scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A natural language fMRI dataset for voxelwise encoding models

Amanda LeBel[1], Lauren Wagner[2], Shailee Jain[3], Aneesh Adhikari-Desai[3,4], Bhavin Gupta[3], Allyson Morgenthal[4], Jerry Tang[3], Lixiang Xu[5] & Alexander G. Huth[3,4]✉

Speech comprehension is a complex process that draws on humans' abilities to extract lexical information, parse syntax, and form semantic understanding. These sub-processes have traditionally been studied using separate neuroimaging experiments that attempt to isolate specific effects of interest. More recently it has become possible to study all stages of language comprehension in a single neuroimaging experiment using narrative natural language stimuli. The resulting data are richly varied at every level, enabling analyses that can probe everything from spectral representations to high-level representations of semantic meaning. We provide a dataset containing BOLD fMRI responses recorded while 8 participants each listened to 27 complete, natural, narrative stories (~6 hours). This dataset includes pre-processed and raw MRIs, as well as hand-constructed 3D cortical surfaces for each participant. To address the challenges of analyzing naturalistic data, this dataset is accompanied by a python library containing basic code for creating voxelwise encoding models. Altogether, this dataset provides a large and novel resource for understanding speech and language processing in the human brain.

Background & Summary

Historically, MRI has been used to study the structural and functional organization of the brain by way of highly controlled paradigms and simplified stimuli. This has also been true in language neuroscience (i.e. using block designs[1,2]) where it is common to use isolated words[1,3] or simple sentences[4–7] as experimental stimuli. While these paradigms have proven useful, they also present several issues. First, their hypothesis-driven design limits the number of scientific questions one can ask using a given dataset. Second, isolated words and sentences are devoid of context, which is a critical component of language understanding in the real world. Thus, results obtained on these simple stimuli may not generalize to natural language perception. And third, small stimulus sets limit the breadth of features sampled by the stimuli. This is problematic for studying complex, high-dimensional feature spaces such as semantics.

An alternative approach is to use natural stimuli that closely approximate language as it is used in everyday life. *Natural language* is language used in real world settings such as conversation, entertainment, and education. Natural language can also include multiple modalities, such as vision for written or signed language and audition for spoken language[8]. Our dataset[9] focuses on one specific subset of natural language: spoken English in the form of complete narrative stories from *The Moth* podcast. This permits detailed study of the auditory speech processing as well as core amodal language systems. While narrow compared to the full breadth of natural language—these stories are largely focused on personal experiences, and contain no dialogue—these stories are still highly varied in semantic and syntactic content. There is also considerable evidence from earlier experiments that such natural stories broadly activate cortex and can be used to study a variety of phenomena[10–17].

One downside of natural language stimuli is the difficulty of analyzing and interpreting the resulting data. Data from controlled, block-design experiments can be analyzed using standard methods such as t-tests, f-tests, and ANOVAs. In natural language, however, the features of interest (e.g. a particular phoneme, or topic) are distributed throughout the stimulus. To model how the brain responds to these features—and account for correlations among them—we use *voxelwise encoding models* that are designed to predict brain responses from the stimuli[18]. The first step in creating encoding models is to extract features of interest from the stimuli. Previous work on this fMRI

[1]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, 94704, USA. [2]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, 90095, USA. [3]Department of Computer Science, The University of Texas at Austin, Austin, TX, 78712, USA. [4]Department of Neuroscience, The University of Texas at Austin, Austin, TX, 78712, USA. [5]Department of Physics, The University of Texas at Austin, Austin, TX, 78712, USA. ✉e-mail: huth@cs.utexas.edu

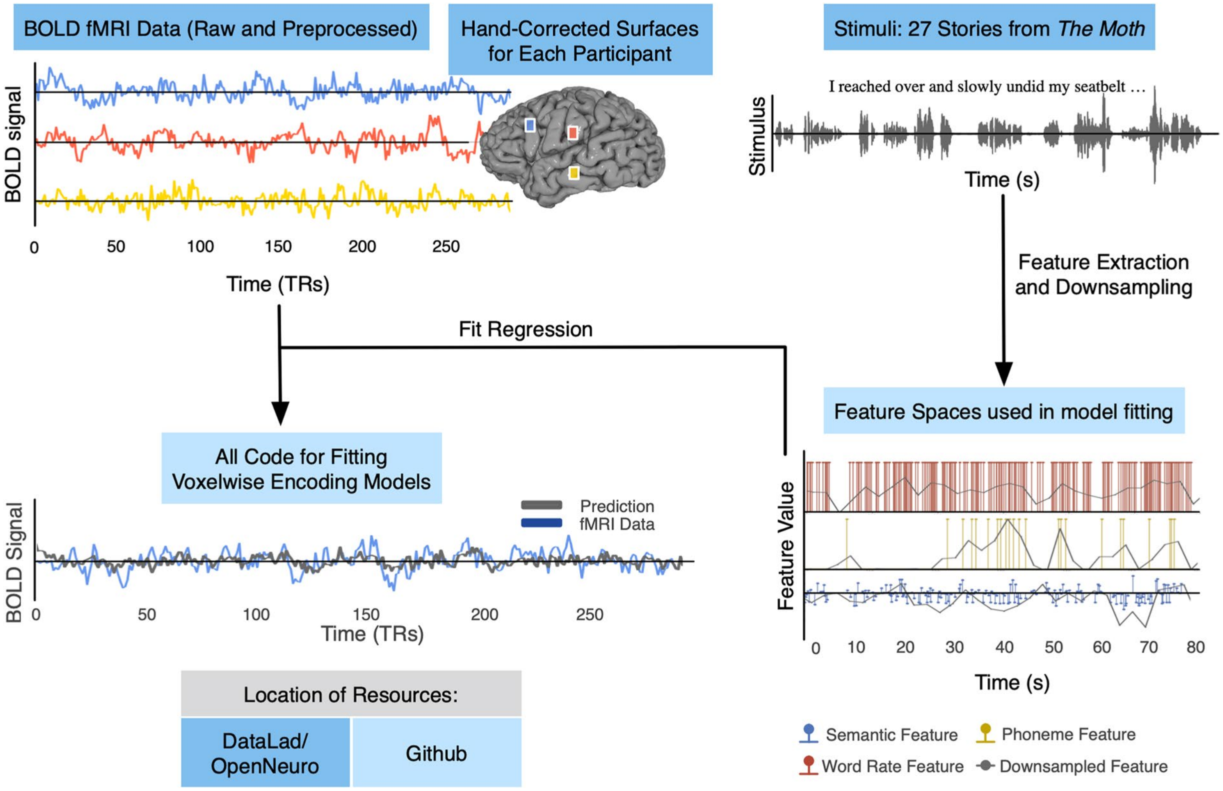
BOLD fMRI Data (Raw and Preprocessed)
Hand-Corrected Surfaces for Each Participant
Stimuli: 27 Stories from The Moth
BOLD signal
0
50
100
150
200
250
Time (TRs)
Stimulus
I reached over and slowly undid my seatbelt …
Time (s)
Feature Extraction and Downsampling
Fit Regression
Feature Spaces used in model fitting
All Code for Fitting Voxelwise Encoding Models
Prediction
fMRI Data
BOLD Signal
Feature Value
0
10
20
30
40
50
60
70
80
Time (s)
Location of Resources:
DataLad/ OpenNeuro
Github
Semantic Feature
Phoneme Feature
Word Rate Feature
Downsampled Feature

Fig. 1 Schematic of naturalistic story-listening paradigm and available resources. 27 unique natural stories from *The Moth* podcast were played for eight participants over five fMRI sessions while they were instructed to passively listen. One of these 27 stories was played in each of the 5 sessions. No other story was repeated. These stimuli can be converted to previously used feature spaces for model fitting, including semantic, phoneme, and word rate feature spaces[10,11,14]. Regularized regression can then be used to fit voxelwise encoding models that use the features to predict BOLD data. Model performance can then be evaluated on a held out dataset. Available resources on OpenNeuro include the stimuli, BOLD data, and hand-corrected surfaces for each of the eight participants. Available resources on GitHub include the feature spaces and code for fitting the encoding models.

dataset[9] has used spectral, articulatory, part-of-speech, and semantic features to predict BOLD responses[10,19–21]. After features are extracted for each word, phoneme, or timepoint, they are downsampled to the rate of the fMRI acquisition, delayed to account for hemodynamic response, and used in a regression model to predict the fMRI data. These models are fit separately for each voxel in each participant, providing high resolution and high fidelity.

For this dataset[9], we conducted an fMRI experiment in which eight participants passively listened to 27 complete, natural, narrative stories (370 minutes) from *The Moth* over the course of five scanning sessions. Three of these participants also listened to a further 57 complete stories (629 minutes). These extra sessions are referred to as the "extended stimulus set" and include stories from *The Moth* and *Modern Love* from *The New York Times*. While only covering a few participants, the large amount of data per participant enables more sophisticated analyses than would be possible with fewer stimuli. Each story was transcribed, aligned, and hand-checked to provide the timing of every word and phoneme. Functional localizer data for known sensorimotor, auditory, and cognitive regions was also collected, as well as high resolution T1-weighted structural scans which were used to create hand-corrected cortical surfaces for each participant. For a summary of available resources see Fig. 1 While there has been a recent boom in naturalistic neuroimaging datasets[22–24], this dataset is unique in that it contains significantly more data per individual participant than others. This allows for better characterization of high dimensional phenomena such as semantics by better covering the available space. Further, this dataset includes all necessary base code to fit standard encoding models. This is the first release of a complete dataset[9] and codebase[25] for encoding models.

Methods

Participants. Data was collected from 8 participants (three female): UTS01 (female, age 24), UTS02 (male, age 34), UTS03 (male, age 21), UTS04 (male, age 31), UTS05 (female, age 24), UTS06 (female, age 23), UTS07 (male, age 25), UTS08 (male, age 24). All participants were healthy and had normal hearing. The experimental protocol (protocol # 2017-07-0030) was approved by the Institutional Review Board at the University of Texas at Austin, and written informed consent was obtained from all participants for both participation in the research and for publicly sharing the data.

Story	Author	Duration	TRs	Words	Phonemes
"Alternate Ithaca Tom"	Tom Weiser	11:47	364	2681	7531
"Souls"	Jen Lee	12:00	375	2481	6819
"Avatar"	Laura Albert	12:35	388	1952	5171
"Legacy"	Kyp Malone	13:40	420	2568	6773
"Ode to Stepfather"	Ethan Hawke	13:48	424	3300	8334
"Under the Influence"	Jeffery Rudell	10:28	324	2087	5932
"How to Draw"	Tricia Rose Burt	12:09	375	2516	6952
"My First Day with the Yankees"	Matt McGough	12:17	378	3180	8866
"Naked"	Catherine Burns	14:25	443	3747	10281
"Life"	Kimberly Reed	14:40	450	2786	7296
"Stagefright"	Suzanne Vega	10:07	314	2469	6669
"Till Death"	Cindy Chupack	11:08	344	2727	7716
"From Boyhood to Fatherhood"	Jonathan Ames	11:57	367	3262	8925
"Sloth"	Todd Hanson	14:55	458	3009	8595
"Exorcism"	Andrew Solomon	15:55	488	3471	9952
"Have You Met Him Yet"	David Litt	16:53	517	3438	10176
"A Doll's House"	Bill Burr	8:24	262	1882	5412
"In a Moment"	Sitawa Wafula	7:10	225	1157	3494
"The Closet that Ate Everything"	Morgan Zipf-Meister	10:49	335	2182	6566
"Adventures in Saying Yes"	Gina Sampaio	13:24	412	2602	7850
"Buck"	Tony Cyprien	11:25	353	1965	5455
"Swimming With Astronauts"	Michael J Massimino	13:11	405	2485	7318
"That Thing on my Arm"	Padma Lakshmi	14:49	454	2449	7080
"Eye Spy"	Michaela Murphy	12:59	399	2742	7920
"It's a Box"	Navreet Chawla	12:11	375	1989	5665
"Hang time"	Brian Gavagan	11:08	344	2226	6423
"Where there's Smoke"	Jennifer Hixson	10:02	310	2308	6068
Total Including Repeats		**6.4 Hours**	**11543 TRs**	**78893 Words**	**219511 Phonemes**

Table 1. Story stimulus from *The Moth*. All story stimuli were from *The Moth* and were hand transcribed and the transcripts were then aligned to the audio The titles and authors for all of the stories in the extended data set can be found in Supplemental Table 1.

Natural language stimulus set. The stimulus set consisted of 26 10–15 minute stories (320 minutes and 24 seconds total duration, see Table 1) from *The Moth* plus one additional 10-minute story played in each session to be used as a test dataset (50 minutes), giving a total of 370 minutes of data per participant. For three of these participants, the stimulus set contains an additional 55 stories for a total of 82 stories (949 minutes) including an additional story to be used as a test dataset. For a complete list of the stimulus including those found in the extended data set see Supplemental Table 1. For stimulus presentation, the audio for each story was filtered to correct for frequency response and phase errors induced by the headphones using calibration data provided by Sensimetrics and custom Python code (https://github.com/alexhuth/sensimetrics_filter). All stimuli were played at 44.1 kHz using the pygame library in Python.

In each story, a single speaker tells an autobiographical story without reading from a prepared script. All stories were manually transcribed by one listener. The Penn Phonetics Lab Forced Aligner (P2FA)[26] was then used to automatically align the audio to the transcript. Certain sounds (for example, laughter and breathing) were also marked to improve the accuracy of the automated alignment (see Table 2). Praat[27] phonetic analysis software was used to manually check and correct the alignment of each word within the transcript.

fMRI data collection. MRI data was collected over 6 scanning sessions (15 scanning sessions for extended dataset) on a 3 T Siemens Skyra scanner at the UT Austin Biomedical Imaging Center using a 64-channel Siemens volume coil. Anatomical data for participant UT-S-02 was collected on a 3 T Siemens TIM Trio at the Berkeley Brain Imaging Center using a 32-channel Siemens volume coil. The first session included an anatomical scan and functional localizers. Each subsequent session consisted of passively listening to 4–5 stories, plus the story used for model testing. Each story was played during one EPI scan that included padding of 10 seconds silence at the beginning and end of each story. Audio was delivered through Sensimetrics S14 in-ear piezoelectric headphones. To minimize head motion, foam headcases (CaseForge, Inc., now defunct) that precisely fill the space between the participant's head and the headcoil were used during data collection. To create the headcases, an RGB Structure.io sensor (Occipital Inc.) was used to collect a 3-dimensional scan of each participant's head while hair was compressed using a swim cap. These scans were then used to mill customized styrofoam headcases for each participant.

Sound Markings	Meaning
{CG}	Cough
{LG}	Laugh
{LS}	Lip smack
{NS}	Misc. Noise
{SL}	Silence

Table 2. Additional sounds marked in scripts. Certain sounds were included in hand aligned transcripts of the stimulus. This was done to improve alignment from the transcripts to the audio.

fMRI parameters.

Functional scans were collected using gradient-echo EPI with repetition time (TR) = 2.00 s, echo time (TE) = 30.8 ms, flip angle = 71°, multi-band factor (simultaneous multi-slice) = 2, voxel size = 2.6 mm × 2.6 mm × 2.6 mm (slice thickness = 2.6 mm), matrix size = (84, 84) and field of view = 220 mm. Field of view covered the entire cortex for all participants. Anatomical data was collected using a T1-weighted multi-echo MPRAGE sequence on the same 3 T scanner with voxel size = 1 mm × 1 mm × 1 mm following the Freesurfer morphometry protocol.

fMRI preprocessing.

fMRI preprocessing was only done on the derivative data. This data was motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FMRIB Software Library (FSL) 5.0[28]. After motion correction, all the volumes within each run were averaged to obtain a single template volume. Cross-run alignment was then performed by using FLIRT to align the template volume from each run to the template volume from the first run in the first story session. These automatic alignments were manually checked. The motion correction and cross-run transformations were then concatenated and used to resample the original data to a motion-corrected and cross-run-aligned space. This process avoids multiple resampling steps, thus minimizing unwanted blurring. Low frequency voxel response drift was then identified using a 2nd order Savitzky-Golay filter with a 120-second window and subtracted from the signal. To avoid artifacts from onset transients and poor detrending performance at the edges of the data, responses were trimmed by removing 20 seconds (10 volumes) at the beginning and end of each scan. This removed the 10-second silent period as well as the first and last 10 seconds of each story. The mean response for each voxel was then subtracted and the remaining response scaled to have unit variance.

Cortical surface reconstruction and visualization.

All anatomical data has been defaced using pydeface (https://github.com/poldracklab/pydeface). For the cortical surfaces, meshes were generated from the T1-weighted anatomical scans using FreeSurfer[29]. Before surface reconstruction, anatomical surface segmentations were hand-checked and corrected. Blender (https://blender.org) was used to remove the corpus callosum and make relaxation cuts for flattening via the interface provided by pycortex[30]. Functional images were aligned to the cortical surface using boundary-based registration (BBR) implemented in FSL. These were checked for accuracy and adjustments were made to the registration parameters as necessary.

Cortical maps of selectivity or model performance were created by projecting the values for each voxel onto the cortical surface using the 'nearest' scheme in pycortex[30]. This projection finds the location of each pixel in the image in 3D space, and assigns the pixel the value associated with the voxel enclosing that location.

Functional localizers and region of interest definitions.

Known regions of interest (ROIs) were defined separately in each participant using three localizer tasks: a visual category localizer, a motor localizer, and an auditory cortex localizer. For the visual category localizer, data was collected in six 4.5 minute scans consisting of 16 blocks of 16 seconds each. During each block, 20 images of places, faces, bodies, household objects, or spatially scrambled objects were displayed. In order to encourage focus, participants were asked to perform a 1-back task where they pressed a button if the same image appeared twice in a row. The corresponding cortical ROIs defined with this localizer were the fusiform face area (FFA)[31], occipital face area (OFA)[31], extrastriate body area (EBA)[32], parahippocampal place area (PPA)[33], retrosplenial cortex (RSC), and the occipital place area (OPA). These ROIs were hand-drawn based on t-value maps from contrasts comparing responses to faces and objects (FFA, OFA), bodies and objects (EBA), and places and objects (PPA, OPA, RSC).

The motor localizer data was collected during two identical 10-minute scans. The participant was cued to perform six different motor tasks in a random order in 20-second blocks. The cues 'hand', 'foot', 'mouth', 'speak', and 'rest' were visually presented at the center of the screen, and the *saccade* cue was presented as a random array of dots. For the *hand* cue, participants were instructed to make small finger-drumming movements for the duration of the cue. For the *foot* cue, participants were instructed to make small foot and toe movements. For the *mouth* cue, participants were instructed to make small nonsense vocalizations (i.e., the syllable string "*balabalabala*"). For the *speak* cue, participants were instructed to self-generate a narrative without vocalization. For the *saccade* cue, participants were instructed to look around for the duration of the task. Simple categorical regression models were fit for each voxel using ordinary least squares (OLS). Beta maps for the *hand*, *foot*, and *mouth* conditions were used to define primary motor and somatosensory areas for the hands, feet, and mouth; supplemental motor areas for the hands and feet; secondary motor areas for the hands, feet, and mouth; and the ventral premotor hand area (PMvh). The beta map for the *saccade* condition was used to define the frontal eye field and intraparietal sulcus visual areas. The beta map for the *speak* condition was used to define Broca's area and the superior premotor ventral (sPMv) speech area[34].

Auditory cortex localizer data was collected in one 10-minute scan. The participants listened to 10 repeats of a 1-minute auditory stimulus containing 20 seconds of music (Arcade Fire), speech (Ira Glass, This American Life), and nature sounds (a babbling brook). To determine whether a voxel was responsive to auditory stimulus, the repeatability of the voxel response across the 10 repeats was calculated using an F-statistic. This map was used to define the auditory cortex (AC).

Stimulus embeddings. The proximal goal of encoding models is to find stimulus features that predict variance in the brain activity. This technique was originally developed for electrophysiology experiments[35], but has been widely adopted for modeling BOLD signals in fMRI. In this framework, participants are presented with a stimulus, in this case stories, while brain activity is recorded. A linear regression model is then fit between some feature space, which is extracted from the stimulus, and the brain activity. The feature space serves as a hypothesis for the kind of information each voxel of brain data is representing. It is important to note that this is not a winner-take-all model and that each voxel is likely representing multiple different components of information[18].

This dataset includes three different representations of the stimulus: an audio waveform, phoneme-level annotations and word-level annotations. It is possible to turn these annotations into many different feature spaces. We have provided python code[26] to generate three of the many possible feature spaces. The first is a word-level semantic feature space called *English1000*. This feature space has previously been used to map semantic representations across the cerebral cortex and cerebellum[10,11,14]. English1000 is a 985-dimensional word embedding feature space based on word co-occurrence in English text[11]. For this feature space we include a saved matrix, "english1000sm.hf5" which can be found in the OpenNeuro dataset[9] under *derivatives*. This matrix includes a vector for each word in the pre-defined vocabulary. The code loads this matrix and assigns the correct vector for each word in the transcript of the story. The second feature space is a phoneme feature space. This is a 1-hot space comprising 44 dimensions, one for each phoneme in American English as defined by the CMU Pronouncing Dictionary[36] as well as a few non-speech sounds. The last feature space is a word rate feature space. This is a 1-dimensional feature space that represents the number of words spoken during each period of time. For both the phoneme feature space and the word rate feature space, the included code generates the feature space directly from annotations, with no need for an external data file. To create a feature space matrix for model fitting, each word (or phoneme) in the stimulus is assigned a vector from the feature space. For example, for the stimulus phrase "I reached over", one would take the embedding vectors for "I", "reached", and "over" from English1000 and concatenate them into a 3 (words) by 985 (features) matrix.

Interpolation of the feature matrix. One challenge in fitting encoding models is that speech and BOLD data are sampled at very different frequencies. Approximately six words are spoken every two seconds, but only one brain image is recorded in that interval. To solve this problem, the stimulus matrix needs to be resampled to the same sampling frequency as the BOLD data. The procedure we provide for downsampling features to the fMRI acquisition rate can be thought of as comprising three steps. First, the discrete features for each word (or phoneme) are transformed into a continuous-time representation N(t) where $t \in [0, T]$ and T indicates the length of the stimulus. This representation is zero at all timepoints except for the exact middle of each word (or phoneme), where it is equal to an infinitesimal-duration spike (Dirac δ-function) that is scaled by the feature value. Next, a low-pass antialiasing Lanczos filter is convolved with N(t) to get $N_{LP}(t)$. The cutoff frequency of this antialiasing filter is selected to match the Nyquist frequency of the fMRI data (half the acquisition rate, or 0.25 Hz). The cutoff frequency and filter roll-off (controlled by the number of lobes: more lobes yield a sharper roll-off, but at the cost of potentially increased noise) can be selected manually, although we recommend using the default values. Finally, $N_{LP}(t)$ is sampled at the fMRI acquisition times t_r where $r \in [1, 2 \ldots n_{TR}]$ corresponds to the volume index in the fMRI acquisition. In practice, these three steps are accomplished simultaneously by way of a single matrix multiplication: the word- (or phoneme-) level stimulus matrix *S* (number of features by number of words/phonemes) is multiplied by a sparse "Lanczos" matrix *L* (number of words/phonemes by number of fMRI volumes). In essence, this assumes that the total brain response is the sum of responses to each word or phoneme. This approach has been widely used for language encoding models with natural stimuli[10,11,14,20,37]. An alternative to this approach would be to simply average the feature vectors for all the word or phonemes that appear within each 2-second period. However, that approach leads to discontinuities since words that fall infinitesimally before or after a boundary wind up in different time bins. The Lanczos method naturally accounts for this issue: if a word falls exactly at the boundary between two time bins, its features contribute equally to both (albeit scaled by 50%).

HRF estimation. The BOLD responses recorded by fMRI are thought to capture delayed and low-pass filtered representations of local neural activity[38]. While most fMRI analyses treat the hemodynamic response function (HRF) as a fixed linear filter[39], the current dataset[9] contains enough data that separate HRFs can be estimated for each feature in each voxel. To efficiently estimate individual HRFs, we use a finite impulse response (FIR) model[40] in which separate model weights are estimated for each feature at several different delays (e.g. 2, 4, 6, and 8 seconds after the stimulus). This is accomplished by concatenating multiple versions of the interpolated stimulus matrix that have been delayed by different amounts.

Model fitting. To fit the linear regression models that predict the response in each voxel, previous work using this dataset[9,11,18,41,42] have used L2-regularized, or, ridge regression. Regularized regression makes assumptions about the size and covariance of the regression weights in order to improve weight estimation in the face of limited and noisy data. This improvement can be measured quantitatively by comparing the prediction performance of models fit using unregularized (ordinary least squares) regression to those fit using regularized regression. Regularized regression methods such as ridge regression work well in cases where there is high collinearity in the feature matrix or the number of features approaches the number of data points, in contrast to ordinary least squares regression, which yields large variance in parameter estimates and low model performance in those settings. Ridge regression solves these issues by adding a penalty on the size of the coefficients. The strength of the penalty is controlled by the alpha parameter, where a larger alpha value results in the model being more robust to collinearity[43]. For a longer discussion on this alpha parameter see the section on using a *single_alpha* parameter below. We provide code for fitting ridge regression models as a part of the voxelwise modeling process.

Model validation. Voxelwise models are typically evaluated on a separate test dataset to avoid overfitting on the training data. To do this, one takes the dot product of the regression weight matrix, consisting of a two-dimensional matrix of voxels by features, and the feature matrix of a new story not used in training the model (features by time). This results in a voxelwise prediction of brain activity in a 2-D matrix of voxels by time. The time-course of predicted brain response for each voxel is then correlated across time with the real brain data to measure model goodness-of-fit. This type of evaluation, which tests how well a model can predict responses to novel natural stimuli, is a good proxy for how well the model captures language representations in the brain[44]. One issue with this approach is that the correlation between predicted and actual responses will be biased downwards due to noise in the fMRI data. The amount of noise can also vary from voxel to voxel depending on factors like proximity to vasculature[45,46], cortical folding[46], or other factors. To better estimate the model performance given the noise ceiling for each voxel, it is common to collect responses to the same test stimulus multiple times and then average them, decreasing the amount of noise in the test data[11,13,14,19,40,47]. Averaging responses across repetitions effectively increases the signal to noise ratio of the BOLD response, providing less biased estimates of model performance. Here, the test dataset comprises one story, which was played once in each of the five scanning sessions.

Another issue that can complicate interpretation of model performance values is that BOLD responses recorded using fMRI are inherently noisy, and the amount of noise can differ across brain areas and between participants. The amount of noise affects the maximum model performance that can be attained, even by a theoretically “perfect” model. The repeated presentations of the test story can also be used to de-bias the prediction performance measure by estimating the noise-ceiling, or, the highest performance that any model can attain[48]. This is done by comparing responses across repeats of the test story, task-wheretheressmoke. We include code for computing the noise ceiling correction using a regularized normalized correlation coefficient (CC_{norm})[48]. This is done by first calculating the absolute product-moment correlation, defined as:

$$CC_{abs}=\frac{Cov\,(X,\ Y)}{\sqrt{Var\,(X)\,Var\,(Y)}}$$

where X are the BOLD responses and Y are the model predictions. Then, to isolate model performance from prediction accuracy, this value is normalized as:

$$CC_{norm}=\frac{CC_{abs}}{\max(CC_{max},\ CC_{floor})}$$

$$CC_{max}=\sqrt{\frac{2}{1+\sqrt{\frac{1}{CC_{half}^{2}}}}}$$

where CC_{max} is the noise ceiling. In comparison to the standard CC_{norm}, we regularize the estimate by limiting the noise ceiling to be greater than $CC_{\mathrm{floor}} = 0.3$. This value was determined to result in the least biased estimates in simulations using realistic noise values. Without this regularization, the estimated correlation after noise ceiling correction is not bounded and often surpasses a correlation of 1.0 for poorly-modeled voxels.

To test for the statistical significance of the model performance for each voxel, one must compare model performance for each voxel to a null distribution. Here it is important to account for temporal auto-correlation in the BOLD signal. Block-wise permutation ensures that the permuted data retains temporal characteristics of the data while breaking the connection to the feature space being tested. This test is performed by shuffling blocks of 10 timepoints in true response time course and then recomputing the correlation between model predictions and the permuted true responses. This process is repeated many times to form an empirical null distribution for each voxel. Additionally, because each voxel is effectively treated as an independent model and each brain contains upwards of 80000 voxels, it is important to correct for multiple comparisons. In previous work, FDR correction was used to account for this[13,14,49].

In prior studies, statistical significance was measured using prediction performance on one test story “task-wheretheressmoke” that has multiple repeats[10,11,13,14]. However, this approach also has a potential drawback: if a voxel is selective for features that are not present in the single test story, then those voxels may falsely be labeled as poorly predicted. To test for significance with multiple stories, an alternative approach is to use a leave-one-out procedure. This can be done by fitting an ensemble of encoding models, each of which excludes one unique training story (or session) in their model estimation. Statistical significance can then be measured for encoding model predictions on all held-out stories and their true BOLD responses, ie., the entire training set. This procedure increases diversity in the test set and improves statistical power.

Data Records

Data. The raw data and derived data are available on OpenNeuro[9], including NIfTI files of all brain data, the story stimuli, derived data, hand-corrected surface reconstructions, and descriptions of paradigms. The data is organized into directories, one for each participant (8 total), chronologically organized by session. The first session includes both anatomical and functional data, broken up into corresponding folders. All other sessions only include functional data. Of note, the initial anatomical scan and functional localizer data for sub-UTS02 was collected previous to this dataset using different sequence parameters. Consequently, the localizer data is not included here. The BOLD data is stored in gzipped NIfTI 4D files under the naming pattern *sub-AA_ses-BB_task-CC_bold.nii.gz*. The story entitled *wheretheressmoke* is repeated in all five of the story sessions and thus also contains the run number in the name (i.e. *sub-AA_ses-BB_task-CC_run-DD_bold.nii.gz)*. The preprocessed cortical BOLD data is contained in HDF5 files for each story, organized by participant. The audio files for all of the story stimuli are stored as WAV files sampled at 44.1 kHz. Hand-corrected TextGrid[27] files contain complete transcripts of each story as well as the temporal boundaries of each word and phoneme, and are stored as a derivative under Textgrids. For more information on TextGrid files, readers can consult Boersma *et al.*[27].

Extended dataset for three participants. In this dataset release, we have also included an extended dataset for three of the participants: UTS01, UTS02, and UTS03. The extended dataset includes 10 extra sessions with the raw BOLD data, preprocessed BOLD data, wav files of the stimulus, and their corresponding textgrids and is organized in the same OpenNeuro dataset[9] as described above. The stimulus for this extended dataset includes stories from *The Moth* and *The New York Times Modern Love* which are listed in Supplemental Table 1. This brings the size of this dataset up to 81 hours of BOLD fMRI across all participants. This extended dataset was originally used in Tang *et al.*[50]. This extended dataset is best used for testing models where high encoding performance is necessary but few participants are required, such as in language decoding experiments. The extended datasets are not evaluated in this paper, but see Tang *et al.*[50] for analyses showing that encoding and decoding model performance scale roughly log-linearly with the amount of training data per subject.

Code. All of the standard code used to fit voxelwise encoding models is available on github[26].

'encoding.py' is the main script to train and evaluate encoding models. It takes 11 arguments, 2 of which are required. The first required argument is the subject code that expects a string for the subject identifier, e.g. "UTS02". The second required argument is the feature space to be used in model fitting. The available options are "articulation", "phonemerate", "wordrate", and "eng1000". Briefly, the "articulation" feature space is a 22-dimesional n-hot binary feature space that carries information about the articulations that are used for the phonemes being spoken in the stimulus. The "phonemerate" and "wordrate" models are 1 dimension each where there is a 1 at each time point where a phoneme or word is spoken. The "phonemerate" and "wordrate" feature spaces are useful for regressing confounds of the auditory stimulation alone. All other arguments are optional.

These other arguments include *sessions* which is the number of training sessions to use with the model as an integer ranging from 1 to 5. Generally more data is better, but training with more data can be memory intensive.

Trim is how many TRs at the beginning and end of both the features and responses have to be removed. TRs at the beginning tend to have more artifacts from sound onset, flip angle stabilization, and poorly-conditioned detrending, so removing a few TRs can improve model performance.

The *ndelays* parameter is the number of delays to use in the finite impulse response function (FIR) that is used to estimate the hemodynamic response (HRF) of each voxel. We have found 4 delays (8 seconds) to be sufficient to cover the HRF. However, there is little harm to increasing the number of delays except for increased memory utilization and potential for overfitting.

The parameters *nboots, chunklen, nchunks*, and *single_alpha* control the cross-validation procedure that is used to select the best regularization parameter for each voxel. Briefly, the cross-validation procedure used to select the ridge regularization parameter (either for each voxel or the whole brain, see the discussion of *single_alpha* below) breaks the stimulus and response data up into pieces of length *chunklen*, and then randomly selects *nchunks* of those pieces to form a validation set on each of the *nboots* cross-validation runs. The chunk length *chunklen* should be several times longer than the longest HRF expected; we use a length of 40 TRs which works well in practice. The number of chunks reserved, *nchunks*, should be set so that roughly 20% of the dataset is reserved for validation at each step. The default parameter values for *chunklen* and *nchunks* are reasonable for data of this size, but at least the *nchunks* parameter should be scaled when using more or less data.

The parameter *use_corr* determines whether to use correlation or variance explained (R^2) to select the optimal ridge parameter during model fitting. For ridge regression this can make a large difference, as highly regularized solutions will have very small norms and explain very little variance while still potentially leading to high correlations.

The last parameter is *single_alpha*; when set to false, each voxel will be assigned its own ridge regularization parameter (alpha) based on cross-validation. This can result in higher model performance but will lead to very differently scaled weights across voxels, which is a problem when the research question requires comparing weights across voxels (e.g. when applying PCA to the voxel weights[11]). When *single_alpha* is set to true, all voxels will be assigned the same alpha parameter, which is selected as the best on average during cross-validation. This will slightly harm model performance (especially in the best voxels), but results in more comparably scaled weights across voxels. The *single_alpha* parameter also interacts with *nboots*: when estimating a single alpha parameter across all voxels, it is usually sufficient to use a small number of cross-validation steps (e.g. *nboots* = 10), as the alpha that is best on average across voxels is quite stable to variation in data sampling. However, when estimating a separate alpha parameter for each voxel (*singla_alpha* = *False*), one should use a larger number of cross-validation steps (e.g. *nboots* = 50), as the cross-validation results within each voxel are noisier than when results are averaged across many voxels.

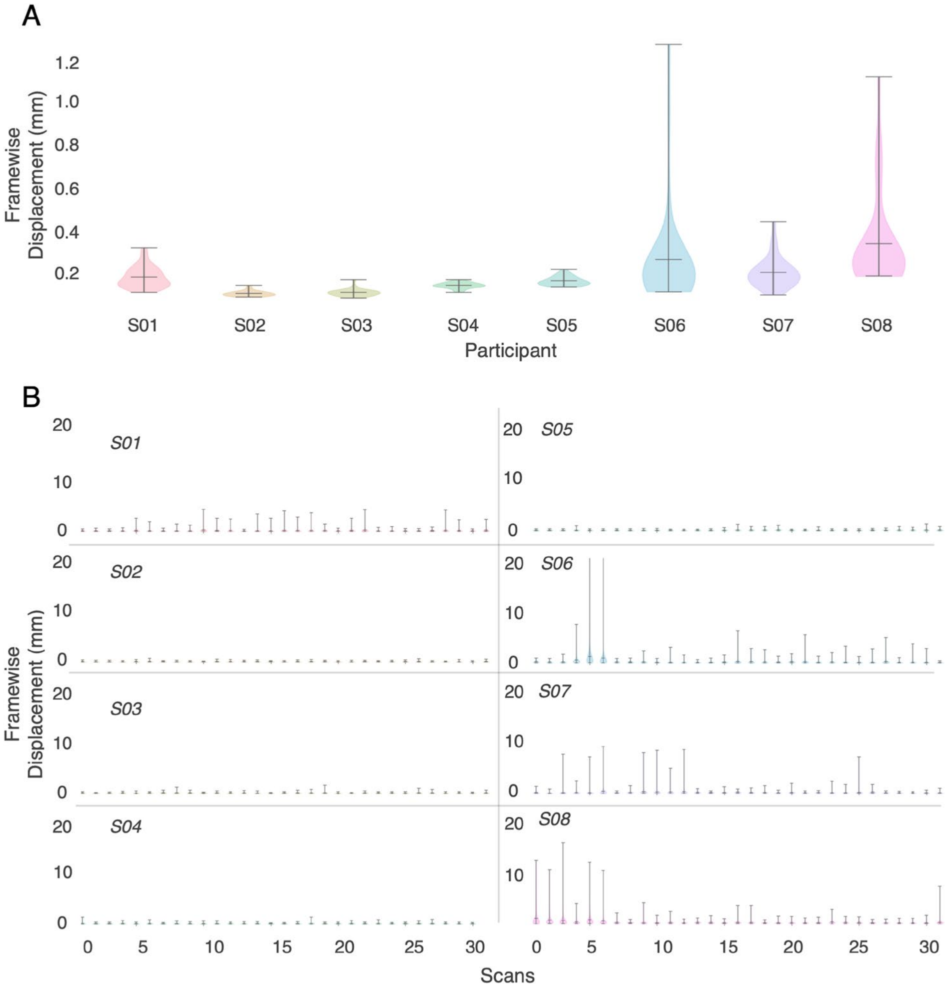

Fig. 2 Head Motion across participants. Framewise displacement is measured as the mean shift needed to align each frame of data to the starting reference frame[52,53]. (**A**) Mean framewise displacement for each participant shows very low motion in participants S02-S05. Participants S06-S08 show the highest framewise displacement as they moved the most during data collection. (**B**) Mean framewise displacement is also assessed at the scale of each individual story for each participant. This similarly shows the lowest displacement for participants S02-S05 and the highest for participants S06-S08. However, these high movement participants had less motion over the course of data collection with later sessions having less movement.

The encoding.py script first finds the train and test stories for the specified fMRI session. Next, it loads the corresponding fMRI responses for the participant. It then obtains down-sampled stimulus features from 'feature_spaces.py' for every story, z-scores them and applies the FIR delays. It then trains and cross-validates a linear regression model on all the specified training stories. The returned encoding model weights, test correlations, ridge parameters and cross-validation splits are all saved in the specified 'save_directory'. Finally, it computes the significance of the voxel correlations by running a blockwise permutation test from 'significance_testing.py'.

Several of the utility functions imported in the script can be found in the 'ridge_utils' directory. 'feature_spaces.py' contains implementations of each feature space—phoneme rate, articulators, word rate, English1000—and their corresponding downsampling functions. To add new feature spaces, one would need to implement a function that takes 'story names' as an argument and returns a dictionary of the downsampled stimulus features per story. Additionally, one would need to update the '_FEATURE_CONFIG' variable which maps between the feature argument as a string in encoding.py and the function that loads that feature space. The 'significance_testing.py' script implements a parallelized block-wise permutation test which takes in a vector of predicted and true voxel responses, the number of permutations to test with and the size of each permutation block. It returns the associated p-value of each voxel. This script also contains a function to control for the false discovery rate using the Benjamini-Hochberg procedure[40,51].

Technical Validation

One important metric for assessing BOLD data quality is head motion. Here we measured motion as framewise displacement[52], which combines both rotation and translation into a single metric. Figure 2A,B show framewise displacement for each participant and each story. Figure 2A shows the mean and range of average framewise displacement across all stories. This functions as a general metric for how still participants tend to be.

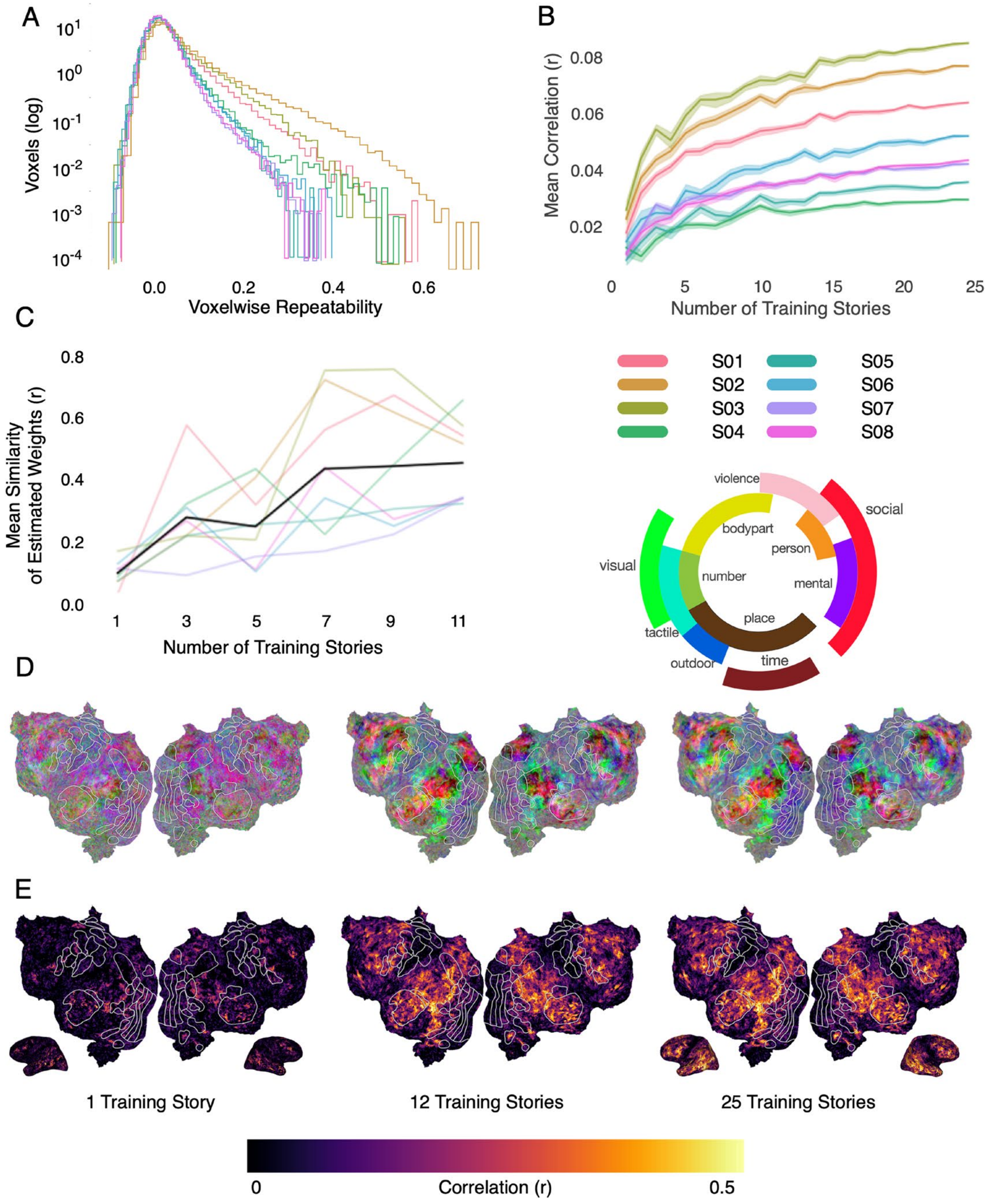

Fig. 3 Model performance variance across participants. (**A**) Voxelwise repeatability across five repeats of *task-wheretheressmoke* was calculated for each participant. Repeatability was calculated as the mean pairwise correlation across each repeat for each participant. Participants S01-S03 had the highest repeatability. (**B**) English1000 encoding models[11] were fit with increasing numbers of stories in the training set. As the dataset grows, so does model performance. Here we show the mean voxelwise model performance (r) for each participant. Participants S01-S03 had the highest model performance.The shaded regions are the standard error of the mean across 15 different training sets created by sampling the stories randomly without replacement. (**C**) Using many stimuli for model training makes encoding weights stable, or, invariant to the exact stimuli that were used. Here we measured weight stability by training encoding models using different stimulus subsets of varying sizes, and then computing the pairwise correlation between the learned weights. To reduce potential correlations between stimulus sets, each pair of models was trained with non-overlapping stimulus sets. Each colored line reflects an individual participant's weight stability and the black line shows the group average. As the training set grows, the estimated model weights for each voxel become more similar across different training subsets. (**D**) Encoding model weights were projected into a lower dimensional space[11] to visualize the semantic map for one example participant (S02). As the training set grows, the semantic map appears to converge. (**E**) Encoding model performance—shown here projected onto a cortical flatmap for one participant—increases with the number of training stories. These increases are particularly evident in temporal, parietal, and prefrontal cortex.

Earlier work in the literature has suggested a threshold of 0.5 mm in framewise displacement as a benchmark for "good" data. Sub-UTS02 and sub-UTS03 have the lowest average displacement across stories while participants *Sub-UTS06-08* have the highest average displacement, with participants UTS01-UTS05 all having all timepoints in the data below the 0.5 mm threshold on average. Figure 2B shows the average displacement for each individual story in each participant. This shows that some of the participants with higher movement improved as the sessions went on and thus their high displacement values are due to outliers, as in *sub-UTS06* and *sub-UTS08*. The participant with the most consistent low framewise displacement is *sub-UTS02*. Out of a total 247 story scans across all participants, the highest mean framewise displacement is only 0.33 mm.

Another important metric of fMRI data quality is functional repeatability, or how similar responses in the same voxel are to the same stimulus. While most of the stimuli were unique, the test story was played once in each of the five story sessions. This was done to enable us to compute the noise ceiling of the data and to increase the SNR of the held out dataset. It also enables us to measure how reliable the BOLD signal is in each voxel. Here, repeatability is calculated in each voxel as the mean pairwise correlation across the five validation story time course for that voxel, where a higher value means more reliable data. Figure 3A shows the repeatability for each participant as a step histogram across all voxels. The participants with the highest mean repeatability are sub-UTS03, sub-UTS02, and sub-UTS01. While repeatability does not have a clear threshold in the literature to compare to like framewise displacement, we chose to include it here so that potential users can make an informed decision about which participants to include in their studies. Repeatability is an estimate of the signal-to-noise ratio of the data.. Since there is no single best value for acceptable noise in data, this is going to vary depending on the research question being posed.

Lastly, one important metric used to assess data quality[13,14,49] is encoding model prediction performance (r). Similar to repeatability, there is not a standard value to compare the model performances here with other datasets. However, this information is provided to give potential users a better sense of the signal and noise in the dataset for each participant. In Fig. 3B, we show the performance of the semantic encoding models (English1000) (mean(r)) as a function of the number of stories used for model training. For each number of training stories, we fit 15 models in which training data were sampled from the full set without replacement. Each point represents the mean prediction performance across all voxels and the cloud around the point represents the standard error across the 15 runs. All participants appear to reach a plateau where increasing the amount of training data does not dramatically improve performance. Note that this metric includes the majority of cortical voxels that are not semantically selective, biasing the result downwards. By this metric, the participants with the best data quality are sub-UTS03, sub-UTS02, and sub-UTS01. These participants are also the participants with the lowest motion and highest repeatability. Figure 3E shows the voxelwise prediction performance (r) for one participant plotted on the flatmaps for an increasing number of stories. This again shows that as the amount of data increases, more voxels are predicted and predicted better.

This improvement in model performance means that the model weights from the fit encoding models become more stable and less affected by attention and noise (Fig. 3C). To demonstrate the importance of large datasets within an individual for interpreting model weights, we fit models using different subsets of stories and then calculate the mean pairwise correlation across estimated model weights. For each pair of models that were compared, we restricted the potential number of training stories to be non-overlapping. Each colored line shows the mean pairwise correlation of weights for each participant and the black line shows the group result averaged across all participants. As the number of stories in the subsets increases, the estimated model weights become more similar regardless of the individual stories used in training. This makes interpretation of the model weights more robust. Figure 3D shows the model weights for each of these models projected into a three-dimensional semantic space that was previously constructed from a group of participants using principal components analysis[11]. This lower-dimensional space is used purely for visualization purposes. Here projections on the first, second, and third principal components are mapped into the red, green, and blue color channels, respectively, for each voxel and then projected onto the cortical surface. The color wheel shows approximately which semantic category each color on the maps represents. The separability and intensity of the weight increases and becomes clearer as dataset size grows. Increasing the size of datasets within individuals thus increases reliability and interpretability of encoding models, and is vital to increasing the reliability of results in the field.

Usage Notes

Anatomical and localizer scans (ses-1) for sub-UTS02 were collected prior to this current dataset, at a different location, and with a different scan protocol than all other data in this project. Consequently, the localizer data for that participant is not included here. However, the hand-defined regions of interest (ROIs) derived from the localizer data can be found for this participant in the pycortex-db in their overlays file. This participant also has additional ROIs from other localizers including retinotopy.

Participant sub-UTS04 has one missing story scan, task-life.

Participant sub-UTS05 was presented with auditory cues for the motor localizer stimuli instead of visual cues. Additionally, this localizer lacked the saccade cue. This difference was due to the participant's visual acuity being too low to successfully read the cue and MRI-safe glasses were not compatible with the headcase being used.

Code availability

All code used for encoding model fitting is publicly available and can be found on github and zenodo[25].

The code used for filtering the audio for each story to correct for frequency response and phase errors induced by the headphones using calibration data provided by Sensimetrics and custom Python code. This code can be found at https://github.com/alexhuth/sensimetrics_filter.

Received: 26 September 2022; Accepted: 2 August 2023;
Published online: 23 August 2023

References

1. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
2. Agarwal, S., Sair, H. I., Gujar, S. & Pillai, J. J. Language Mapping With fMRI: Current Standards and Reproducibility. *Top. Magn. Reson. Imaging* **28**, 225–233 (2019).
3. Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M. & Baroni, M. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *Neuroimage* **120**, 309–322 (2015).
4. Kuperberg, G. R. *et al.* Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI. *study. J. Cogn. Neurosci.* **12**, 321–341 (2000).
5. Humphries, C., Binder, J. R., Medler, D. A. & Liebenthal, E. Time course of semantic processes during sentence comprehension: an fMRI study. *Neuroimage* **36**, 924–932 (2007).
6. Anderson, A. J. *et al.* Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation. *Cereb. Cortex* **27**, 4379–4395 (2017).
7. Brennan, J. & Pylkkänen, L. The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage* **60**, 1139–1148 (2012).
8. Grall, C. & Finn, E. S. Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. *Soc. Cogn. Affect. Neurosci.* **17**, 598–608 (2022).
9. LeBel, A. *et al.* An fMRI dataset during a passive natural language listening task. *OpenNeuro* https://doi.org/10.18112/openneuro.ds003020.v2.0.0 (2020).
10. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The Hierarchical Cortical Organization of Human Speech Processing. *J. Neurosci.* **37**, 6539–6557 (2017).
11. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
12. Deniz, F., Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *J. Neurosci.* **39**, 7722–7736 (2019).
13. Jain, S. *et al.* Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. Advances in Neural Information Processing Systems 34 (2020).
14. LeBel, A., Jain, S. & Huth, A. G. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *Journal of Neuroscience* https://doi.org/10.1523/JNEUROSCI.0118-21.2021 (2021).
15. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
16. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
17. Millet, J. *et al.* Toward a realistic model of speech processing in the brain with self-supervised learning. arXiv [q-bio.NC] (2022).
18. Dupre la Tour, T., Eickenberg, M. & Gallant, J. L. Feature-space selection with banded ridge regression. bioRxiv 2022.05.05.490831, https://doi.org/10.1101/2022.05.05.490831 (2022).
19. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
20. Jain, S. & Huth, A. G. Incorporating context into language encoding models for fMRI. vol. 2018-Decem 6628–6637, http://papers.nips.cc/paper/7897-incorporating-context-into-language-encoding-models-for-fmri.pdf (2018).
21. Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One* **9**, e112575 (2014).
22. Visconti di Oleggio Castello, M., Chauhan, V., Jiahui, G. & Gobbini, M. I. An fMRI dataset in response to 'The Grand Budapest Hotel', a socially-rich naturalistic movie. Scientific Data 7 (2020).
23. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Sci Data* **7**, 347 (2020).
24. Nastase, S. A. *et al.* The 'Narratives' fMRI dataset for evaluating models of naturalistic language comprehension. *Sci Data* **8**, 250 (2021).
25. Jain, S., LeBel, A., Huth, A. & Singh, C. *HuthLab/deep-fMRI-dataset: Zenodo integration* https://doi.org/10.5281/zenodo.8208839 (2023).
26. Jiahong Yuan, M. L. Speaker identification on the SCOTUS corpus. in In Proceedings of Acoustics 2008 (2008).
27. Boersma, P. & Weenink, D. Praat: doing phonetics by computer. (2014).
28. Woolrich, M. W. *et al.* Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45**, S173–86 (2009).
29. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
30. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23 (2015).
31. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
32. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
33. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
34. Chang, S.-E., Horwitz, B., Ostuni, J., Reynolds, R. & Ludlow, C. L. Evidence of left inferior frontal-premotor structural and functional connectivity deficits in adults who stutter. *Cereb. Cortex* **21**, 2507–2518 (2011).
35. Holdgraf, C. R. *et al.* Encoding and Decoding Models in Cognitive Electrophysiology. *Front. Syst. Neurosci.* **11**, 61 (2017).
36. Carnegie Mellon Pronouncing Dictionary. (1993).
37. Antonello, R., Turek, J. S., Vo, V. & Huth, A. Low-Dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. Adv. Neural Inf. Process. Syst. 34 (2021).
38. Logothetis, N. K. The underpinnings of the BOLD functional magnetic resonance imaging signal. *J. Neurosci.* **23**, 3963–3971 (2003).
39. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience* **16**, 4207–4221 (1996).
40. Nishimoto, S. *et al.* Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
41. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).
42. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
43. Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage* **197**, 482–492 (2019).
44. Drew, P. J., Mateo, C., Turner, K. L., Yu, X. & Kleinfeld, D. Ultra-slow Oscillations in fMRI and Resting-State Connectivity: Neuronal and Vascular Contributions and Technical Confounds. *Neuron* **107**, 782–804 (2020).
45. Kay, K. *et al.* A critical assessment of data quality and venous effects in sub-millimeter fMRI. *Neuroimage* **189**, 847–869 (2019).

46. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
47. Hamilton, L. S. & Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang Cogn Neurosci* **35**, 573–582 (2020).
48. Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J. & Schnupp, J. W. H. Measuring the Performance of Neural Models. *Front. Comput. Neurosci.* **10**, 10 (2016).
49. Tang, J., LeBel, A. & Huth, A. G. Cortical Representations of Concrete and Abstract Concepts in Language Combine Visual and Linguistic Representations. https://doi.org/10.1101/2021.05.19.444701 (2021).
50. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* **26**, 858–866 (2023).
51. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
52. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage* **84**, 320–341 (2014).
53. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).

Acknowledgements

This work was supported by the Whitehall Foundation, Alfred P. Sloan Foundation, Burroughs-Wellcome Fund, the Texas Advanced Computing Center (TACC), William Orr Dingwall Foundation (S.J.), and the National Science Foundation (A.L).

Author contributions

A.L. and A.G.H. designed the research. A.L., L.W., S.J., A.M., J.T. and L.X. performed the research. A.L. analysed the data. A.L. and L.W. wrote the first draft. A.L. and A.G.H. edited the paper. A.A. and B.G. contributed unpublished reagents/analytic tools.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-02437-z.

Correspondence and requests for materials should be addressed to A.G.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.