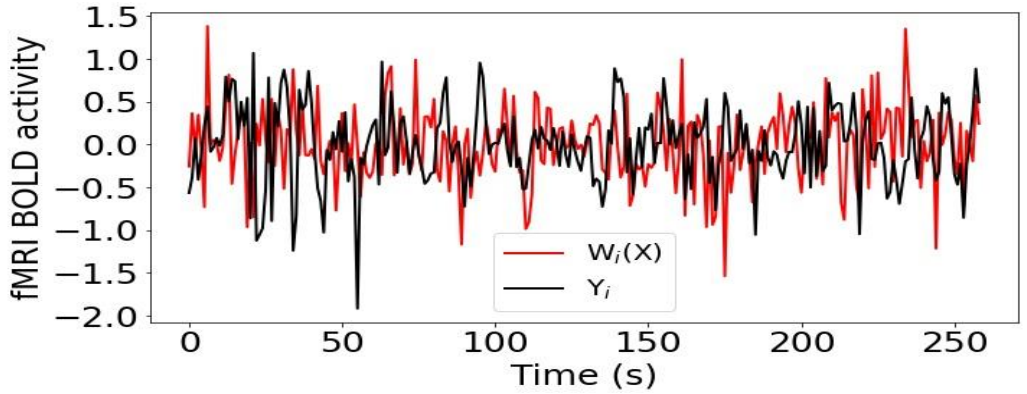
Models



Internal Neural Representation
Hidden unit activations



Pearson Correlation (R) = Corr(Y, W(X))

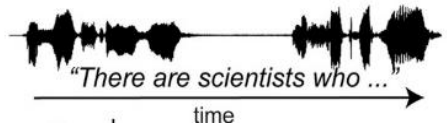


X Input

W Ridge Regression

Y Output

Internal Neural Representations
fMRI

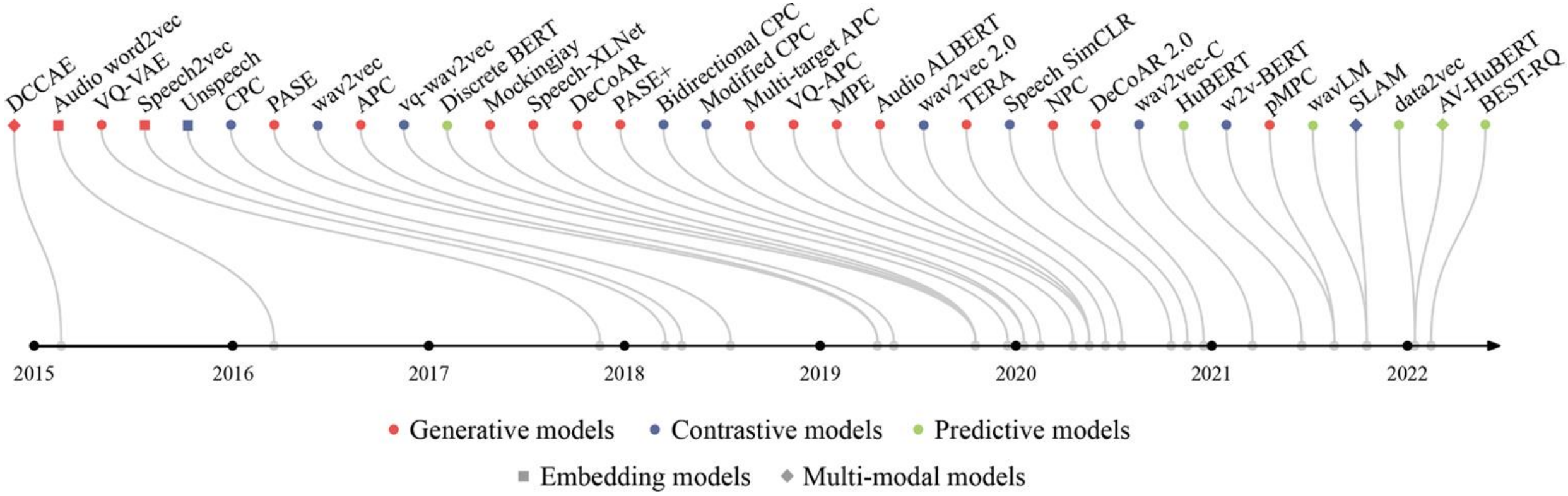


Stimulus

Present

Humans

Speech representation learning methods



Abdelrahman Mohamed, Hung-yi Lee. Self-Supervised Representation Learning for Speech Processing. Tutorial, ICASSP 2022.

Limitations of Earlier Studies

- Speech stimuli have mostly been represented using encodings of text transcription.
- But, text transcription-based methods ignore the raw audio-sensory information completely.
- The best models of the auditory system are still either hand-engineered or supervised (i.e. used basic features like phoneme rate, the sum of squared FFT coefficients)

Self-Supervised speech models accurately predict brain activity

Toward a realistic model of speech processing in the brain with self-supervised learning

Self-supervised models of audio effectively explain human cortical responses to speech

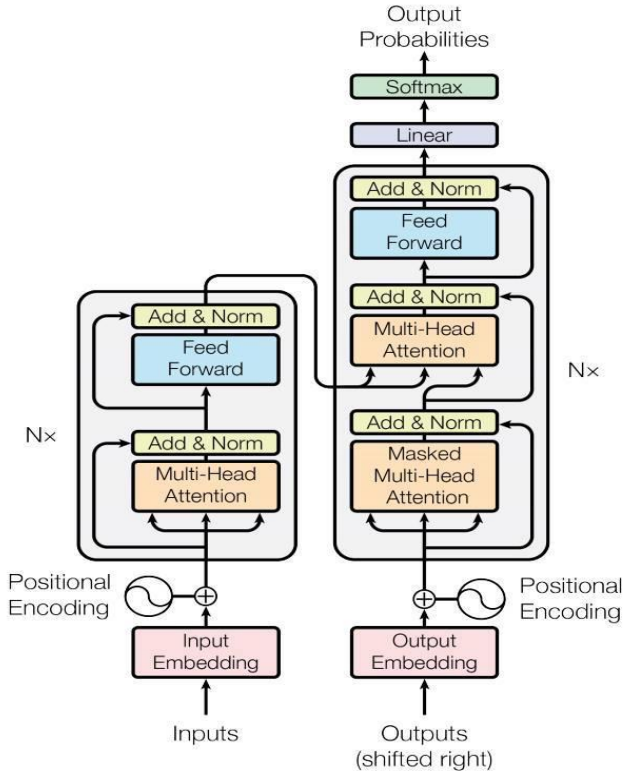
Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence

Greta Tuckute*[1,2], Jenelle Feather*[1,2], Dana Boebinger[1,2,3,4], Josh H. McDermott[1,2,3]
*co-first authors

[1]Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research MIT, Cambridge, MA, USA
[2]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA
[3]Program in Speech and Hearing Biosciences and Technology, Harvard, Cambridge, MA, USA
[4]University of Rochester Medical Center, Rochester, NY, USA

Abstract

Deep neural networks are commonly used as models of the visual system, but are less explored in audition. Prior work provided examples of audio-trained neural networks that produced good predictions of auditory cortical fMRI responses and exhibited correspondence between model stages and brain regions, but left it unclear whether these results generalize to other neural network models. We evaluated brain-model correspondence for publicly available audio neural



Self-Supervised Speech Models (Wav2Vec2.0, HuBERT, APC,…)

Vaidya, Aditya R., Shailee Jain, and Alexander G. Huth. "Self-supervised models of audio effectively explain human cortical responses to speech." ICML (2022).

Millet, Juliette, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. "Toward a realistic model of speech processing in the brain with self-supervised learning." arXiv preprint arXiv:2206.01685 (2022).

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. "Many but not all deep neural network audio models capture brain".

Key Contributions

- We perform an extensive study for brain encoding using DL-based speech models.
- We evaluate 30 speech models grouped into four types against a popular BOLD fMRI dataset (Moth-Radio-Hour).

Traditional approaches

Contrastive approaches

Predictive approaches

Generative approaches

Abdelrahman Mohamed, Hung-yi Lee. Self-Supervised Representation Learning for Speech Processing. Tutorial, ICASSP 2022.

Speech Models

Traditional approaches

Spectrogram
Filter bank
Mel
MFCC
VGGish

Contrastive approaches

PASE+
DeCoAR
DeCoAR2.0
NPC
TERA
Mockingjay
APC
VQ-APC
Audio ALBERT
MAE-AST
SS-AST

Predictive approaches

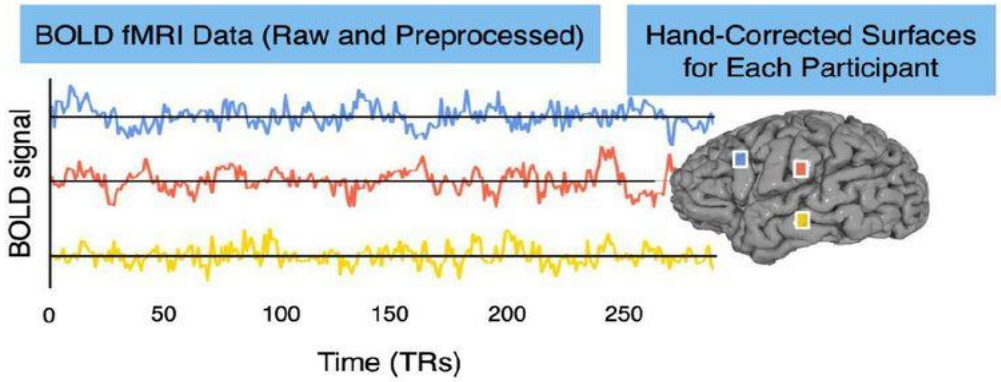
Modified CPC
Wav2Vec
VQ-Wav2Vec2.0
Wav2Vec2.0
Wav2Vec2.0-Large
Wav2Vec2.0-C
Discrete BERT
BYOL-A
Unispeech

Generative approaches

WavLM
HuBERT
Data2Vec
DistilHuBERT
LightHuBERT

Listening data target: human brain recordings

- We use Moth-Radio-Hour story listening dataset:
  - 6 subjects,
  - 27-stories
  - 9737 TRs (TR: repetition time)
  - Each TR is 2.0045 sec.

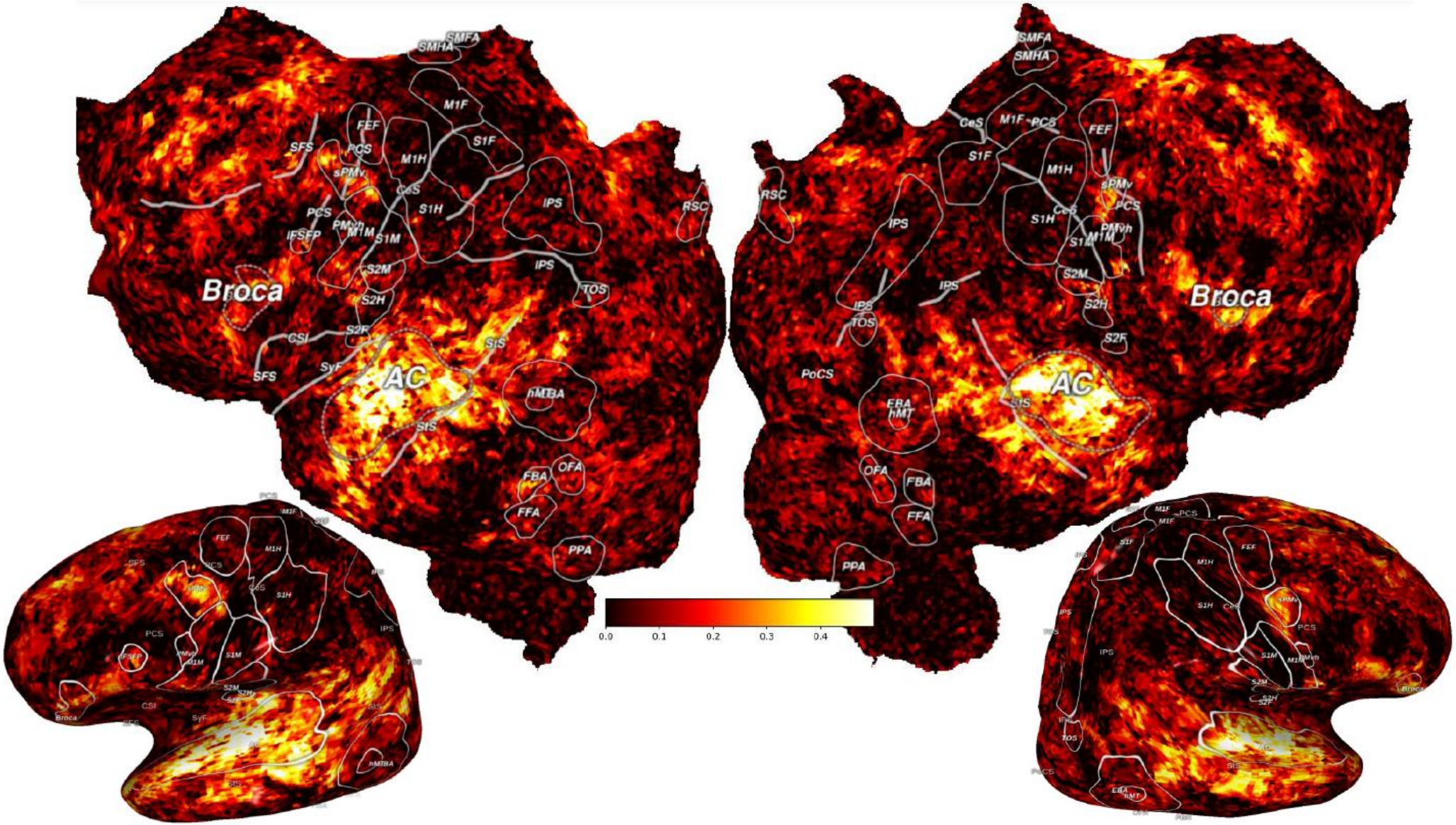


Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, Alexander G. Huth "A natural language fMRI dataset for voxelwise encoding models" Nature (2022).

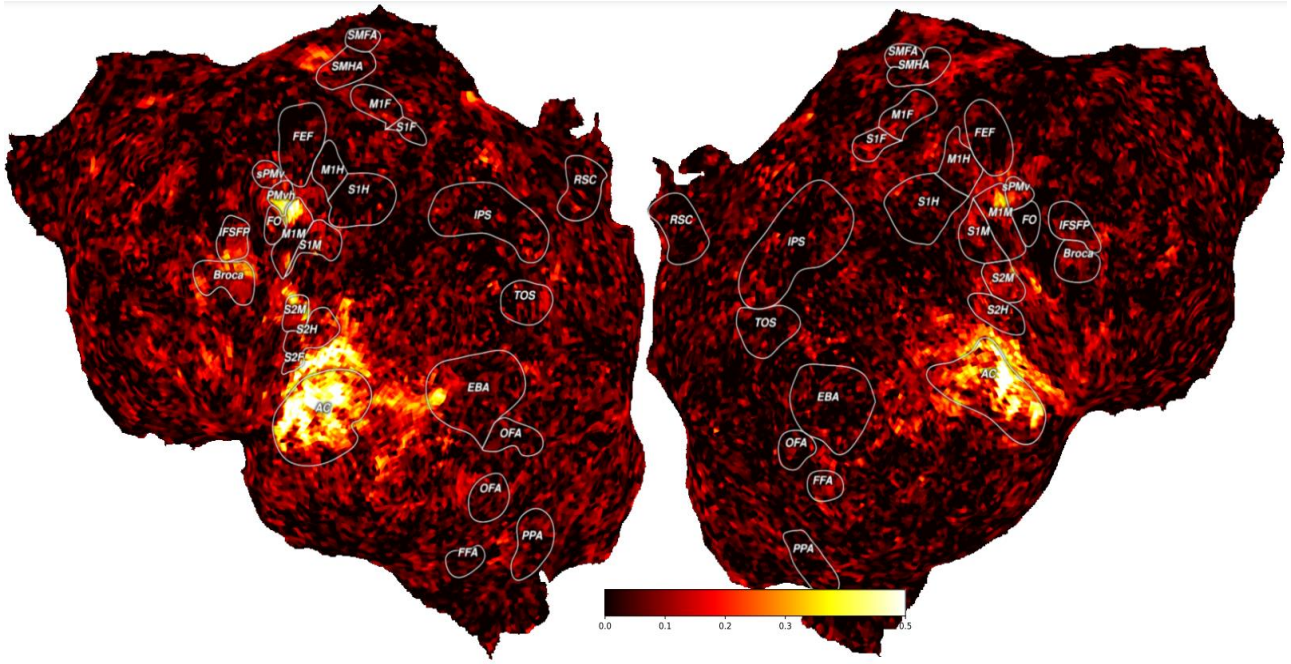
Encoding Performance of Speech Models

Category	Model	AC	Broca	Whole Brain
Traditional non-DL & non-SS DL Methods	Spectrogram	0.0545	0.0511	0.0495
	Filter bank	0.0477	0.0450	0.0498
	Mel	0.0489	0.0515	0.0511
	MFCC	0.0495	0.0520	0.0517
	VGGish	**0.1612**	**0.0785**	**0.0605**
Generative Self-Supervised Methods	PASE+	0.1272	0.0719	0.0601
	DeCoAR	0.2332	0.1017	0.0695
	DeCoAR2.0	0.2293	**0.1142**	0.0722
	NPC	0.2123	0.0995	0.0678
	TERA	0.2332	0.1052	0.0718
	Mockingjay	0.1812	0.0946	0.0624
	APC	**0.2382**	0.0991	0.0710
	VQ-APC	0.2085	0.0891	0.0658
	Audio ALBERT	0.2184	0.0992	0.0688
	MAE-AST	0.2355	0.1132	**0.0729**
	SS-AST	0.2193	0.1023	0.0673
Contrastive Self-Supervised Methods	Modified CPC	0.2128	0.1019	0.0671
	Wav2Vec	0.2209	0.1044	0.0719
	VQ-Wav2Vec2.0	0.2307	0.1167	0.0754
	Wav2Vec2.0	0.2662	0.1741	0.0861
	Wav2Vec2.0-Large	**0.2676**	**0.1750**	**0.0882**
	Wav2Vec2.0-C	0.2655	0.1740	0.0860
	Discrete BERT	0.2277	0.1065	0.0715
	BYOL-A	0.1302	0.0784	0.0566
	Unispeech	0.2378	0.1356	0.0738
Predictive Self-Supervised Methods	WavLM	0.2356	0.1116	0.0727
	HuBERT	0.2298	0.1088	0.0730
	Data2Vec	**0.2683**	**0.1756**	**0.0886**
	DistilHuBERT	0.2323	0.1101	0.0738
	LightHuBERT	0.2328	0.1102	0.0737

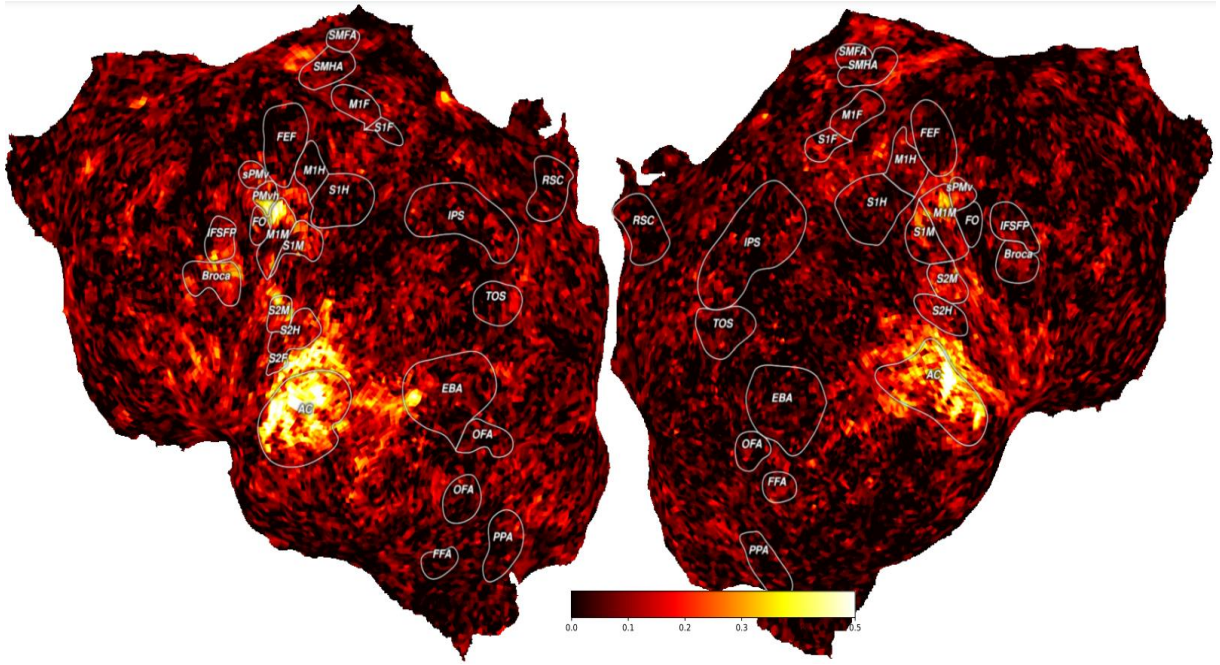
Model Encoding Performance (Data2Vec)



Model Encoding Performance

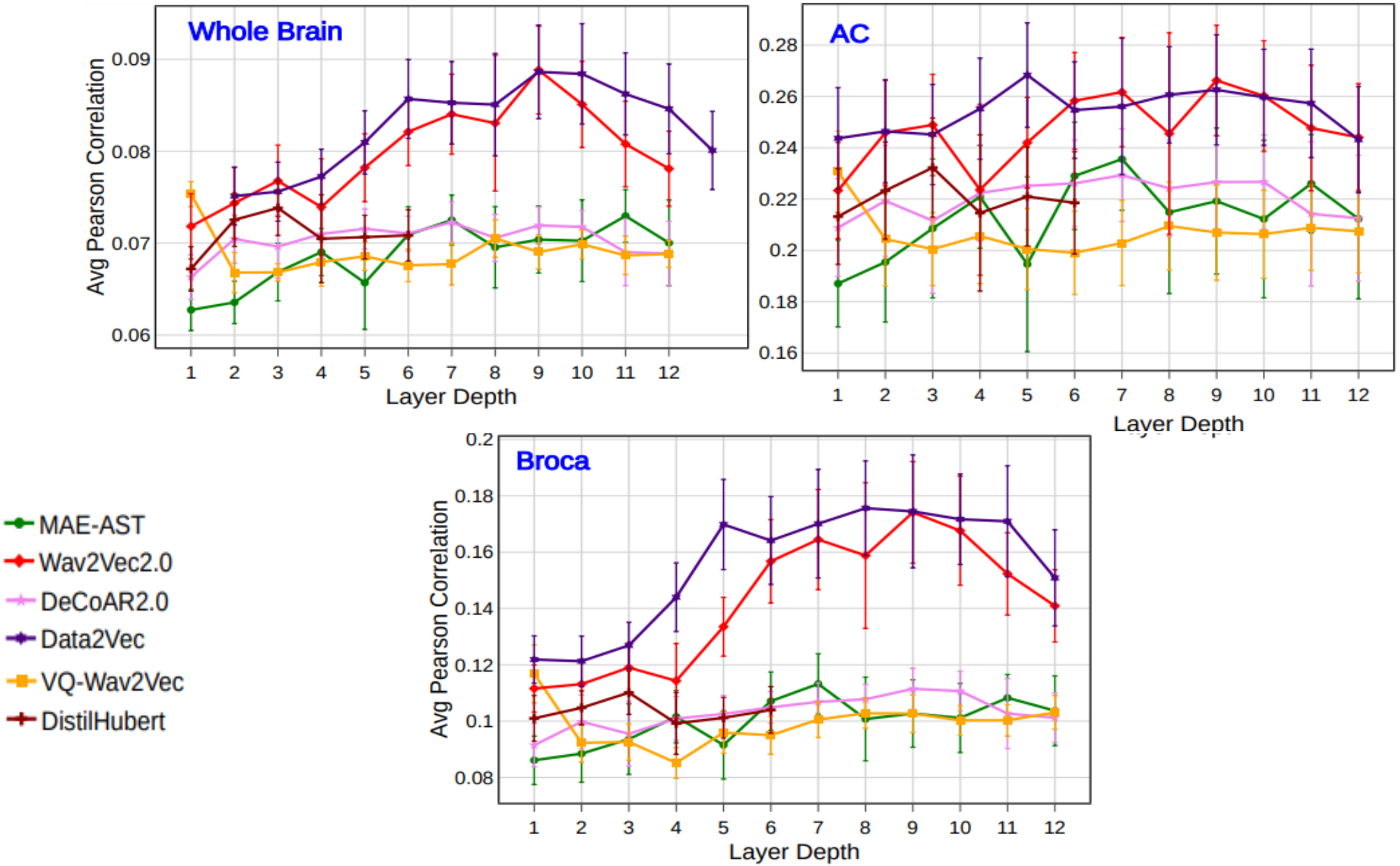


Tera

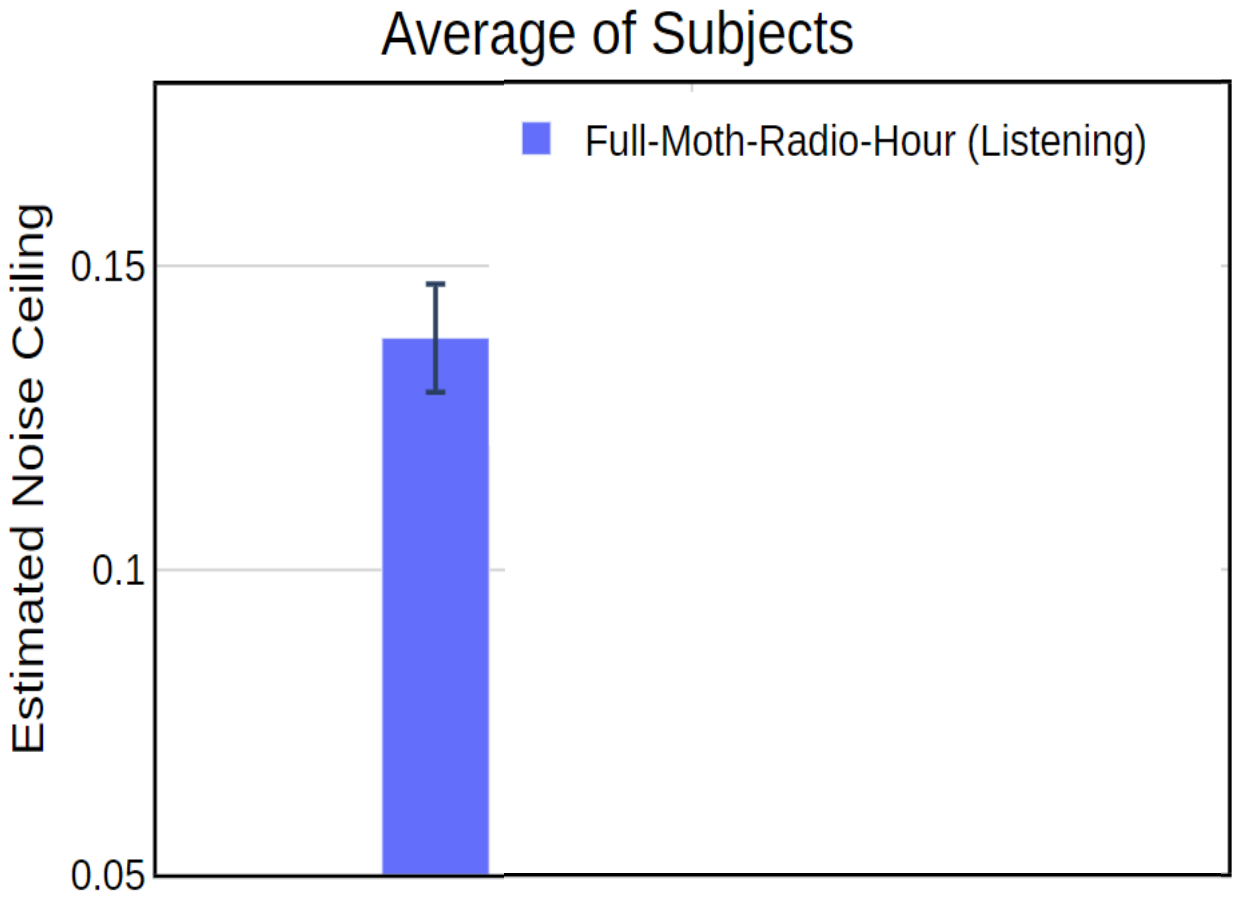


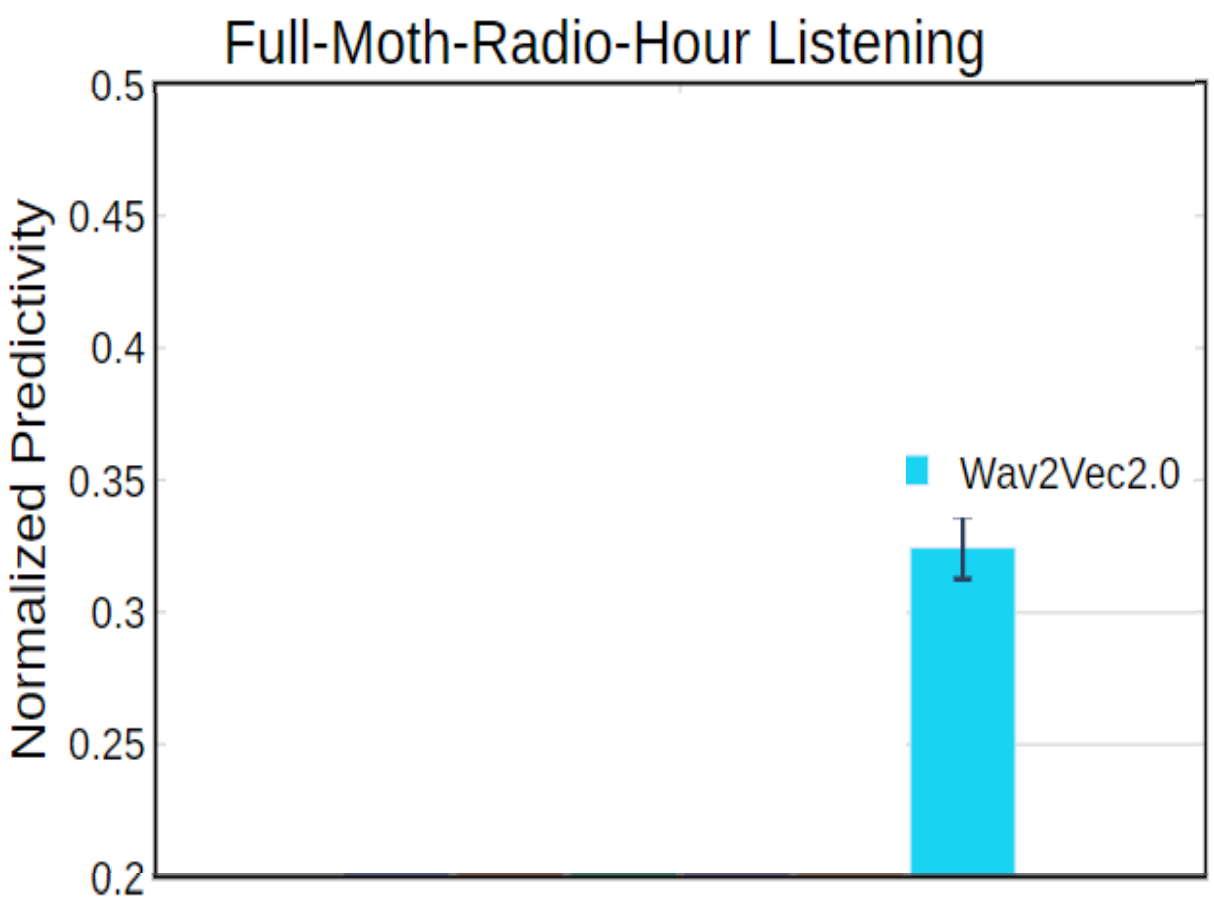
Unispeech

Layer Selectivity

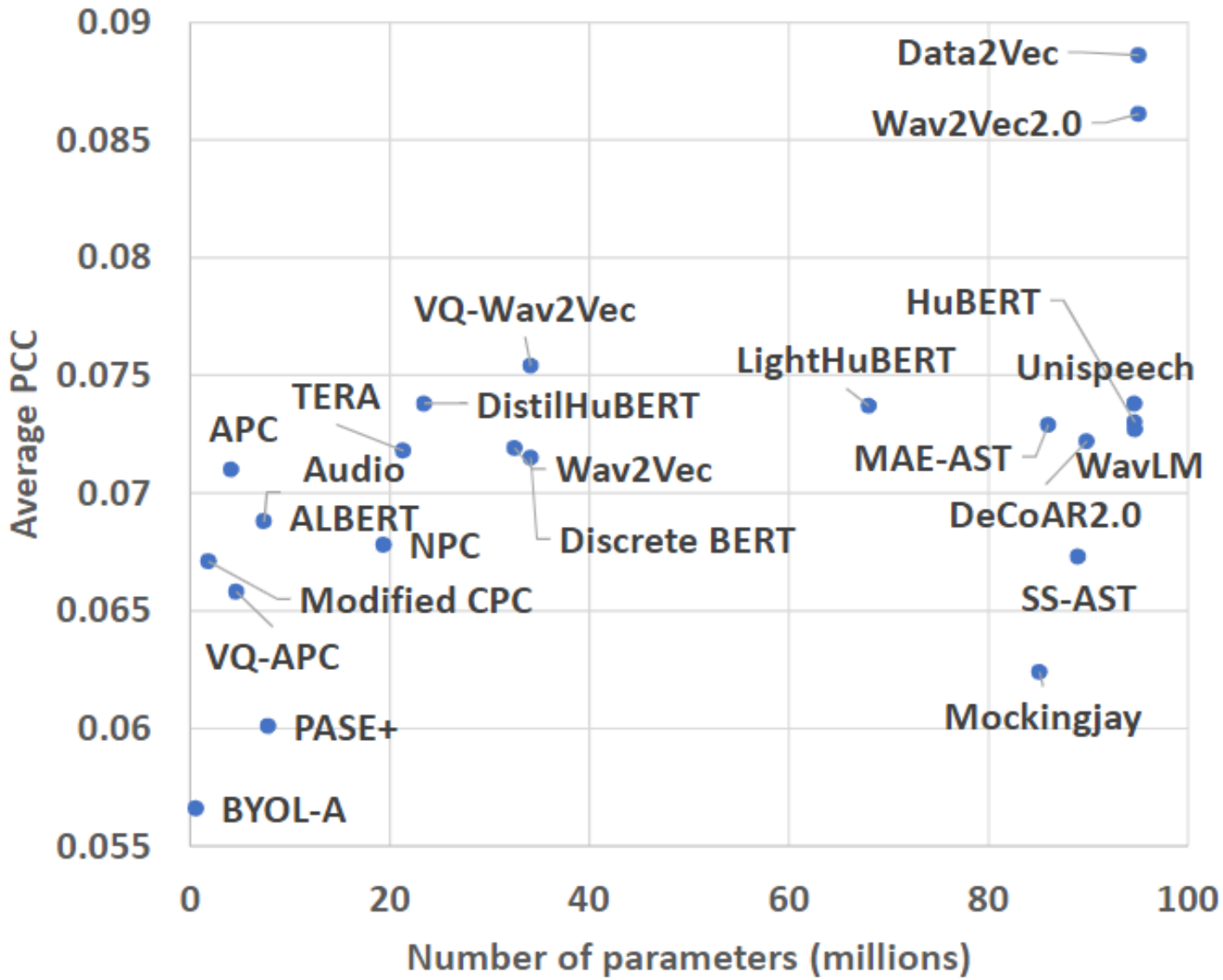


How do we assess models' performance?





Model Size vs. PCC



Conclusion & Future Works

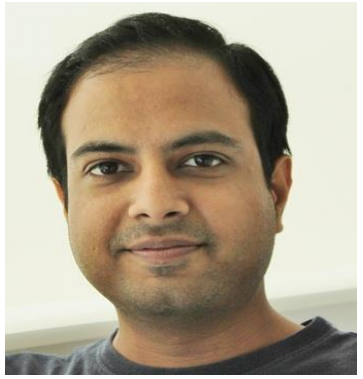
- We comprehensively evaluated different categories of encoding models to evaluate their efficacy in learning brain-like representations –
  - traditional DL and non-DL,
  - generative,
  - contrastive, and
  - predictive self-supervised (SS) models- to evaluate their efficacy in learning brain-
- Contrastive and predictive models encode the information better than the generative and the traditional low-level acous-tic baselines, and VGGish models.
- We plan to explore the effect of finetuning these speech models (SUPERB benchmark) rather than using them in probe mode only.
- Also, we plan to extend the experiments to multi-modal models that encode audio and text together

Collaborators

Subba Reddy Oota

Khushbu Pahwa

Mounika Marreddy



Manish Gupta

Bapi Raju Surampudi

Acknowledgement

- We thank Dr. Mariya Toneva for her very helpful comments on this paper.