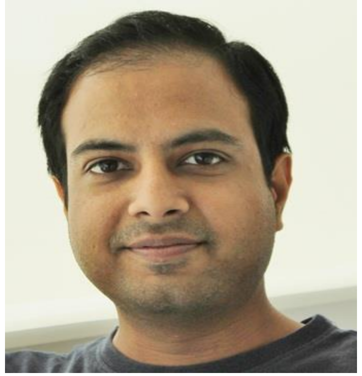
INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

Speech Taskonomy: Which Speech Tasks are the most Predictive of fMRI Brain Activity?




Subba Reddy Oota[1] Veeral Agarwal[2] Mounika Marreddy[2] Manish Gupta[2,3] Bapi Raju Surampudi[2]

[1]Inria Bordeaux, France [2]IIIT-Hyderabad, India [3]Microsoft, India

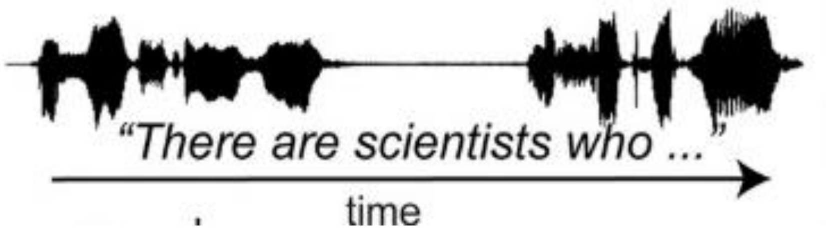
inria
inventeurs du monde numérique

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

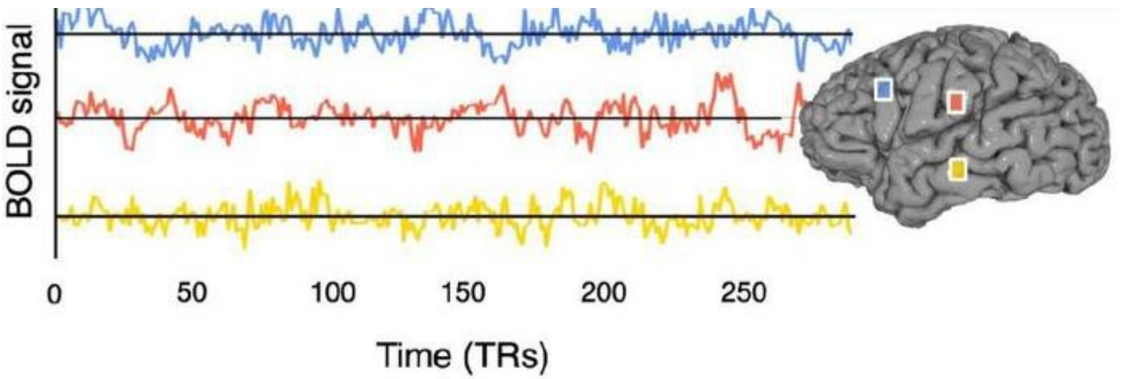
Microsoft

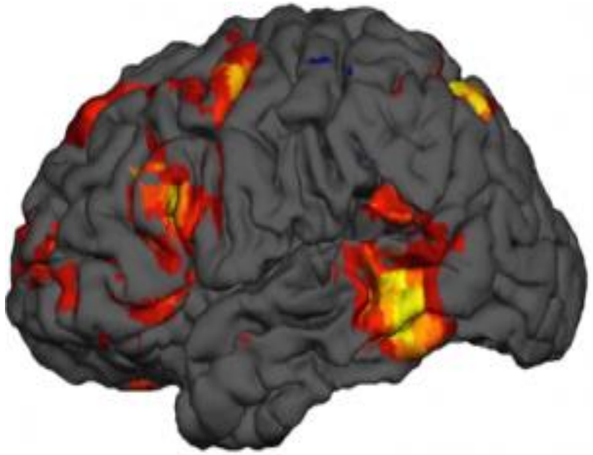
What is fMRI?

Narrative Story



A listening task in the scanner



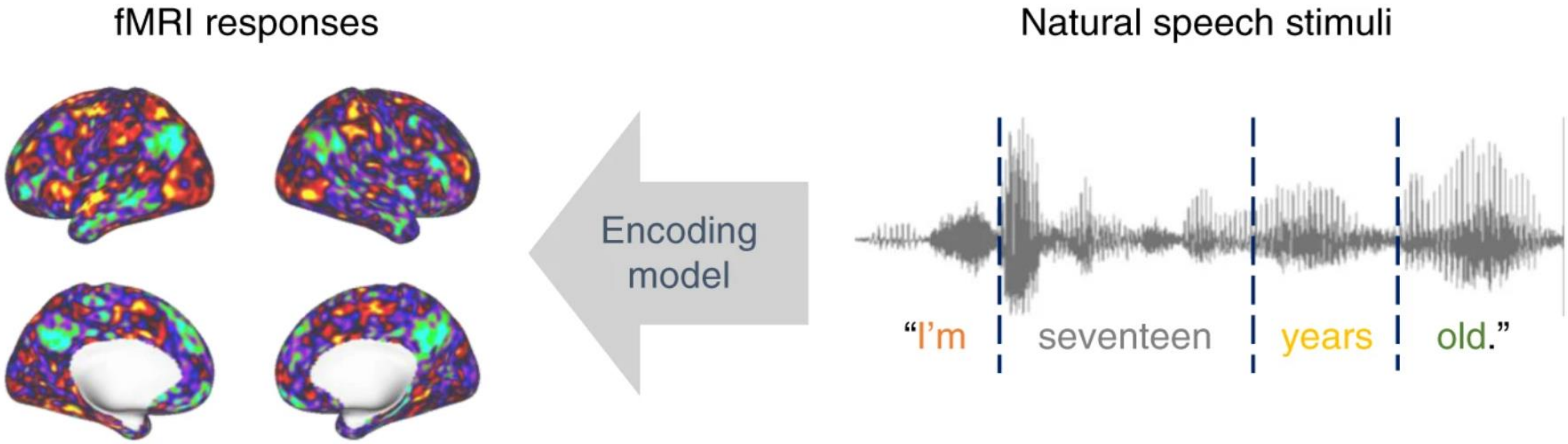


fMRI Brain Activity

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

Brain Encoding vs Decoding

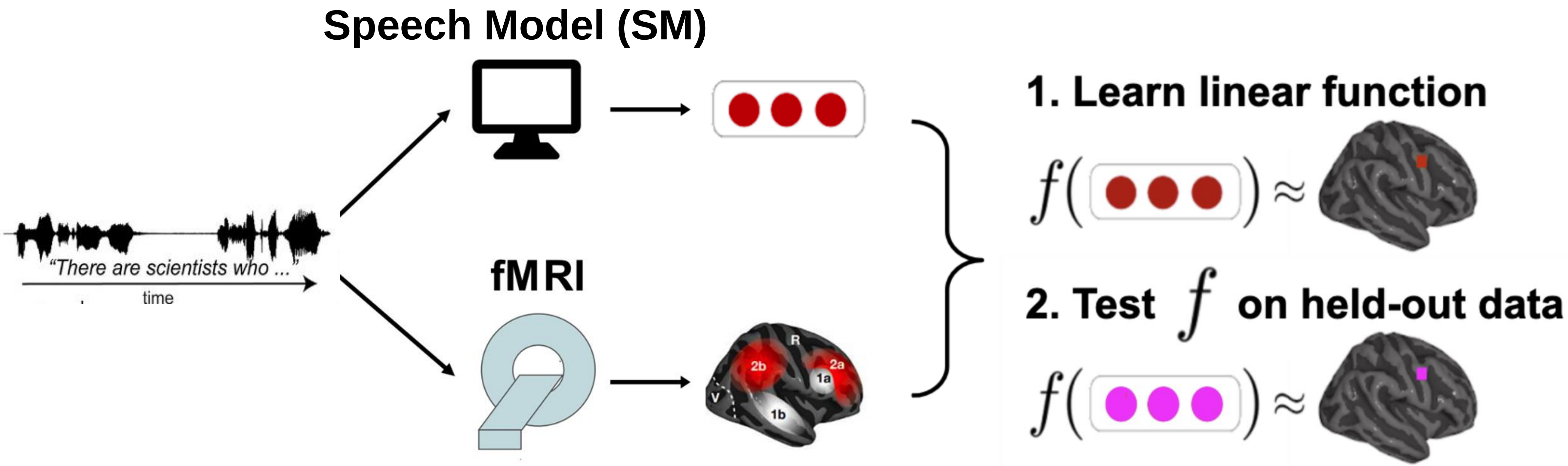
fMRI → Decoding → Stimulus Representation



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).

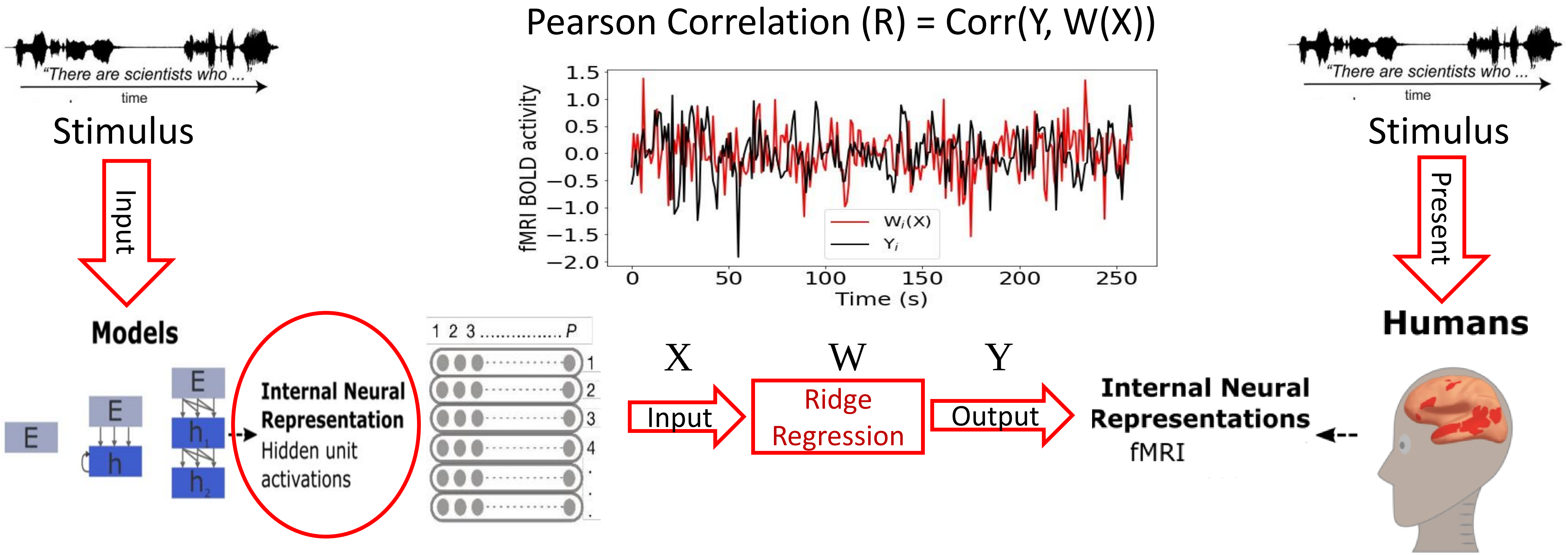
Method to study alignment of SM & brain representations



Brain alignment of a SM ⇒ how similar its representations are to a human brain's

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.

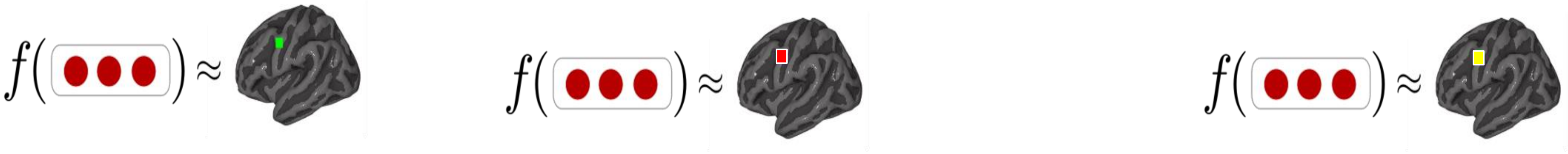
Brain Encoding?



Encoding: training **independent** models

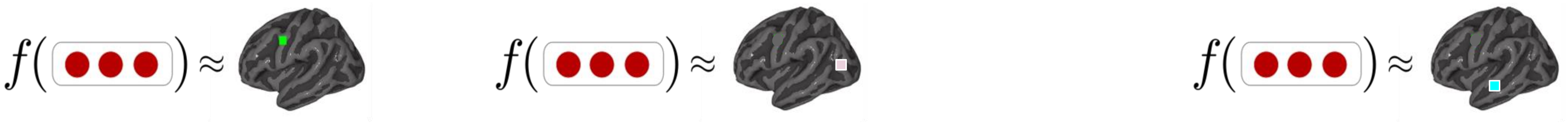
- Independent model per participant

P1 **P2** … **PN**



- Independent model per voxel / sensor-timepoint

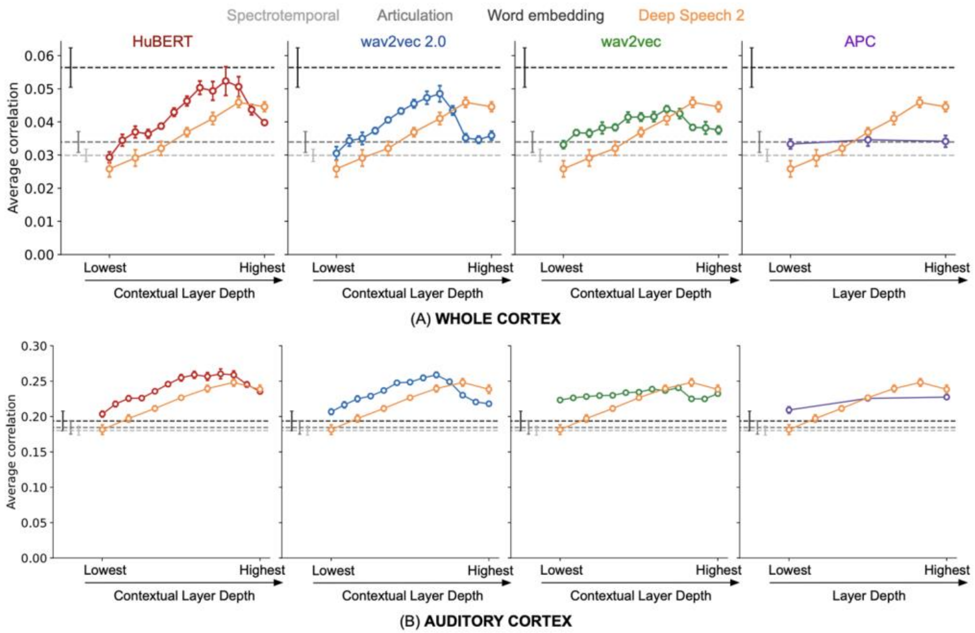
P1, v1 **P1, v2** … **P1, vm**

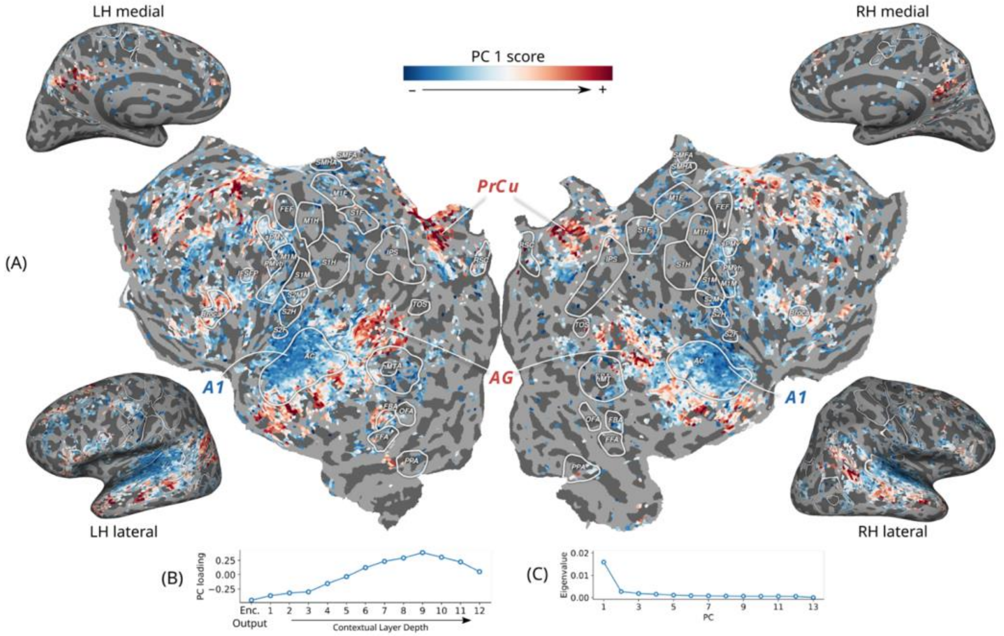


Recent work utilizing progress in self-supervised speech models for encoding

- Stimuli: Moth Radio Hour
- Stimulus representation: derived from pretrained **self-supervised speech models (HuBERT, Wav2Vec2.0, APC)**
- Brain recording & modality: fMRI, listening

Middle layers of self-supervised speech models predict auditory cortex the best



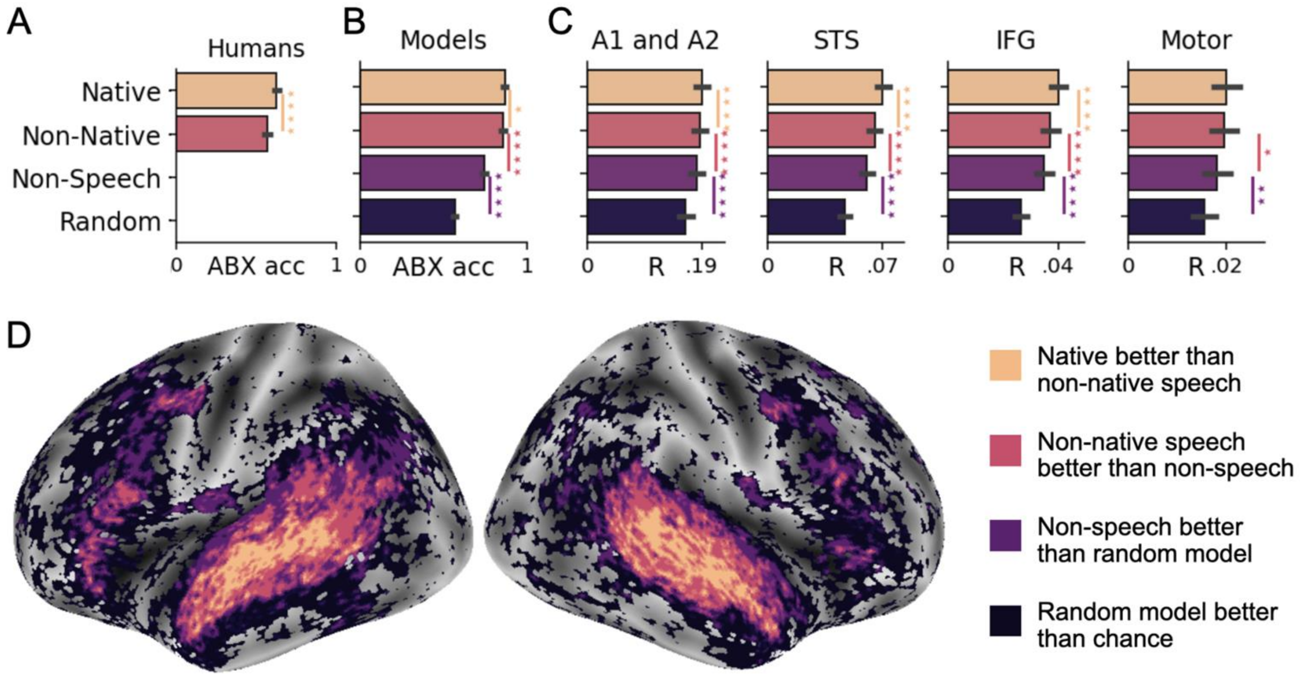


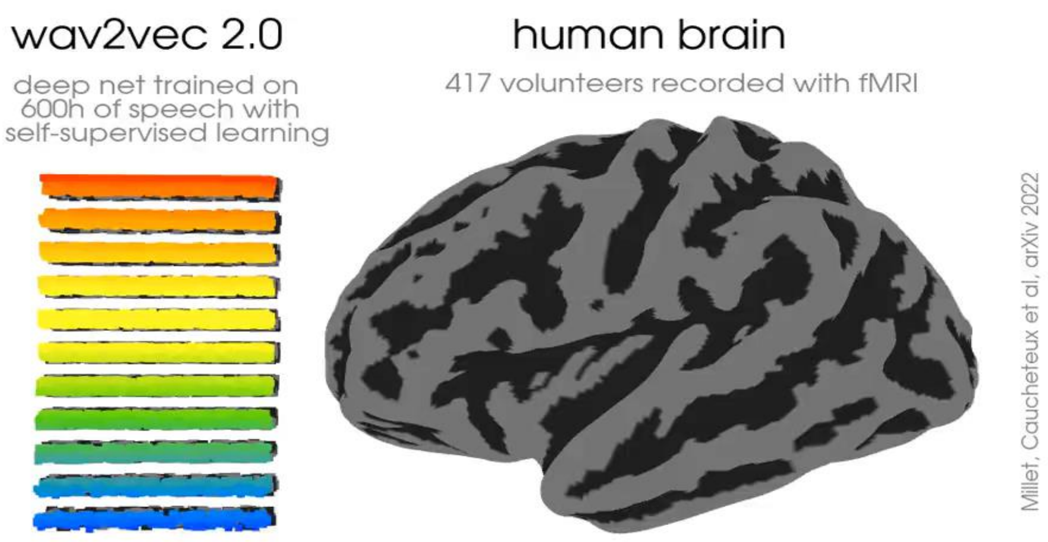
Vaidya, Aditya R., Shailee Jain, and Alexander G. Huth. "Self-supervised models of audio effectively explain human cortical responses to speech." ICML (2022).

Audio: work utilizing DL progress

Self-supervised speech models exhibit specialization for native sounds in the STS and MTG;
IFG and AG show more general specialization for speech rather than native-language

- Stimuli: audio books
- Stimulus representation: derived from pretrained self-supervised speech model (Wav2Vec2.0)
- Brain recording & modality: fMRI, listening in 3 languages (Eng, Fr, Mandarin)



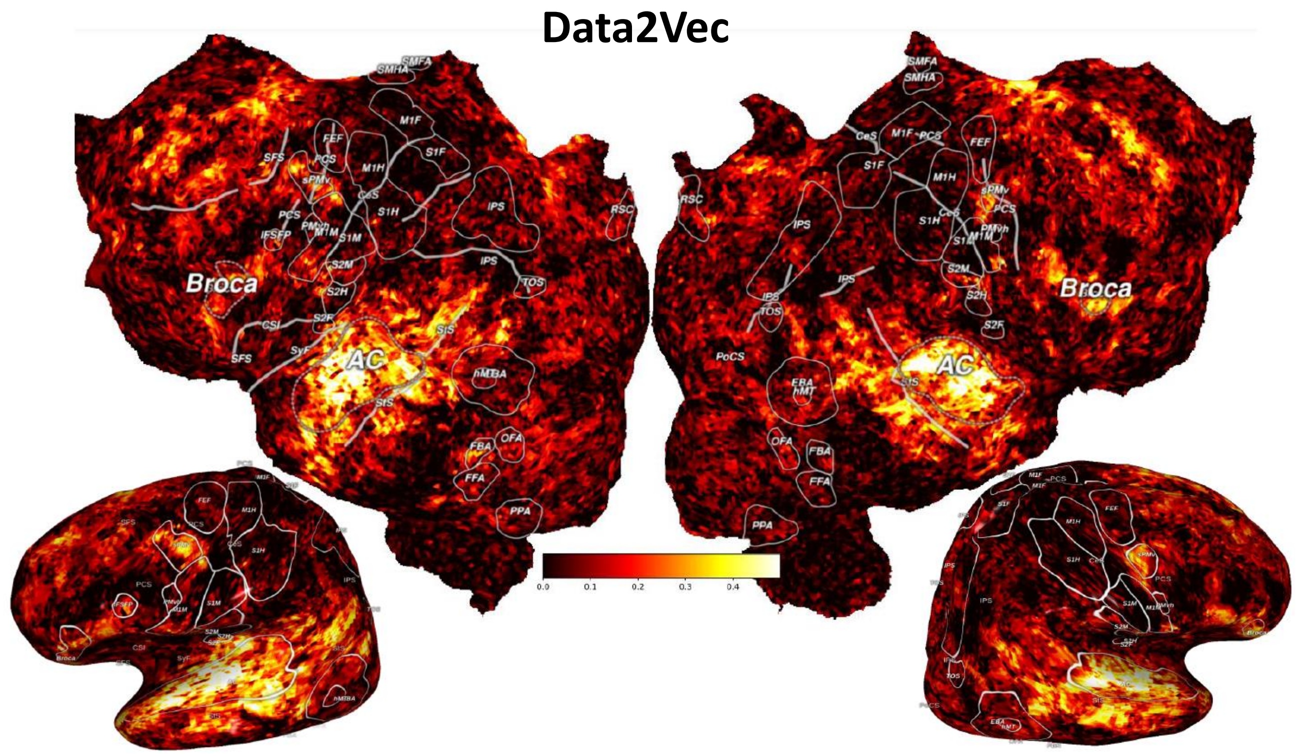


Millet, Juliette, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. "Toward a realistic model of speech processing in the brain with self-supervised learning." arXiv preprint arXiv:2206.01685 (2022).

Audio: work utilizing DL progress

Contrastive and predictive models encode the information better than the generative and the traditional low-level acoustic baselines, and VGGish models.

- Stimuli: Moth-Radio-Hour
- Stimulus representation: derived from 5 basic + 25 pretrained self-supervised speech models
- Brain recording & modality: fMRI



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

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Bapi S. Raju. "Neural architecture of speech" ICASSP-2023

Challenges in using DL for cognitive science

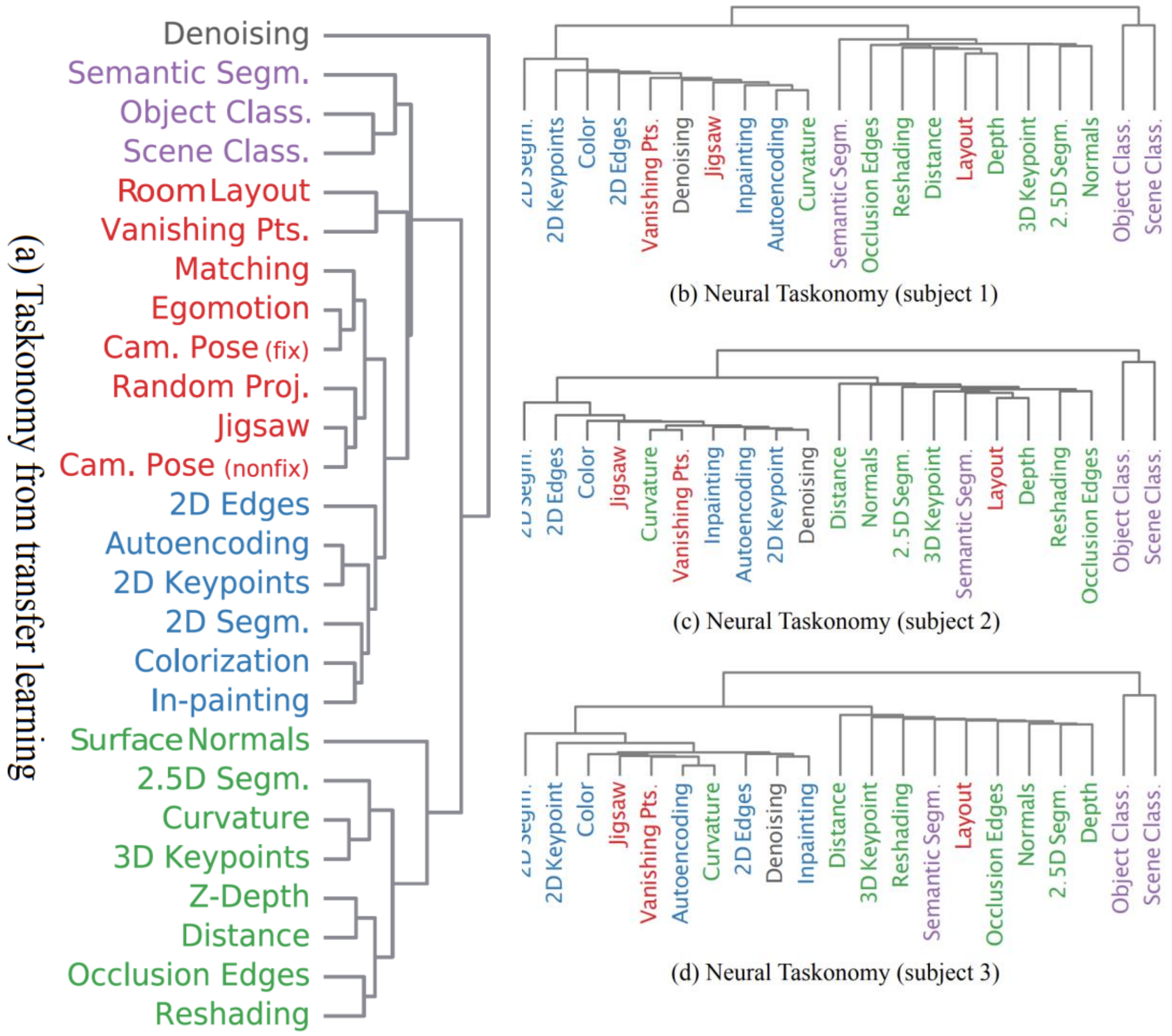
- Not designed to specifically model brain processing
  - Training DL models using brain recordings
  - **Task-based modeling**
- Can be difficult to interpret due to multiple sources of information
  - Disentangling contributions of different info sources to brain predictions

Tasks affect processing

- Stimuli: images of natural scenes
- Stimulus representation: task-optimized CNNs for a range of tasks
- Brain recording & modality: fMRI, vision

Vision tasks with higher transferability make similar predictions for brain responses from different regions

Semantic Low-dim. Geometric 2D 3D

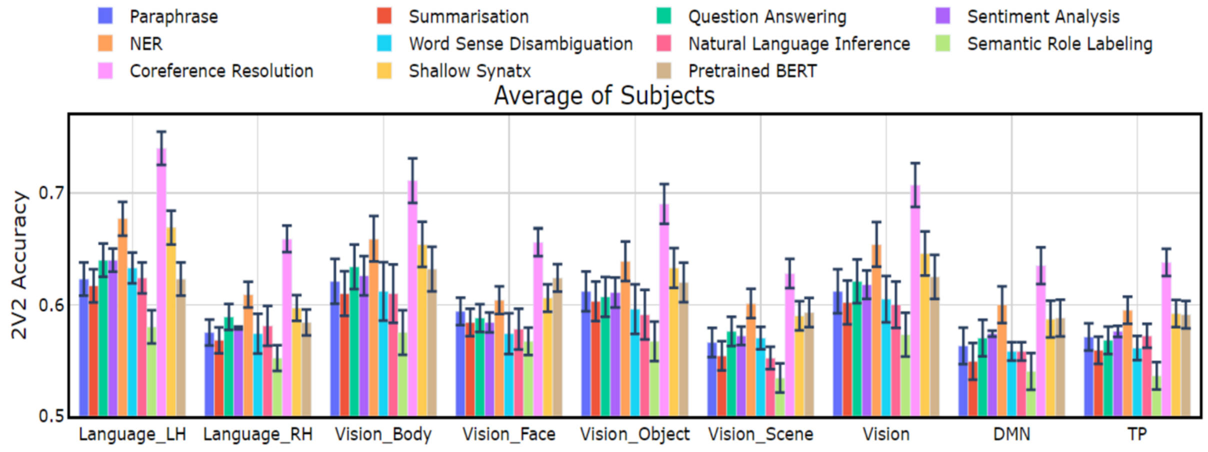


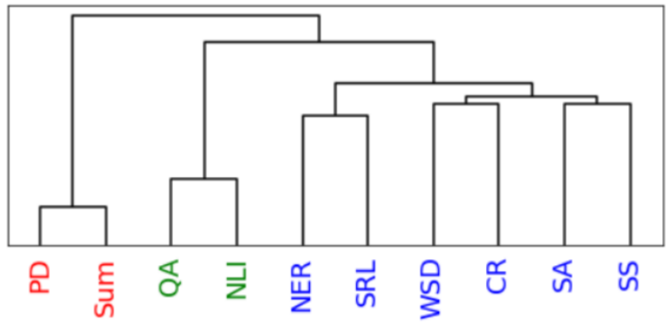
Wang, Aria, Michael Tarr, and Leila Wehbe. "Neural taskonomy: Inferring the similarity of task-derived representations from brain activity." Advances in Neural Information Processing Systems 32 (2019).

Tasks affect processing

- Stimuli: passages and narratives
- Stimulus representation: task-optimized NLP models for a range of tasks
- Brain recording & modality: fMRI, reading & listening of different stimuli

Reading fMRI best explained by coref. resolution, NER, shallow syntax parsing
Listening fMRI best explained by paraphrasing, summarization, NLI

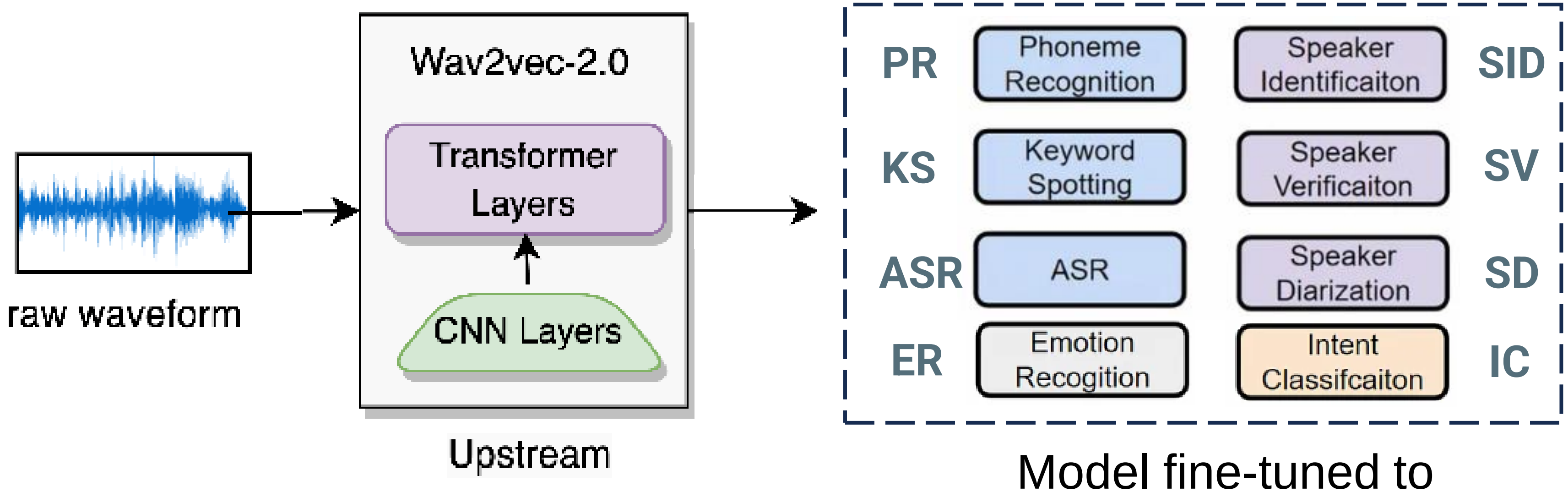




Oota, Subba Reddy, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?." *arXiv preprint arXiv:2205.01404* (2022).

Can task-specific speech models better predict fMRI brain activity?

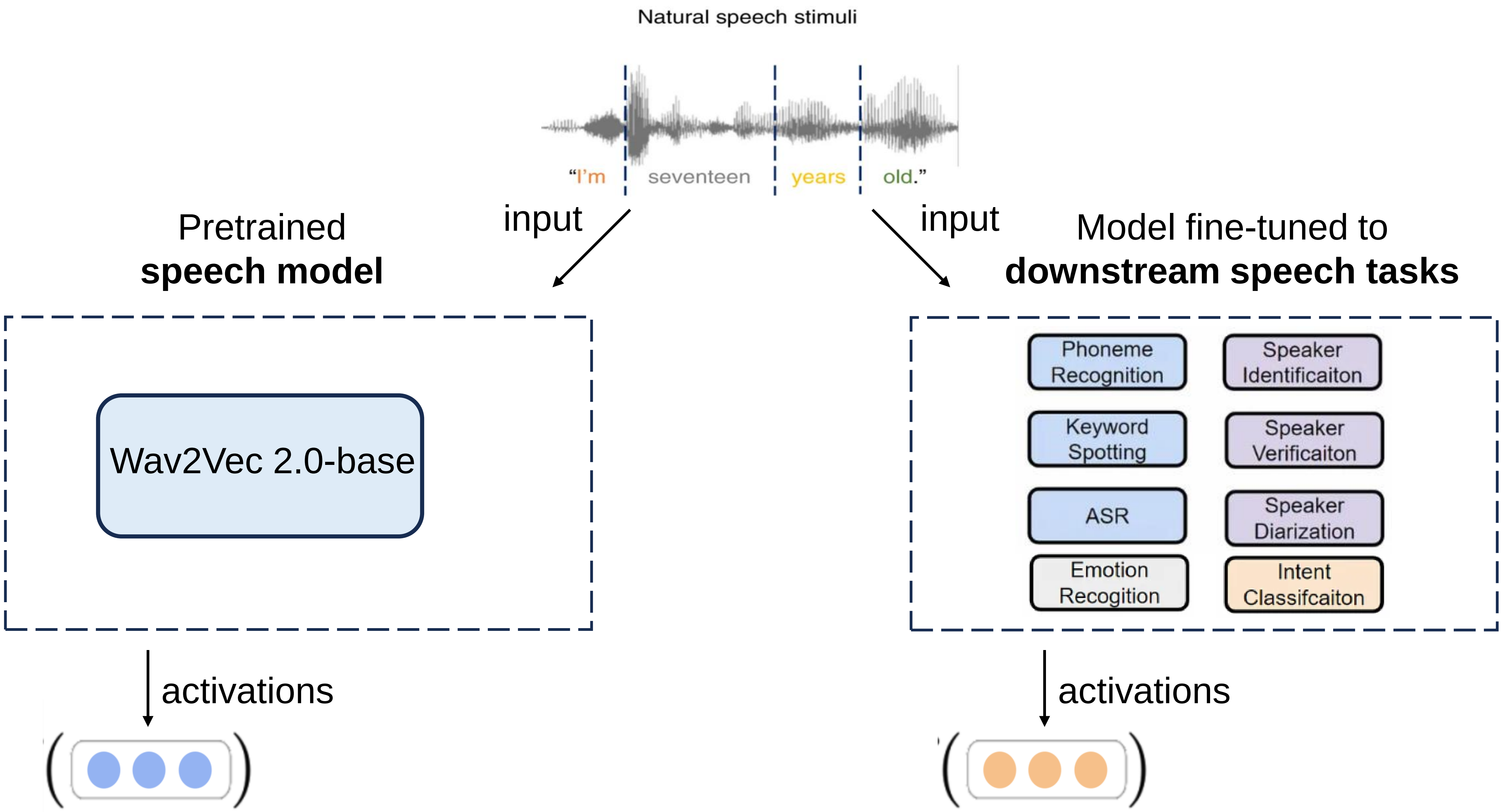
SUPERB (Speech Processing Universal PERformance Benchmark)

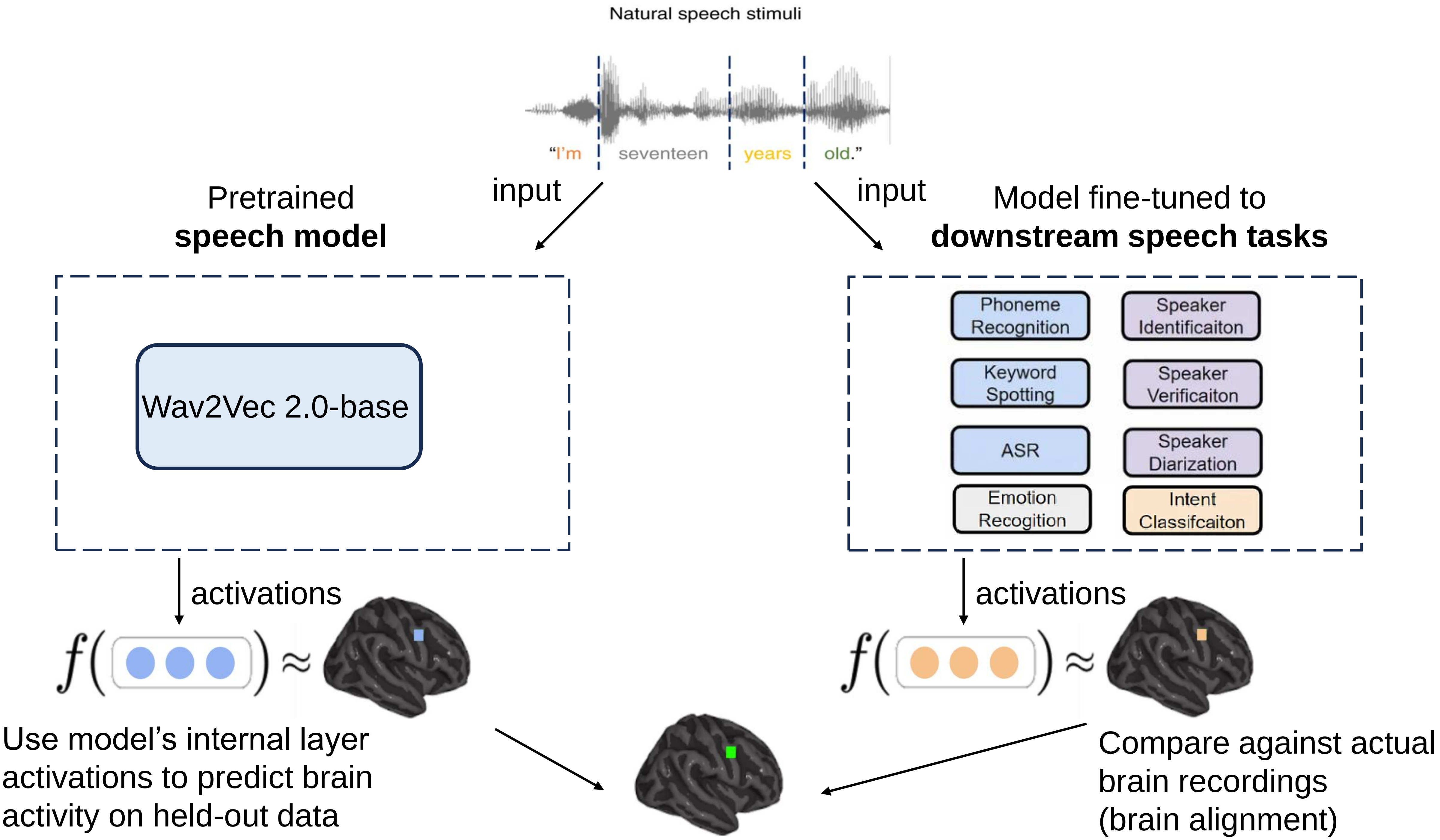


Model fine-tuned to
downstream speech tasks

Ashish Seth, Vasista Sai Lodagala, Sreyan Ghosh, and S. Umesh. "ANALYZING THE FACTORS AFFECTING USEFULNESS OF SELF-SUPERVISED" https://arxiv.org/pdf/2203.16973.pdf

https://superbbenchmark.org


Natural speech stimuli
“I’m
seventeen
years
old.”
input
input
Pretrained
speech model
Model fine-tuned to
downstream speech tasks
Wav2Vec 2.0-base
Phoneme Recognition
Speaker Identificaiton
Keyword Spotting
Speaker Verificaiton
ASR
Speaker Diarization
Emotion Recogition
Intent Classifcaiton
activations
activations

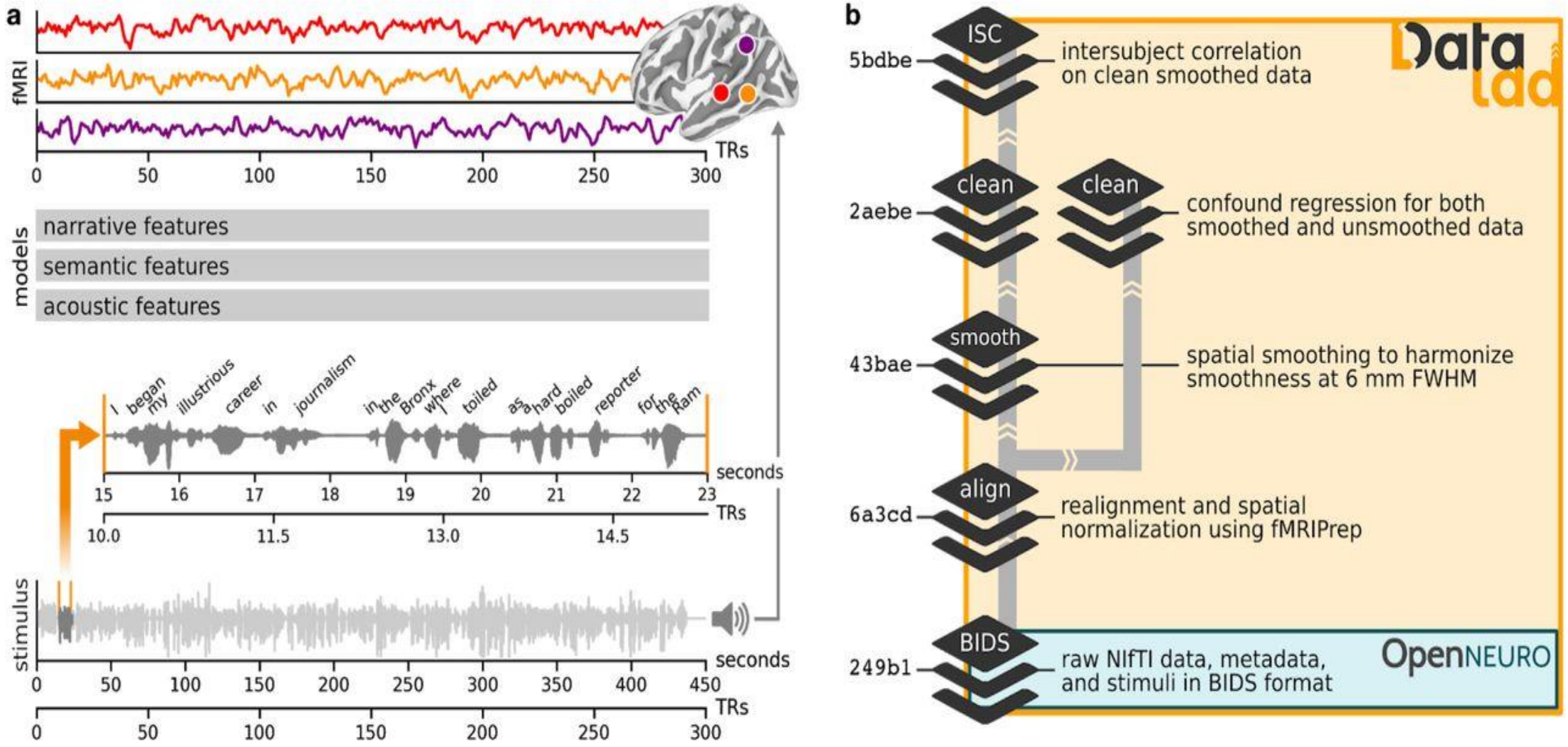

Natural speech stimuli
"I'm
seventeen
years
old."
input
input
Pretrained
speech model
Model fine-tuned to
downstream speech tasks
Wav2Vec 2.0-base
Phoneme Recognition
Speaker Identificaiton
Keyword Spotting
Speaker Verificaiton
ASR
Speaker Diarization
Emotion Recogition
Intent Classifcaiton
activations
activations
Use model's internal layer activations to predict brain activity on held-out data
Compare against actual brain recordings (brain alignment)

Hypothesis: If the pretrained model fine-tuned to downstream speech task has greater brain alignment than pretrained model, the downstream task is capturing more brain-relevant information

Listening data target: human brain recordings

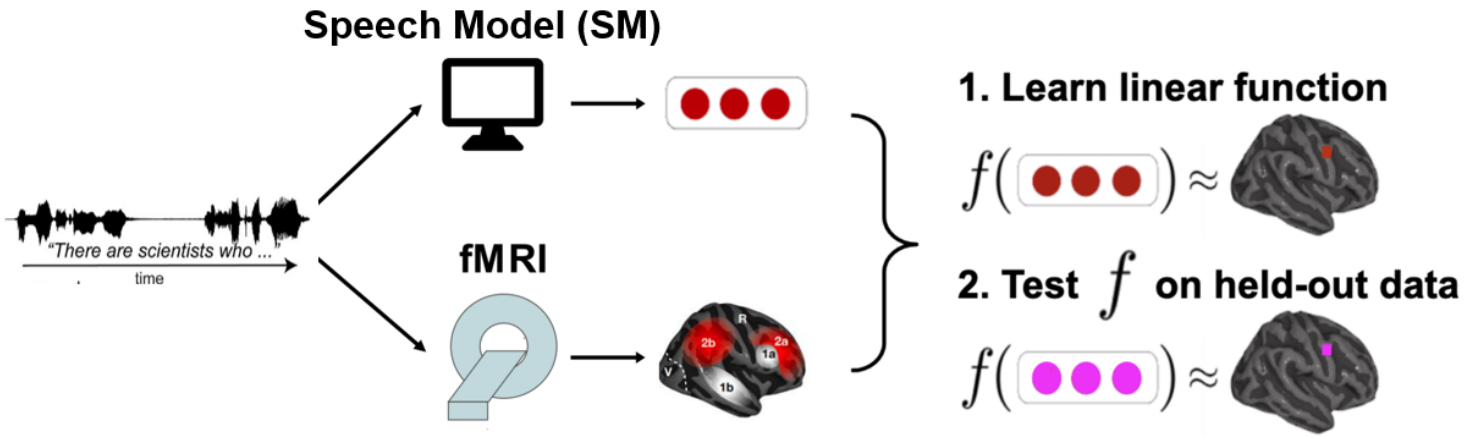
- We use Pieman story listening:
  - 82 subjects,
  - 282 TRs (repetition time)
  - here it is 1.5 sec.

Example: "I began my illustrious carrier in journalism…"

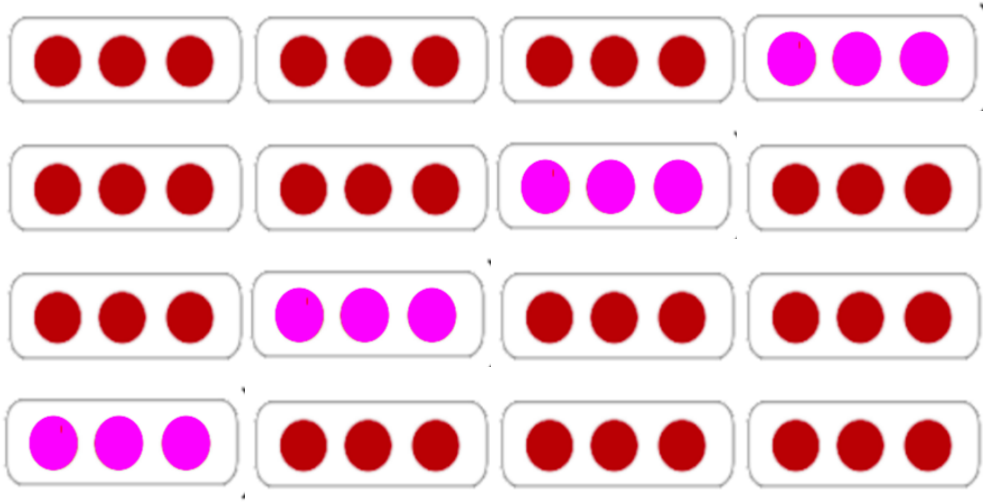


[Samuel A. Nastase. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. Nature, 2021.]

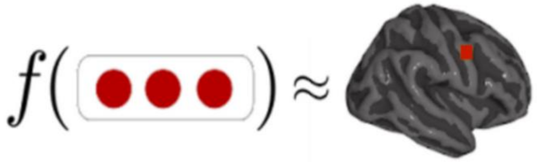
Brain alignment – 4-fold Cross-Validation + Ridge regression



(1) 4-fold Cross-Validation



(2) Linear regression regularized with ridge penalty



Input: $\mathbf{R}^{282 \times d}$

d = embedding size, e.g. 4608

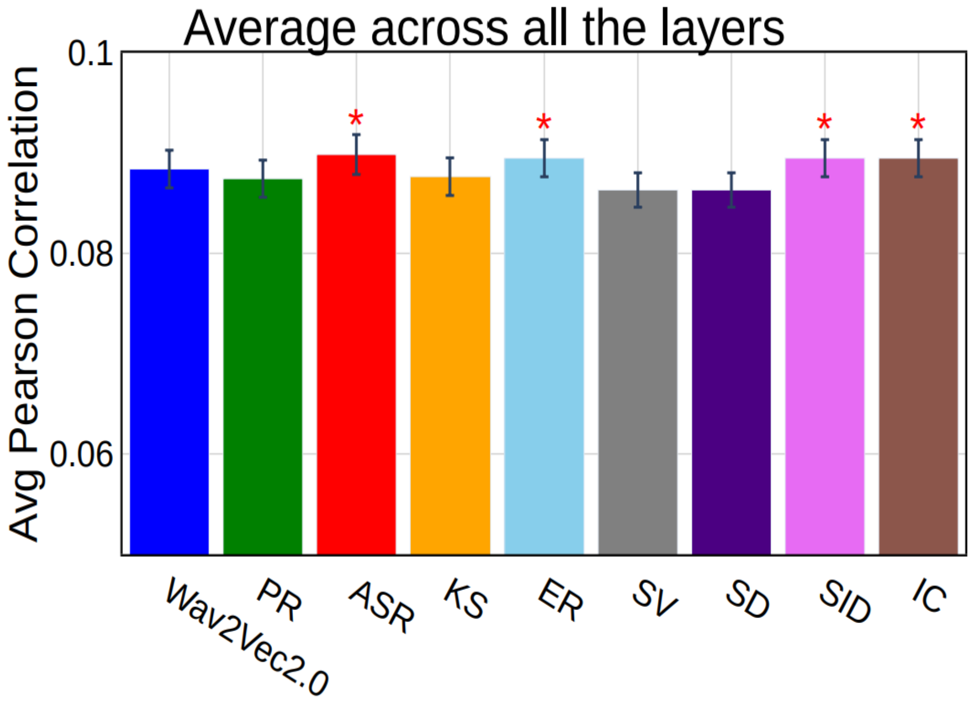
Prediction: $\mathbf{R}^{n \times v}$

n = number of fMRI intervals
v = number of voxels in participant's brain

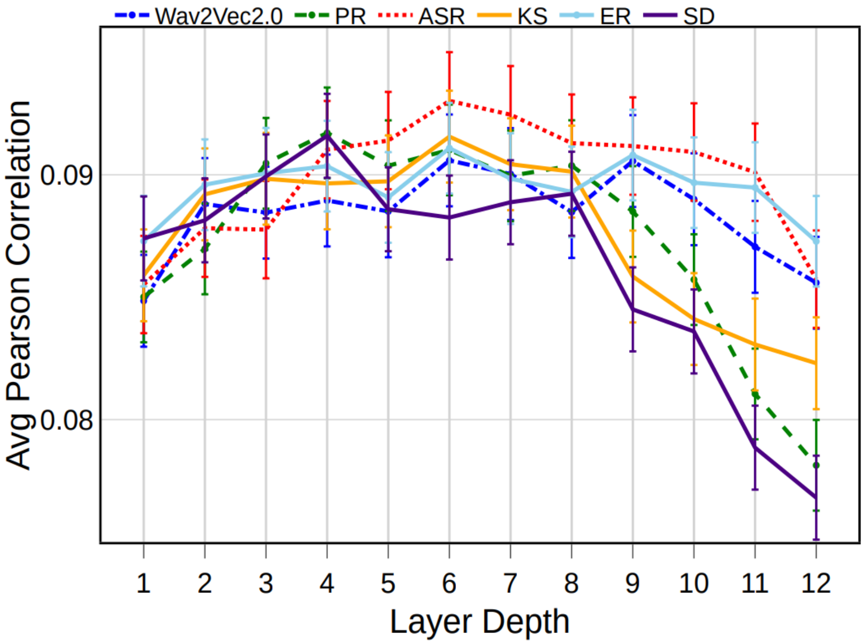
Brain alignment – Methodology

- 300 TRs x 768 ⇒ SM representations
  - Wav2Vec 2.0-base
  - SUPERB Benchmark downstream tasks : eight tasks
- 282 TRs x 768
  - Remove 18 TRs ⇒ 10 TRs in the beginning and 8 TRs in the ending (silent music)
- 282 fMRI time intervals x 4608
  - Concatenate SM representations for previous 6 TRs ⇒ fMRI response from brain activity peaks about 8-10 seconds after stimulus onset
- 282 fMRI intervals x 4608
  - Ridge regression (RR) ⇒ for each voxel, 282 data rows of 4608 parameters to predict 1 output
  - 4-fold Cross-Validation to improve reliability
- 282 fMRI intervals x 52400 voxels ⇒ fMRI predictions (same dimensions as actual brain activity)

Results: Whole-Brain



ASR best encodes speech stimuli for brain response prediction.

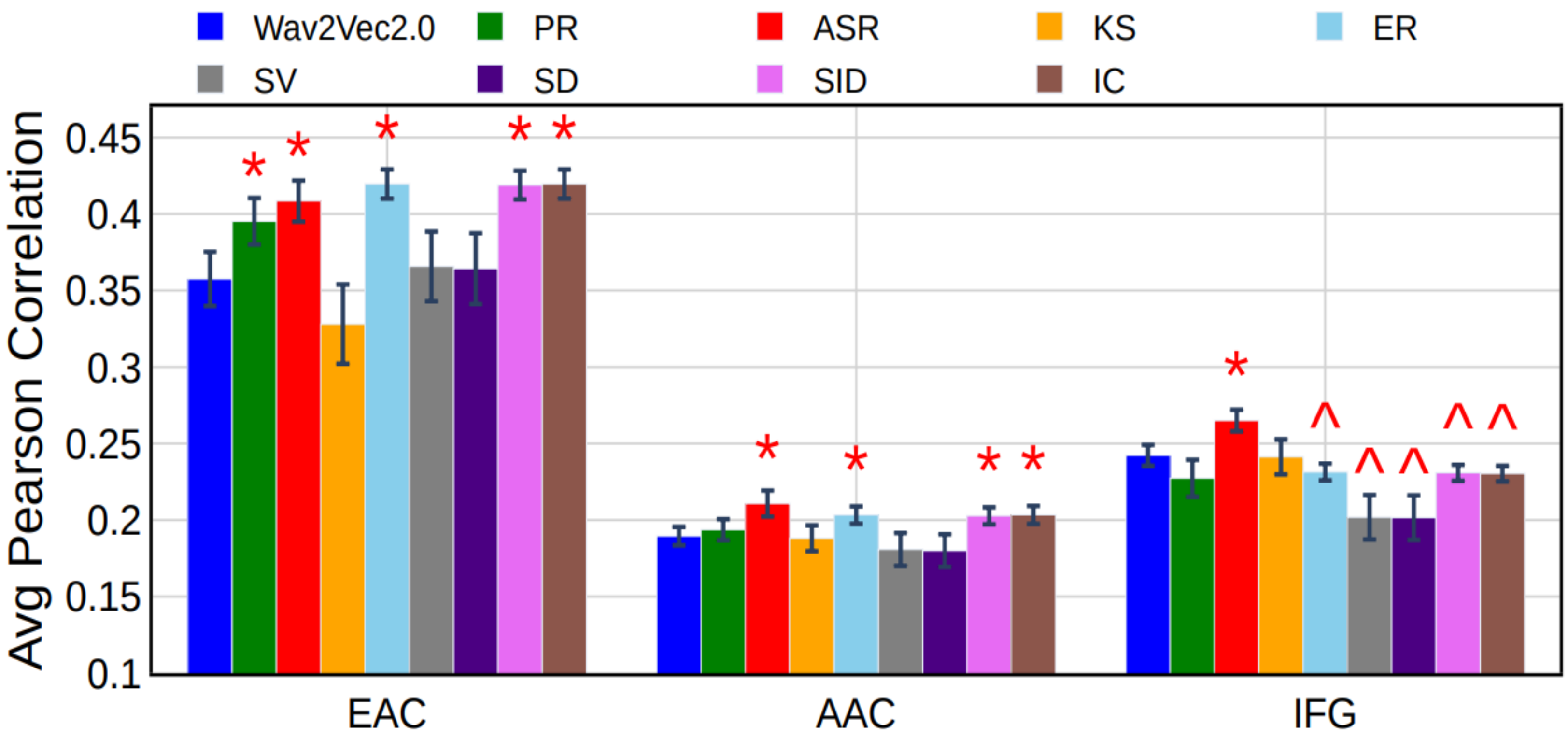


ASR task has the best brain alignment in the middle layers.

- Certain speech tasks (ASR, ER, SID and IC) that are important for improved brain alignment over pretrained Wav2Vec2.0.
- SD and SV are not important in listening to stories.

Region level alignments

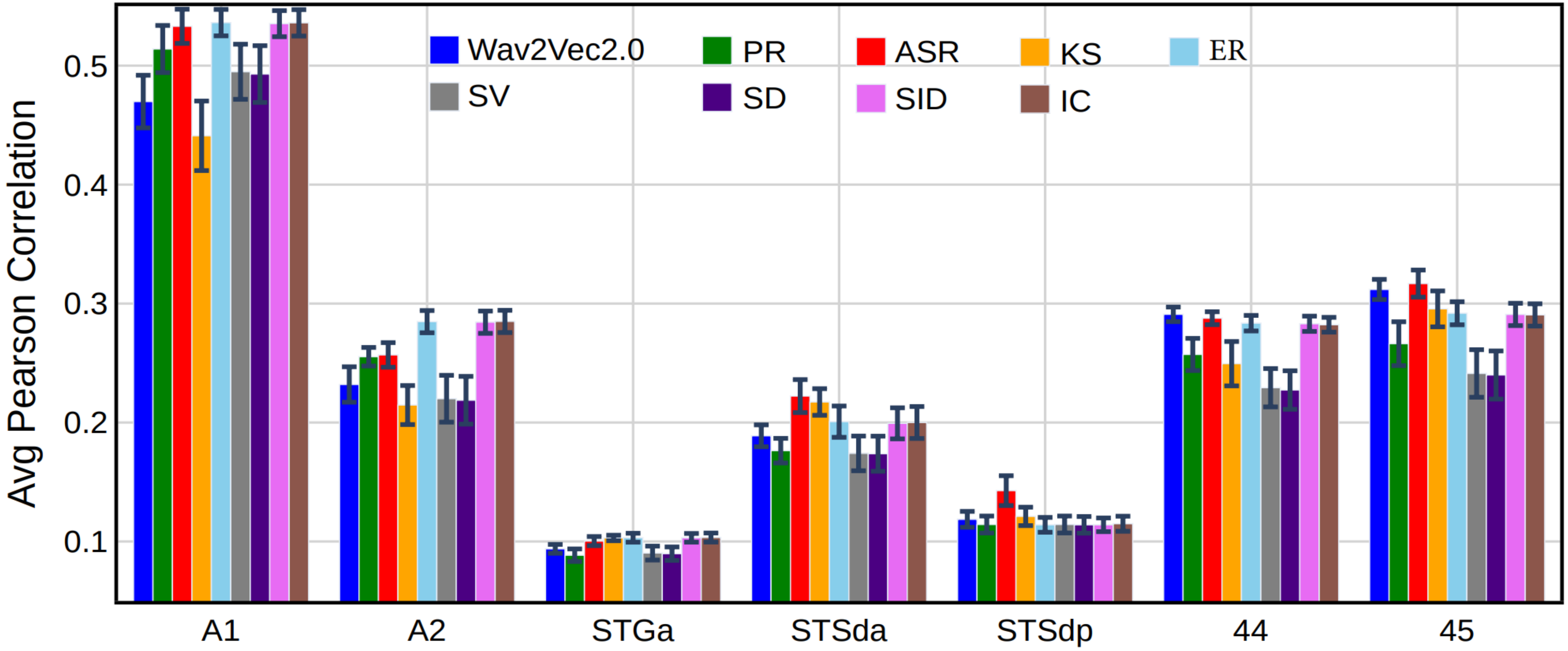
- All speech tasks are better aligned with EAC compared to AAC and IFG regions.
- Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex
- Finetuning on ASR provides the best encoding for the auditory associative cortex and language regions.



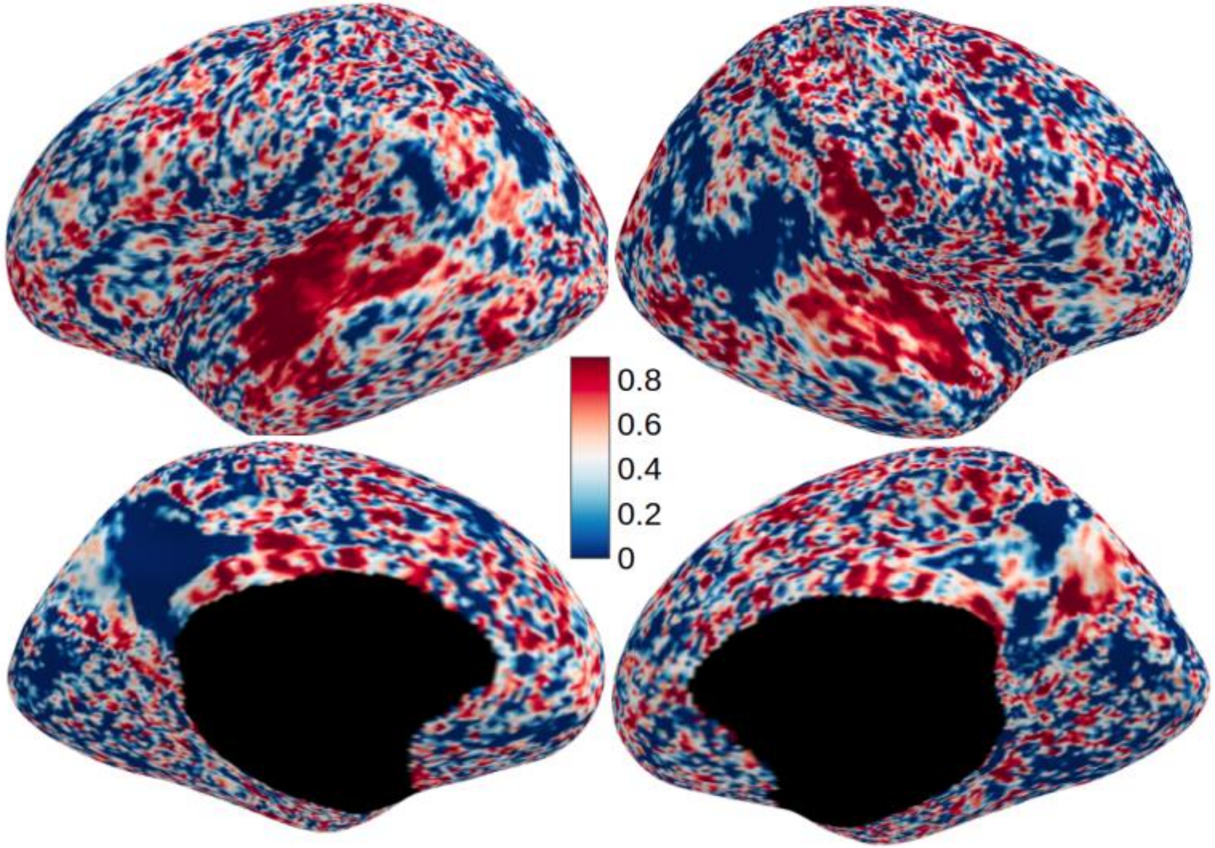
Speech Taskonomy: Which Speech Tasks are the most Predictive of fMRI Brain Activity? Subba Reddy Oota, Veeral Agarwal, Mounika Marreddy, Manish Gupta, Raju Bapi. InterSpeech 2023

Sub-region level alignments

- EAC: A1 has a higher Pearson correlation than other sub-ROIs.
- Language ROIs 44 and 45, together with STSda and STSdp in AAC, are part of the well-known language network associated with narrative comprehension; ASR finetuned model performs best in these regions.



Qualitative Analysis: Brain Maps



Voxel-wise correlation values for the brain alignment of pretrained Wav2Vec2.0 and ASR

- Correlation is high in temporal lobes but not in language and parietal regions.
- Low correlations in some regions indicate that finetuning changes predictions for those regions.
- Perhaps that is why, like language models, the ASR model also has the best performance for middle layers

Limitations & Future Works

- We leveraged models finetuned using datasets of different sizes across tasks.
- While a fair comparison of dataset sizes across tasks is impossible,
  - we understand that this could have resulted in some bias in our results.

Thank You

Questions?

Inria
inventeurs du monde numérique

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Microsoft