Intracranial EEG reveals multiplexed encoding of auditory, speech, and language embeddings in the human temporal lobe

Atlas Kazemian (atlaskaz@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305, United States

Josef Parvizi (jparvizi@stanford.edu)
Department of Neurology, Stanford School of Medicine
Stanford, CA 94305

Daniel K. Yamins (yamins@stanford.edu)
Department of Psychology and Computer Science, Stanford University
Stanford, CA 94305

Laura Gwilliams (laura.gwilliams@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305

Abstract

Speech understanding in the human brain involves several representational transformations: Air pressure fluctuations become a time-frequency representation in the cochlear; auditory cortex extracts speech-relevant units; distributed networks extract amodal representations of meaning and structure. Recent advances in speech and language models have led to a series of studies using text-based large language model (LLM) representations to model these neural transformations. Here we examine the brain's end-to-end processing of language by extending the investigation to include biophysical models of the cochlear and auditory cortex, as well as performance-optimized models of speech (Whisper) and text (GPT-2, Llama-3). We use these models to predict the activity of spatially precise neural populations recorded via intracranial EEG (iEEG) as participants listened to audiobooks. Our findings are twofold. First, we observe clear differences in encoding performance within both auditory and language model families. Second, each model type captures distinct aspects of the signal in temporal lobe electrodes, suggesting that these regions encode a mixture of intermediate auditory and higher-level semantic features. Together, these results highlight the importance of examining model–brain alignment with fine-grained temporal precision.

Keywords: Temporal Dynamics; Encoding Models; Speech Models; Large Language Models; Speech Processing; Intracranial EEG

Introduction

Large Language Models (LLMs) have shown high representational similarity to language-evoked neural responses (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Tuckute et al., 2024). Likewise, state-of-the-art speech-based models have demonstrated high encoding performance in auditory and speech-selective brain regions (Millet et al., 2022; Goldstein et al., 2023). Yet, questions remain about how the brain encodes features of varying complexity during speech processing—from low-level acoustics to high-level semantics. Building on prior research, we examine how models with different levels of representational complexity predict time-resolved neural data during speech processing.

Methods

We use intracranial EEG (iEEG) recordings of five subjects listening to 30 minutes of audiobook snippets. We focused on 344 electrode contacts, which broadly sample brain regions considered to be involved in speech and language processing (Fedorenko, Ivanova, & Regev, 2024). We extracted LFP response amplitude, and applied a low-pass filter of 20 Hz. The data were then downsampled to 40 kHz, and epoched from -500 ms to +1000 ms around word onset.

Each audio segment was repeated twice, and the repeats were used to estimate a time-resolved noise ceiling using Pearson correlation between the response at a given electrode, at each time lag, across repetitions of the same words. For the analysis of the encoding model, we compared several acoustic, speech, and language encoding models to examine how neural responses encode representations of varying sensory and semantic complexity, at distinct moments in time. The acoustic embeddings were derived from a cochleagram model (Feather, Leclerc, Madry, & McDermott, 2023). For speech embeddings, we use a supervised auditory model with a ResNet50 (He, Zhang, Ren, & Sun, 2015) backbone and a cochleagram front end (CochDNN) trained for word recognition (Tuckute, Feather, Boebinger, & McDermott, 2023). In addition, we use embeddings obtained from the encoding layers of the Whisper model, which are trained for automatic speech recognition using large-scale, weakly supervised audio-text pairs (Radford et al., 2022). Language embeddings were extracted from GPT-2 (Radford et al., 2019) and Llama3 (Touvron et al., 2024), as well as from the decoder layers of Whisper, which are optimized for generating text.

For each electrode and each time sample, we fit an L2-regularized Ridge regression to predict neural activity from model embeddings across all words in the audiobooks. Performance was evaluated as the Pearson correlation between the actual and predicted neural responses as a function of time relative to word onset.

Results

Figure 1 shows heatmap plots of encoding performance for all models across time. Each row corresponds to one recording site in one subject, and the rows are ordered by region and sorted based on score.

In line with previous findings (Tuckute et al., 2023), we show that CochDNN substantially outperforms the cochleagram model, suggesting that it captures non-trivial auditory features beyond the basic acoustic structure of the signal. Within the language models, LLaMA-3 (8B) reduces the gap with the noise ceiling prior to word onset, compared to GPT-2, suggesting improved encoding of predictive context. Notably, the Whisper encoder exhibits a hybrid performance profile, showing characteristics of both auditory and language models, while the Whisper decoder more closely resembles the temporal dynamics of language models.

To assess whether the auditory and language models capture shared or distinct aspects of the neural response, we first build a joint encoding model by concatenating CochDNN and LLaMA-3 (8B) features and fitting them simultaneously (Figure 2). We then performed variance partitioning to isolate each model's unique contribution. As shown in Figure 2, CochDNN and LLaMA-3 each explain non-overlapping variance across different time lags, demonstrating that they capture distinct components of the neural signal.

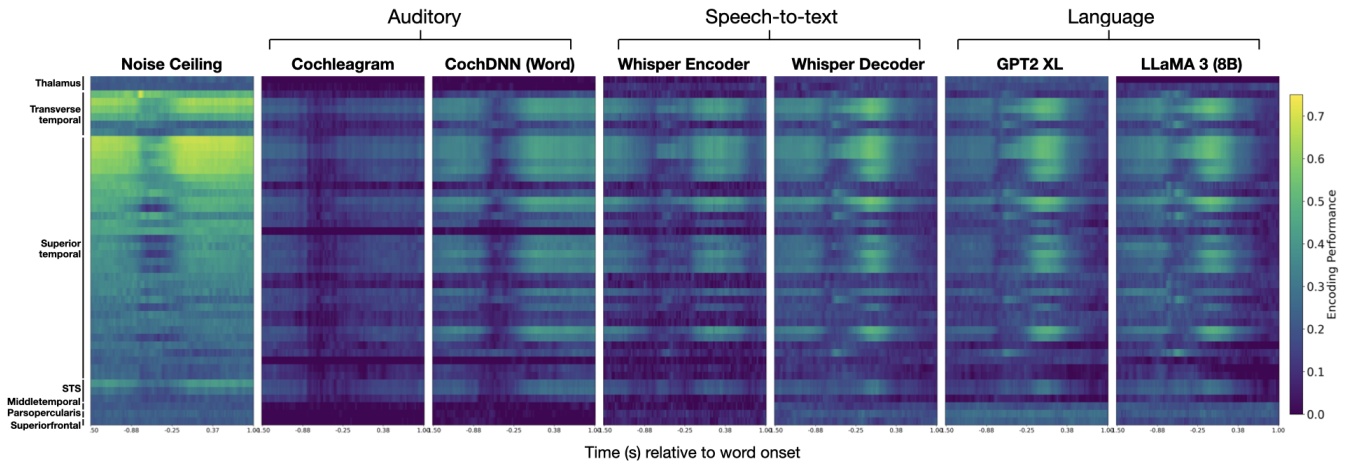

Figure 1: Encoding performance over time relative to word onset for each model, measured as the Pearson correlation between model activations and neural responses at the best-performing layer per electrode. Each row in the heatmap represents the encoding performance for an electrode in a subject. Rows are ordered by region and sorted by noise ceiling value within region. Heatmap colors represent Pearson Correlation values.

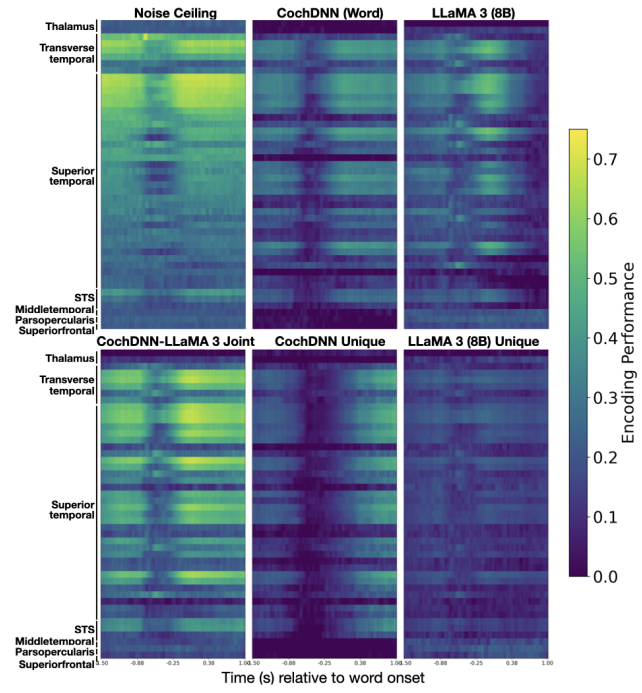

Figure 2: Variance partitioning for CochDNN and LLaMA-3 (8B). The encoding performance of the two models is compared against their joint performance as well as the unique score from each model. The rows correspond to the same electrode as in Figure 1.

Discussion

We aimed to track the evolution of auditory, speech, and language representations in the human brain by leveraging models optimized for each of these domains. By comparing the temporal structure of model performance relative to the neural noise ceiling, we find clear distinctions even among models of the same type. Importantly, we observe a substantial difference in encoding performance prior to word onset between an older generation language model, GPT-2, and a more recent model, LLaMA-3. This effect may be obscured in datasets with limited temporal resolution. Further, our fig$_2$*showthatauditoryandlanguagemodelscapturedistinctcomponentsofth levelsemanticinformation.Leveragingmodelsthatspanarangeoffeaturety*

Taken together, our approach provides a more comprehensive account of how encoding models capture speech and language processing during natural language comprehension, highlighting the importance of fine-grained temporal analysis for understanding model–brain alignment.

References

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

Feather, J., Leclerc, G., Madry, A., & McDermott, J. H. (2023, October 16). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*, 2017–2034. (Open access)

Fedorenko, E., Ivanova, A. A., & Regev, T. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*.

Goldstein, A., Wang, H., Niekerken, L., Zada, Z., Aubrey, B., Sheffer, T., ... Hasson, U. (2023). Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations. *bioRxiv*. (Preprint) doi: 10.1101/2023.06.26.546557

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

He, K., Zhang, X., Ren, S., & Sun, J. (2015, Dec). *Deep residual learning for image recognition.* ArXiv:1512.03385 [cs] [Internet]. (Cited 2022 Feb 8. Available from: http://arxiv.org/abs/1512.03385)

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., ... King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv preprint arXiv:2202.01234*. Retrieved from https://arxiv.org/abs/2202.01234

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* https://arxiv.org/abs/2212.04356. (Submitted on 6 Dec 2022; arXiv preprint)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* (OpenAI Blog)

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Touvron, H., Albert, P., Bojanowski, P., Douze, M., Goyal, N., Izacard, G., ... Others (2024). *The llama 3 herd of models.* https://arxiv.org/abs/2407.21783. (arXiv preprint arXiv:2407.21783)

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Computational Biology*. (Research article; authors Greta Tuckute and Jenelle Feather contributed equally to this work)

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., ... Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*, 544–561.