Improving Neural Decoding by Integrating Information Over Time

Zhuoyang (Gio) Li (zli230@jhu.edu)
Department of Neuroscience, Johns Hopkins University

Kristijan Armeni (karmeni1@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Christopher J. Honey (chris.honey@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Abstract

Deep language models (DLMs) provide a powerful basis for building encoding models and decoding models of neural responses. However, DLMs are usually only used to predict a single moment in the neural signal. We reasoned that decoding performance might be improved by combining information across time, and leveraging the temporal dependencies of neural activity. To test this, we analyzed word-level decoding from electrocorticography (ECoG) recordings of 9 participants who listened to a 7-minute narrative. We found that DLM-based encoding models could predict neural responses seconds before and after a word onset, and that the predictions did not generalize across time intervals around word onset. Moreover, we were able to boost decoder performance by integrating information across distinct time intervals. Thus, human brains represent diverse word-related information for hundreds of milliseconds before and after word onset, and ensembling information over time is a promising approach for naturalistic neural decoding.

Keywords: ECoG; language models; neural decoding; temporal integration

Introduction

How can we best decode mental states from neural signals? Solving this problem is not only of theoretical interest, but also has wide-ranging practical and therapeutic implications (Anumanchipalli, Chartier, & Chang, 2019). Contextual embeddings extracted from autoregressive deep language models (DLMs) were used to model human brain's responses to naturalistic narratives (Goldstein et al., 2022; Caucheteux, Gramfort, & King, 2022). Notably, these encoding models can predict neural responses for several seconds before and after the onset of a word. Although some recent models aggregate neural data across time points to decode linguistic content (Goldstein et al., 2022; Tang, LeBel, Jain, & Huth, 2023), it remains unclear (i) whether neural signals at different time lags surrounding word onsets encode mostly redundant or distinct information and (ii) which aspects of language decoding are boosted by integrating across time. To fill this gap, here we apply encoding and decoding analyses to human electrocorticography (ECoG) responses to a spoken narrative, in order to characterize the time-varying nature of language responses.

Methods

Preprocessing

Raw ECoG recordings from 9 participants were re-referenced (common average) and high-pass filtered at 0.1 Hz, following exclusion of electrodes with visually apparent artifact or noise. We used six-cycle Morlet wavelets to estimate the spectral power at each of [70, 75, 80, 85, ..., 200] Hz. The power time-series were log-transformed, z-scored and then mean-averaged to obtain a single ”broadband” high frequency power estimate (Goldstein et al., 2022). To focus on electrodes with robust stimulus-locked signal, we selected from each participant the 3 electrodes with highest repeat reliability (Pearson correlation of ECoG responses to two repeats of the same narrative; Fig. 1, A), for a total of 27 ECoG electrodes.

Encoding analysis

To predict neural responses ($\mathbf{r}$) from contextual embeddings ($\mathbf{s}$), we implemented a linear ridge regression model ($f(\mathbf{s})$) for each time lag relative to word onset (Goldstein et al., 2022). Word embeddings were extracted from the hidden representations in the 7th layer (Caucheteux et al., 2022) of a 12-layer transformer language model (GPT-2, Radford et al., 2019). We reduced the dimensionality of embeddings from 768 to 30 using principal component analysis. The neural responses were averaged across a 200-ms window for 41 lags from -4,000 to 4,000 ms in 200-ms steps relative to word onset. We used 10-fold cross-validation to split the data ($N^{samples} = 1002$) into training and test sets. The correlation between actual and predicted neural responses in test data was averaged across the 10 test-folds. We evaluated encoding models' generalization performance across time lags by computing the Pearson correlation between the actual neural signal at one lag and the predicted signal based on encoding weights obtained at another lag (King & Dehaene, 2014). The encoding performance and temporal generalization performance were evaluated in each electrode and then averaged across the 3 electrodes for each participant. Finally, we computed the grand mean correlation across the 9 participants.

Decoding analysis

We leveraged the contextual encoding models to decode word identities from neural responses surrounding word onset. We computed the likelihood of observing the neural responses $\mathbf{r}$ given a word embedding $\mathbf{s}$ as $P(\mathbf{r}|\mathbf{s}) \propto \exp\{-[\mathbf{r} -$

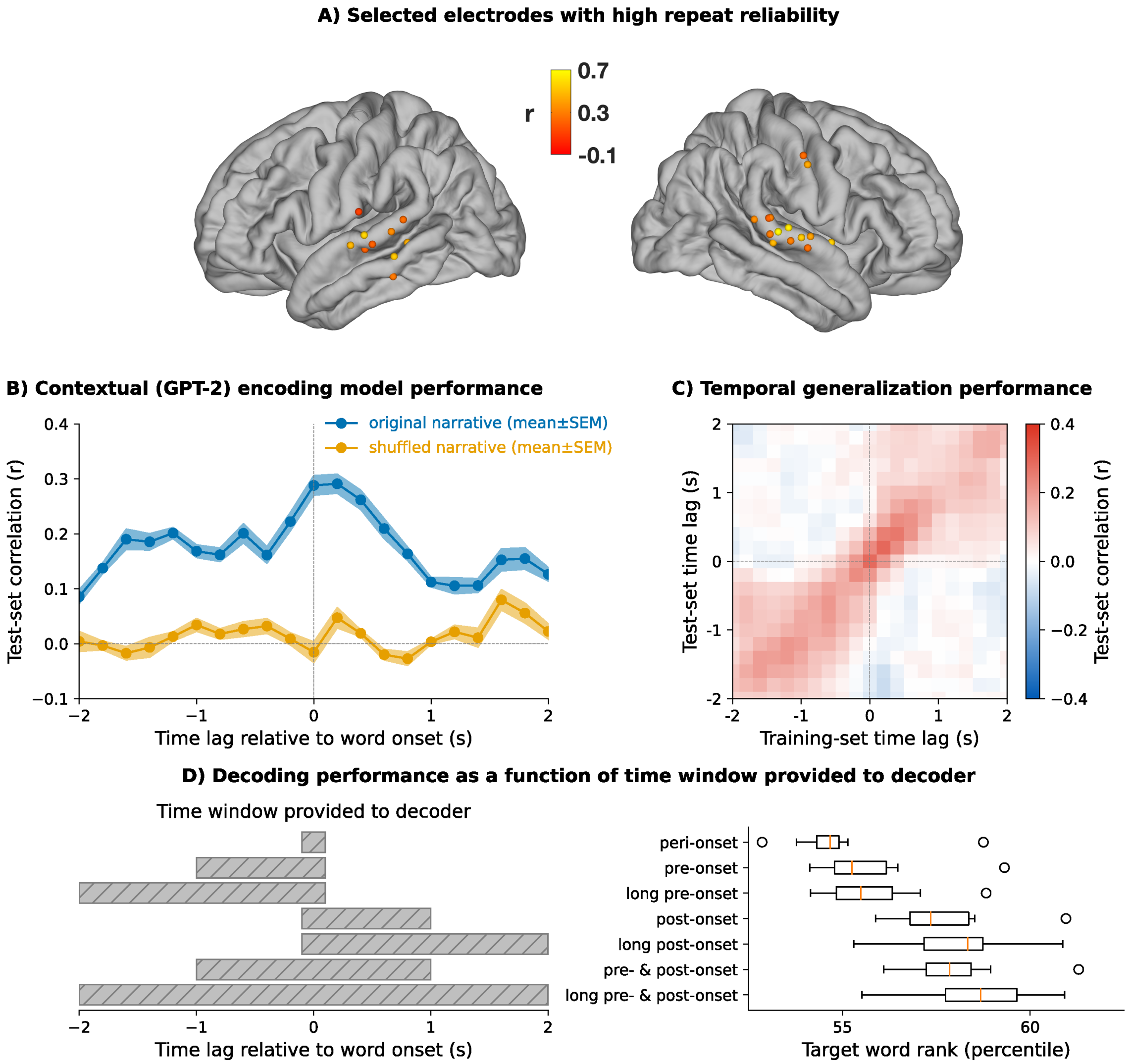

Figure 1: A) Selected electrodes with high repeat reliability; B) Cross-validated encoding performance as a function of time lag relative to word onset. The shaded regions indicate the s.e.m. of encoding model performance across participants; C) Temporal generalization performance of the encoding models; D) Decoding performance as a function of the time window provided.

$f(\mathbf{s})]\Sigma^{-1}[\mathbf{r} - f(\mathbf{s})]^T\}$, where Σ is the noise covariance matrix (Nishimoto et al., 2011; Tang et al., 2023). For each unique word in the story we generated a likelihood using the model. We then calculated the decoding score for each target word as the fraction of words with a lower likelihood than that assigned to the target word. The decoding analysis was performed for windows of varying lengths relative to word onset, providing a measurement for how well the word could be decoded by aggregating ECoG signals across lags around word onset.

Results and Discussion

Using the contextual embeddings from GPT-2, the model predicted neural responses with a correlation greater than 0.3 at word onset. Lower, but still statistically reliable prediction was obtained out to 2 seconds before and after the onset of words (Fig. 1, B). The encoding models trained on neural data before word onsets did not generalize to neural responses after word onsets, and vice-versa (Fig. 1, C, off-diagonal quadrants versus diagonal quadrants, $p < .001$). Thus, pre- and post-onset time periods encode distinct and non-redundant information. This is consistent with the observation that decoding performance was improved by including neural responses both before and after word onset (Fig. 1, D). The decoder with access to neural signals out to 2 seconds before and after word onset significantly outperformed the decoder only using signals 100ms before and after word onset, $t(8) = 5.88, p < .001$.

During naturalistic listening, the human brain encodes distinct word-related information before and after word onset. Aggregating information over time significantly improved the ability to decode words from neural responses. We are now investigating the distinct lexical and semantic properties that are available before, around, and after word onsets.

Acknowledgments

This work was supported by the National Institute of Mental Health (P50-MH109429).

References

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493–498.

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), 16327.

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.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, *18*(4), 203–210.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866.