MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

SIC Saarland Informatics Campus

Improving Semantic Understanding in Speech Language Models via Brain-tuning

Omer Moussa, Dietrich Klakow, Mariya Toneva

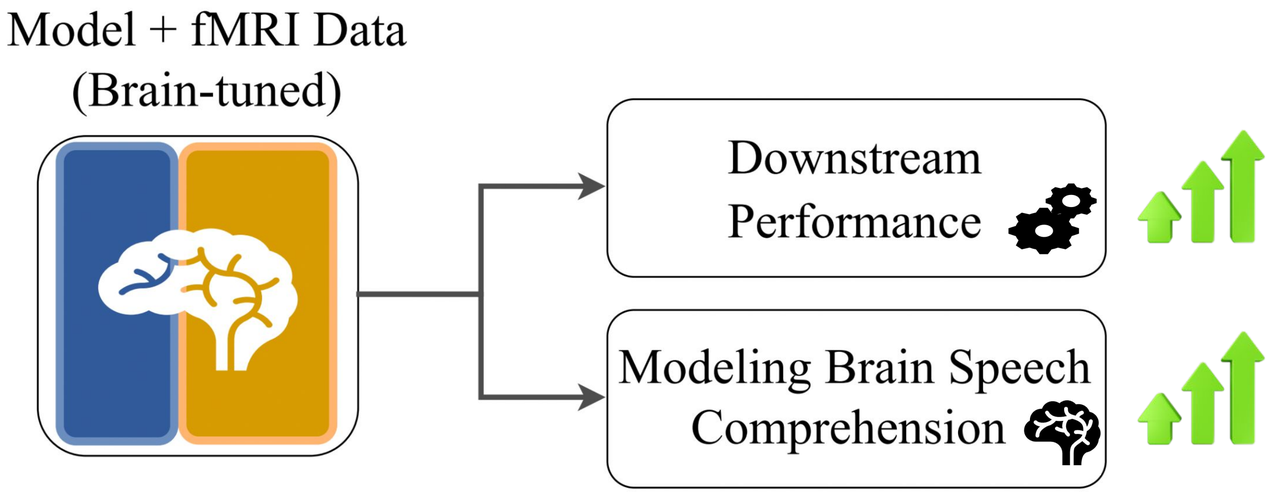
ICLR
Apr 24 – 28, 2025

https://iclr.cc/virtual/2025/poster/30063

Contributions

First work to show that incorporating **brain signals** into the training of language models **(Brain-tuning)** improves their **semantic understanding.**

Training to increase **alignment with the human brain** enhances **downstream** performance and leads to improved model of speech comprehension in the **brain.**

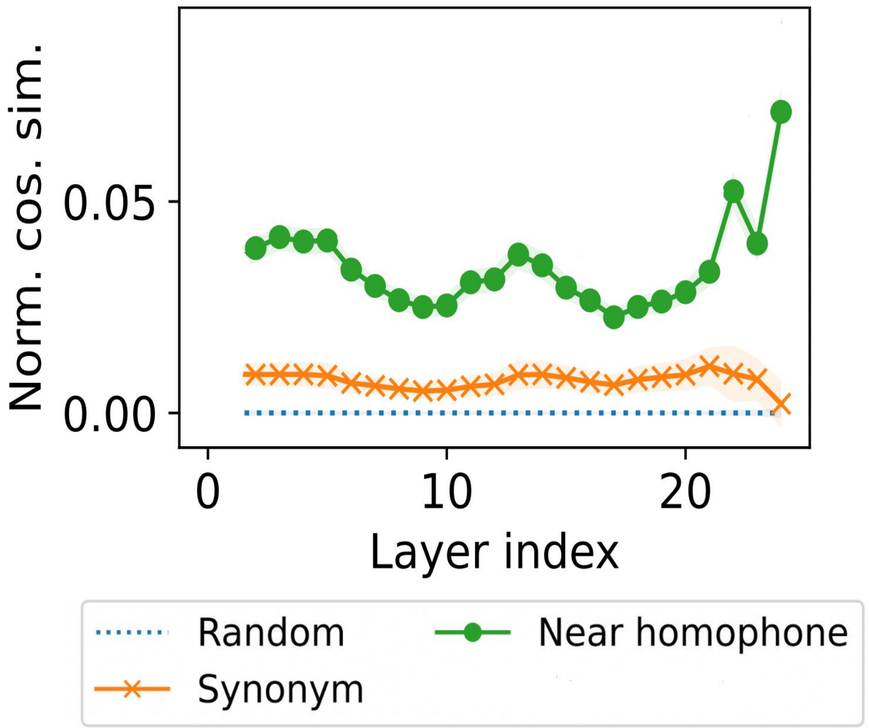


Current Speech Models Lack Semantics

ICLR
Apr 24 – 28, 2025

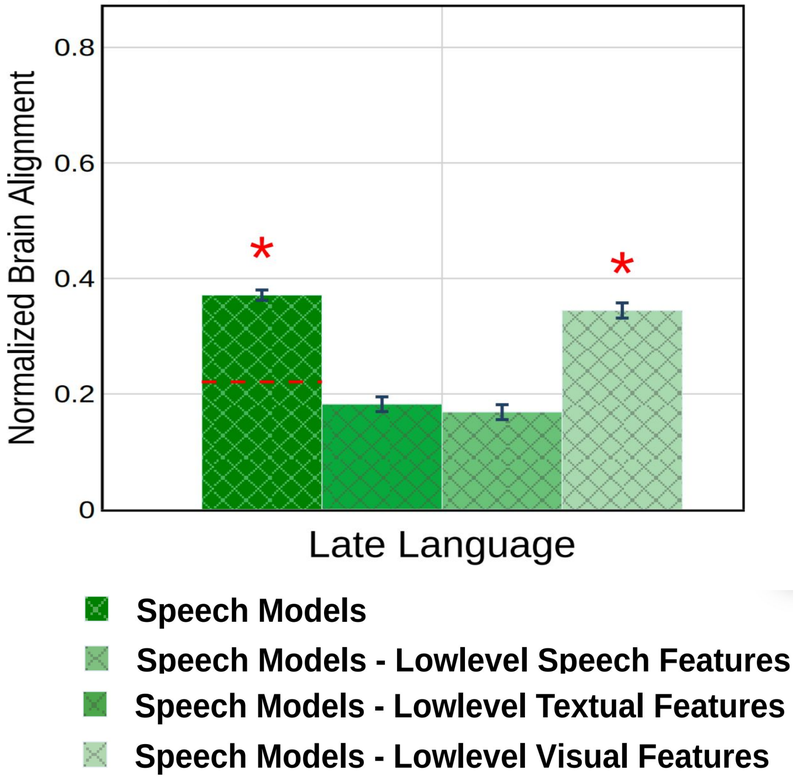
Representations are phonetic not semantic

Word representation is more similar to homophones[1]
(prefer closer sound over closer meaning)



Models lack brain-relevant semantics

Brain alignment vanishes when removing low-level info[2]
(High reliance on low-level features)



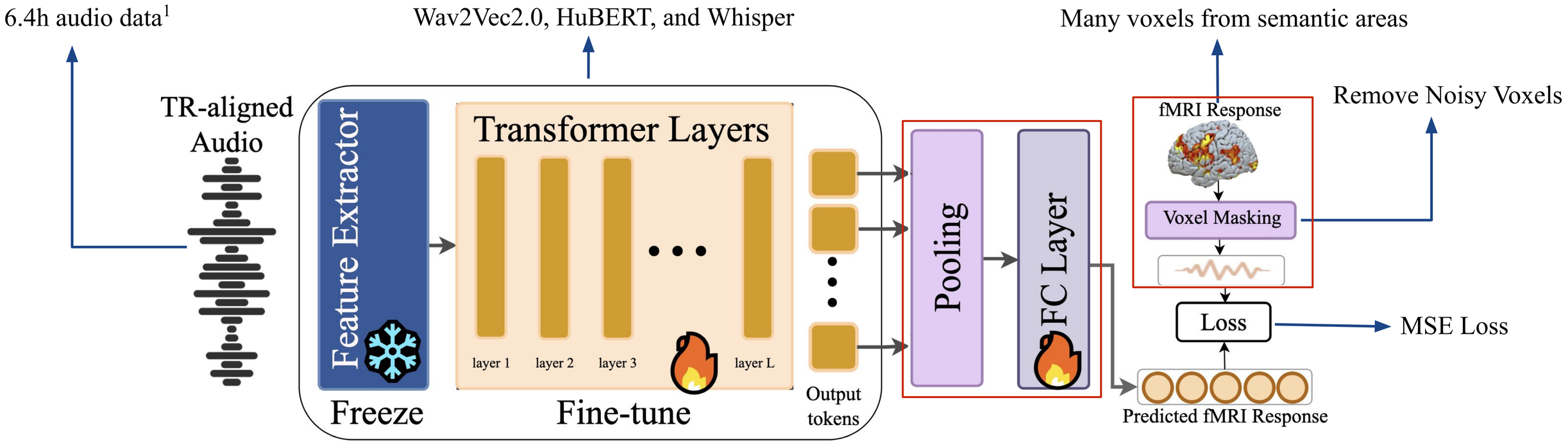
[1]Choi, Kwanghee, et al. 2024

[2]Oota, et al. ACL 2024

Proposed Brain-tuning Approach

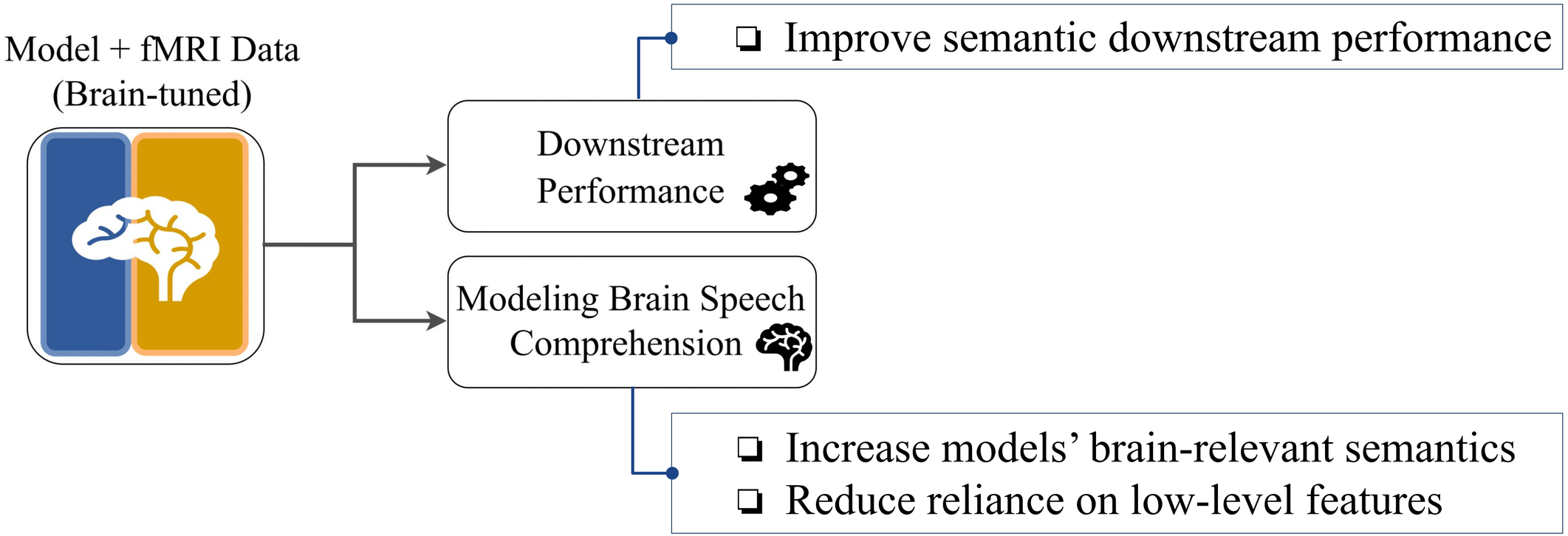
ICLR
Apr 24 – 28, 2025

Brain-tuning: inducing brain relevant bias directly into the model by fine-tuning with brain **fMRI data**.



[1]LeBel et al., 2024

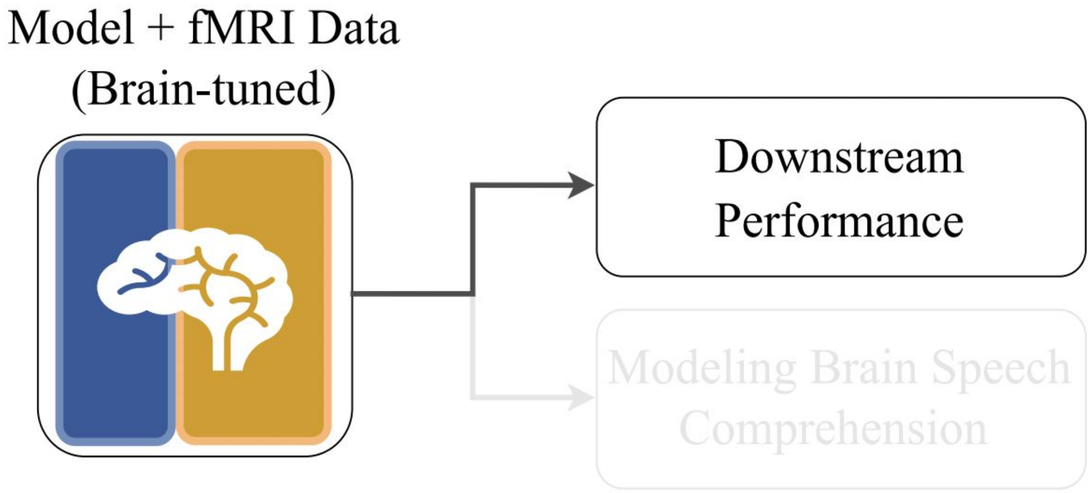
Goals of Brain-tuning

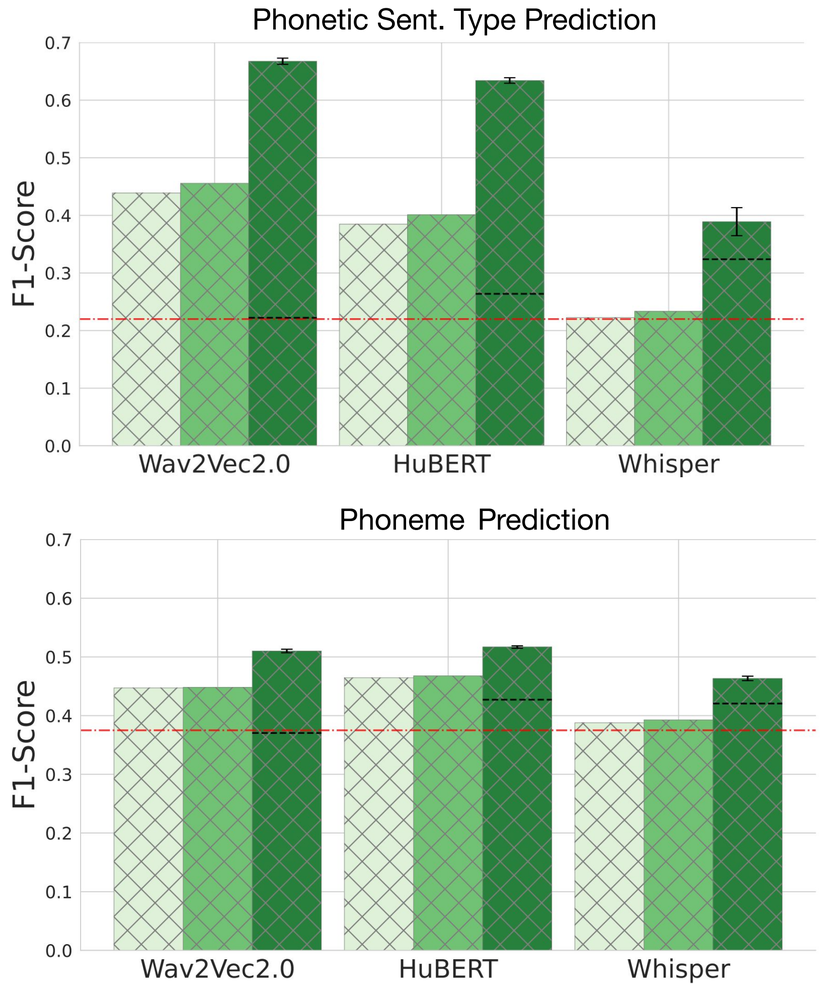


[1]LeBel et al., 2024

Brain-tuning Downstream Results

☑ Improve semantic downstream performance



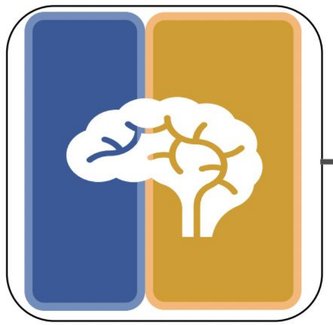


Brain-tuning Brain Alignment Results

ICLR
Apr 24 – 28, 2025

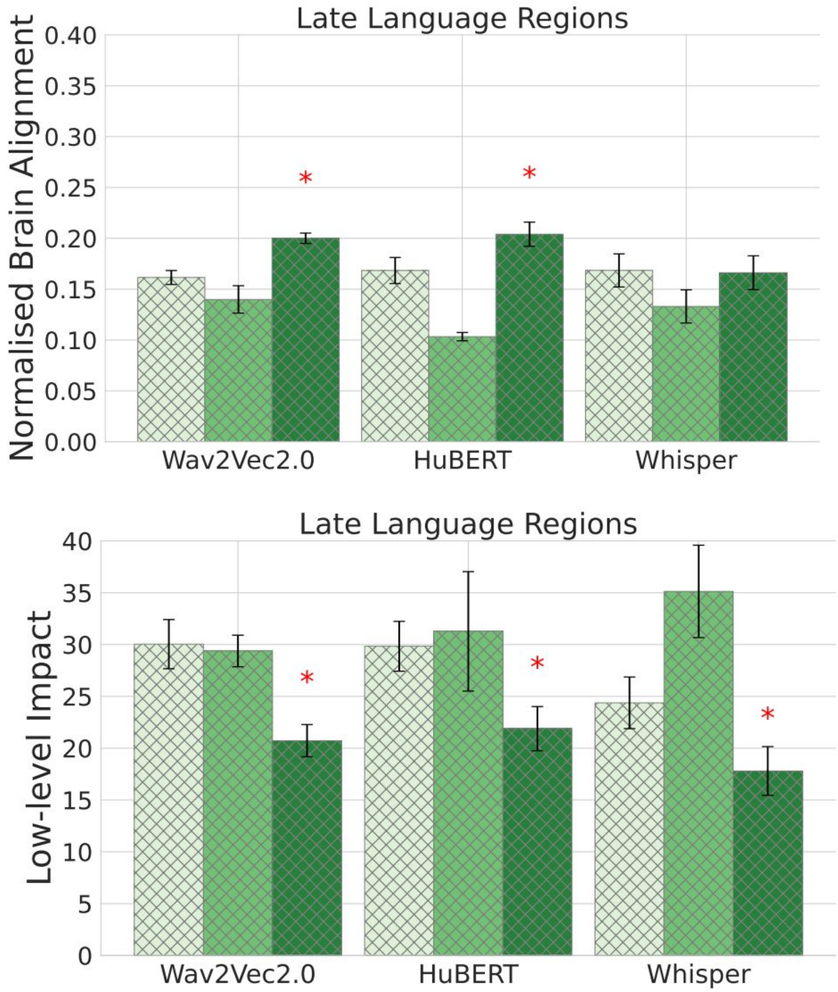
- ☑ Increase models' brain-relevant semantics
- ☑ Reduce reliance on low-level features

Model + fMRI Data
(Brain-tuned)



Downstream Performance

Modeling Brain Speech Comprehension



Conclusion and Future Directions

ICLR
Apr 24 – 28, 2025

Clear increase in semantic understanding

- Only **0.7%** more data (relative to pretraining data)
- Consistent across Model Families

First work to show substantial semantic downstream gains with brain data

Leads to better model organisms for auditory language processing in the brain

Future work: scaling to more data, bigger models, and different loss functions

Thank you!

ICLR
Apr 24 – 28, 2025

Questions?

ICLR Poster: Sat 26 Apr 10 a.m. — 12:30 p.m.

Feel Free to reach us at:

omoussa@mpi-sws.org
mtoneva@mpi-sws.org

https://iclr.cc/virtual/2025/poster/30063

References

- Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. ”an fmri dataset during a passive natural language listening task”. 2024. doi: doi:10.18112/openneuro.ds003020.v2.2.0.
- Subba Reddy Oota, Emin C¸ elik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. ACL, 2024a.
- Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. Advances in Neural Information Processing Systems, 36, 2024b
- Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. Self-supervised speech representations are more phonetic than semantic, 2024. URL https://arxiv.org/abs/2406.08619.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460, 2020.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pp. 28492–28518. PMLR, 2023
- Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. Advances in neural information processing systems, 32, 2019.