nature neuroscience

Article

https://doi.org/10.1038/s41593-023-01304-9

Semantic reconstruction of continuous language from non-invasive brain recordings

Received: 1 April 2022

Accepted: 15 March 2023

Published online: 1 May 2023

Check for updates

Jerry Tang[1], Amanda LeBel[2], Shailee Jain[1] & Alexander G. Huth[1,2]

A brain–computer interface that decodes continuous language from non-invasive recordings would have many scientific and practical applications. Currently, however, non-invasive language decoders can only identify stimuli from among a small set of words or phrases. Here we introduce a non-invasive decoder that reconstructs continuous language from cortical semantic representations recorded using functional magnetic resonance imaging (fMRI). Given novel brain recordings, this decoder generates intelligible word sequences that recover the meaning of perceived speech, imagined speech and even silent videos, demonstrating that a single decoder can be applied to a range of tasks. We tested the decoder across cortex and found that continuous language can be separately decoded from multiple regions. As brain–computer interfaces should respect mental privacy, we tested whether successful decoding requires subject cooperation and found that subject cooperation is required both to train and to apply the decoder. Our findings demonstrate the viability of non-invasive language brain–computer interfaces.

Previous brain–computer interfaces have demonstrated that speech articulation[1] and other signals[2] can be decoded from intracranial recordings to restore communication to people who have lost the ability to speak[3,4]. Although effective, these decoders require invasive neurosurgery, making them unsuitable for most other uses. Language decoders that use non-invasive recordings could be more widely adopted and have the potential to be used for both restorative and augmentative applications. Non-invasive brain recordings can capture many kinds of linguistic information[5–8], but previous attempts to decode this information have been limited to identifying one output from among a small set of possibilities[9–12], leaving it unclear whether current non-invasive recordings have the spatial and temporal resolution required to decode continuous language.

Here we introduce a decoder that takes non-invasive brain recordings made using functional magnetic resonance imaging (fMRI) and reconstructs perceived or imagined stimuli using continuous natural language. To accomplish this, we needed to overcome one major obstacle: the low temporal resolution of fMRI. Although fMRI has excellent spatial specificity, the blood-oxygen-level-dependent (BOLD) signal that it measures is notoriously slow—an impulse of neural activity causes BOLD to rise and fall over approximately 10 s (ref. 13). For naturally spoken English (over two words per second), this means that each brain image can be affected by over 20 words. Decoding continuous language thus requires solving an ill-posed inverse problem, as there are many more words to decode than brain images. Our decoder accomplishes this by generating candidate word sequences, scoring the likelihood that each candidate evoked the recorded brain responses and then selecting the best candidate.

To compare word sequences to a subject's brain responses, we used an encoding model[5] that predicts how the subject's brain responds to natural language. We recorded brain responses while the subject listened to 16 h of naturally spoken narrative stories, yielding over five times more data than the typical language fMRI experiment. We trained the encoding model on this dataset by extracting semantic features that capture the meaning of stimulus phrases[8,14–17] and using linear regression to model how the semantic features influence brain responses (Fig. 1a). Given any word sequence, the encoding model predicts how the subject's brain would respond when hearing the sequence with considerable accuracy (Extended Data Fig. 1). The encoding model can then score the likelihood that the word sequence evoked the recorded

[1]Department of Computer Science, The University of Texas at Austin, Austin, TX, USA. [2]Department of Neuroscience, The University of Texas at Austin, Austin, TX, USA. e-mail: huth@cs.utexas.edu

https://doi.org/10.1038/s41593-023-01304-9

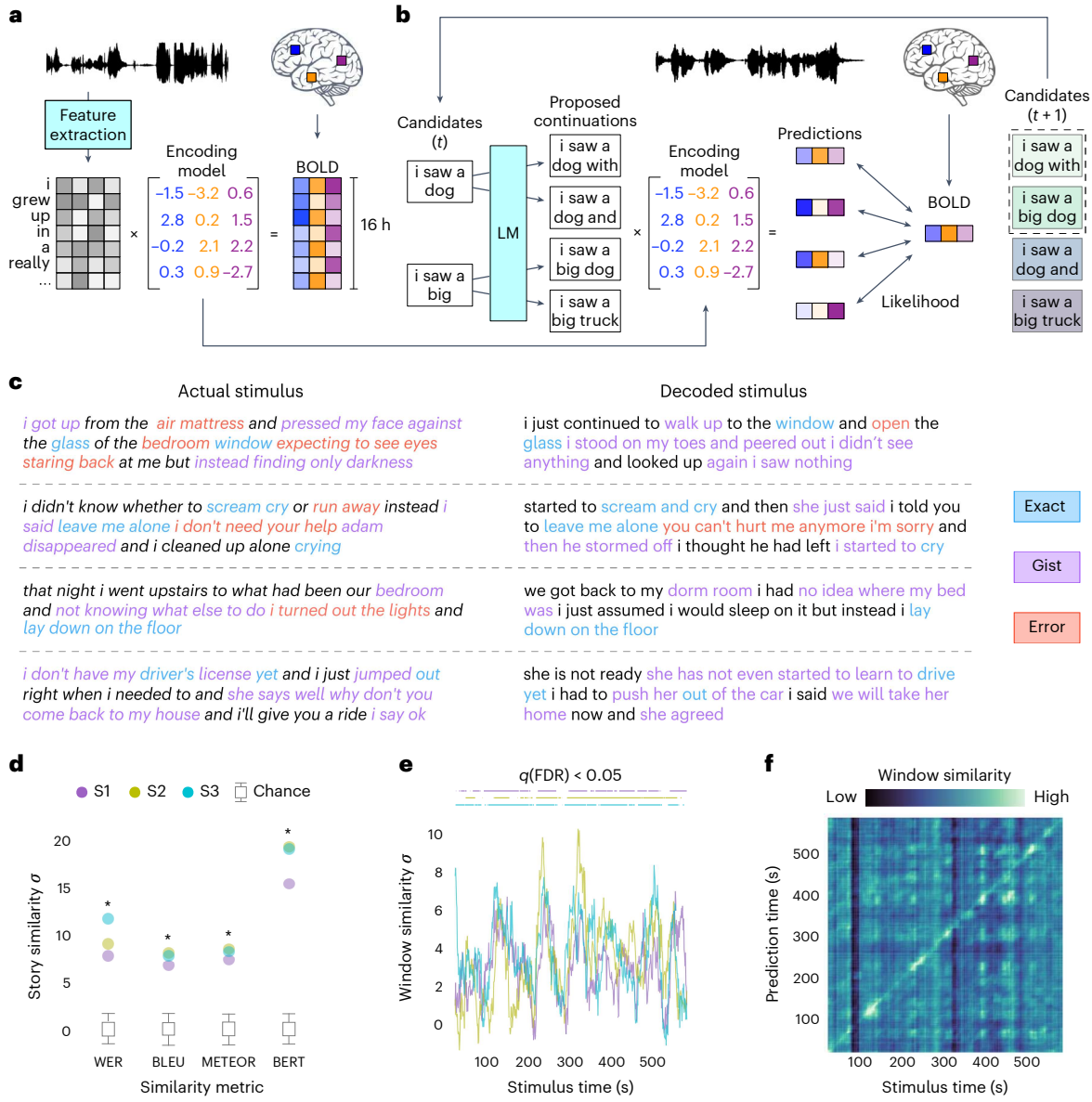


Fig. 1 | Language decoder. a, BOLD fMRI responses were recorded while three subjects listened to 16 h of narrative stories. An encoding model was estimated for each subject to predict brain responses from semantic features of stimulus words. **b**, To reconstruct language from novel brain recordings, the decoder maintains a set of candidate word sequences. When new words are detected, a language model (LM) proposes continuations for each sequence, and the encoding model scores the likelihood of the recorded brain responses under each continuation. The most likely continuations are retained. **c**, Decoders were evaluated on single-trial brain responses recorded while subjects listened to test stories that were not used for model training. Segments from four test stories are shown alongside decoder predictions for one subject. Examples were manually selected and annotated to demonstrate typical decoder behaviors. The decoder exactly reproduces some words and phrases and captures the gist of many more. **d**, Decoder predictions for a test story were significantly more similar to the actual stimulus words than expected by chance under a range of language similarity metrics (* indicates $q(\text{FDR}) < 0.05$ for all subjects, one-sided non-parametric test). To compare across metrics, results are shown as standard deviations away from the mean of the null distribution (Methods). Boxes indicate the interquartile range of the null distribution ($n = 200$ samples); whiskers indicate the 5th and 95th percentiles. **e**, For most timepoints, decoding scores were significantly higher than expected by chance ($q(\text{FDR}) < 0.05$, one-sided non-parametric test) under the BERTScore metric. **f**, Identification accuracy for one subject. The color at (i, j) reflects the similarity between the i-th second of the prediction and the j-th second of the actual stimulus. Identification accuracy was significantly higher than expected by chance ($P < 0.05$, one-sided permutation test).

brain responses by measuring how well the recorded brain responses match the predicted brain responses[18,19].

In theory, we could identify the most likely stimulus words by comparing the recorded brain responses to encoding model predictions for every possible word sequence[18,19]. However, the number of possible word sequences is far too large for this approach to be practical, and the vast majority of those sequences do not resemble natural language. To restrict the candidate sequences to well-formed English, we used a generative neural network language model[20] that was trained on a large dataset of natural English word sequences. Given any word sequence, the language model predicts the words that could come next.

However, even with the constraints imposed by the language model, it is computationally infeasible to generate and score all candidate sequences. To efficiently search for the most likely word sequences, we used a beam search algorithm[21] that generates candidate sequences word by word. In beam search, the decoder maintains

https://doi.org/10.1038/s41593-023-01304-9

Table 1 | Language similarity scores

	WER	BLEU-1	METEOR	BERTScore
Null	0.9637	0.1908	0.1323	0.7899
Subject 1	0.9407	0.2331	0.1621	0.8077
Subject 2	0.9354	0.2426	0.1677	0.8104
Subject 3	0.9243	0.2470	0.1703	0.8116
Translation	0.7459	0.4363	0.3991	0.8797

Decoder predictions for a perceived story were compared to the actual stimulus words using a range of language similarity metrics. A floor for each metric was computed by scoring the mean similarity between the actual stimulus words and 200 null sequences generated from a language model without using any brain data. A ceiling for each metric was computed by manually translating the actual stimulus words into Mandarin Chinese, automatically translating the words back into English using a state-of-the-art machine translation system and scoring the similarity between the actual stimulus words and the output of the machine translation system. Under the BERTScore metric, the decoder—which was trained on far less paired data and used far noisier input—performed around 20% as well as the machine translation system relative to the floor.

a beam containing the k most likely candidate sequences at any given time. When new words are detected based on brain activity in auditory and speech areas (Methods and Extended Data Fig. 1), the language model generates continuations for each sequence in the beam using the previously decoded words as context. The encoding model then scores the likelihood that each continuation evoked the recorded brain responses, and the k most likely continuations are retained in the beam for the next timestep (Fig. 1b). This process continually approximates the most likely stimulus words across an arbitrary amount of time.

Results

We trained decoders for three subjects and evaluated each subject's decoder on separate, single-trial brain responses that were recorded while the subject listened to novel test stories that were not used for model training. Because our decoder represents language using semantic features rather than motor or auditory features, the decoder predictions should capture the meaning of the stimuli. Results show that the decoded word sequences captured not only the meaning of the stimuli but often even exact words and phrases, demonstrating that fine-grained semantic information can be recovered from the BOLD signal (Fig. 1c and Supplementary Table 1). To quantify decoding performance, we compared decoded and actual word sequences for one test story (1,839 words) using several language similarity metrics (Methods). Standard metrics such as word error rate (WER), BLEU and METEOR measure the number of words shared by two sequences. However, because different words can convey the same meaning—for instance, 'we were busy' and 'we had a lot of work'—we also used BERTScore, a newer method that uses machine learning to quantify whether two sequences share a meaning. Story decoding performance was significantly higher than expected by chance under each metric but particularly BERTScore (q(false discovery rate (FDR)) < 0.05, one-sided non-parametric test; Fig. 1d; see Table 1 for raw values). Most timepoints in the story (72–82%) had a significantly higher BERTScore than expected by chance (Fig. 1e) and could be identified from other timepoints (mean percentile rank = 0.85–0.91) based on BERTScore similarities between the decoded and actual words (Fig. 1f and Extended Data Fig. 2a). We also tested whether the decoded words captured the original meaning of the story using a behavioral experiment, which showed that nine of 16 reading comprehension questions could be answered by subjects who had only read the decoded words (Extended Data Fig. 3).

Decoding across cortical regions

The decoding results shown in Fig. 1 used responses from multiple cortical regions to achieve good performance. We next used the decoder to study how language is represented within each of these regions. Although previous studies have demonstrated that most parts of cortex are active during language processing[5,22–24], it is unclear which regions represent language at the granularity of words and phrases[25], which regions are consistently engaged in language processing[26] and whether different regions encode complementary[27] or redundant[28] language representations. To answer these questions, we partitioned brain data into three macro-scale cortical regions previously shown to be active during language processing—the speech network[29], the parietal-temporal-occipital association region[23] and the prefrontal region[5]—and separately decoded from each region in each hemisphere (Fig. 2a and Extended Data Fig. 4a).

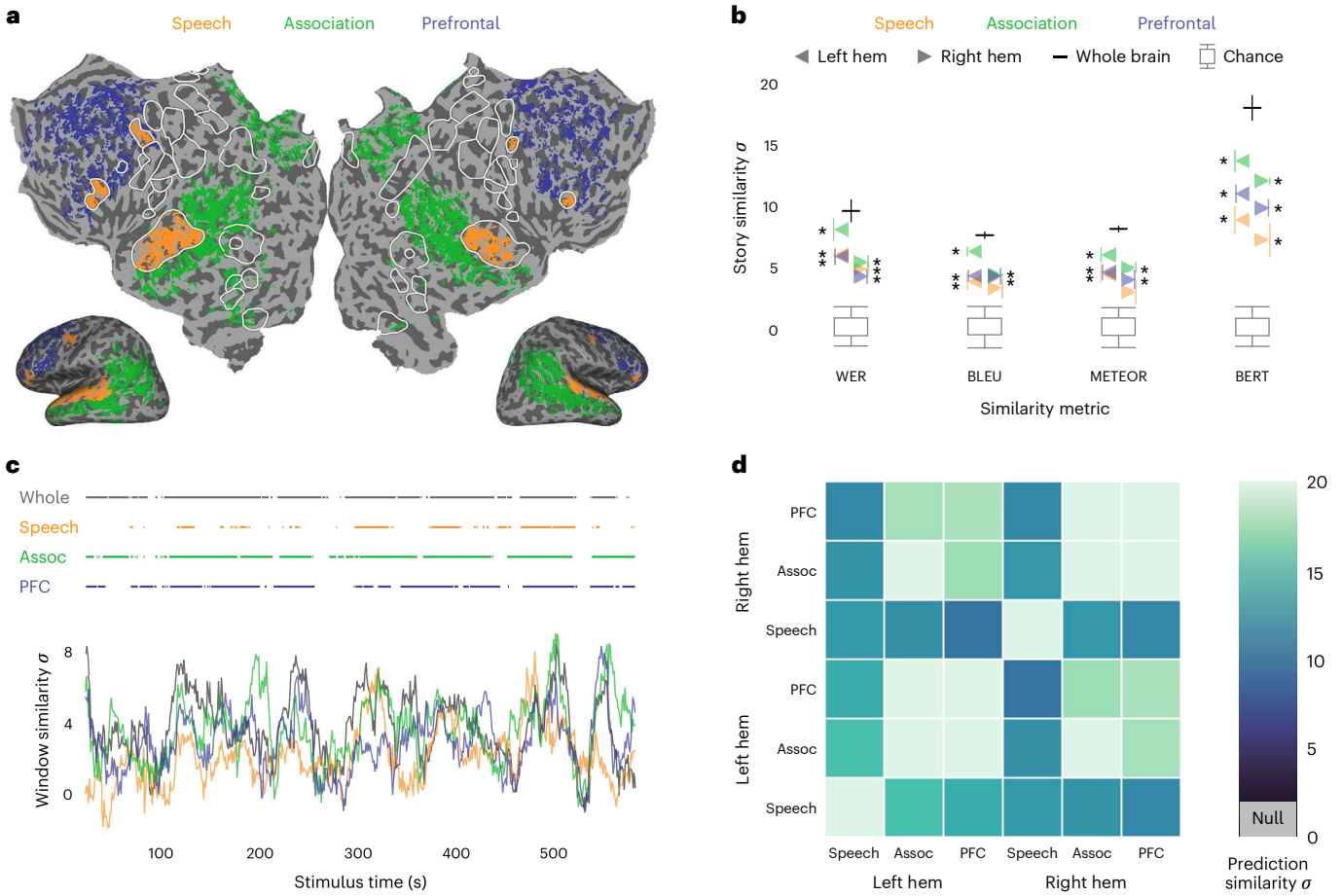
To test whether a region encodes semantic information at the granularity of words and phrases, we evaluated decoder predictions from the region using multiple language similarity metrics. Previous studies have decoded semantic features from BOLD responses in different regions[11], but the distributed nature of the semantic features and the low temporal resolution of the BOLD signal make it difficult to evaluate whether a region represents fine-grained words or coarser-grained categories[25]. Because our decoder produces interpretable word sequences, we can directly assess how precisely each region represents the stimulus words (Fig. 2b). Under the WER and BERTScore metrics, decoder predictions were significantly more similar to the actual stimulus words than expected by chance for all regions (q(FDR) < 0.05, one-sided non-parametric test). Under the BLEU and METEOR metrics, decoder predictions were significantly more similar to the actual stimulus words than expected by chance for all regions except the right hemisphere speech network (q(FDR) < 0.05, one-sided non-parametric test). These results demonstrate that multiple cortical regions represent language at the granularity of individual words and phrases.

Although the previous analysis quantifies how well a region represents the stimulus as a whole, it does not specify whether the region is consistently engaged throughout the stimulus or only active at certain times[26]. To identify regions that are consistently engaged in language processing, we next computed the fraction of timepoints that were significantly decoded from each region. We found that most of the timepoints that were significantly decoded from the whole brain could be separately decoded from the association (80–86%) and prefrontal (46–77%) regions (Fig. 2c and Extended Data Fig. 4b), suggesting that these regions consistently represent the meaning of words and phrases in language. Notably, only 28–59% of the timepoints that were significantly decoded from the whole brain could be decoded from the speech network. This is likely a consequence of our decoding framework—the speech network is known to be consistently engaged in language processing, but it tends to represent lower-level articulatory and auditory features[6], whereas our decoder operates on higher-level semantic features of entire word sequences.

Finally, we assessed the relationship between language representations encoded in different regions. One possible explanation for our successful decoding from multiple regions is that different regions encode complementary representations—such as different parts of speech—in a modular organization[27]. If this were the case, different aspects of the stimulus may be decodable from individual regions, but the full stimulus should only be decodable from the whole brain. Alternatively, different regions might encode redundant representations of the full stimulus[28]. If this were the case, the same information may be separately decodable from multiple individual regions. To differentiate these possibilities, we directly compared decoded word sequences across regions and hemispheres and found that the similarity between each pair of predictions was significantly higher than expected by chance (q(FDR) < 0.05, two-sided non-parametric test; Fig. 2d). This suggests that different cortical regions encode redundant word-level language representations. However, the same words could be encoded in different regions using different features[23,30], and understanding the nature of these features remains an open question with important scientific and practical implications.

Together, our results demonstrate that the word sequences that can be decoded from the whole brain can also be consistently decoded

https://doi.org/10.1038/s41593-023-01304-9



e

Actual stimulus	Left speech	Left assoc	Left PFC	Right speech	Right assoc	Right PFC
i was like no i'm out of here this is great and i went and hid behind a cabana and he left	they drove off they didn't even look back as i sat there thinking what the hell i should do	i ran outside and told them to leave me alone and go home i walked out the back	i told them to leave but they kept saying i can't stay here so i left and got my keys	ran away and didn't look back at me and said you can go on without me i'm leaving now	drop me off and go home but i could not leave without saying good bye i walked in the door	let me through i don't know where he is right now but i will get there soon enough
i drew out this map for you and you're really you're like a mile and a half from home	i try to keep track of how many miles it is before the number stops at the exact spot	up the street and make a left turn that would take me down to the first house on the right	the number on a map and find out how far they had to drive to reach the address	let me just walk away from the case and start the process of making sure	write down how many miles per hour you are traveling to the closest city	just to see how long it takes so i drove down the hill and over to the bank

Fig. 2 | Decoding across cortical regions. a, Cortical regions for one subject. Brain data used for decoding (colored regions) were partitioned into the speech network, the parietal-temporal-occipital association region and the prefrontal cortex (PFC) region. **b**, Decoder predictions from each region in each hemisphere were significantly more similar to the actual stimulus words than expected by chance under most metrics (* indicates q(FDR) < 0.05 for all subjects, one-sided non-parametric test). Error bars indicate the standard error of the mean (n = 3 subjects). Boxes indicate the interquartile range of the null distribution (n = 200 samples); whiskers indicate the 5th and 95th percentiles. **c**, Decoding performance timecourse from each region for one subject. Horizontal lines indicate when decoding performance was significantly higher than expected by chance under the BERTScore metric (q(FDR) < 0.05, one-sided non-parametric test). Most of the timepoints that were significantly decoded from the whole brain were also significantly decoded from the association and prefrontal regions. **d**, Decoder predictions were compared across regions. Decoded word sequences from each pair of regions were significantly more similar than expected by chance (q(FDR) < 0.05, two-sided non-parametric test). **e**, Segments from a test story are shown alongside decoder predictions from each region in each hemisphere for one subject. Examples were manually selected and annotated to demonstrate typical decoder behaviors. Colors indicate corresponding phrases. These results demonstrate that multiple cortical regions encode fine-grained, consistent and redundant representations of natural language. Assoc, association; hem, hemisphere.

from multiple individual regions (Fig. 2e). A practical implication of this redundant coding is that future brain–computer interfaces may be able to attain good performance even while selectively recording from regions that are most accessible or intact.

Decoder applications and privacy implications

In the previous analyses, we trained and tested language decoders on brain responses to perceived speech. Next, to demonstrate the range of potential applications for our semantic language decoder, we assessed

whether language decoders trained on brain responses to perceived speech could be used to decode brain responses to other tasks.

Imagined speech decoding. A key task for brain–computer interfaces is decoding covert imagined speech in the absence of external stimuli. To test whether our language decoder can be used to decode imagined speech, subjects imagined telling five 1-min stories while being recorded with fMRI and separately told the same stories outside of the scanner to provide reference transcripts. For each 1-min scan, we correctly identified the story that the subject was imagining by decoding the scan, normalizing the similarity scores between the decoder prediction and the reference transcripts into probabilities and choosing the most likely transcript (100% identification accuracy; Fig. 3a and Extended Data Fig. 2b). Across stories, decoder predictions were significantly more similar to the corresponding transcripts than expected by chance ($P < 0.05$, one-sided non-parametric test). Qualitative analysis shows that the decoder can recover the meaning of imagined stimuli (Fig. 3b and Supplementary Table 2).

For the decoder to transfer across tasks, the target task must share representations with the training task[1,31–33]. Our encoding model is trained to predict how a subject's brain would respond to perceived speech, so the explicit goal of our decoder is to generate words that would evoke the recorded brain responses when heard by the subject. The decoder successfully transfers to imagined speech because the semantic representations that are activated when the subject imagines a story are similar to the semantic representations that would have been activated had the subject heard the story. Nonetheless, decoding performance for imagined speech was lower than decoding performance for perceived speech (Extended Data Fig. 5a), which is consistent with previous findings that speech production and speech perception involve partially overlapping brain regions[34]. We may be able to achieve more precise decoding of imagined speech by replacing our encoding model trained on perceived speech data with an encoding model trained on attempted or imagined speech data[4]. This would give the decoder the explicit goal of generating words that would evoke the recorded brain responses when imagined by the subject.

Cross-modal decoding. Semantic representations are also shared between language perception and a range of other perceptual and conceptual processes[23,35,36], suggesting that, unlike previous language decoders that used mainly motor[1,3] or auditory[2] signals, our semantic language decoder may be able to reconstruct language descriptions from brain responses to non-linguistic tasks. To test this, subjects watched four short films without sound while being recorded with fMRI, and the recorded responses were decoded using the semantic language decoder. We compared the decoded word sequences to language descriptions of the films for the visually impaired (Methods) and found that they were significantly more similar than expected by chance (q(FDR) < 0.05, one-sided non-parametric test; Extended Data Fig. 5a). Qualitatively, the decoded sequences accurately described events from the films (Fig. 3c, Supplementary Table 3 and Supplementary Video 1). This suggests that a single semantic decoder trained during language perception could be used to decode a range of semantic tasks.

Attention effects on decoding. Because semantic representations are modulated by attention[37,38], our semantic decoder should selectively reconstruct attended stimuli[39,40]. To test the effects of attention on decoding, subjects listened to two repeats of a multi-speaker stimulus that was constructed by temporally overlaying a pair of stories told by female and male speakers. On each presentation, subjects were cued to attend to a different speaker. Decoder predictions were significantly more similar to the attended story than to the unattended story (q(FDR) < 0.05 across subjects, one-sided paired t-test; $t(2) = 12.76$ for the female speaker and $t(2) = 7.26$ for the male speaker), demonstrating that the decoder selectively reconstructs attended stimuli (Fig. 3d and Extended Data Fig. 5b). These results suggest that semantic decoders could perform well in complex environments with multiple sources of information. Moreover, these results demonstrate that subjects have conscious control over decoder output and suggest that semantic decoders can reconstruct only what subjects are actively attending to.

Privacy implications. An important ethical consideration for semantic decoding is its potential to compromise mental privacy[41]. To test if decoders can be trained without a person's cooperation, we attempted to decode perceived speech from each subject using decoders trained on data from other subjects. For this analysis, we collected data from seven subjects as they listened to 5 h of narrative stories. These data were anatomically aligned across subjects using volumetric and surface-based methods (Methods). Decoders trained on cross-subject data (Extended Data Fig. 6) performed barely above chance and significantly worse than decoders trained on within-subject data (q(FDR) < 0.05, two-sided t-test). This suggests that subject cooperation remains necessary for decoder training (Fig. 3e, Extended Data Fig. 5c and Supplementary Table 4).

To test if a decoder trained with a person's cooperation can later be consciously resisted, subjects silently performed three cognitive tasks—calculation ('count by sevens'), semantic memory ('name and imagine animals') and imagined speech ('tell a different story')—while listening to segments from a narrative story. We found that performing the semantic memory ($t(2) = 6.95$ for the whole brain, $t(2) = 4.93$ for the speech network, $t(2) = 6.93$ for the association region, $t(2) = 4.70$ for the prefrontal region) and imagined speech ($t(2) = 4.79$ for the whole brain, $t(2) = 4.25$ for the speech network, $t(2) = 3.75$ for the association region, $t(2) = 5.73$ for the prefrontal region) tasks significantly lowered decoding performance relative to a passive listening baseline for each cortical region (q(FDR) < 0.05 across subjects, one-sided paired t-test). This demonstrates that semantic decoding can be consciously resisted in an adversarial scenario and that this resistance cannot be overcome by focusing the decoder only on specific brain regions (Fig. 3f and Extended Data Fig. 5d).

Sources of decoding error

To identify potential avenues for improvement, we assessed whether decoding error during language perception reflects limitations of the fMRI recordings, our models or both (Fig. 4a).

BOLD fMRI recordings typically have a low signal-to-noise ratio (SNR). During model estimation, the effects of noise in the training data can be reduced by increasing the size of the dataset. To evaluate if decoding performance is limited by the size of our training dataset, we trained decoders using different amounts of data. Decoding scores were significantly higher than expected by chance with just a single session of training data, but substantially more training data were required to consistently decode the different parts of the test story (Extended Data Fig. 7 and Supplementary Table 5). Decoding scores appeared to increase by an equal amount each time the size of the training dataset was doubled (Fig. 4b). This suggests that training on more data will improve decoding performance, albeit with diminishing returns for each successive scanning session[42].

Low SNR in the test data may also limit the amount of information that can be decoded. To evaluate whether future improvements to single-trial fMRI SNR might improve decoding performance, we artificially increased SNR by averaging brain responses collected during different repeats of the test story. Decoding performance slightly increased with the number of averaged responses (Fig. 4c), suggesting that some component of the decoding error reflects noise in the test data.

Another limitation of fMRI is that current scanners are too large and expensive for most practical decoder applications. Portable techniques such as functional near-infrared spectroscopy (fNIRS) measure the same hemodynamic activity as fMRI, albeit at a lower spatial resolution[43,44]. To test whether our decoder relies on the high spatial resolution of

https://doi.org/10.1038/s41593-023-01304-9

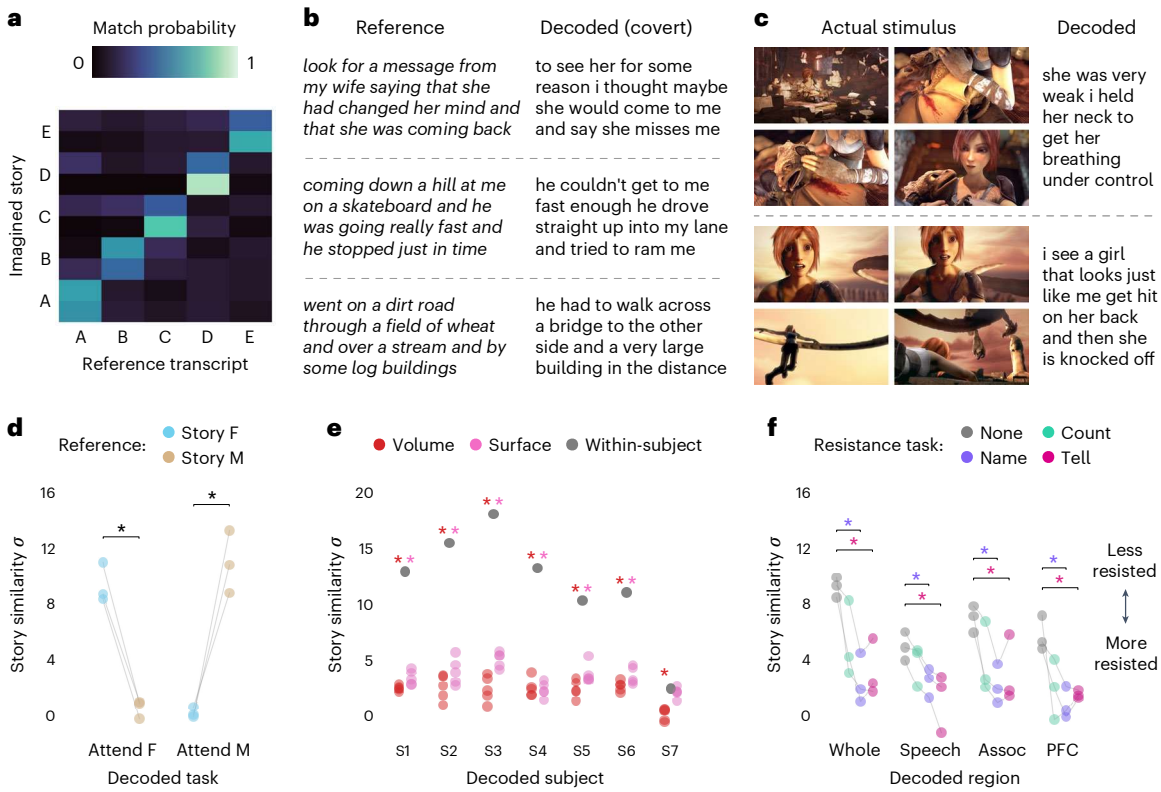


Fig. 3 | Decoder applications and privacy implications. a, To test whether the language decoder can transfer to imagined speech, subjects were decoded while they imagined telling five 1-min test stories twice. Decoder predictions were compared to reference transcripts that were separately recorded from the same subjects. Identification accuracy is shown for one subject. Each row corresponds to a scan, and the colors reflect the similarities between the decoder prediction and all five reference transcripts (100% identification accuracy). **b**, Reference transcripts are shown alongside decoder predictions for three imagined stories for one subject. **c**, To test whether the language decoder can transfer across modalities, subjects were decoded while they watched four silent short films. Decoder predictions were significantly related to the films (q(FDR) < 0.05, one-sided non-parametric test). Frames from two scenes are shown alongside decoder predictions for one subject (Blender Foundation; https://www.sintel.org (ref. 48)). **d**, To test whether the decoder is modulated by attention, subjects attended to the female speaker or the male speaker in a multi-speaker stimulus. Decoder predictions were significantly more similar to the attended story than to the unattended story (* indicates q(FDR) < 0.05 across n = 3 subjects, one-sided paired t-test). Markers indicate individual subjects. **e**, To test whether decoding can succeed without training data from a particular subject, decoders were trained on anatomically aligned brain responses from five sets of other subjects (indicated by markers). Cross-subject decoders performed barely above chance and substantially worse than within-subject decoders (* indicates q(FDR) < 0.05, two-sided t-test), suggesting that within-subject training data are critical. **f**, To test whether decoding can be consciously resisted, subjects silently performed three resistance tasks: counting, naming animals and telling a different story. Decoding performance was compared to a passive listening task (* indicates q(FDR) < 0.05 across n = 3 subjects, one-sided paired t-test). Naming animals and telling a different story significantly lowered decoding performance in each cortical region, demonstrating that decoding can be resisted. Markers indicate individual subjects. Different experiments cannot be compared based on story decoding scores, which depend on stimulus length; see Extended Data Fig. 5 for a comparison based on the fraction of significantly decoded timepoints. Assoc, association; PFC, prefrontal cortex.

fMRI, we smoothed our fMRI data to the estimated spatial resolution of current fNIRS systems and found that around 50% of the stimulus timepoints could still be decoded (Extended Data Fig. 8). This suggests that our decoding approach could eventually be adapted for portable systems.

Finally, to evaluate if decoding performance is limited by model mis-specification—such as using suboptimal features to represent language stimuli—we tested whether the decoding error follows systematic patterns. We scored how well each individual word was decoded across six test stories (Methods) and compared the scores to behavioral word ratings and dataset statistics. If the decoding error were caused solely by noise in the test data, all words should be equally affected. However, we found that decoding performance was significantly correlated with behavioral ratings of word concreteness (rank correlation ρ = 0.14–0.27, q(FDR) < 0.05), suggesting that the decoder is worse at recovering words with certain semantic properties (Fig. 4d). Notably, decoding performance was not significantly correlated with word frequency in the training stimuli, suggesting that model mis-specification is not primarily caused by noise in the training data (Fig. 4e).

Our results indicate that model mis-specification is a major source of decoding error separate from random noise in the training and test data. Assessing how the different components of the decoder contribute to this mis-specification, we found that the decoder continually relies on the encoding model to achieve good performance (Extended Data Fig. 9), and poorly decoded timepoints tend to reflect errors in the encoding model (Extended Data Fig. 10). We expect computational advances that reduce encoding model mis-specification—such as the development of better semantic feature extractors—to substantially improve decoding performance.

Discussion

This study demonstrates that the meaning of perceived and imagined stimuli can be decoded from the BOLD signal into continuous language, marking an important step for non-invasive brain–computer interfaces. Although previous studies have shown that the BOLD signal contains rich semantic information[5,11], our results show that this information is captured at the granularity of individual words and phrases. To reconstruct this information, our decoder relies on two innovations that

https://doi.org/10.1038/s41593-023-01304-9

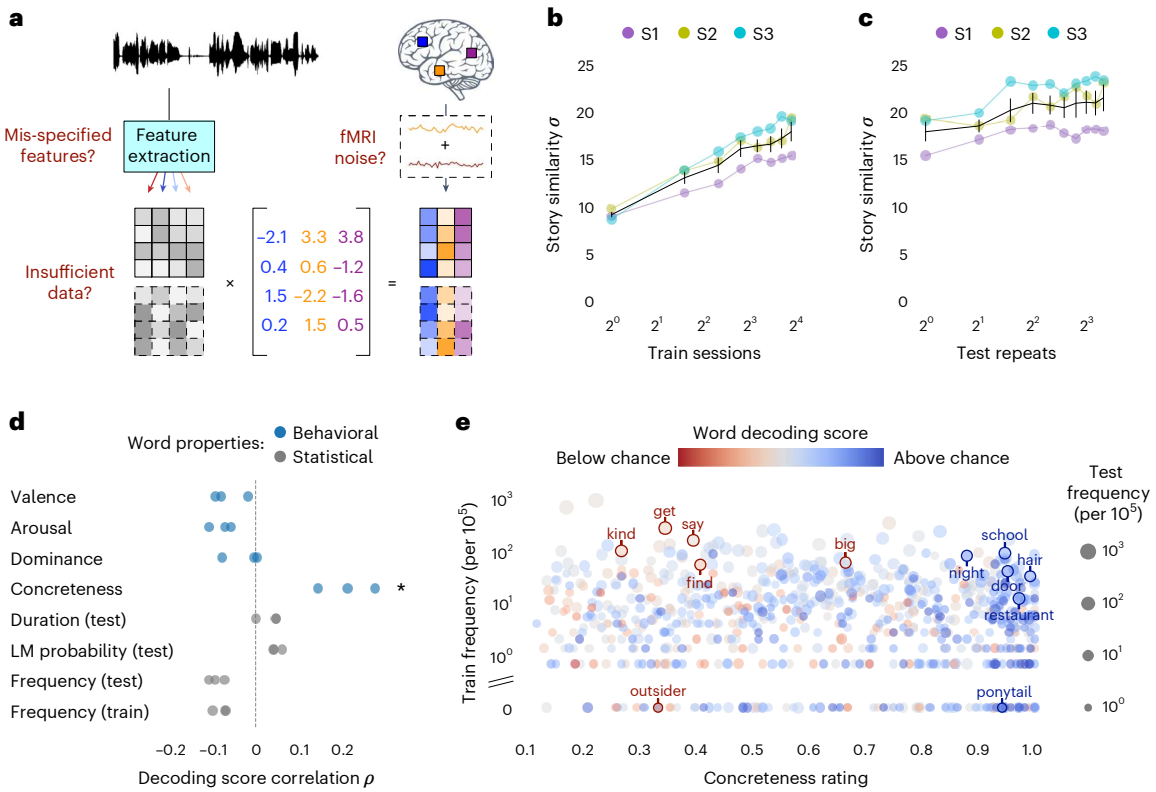


Fig. 4 | Sources of decoding error. a, Potential factors limiting decoding performance. **b**, To test if decoding performance is limited by the size of the training dataset, decoders were trained on different amounts of data. Decoding scores appeared to increase by an equal amount each time the size of the training dataset was doubled. **c**, To test if decoding performance is limited by noise in the test data, the SNR of the test responses was artificially raised by averaging across repeats of the test story. Decoding performance slightly increased with the number of averaged responses. **d**, To test if decoding performance is limited by model mis-specification, word-level decoding scores were compared to behavioral ratings and dataset statistics (* indicates q(FDR) < 0.05 for all subjects, two-sided permutation test). Markers indicate individual subjects. **e**, Decoding performance was significantly correlated with word concreteness, suggesting that model mis-specification contributes to decoding error. Decoding performance was not significantly correlated with word frequency in the training stimuli, suggesting that model mis-specification is not caused by noise in the training data. For all results, black lines indicate the mean across subjects, and error bars indicate the standard error of the mean ($n = 3$). LM, language model.

account for the combinatorial structure of language: an autoregressive prior is used to generate novel sequences, and a beam search algorithm is used to efficiently search for the best sequences. Together, these innovations enable the decoding of structured sequential information from relatively slow brain signals.

Most existing language decoders map brain activity into explicit motor features[1] or record data from regions that encode motor representations during overt or attempted language production[3]. In contrast, our decoder represents language using semantic features and primarily uses data from regions that encode semantic representations[5] during language perception[2]. Unlike motor representations, which are only accessible during attempted speech[1,4], semantic representations are accessible during both attempted and imagined speech. Moreover, semantic representations are shared between language and a range of other cognitive tasks, and our analyses demonstrate that semantic decoders trained during language perception can be used to decode some of these other tasks. This cross-task transfer could enable novel decoder applications, such as covert speech translation, while reducing the need to collect separate training data for different decoder applications.

However, there are also advantages to decoding using motor features. Although our decoder successfully reconstructs the meaning of language stimuli, it often fails to recover exact words (WER 0.92–0.94 for the perceived speech test story). This high WER for novel stimuli is similar to out-of-set performance for existing invasive decoders[45]—which require training on multiple repeats of the test stimuli before attaining a WER below 0.8—indicating that loss of specificity is not unique to non-invasive decoding. In our decoder, loss of specificity occurs when different word sequences with similar meanings share semantic features, causing the decoder to paraphrase the actual stimulus. Motor features are better able to differentiate between the actual stimulus and its paraphrases, as they are directly related to the surface form of the stimulus. Motor features may also give users more control over decoder output, as they are less likely to be correlated with semantic processes such as perception and memory. We may be able to improve the performance of our decoder by modeling language using a combination of semantic features and motor features. This could make use of complementary recording methods such as electroencephalography (EEG) or magnetoencephalography (MEG), which capture precise timing information that is not captured by fMRI[7,8].

One other important factor that may improve decoding performance is subject feedback. Previous invasive studies have employed a closed-loop decoding paradigm, where decoder predictions are shown to the subject in real time[3,4]. This feedback allows the subject to adapt to the decoder, providing them more control over decoder output[46]. Although fMRI has lower temporal resolution than invasive methods, closed-loop decoding may still provide many benefits for imagined speech decoding.

Finally, our privacy analysis suggests that subject cooperation is currently required both to train and to apply the decoder. However, future developments might enable decoders to bypass

these requirements. Moreover, even if decoder predictions are inaccurate without subject cooperation, they could be intentionally misinterpreted for malicious purposes. For these and other unforeseen reasons, it is critical to raise awareness of the risks of brain decoding technology and enact policies that protect each person's mental privacy[47].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-023-01304-9.

References

1. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
2. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol*. **10**, e1001251 (2012).
3. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
4. Moses, D. A. et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
5. 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).
6. 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).
7. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol*. **28**, 803–809 (2018).
8. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
9. Farwell, L. A. & Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol*. **70**, 510–523 (1988).
10. Mitchell, T. M. et al. Predicting human brain activity associated with the meanings of nouns. *Science* **320**, 1191–1195 (2008).
11. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
12. Dash, D., Ferrari, P. & Wang, J. Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Front. Neurosci*. **14**, 290 (2020).
13. Logothetis, N. K. The underpinnings of the BOLD functional magnetic resonance imaging signal. *J. Neurosci.* **23**, 3963–3971 (2003).
14. Jain, S. & Huth, A. G. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems 31* 6629–6638 (NeurIPS, 2018).
15. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems 32* 14928–14938 (NeurIPS, 2019).
16. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
17. LeBel, A., Jain, S. & Huth, A. G. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *J. Neurosci.* **41**, 10341–10355 (2021).
18. Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. & Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902–915 (2009).
19. Nishimoto, S. et al. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol*. **21**, 1641–1646 (2011).
20. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at *OpenAI* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
21. Tillmann, C. & Ney, H. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguist.* **29**, 97–133 (2003).
22. 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).
23. Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends Cogn. Sci.* **15**, 527–536 (2011).
24. 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).
25. Gauthier, J. & Ivanova, A. Does the brain represent words? An evaluation of brain decoding studies of language understanding. In *2018 Conference on Cognitive Computational Neuroscience* 1–4 (CCN, 2018).
26. Fedorenko, E. & Thompson-Schill, S. L. Reworking the language network. *Trends Cogn. Sci.* **18**, 120–126 (2014).
27. Fodor, J. A. *The Modularity of Mind* (MIT Press, 1983).
28. Keller, T. A., Carpenter, P. A. & Just, M. A. The neural bases of sentence comprehension: a fMRI examination of syntactic and lexical processing. *Cereb. Cortex* **11**, 223–237 (2001).
29. Geschwind, N. The organization of language and the brain. *Science* **170**, 940–944 (1970).
30. Barsalou, L. W. Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
31. Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H.-J. & Jancke, L. Scanning silence: mental imagery of complex sounds. *Neuroimage* **26**, 1119–1127 (2005).
32. Martin, S. et al. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* **7**, 14 (2014).
33. Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K. & Gallant, J. L. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage* **105**, 215–228 (2015).
34. Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D. & Hasson, U. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proc. Natl Acad. Sci. USA* **111**, E4687–E4696 (2014).
35. Fairhall, S. L. & Caramazza, A. Brain regions that represent amodal conceptual knowledge. *J. Neurosci.* **33**, 10552–10558 (2013).
36. Popham, S. F. et al. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.* **24**, 1628–1636 (2021).
37. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* **16**, 763–770 (2013).
38. Kiremitçi, I. et al. Attentional modulation of hierarchical speech representations in a multitalker environment. *Cereb. Cortex* **31**, 4986–5005 (2021).
39. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).

40. Horikawa, T. & Kamitani, Y. Attention modulates neural representation to render reconstructions according to subjective appearance. *Commun. Biol.* **5**, 34 (2022).
41. Rainey, S., Martin, S., Christen, A., Mégevand, P. & Fourneret, E. Brain recording, mind-reading, and neurotechnology: ethical issues from consumer devices to brain-based speech decoding. *Sci. Eng. Ethics* **26**, 2295–2311 (2020).
42. Kaplan, J. et al. Scaling laws for neural language models. Preprint at *arxiv* https://doi.org/10.48550/arXiv.2001.08361 (2020).
43. White, B. R. & Culver, J. P. Quantitative evaluation of high-density diffuse optical tomography: in vivo resolution and mapping performance. *J. Biomed. Opt.* **15**, 026006 (2010).
44. Eggebrecht, A. T. et al. A quantitative spatial comparison of high-density diffuse optical tomography and fMRI cortical mapping. *Neuroimage* **61**, 1120–1128 (2012).
45. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nat. Neurosci.* **23**, 575–582 (2020).
46. Orsborn, A. L. et al. Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. *Neuron* **82**, 1380–1393 (2014).
47. Goering, S. et al. Recommendations for responsible development and application of neurotechnologies. *Neuroethics* **14**, 365–386 (2021).
48. Levy, C. *Sintel* (Blender Foundation, 2010).

Methods

Subjects

Data were collected from three female subjects and four male subjects: S1 (female, age 26 years at time of most recent scan), S2 (male, age 36 years), S3 (male, age 23 years), S4 (female, age 23 years), S5 (female, age 23 years), S6 (male, age 25 years) and S7 (male, age 24 years). Data from S1, S2 and S3 were used for the main decoding analyses. Data from all subjects were used to estimate and evaluate cross-subject decoders (Fig. 3e). No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[1,3,4,18,19]. No blinding was performed as there were no experimental groups in the fMRI analyses. All subjects were healthy and had normal hearing and normal or corrected-to-normal vision. To stabilize head motion, subjects wore a personalized head case that precisely fit the shape of each subject's head. The experimental protocol was approved by the institutional review board at The University of Texas at Austin. Written informed consent was obtained from all subjects. Subjects were compensated at a rate of $25 per hour. No data were excluded from analysis.

MRI data collection

MRI data were collected on a 3T Siemens Skyra scanner at the UT Austin Biomedical Imaging Center using a 64-channel Siemens volume coil. Functional scans were collected using gradient echo planar imaging (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.

Anatomical data for all subjects except S2 were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner with voxel size = 1 mm × 1 mm × 1 mm following the FreeSurfer morphometry protocol. Anatomical data for subject S2 were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center with a 32-channel Siemens volume coil using the same sequence.

Cortical regions

Whole-brain MRI data were partitioned into three cortical regions: the speech network, the parietal-temporal-occipital association region and the prefrontal region.

The speech network was functionally localized in each subject using an auditory localizer and a motor localizer. Auditory localizer data were collected in one 10-min scan. The subject listened to 10 repeats of a 1-min auditory stimulus containing 20 s of music (Arcade Fire), speech (Ira Glass, *This American Life*) and natural sound (a babbling brook). To determine whether a voxel was responsive to the auditory stimulus, the repeatability of the voxel response was quantified using an F statistic, which was computed by taking the mean response across the 10 repeats, subtracting this mean response from each single-trial response to obtain single-trial residuals and dividing the variance of the single-trial residuals by the variance of the single-trial responses. This metric directly quantifies the amount of variance in the voxel response that can be explained by the mean response across repeats. The repeatability map was used by a human annotator to define the auditory cortex (AC). Motor localizer data were collected in two identical 10-min scans. The subject was cued to perform six different tasks ('hand', 'foot', 'mouth', 'speak', 'saccade' and 'rest') in a random order in 20-s blocks. For the 'speak' cue, subjects were instructed to self-generate a narrative without vocalization. Linear models were estimated to predict the response in each voxel using the six cues as categorical features. The weight map for the 'speak' feature was used by a human annotator to define Broca's area and the superior ventral premotor (sPMv) speech area. Unlike the parietal-temporal-occipital association and prefrontal regions, there is broad agreement that these speech areas are necessary for speech perception and production. Most existing invasive language decoders record brain activity from these speech areas[1,4,45].

The parietal-temporal-occipital association region and the prefrontal region were anatomically localized in each subject using FreeSurfer regions of interest (ROIs). The parietal-temporal-occipital association region was defined using the *superiorparietal*, *inferiorparietal*, *supramarginal*, *postcentral*, *precuneus*, *superiortemporal*, *middletemporal*, *inferiortemporal*, *bankssts*, *fusiform*, *transversetemporal*, *entorhinal*, *temporalpole*, *parahippocampal*, *lateraloccipital*, *lingual*, *cuneus*, *pericalcarine*, *posteriorcingulate* and *isthmuscingulate* labels. The prefrontal region was defined using the *superiorfrontal*, *rostralmiddlefrontal*, *caudalmiddlefrontal*, *parsopercularis*, *parstriangularis*, *parsorbitalis*, *lateralorbitofrontal*, *medialorbitofrontal*, *precentral*, *paracentral*, *frontalpole*, *rostralanteriorcingulate* and *caudalanteriorcingulate* labels. Voxels identified as part of the speech network (AC, Broca's area and sPMv speech area) were excluded from the parietal-temporal-occipital association region and the prefrontal region. We used a functional definition for the speech network because previous studies have shown that the anatomical location of the speech network varies across subjects[49], whereas we used anatomical definitions for the parietal-temporal-occipital association region and the prefrontal region because these regions are broad and functionally diverse.

To quantify the signal quality in a region, brain responses were recorded while subjects listened to 10 repeats of the test story 'Where There's Smoke' by Jenifer Hixson from *The Moth Radio Hour*. We computed a repeatability score for each voxel by taking the mean response across the 10 repeats, subtracting this mean response from each single-trial response to obtain single-trial residuals and dividing the variance of the single-trial residuals by the variance of the single-trial responses. This metric directly quantifies the amount of variance in the voxel response that can be explained by the mean response across repeats. The speech network had 1,106–1,808 voxels with a mean repeatability score of 0.123–0.245; the parietal-temporal-occipital association region had 4,232–4,698 voxels with a mean repeatability score of 0.070–0.156; and the prefrontal region had 3,177–3,929 voxels with a mean repeatability score of 0.051–0.140.

Experimental tasks

The model training dataset consisted of 82 5–15-min stories taken from *The Moth Radio Hour* and *Modern Love* (Supplementary Table 6). In each story, a single speaker tells an autobiographical narrative. Each story was played during a separate fMRI scan with a buffer of 10 s of silence before and after the story. These data were collected during 16 scanning sessions, with the first session consisting of the anatomical scan and localizers, and the 15 subsequent sessions each consisting of five or six stories. All 15 story sessions were collected for subjects S1, S2 and S3. The first five story sessions were collected for the remaining subjects.

Stories were played over Sensimetrics S14 in-ear piezoelectric headphones. The audio for each stimulus was converted to mono and 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.

Each story was manually transcribed by one listener. Certain sounds (for example, laughter and breathing) were also marked to improve the accuracy of the automated alignment. The audio of each story was then downsampled to 11 kHz, and the Penn Phonetics Lab Forced Aligner (P2FA)[50] was used to automatically align the audio to the transcript. After automatic alignment was complete, Praat and colleagues[51] was used to check and correct each aligned transcript manually.

The model testing dataset consisted of five different fMRI experiments: perceived speech, imagined speech, perceived movie, multi-speaker and decoder resistance. In the perceived speech experiment, subjects listened to 5–15-min stories from *The Moth Radio Hour*, *Modern Love* and *The Anthropocene Reviewed*. These test stories were held out from model training. Each story was played during a single

fMRI scan with a buffer of 10 s of silence before and after the story. For all quantitative perceived speech analyses, we used the test story 'Where There's Smoke' by Jenifer Hixson from *The Moth Radio Hour*.

In the imagined speech experiment, subjects imagined telling 1-min segments from five *Modern Love* stories that were held out from model training. Subjects learned an ID associated with each segment ('alpha', 'bravo', 'charlie', 'delta' and 'echo'). Subjects were cued with each ID over headphones and imagined telling the corresponding segment from memory. Each story segment was cued twice in a single 14-min fMRI scan, with 10 s of preparation time after each cue and 10 s of rest time after each segment.

In the perceived movie experiment, subjects viewed four 4–6-min movie clips from animated short films: 'La Luna' (Pixar Animation Studios)[52], 'Presto' (Pixar Animation Studios)[53], 'Partly Cloudy' (Pixar Animation Studios)[54] and 'Sintel' (Blender Foundation)[48]. The movie clips were self-contained and almost entirely devoid of language. The original high-definition movie clips were cropped and downsampled to 727 × 409 pixels. Subjects were instructed to pay attention to the movie events. Notably, subjects were not instructed to generate an internal narrative. Each movie clip was presented without sound during a single fMRI scan, with a 10-s black screen buffer before and after the movie clip.

In the multi-speaker experiment, subjects listened to two repeats of a 6-min stimulus constructed by temporally overlaying a pair of stories from *The Moth Radio Hour* told by a female and a male speaker. Both stories were held out from model training. The speech waveforms of the two stories were converted to mono and temporally overlaid. Subjects attended to the female speaker for one repeat and the male speaker for the other, with the order counterbalanced across subjects. Each repeat was played during a single fMRI scan with a buffer of 10 s of silence before and after the stimulus.

In each trial of the decoder resistance experiment, subjects were played one of four 80-s segments from a test story over headphones. Before the segment, subjects were cued to perform one of four cognitive tasks ('listen', 'count', 'name' and 'tell'). For the 'listen' cue, subjects were instructed to passively listen to the story segment. For the 'count' cue, subjects were instructed to count by sevens in their heads. For the 'name' cue, subjects were instructed to name and imagine animals in their heads. For the 'tell' cue, subjects were instructed to tell different stories in their heads. For all cues, subjects were instructed not to speak or make any other movements. Trials were balanced such that (1) each task was the first to be cued for some segment and (2) each task was cued exactly once for every segment, resulting in a total of 16 trials. We conducted two 14-min fMRI scans each comprising eight trials, with 10 s of preparation time after each cue and 10 s of rest time after each trial.

fMRI data pre-processing

Each functional run was motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0 (ref. 55). All volumes in the run were then averaged to obtain a high-quality template volume. FLIRT was then used to align the template volume for each run to the overall template, which was chosen to be the template for the first functional run for each subject. These automatic alignments were manually checked.

Low-frequency voxel response drift was identified using a second-order Savitzky–Golay filter with a 120-s window and then subtracted from the signal. The mean response for each voxel was then subtracted, and the remaining response was scaled to have unit variance.

Cortical surface reconstruction and visualization

Cortical surface meshes were generated from the T1-weighted anatomical scans using FreeSurfer[56]. Before surface reconstruction, anatomical surface segmentations were hand-checked and corrected. Blender was used to remove the corpus callosum and make relaxation cuts for flattening. Functional images were aligned to the cortical surface using boundary based registration (BBR) implemented in FSL. These alignments were manually checked for accuracy, and adjustments were made as necessary.

Flatmaps were created by projecting the values for each voxel onto the cortical surface using the 'nearest' scheme in pycortex[57]. This projection finds the location of each pixel in the flatmap in three-dimensional (3D) space and assigns that pixel the associated value.

Language model

Generative Pre-trained Transformer (GPT, also known as GPT-1) is a 12-layer neural network that uses multi-head self-attention to combine representations of each word in a sequence with representations of previous words[20]. GPT was trained on a large corpus of books to predict the probability distribution over the next word s_n in a sequence $(s_1, s_2, \ldots, s_{n-1})$.

We fine-tuned GPT on a corpus comprising Reddit comments (over 200 million total words) and 240 autobiographical stories from *The Moth Radio Hour* and *Modern Love* that were not used for decoder training or testing (over 400,000 total words). The model was trained for 50 epochs with a maximum context length of 100.

GPT estimates a prior probability distribution $P(S)$ over word sequences. Given a word sequence $S = (s_1, s_2, \ldots, s_n)$, GPT computes the probability of observing S in natural language by multiplying the probabilities of each word conditioned on the previous words: $P(S) = \prod_1^n P(s_i|s_{1:i-1})$ where $s_{1:0}$ is the empty sequence $\emptyset$.

GPT is also used to extract semantic features from language stimuli. To successfully perform the next word prediction task, GPT learns to extract quantitative features that capture the meaning of input sequences. Given a word sequence $S = (s_1, s_2, \ldots, s_n)$, the GPT hidden layer activations provide vector embeddings that represent the meaning of the most recent word s_n in context.

Encoding model

In voxel-wise modeling, quantitative features are extracted from stimulus words, and regularized linear regression is used to estimate a set of weights that predict how each feature affects the BOLD signal in each voxel.

A stimulus matrix was constructed from the training stories. For each word–time pair (s_i, t_i) in each story, we provided the word sequence $(s_{i-5}, s_{i-4}, \ldots, s_{i-1}, s_i)$ to the GPT language model and extracted semantic features of s_i from the ninth layer. Previous studies have shown that middle layers of language models extract the best semantic features for predicting brain responses to natural language[8,14,15,17]. This yields a new list of vector–time pairs $(\mathbf{M}_i, t_i)$ where $\mathbf{M}_i$ is a 768-dimensional semantic embedding for s_i. These vectors were then resampled at times corresponding to the fMRI acquisitions using a three-lobe Lanczos filter[5].

A linearized finite impulse response (FIR) model was fit to every cortical voxel in each subject's brain[5]. A separate linear temporal filter with four delays ($t-1$, $t-2$, $t-3$ and $t-4$ timepoints) was fit for each of the 768 features, yielding a total of 3,072 features. With a TR of 2 s, this was accomplished by concatenating the feature vectors from 2 s, 4 s, 6 s and 8 s earlier to predict responses at time t. Taking the dot product of this concatenated feature space with a set of linear weights is functionally equivalent to convolving the original stimulus vectors with linear temporal kernels that have non-zero entries for 1-, 2-, 3- and 4-timepoint delays. Before doing regression, we first z-scored each feature channel across the training matrix. This was done to match the features to the fMRI responses, which were z-scored within each scan.

The 3,072 weights for each voxel were estimated using L2-regularized linear regression[5]. The regression procedure has a single free parameter that controls the degree of regularization. This regularization coefficient was found for each voxel in each subject by repeating a regression and cross-validation procedure 50 times. In each iteration, approximately a fifth of the timepoints were removed

from the model training dataset and reserved for validation. Then, the model weights were estimated on the remaining timepoints for each of 10 possible regularization coefficients (log spaced between 10 and 1,000). These weights were used to predict responses for the reserved timepoints, and then R^2 was computed between actual and predicted responses. For each voxel, the regularization coefficient was chosen as the value that led to the best performance, averaged across bootstraps, on the reserved timepoints. The 10,000 cortical voxels with the highest cross-validation performance were used for decoding.

The encoding model estimates a function $\hat{R}$ that maps from semantic features S to predicted brain responses $\hat{R}(S)$. Assuming that BOLD signals are affected by Gaussian additive noise, the likelihood of observing brain responses R given semantic features S can be modeled as a multivariate Gaussian distribution $P(R|S)$ with mean $\mu = \hat{R}(S)$ and covariance $\sum = \left\langle (R - \hat{R}(S))^T (R - \hat{R}(S)) \right\rangle$ (ref. 19). Previous studies estimated the noise covariance $\sum$ using the residuals between the predicted responses and the actual responses to the training dataset[19]. However, this underestimates the actual noise covariance, because the encoding model learns to predict some of the noise in the training dataset during model estimation. To avoid this issue, we estimated $\sum$ using a bootstrap procedure. Each story was held out from the model training dataset, and an encoding model was estimated using the remaining data. A bootstrap noise covariance matrix for the held-out story was computed using the residuals between the predicted responses and the actual responses to the held-out story. We estimated $\sum$ by averaging the bootstrap noise covariance matrices across held-out stories.

All model fitting and analysis was performed using custom software written in Python, making heavy use of NumPy[58], SciPy[59], PyTorch[60], Transformers[61] and pycortex[57].

Word rate model

A word rate model was estimated for each subject to predict when words were perceived or imagined. The word rate at each fMRI acquisition was defined as the number of stimulus words that occurred since the previous acquisition. Regularized linear regression was used to estimate a set of weights that predict the word rate w from the brain responses R. To predict word rate during perceived speech, brain responses were restricted to the auditory cortex. To predict word rate during imagined speech and perceived movies, brain responses were restricted to Broca's area and the sPMv speech area. A separate linear temporal filter with four delays ($t+1$, $t+2$, $t+3$ and $t+4$) was fit for each voxel. With a TR of 2 s, this was accomplished by concatenating the responses from 2 s, 4 s, 6 s and 8 s later to predict the word rate at time t. Given novel brain responses, this model predicts the word rate at each acquisition. The time between consecutive acquisitions (2 s) is then evenly divided by the predicted word rates (rounded to the nearest non-negative integers) to predict word times.

Beam search decoder

Under Bayes' theorem, the distribution $P(S|R)$ over word sequences given brain responses can be factorized into a prior distribution $P(S)$ over word sequences and an encoding distribution $P(R|S)$ over brain responses given word sequences. Given novel brain responses R_{test}, the most likely word sequence S_{test} could theoretically be identified by evaluating $P(S)$—with the language model—and $P(R_{\text{test}}|S)$—with the subject's encoding model—for all possible word sequences S. However, the combinatorial structure of natural language makes it computationally infeasible to evaluate all possible word sequences. Instead, we approximated the most likely word sequence using a beam search algorithm[21].

The decoder maintains a beam containing the k most likely word sequences. The beam is initialized with an empty word sequence. When new words are detected by the word rate model, the language model generates continuations for each candidate S in the beam. The language model uses the last 8 s of predicted words $(s_{n-i}, \ldots, s_{n-1})$ in the candidate to predict the distribution $P(s_n|s_{n-i}, \ldots, s_{n-1})$ over the next word. The decoder does not have access to the actual stimulus words. The probability distribution over the decoder vocabulary—which consists of the 6,867 unique words that occurred at least twice in the encoding model training dataset—was rescaled to sum to 1. Nucleus sampling[62] is used to identify words that belong to the top p percent of the probability mass and have a probability within a factor r of the most likely word. Content words that occur in the language model input $(s_{n-i}, \ldots, s_{n-1})$ are filtered out, as language models have been shown to be biased toward such words. Each word in the remaining nucleus is appended to the candidate to form a continuation C.

The encoding model scores each continuation by the likelihood $P(R_{\text{test}}|C)$ of observing the recorded brain responses. The k most likely continuations across all candidates are retained in the beam. To increase beam diversity, we accept a maximum of five continuations for each candidate. To increase linguistic coherence, the number of accepted continuations for a candidate is determined by the probability of the candidate under the language model. Candidates in the top quintile under $P(S)$ are permitted the maximum five continuations. Candidates in the next quintile are permitted four continuations and so on, with candidates in the bottom quintile permitted one continuation. After iterating through all of the predicted word times, the decoder outputs the candidate sequence with the highest likelihood.

Bayesian decoders have previously been used to decode perceived images and videos[18,19]. Our decoder differs from existing Bayesian decoders in two important ways. First, existing Bayesian decoders collect a large empirical prior of images or videos and only compute $P(R|S)$ for stimuli in the empirical prior. The decoder prediction is obtained by choosing the most likely stimulus or taking a weighted combination of the stimuli. In contrast, our decoder uses a generative language model prior, which can produce completely novel sequences. Second, existing Bayesian decoders evaluate all stimuli in the empirical prior. In contrast, our decoder uses a beam search algorithm to efficiently search the combinatorial space of possible sequences, so the words that are evaluated at each point in time depend on the words that were previously decoded. Together, these two innovations enable our decoder to efficiently reconstruct structured sequential information.

Decoder parameters

The decoder has several parameters that affect model performance. The beam search algorithm is parameterized by the beam width k. The encoding model is parameterized by the number of context words provided when extracting GPT embeddings. The noise model is parameterized by a shrinkage factor α that regularizes the covariance $\sum$. Language model parameters include the length of the input context, the nucleus mass p and ratio r and the set of possible output words.

In preliminary analyses, we found that decoding performance increased with the beam width but plateaued after $k = 200$, so we used a beam width of 200 sequences for all analyses. All other parameters were tuned by grid search and by hand on data collected as subject S3 listened to a calibration story separate from the training and test stories ('From Boyhood to Fatherhood' by Jonathan Ames from *The Moth Radio Hour*). We decoded the calibration story using each configuration of parameters. The best-performing parameter values were validated and adjusted through qualitative analysis of decoder predictions. The parameters that had the largest effect on decoding performance were the nucleus ratio r and the noise model shrinkage α. Setting r to be too small makes the decoder less linguistically coherent, whereas setting r to be too large makes the decoder less semantically correct. Setting α to be too small overestimates the actual noise covariance, whereas setting α to be too large underestimates the actual noise covariance; both make the decoder less semantically correct. The parameter values used in this study provide a default decoder configuration but, in practice, can be tuned separately and continually for each subject to improve performance.

To ensure that our results generalize to new subjects and stimuli, we restricted all pilot analyses to data collected as subject S3 listened to the test story 'Where There's Smoke' by Jenifer Hixson from *The Moth Radio Hour*. All pilot analyses on the test story were qualitative. We froze the analysis pipeline before we viewed any results for the remaining subjects, stimuli and experiments.

Language similarity metrics

Decoded word sequences were compared to reference word sequences using a range of automated metrics for evaluating language similarity. WER computes the number of edits (word insertions, deletions or substitutions) required to change the predicted sequence into the reference sequence. BLEU[63] computes the number of predicted n-grams that occur in the reference sequence (precision). We used the unigram variant BLEU-1. METEOR[64] combines the number of predicted unigrams that occur in the reference sequence (precision) with the number of reference unigrams that occur in the predicted sequence (recall) and accounts for synonymy and stemming using external databases. BERTScore[65] uses a bidirectional transformer language model to represent each word in the predicted and reference sequences as a contextualized embedding and then computes a matching score over the predicted and reference embeddings. We used the recall variant of BERTScore with inverse document frequency (IDF) importance weighting computed across stories in the training dataset. BERTScore was used for all analyses where the language similarity metric is not specified.

For the perceived speech, multi-speaker and decoder resistance experiments, stimulus transcripts were used as reference sequences. For the imagined speech experiment, subjects told each story segment out loud outside of the scanner, and the audio was recorded and manually transcribed to provide reference sequences. For the perceived movie experiment, official audio descriptions from Pixar Animation Studios were manually transcribed to provide reference sequences for three movies. To compare word sequences decoded from different cortical regions (Fig. 2d), each sequence was scored using the other as reference, and the scores were averaged (prediction similarity).

We scored the predicted and reference words within a 20-s window around every second of the stimulus (window similarity). Scores were averaged across windows to quantify how well the decoder predicted the full stimulus (story similarity).

To estimate a ceiling for each metric, we had the perceived speech test story 'Where There's Smoke' translated into Mandarin Chinese by a professional translator. The translator was instructed to preserve all of the details of the story in the correct order. We then translated the story back into English using a state-of-the-art machine translation system. We scored the similarity between the original story words and the output of the machine translation system. These scores provide a ceiling for decoding performance, because modern machine translation systems are trained on large amounts of paired data, and the Mandarin Chinese translation contains virtually the same information as the original story words.

To test whether perceived speech timepoints can be identified using decoder predictions, we performed a post hoc identification analysis using similarity scores between the predicted and reference sequences. We constructed a matrix M where M_{ij} reflects the similarity between the i-th predicted window and the j-th reference window. For each timepoint i, we sorted all of the reference windows by their similarity to the i-th predicted window and scored the timepoint by the percentile rank of the i-th reference window. The mean percentile rank for the full stimulus was obtained by averaging percentile ranks across timepoints.

To test whether imagined speech scans can be identified using decoder predictions, we performed a post hoc identification analysis using similarity scores between the predicted and reference sequences. For each scan, we normalized the similarity scores between the decoder prediction and the five reference transcripts into probabilities. We computed top-1 accuracy by assessing whether the decoder prediction for each scan was most similar to the correct transcript. We observed 100% top-1 accuracy for each subject. We computed cross-entropy for each scan by taking the negative logarithm (base 2) of the probability of the correct transcript. We observed a mean cross-entropy of 0.23–0.83 bits. A perfect decoder would have a cross-entropy of 0 bits, and a chance-level decoder would have a cross-entropy of $\log_2(5) = 2.32$ bits.

Statistical testing

To test statistical significance of the word rate model, we computed the linear correlation between the predicted and the actual word rate vectors across a test story and generated 2,000 null correlations by randomly shuffling 10-TR segments of the actual word rate vector. We compared the observed linear correlation to the null distribution using a one-sided permutation test; P values were computed as the fraction of shuffles with a linear correlation greater than or equal to than the observed linear correlation.

To test statistical significance of the decoding scores, we generated null sequences by sampling from the language model without using any brain data except to predict word times. We separately evaluated the word rate model and the decoding scores because the language similarity metrics used to compute the decoding scores are affected by the number of words in the predicted sequences. By generating null sequences with the same word times as the predicted sequence, our test isolates the ability of the decoder to extract semantic information from the brain data. To generate null sequences, we followed the same beam search procedure as the actual decoder. The null model maintains a beam of 10 candidate sequences and generates continuations from the language model nucleus[62] at each predicted word time. The only difference between the actual decoder and the null model is that, instead of ranking the continuations by the likelihood of the fMRI data, the null model randomly assigns a likelihood to each continuation. After iterating through all of the predicted word times, the null model outputs the candidate sequence with the highest likelihood. We repeated this process 200 times to generate 200 null sequences. This process is as similar as possible to the actual decoder without using any brain data to select words, so these sequences reflect the null hypothesis that the decoder does not recover meaningful information about the stimulus from the brain data. We scored the null sequences against the reference sequence to produce a null distribution of decoding scores. We compared the observed decoding scores to this null distribution using a one-sided non-parametric test; P values were computed as the fraction of null sequences with a decoding score greater than or equal to the observed decoding score.

To check that the null scores are not trivially low, we compared the similarity scores between the reference sequence and the 200 null sequences to the similarity scores between the reference sequence and the transcripts of 62 other narrative stories. We found that the mean similarity between the reference sequence and the null sequences was higher than the mean similarity between the reference sequence and the other story transcripts, indicating that the null scores are not trivially low.

To test statistical significance of the post hoc identification analysis, we randomly shuffled 10-row blocks of the similarity matrix M before computing mean percentile ranks. We evaluated 2,000 shuffles to obtain a null distribution of mean percentile ranks. We compared the observed mean percentile rank to this null distribution using a one-sided permutation test; P values were computed as the fraction of shuffles with a mean percentile rank greater than or equal to than the observed mean percentile rank.

Unless otherwise stated, all tests were performed within each subject and then replicated across all subjects ($n = 7$ for the cross-subject decoding analysis shown in Fig. 3e, $n = 3$ for all other analyses). All tests were corrected for multiple comparisons when necessary using

the FDR[66]. Data distributions were assumed to be normal, but this was not formally tested due to our small-*n* study design. Distributions of individual data points used in *t*-tests are shown in Fig. 3d–f. The range across subjects was reported for all quantitative results.

Behavioral comprehension assessment

To assess the intelligibility of decoder predictions, we conducted an online behavioral experiment to test whether other people could answer multiple-choice questions about a stimulus story using just a subject's decoder predictions (Extended Data Fig. 3). We chose four 80-s segments of the perceived speech test story on the basis of being relatively self-contained. For each segment, we wrote four multiple-choice questions about the actual stimulus without looking at the decoder predictions. To further ensure that the questions were not biased toward the decoder predictions, the multiple-choice answers were written by a separate researcher who had never seen the decoder predictions.

The experiment was presented as a Qualtrics questionnaire. We recruited 100 online subjects (50 female, 49 male and 1 non-binary) between the ages of 19 years and 70 years over Prolific and randomly assigned them to experimental and control groups. Researchers and participants were blinded to group assignment. For each segment, the experimental group subjects were shown the decoded words from subject S3, whereas the control group subjects were shown the actual stimulus words. Control group participants were expected to perform close to ceiling accuracy, so we determined a priori that a sample size of 100 provides sufficient power to detect significance differences with test accuracies as high as 70% (G*Power[67], exact test of proportions with independent groups). The words for each segment and the corresponding multiple-choice questions were shown together on a single page of the Qualtrics questionnaire. Segments were shown in story order. Back button functionality was disabled, so subjects were not allowed to change their answers for previous segments after seeing a new segment. The experimental protocol was approved by the institutional review board at The University of Texas at Austin. Informed consent was obtained from all subjects. Participants were paid $4 to complete the questionnaire, corresponding to an average rate of $24 per hour. No data were excluded from analysis.

Sources of decoding error

To test if decoding performance is limited by the size of our training dataset, we trained decoders on different amounts of data. Decoding scores appeared to linearly increase each time the size of the training dataset was doubled. To test if the diminishing returns of adding training data are due to the fact that decoders were trained on overlapping samples of data, we used a simulation to compare how decoders would perform when trained on non-overlapping and overlapping samples of data. We used the actual encoding model and the actual noise model to simulate brain responses to 36 sessions of training stories. We obtained non-overlapping samples of 3, 7, 11 and 15 sessions by taking sessions 1 through 3, 4 through 10, 11 through 21 and 22 through 36. We obtained overlapping samples of 3, 7, 11 and 15 sessions by taking sessions 1 through 3, 1 through 7, 1 through 11 and 1 through 15. We trained decoders on these simulated datasets and found that the relationship between decoding scores and the number of training sessions was very similar for the non-overlapping and overlapping datasets (Supplementary Fig. 1). This suggests that the observed diminishing returns of adding training data are not due to the fact that decoders were trained on overlapping samples of data.

To test if decoding performance relies on the high spatial resolution of fMRI, we spatially smoothed the fMRI data by convolving each image with a 3D Gaussian kernel (Extended Data Fig. 8). We tested Gaussian kernels with standard deviations of 1, 2, 3, 4 and 5 voxels, corresponding to 6.1 mm, 12.2 mm, 18.4 mm, 24.5 mm and 30.6 mm full width at half maximum (FWHM). We estimated the encoding model, noise model and word rate model on spatially smoothed perceived speech training data and evaluated the decoder on spatially smoothed perceived speech test data.

To test if decoding performance is limited by noise in the test data, we artificially raised the SNR of the test responses by averaging across repeats of a test story.

To test if decoding performance is limited by model misspecification, we quantified word-level decoding performance by representing words using 300-dimensional GloVe embeddings[68]. We considered a 10-s window centered around each stimulus word. We computed the maximum linear correlation between the stimulus word and the predicted words in the window. Then, for each of the 200 null sequences, we computed the maximum linear correlation between the stimulus word and the null words in the window. The match score for the stimulus word was defined as the number of null sequences with a maximum correlation less than the maximum correlation of the predicted sequence. Match scores above 100 indicate higher decoding performance than expected by chance, whereas match scores below 100 indicate lower decoding performance than expected by chance. Match scores were averaged across all occurrences of a word in six test stories. The word-level match scores were compared to behavioral ratings of valence (pleasantness), arousal (intensity of emotion), dominance (degree of exerted control) and concreteness (degree of sensory or motor experience)[69,70]. Each set of behavioral ratings was linearly rescaled to be between 0 and 1. The word-level match scores were also compared to word duration in the test dataset, language model probability in the test dataset (which corresponds to the information conveyed by a word)[71], word frequency in the test dataset and word frequency in the training dataset.

Decoder ablations

When the word rate model detects new words, the language model proposes continuations using the previously predicted words as autoregressive context, and the encoding model ranks the continuations using the fMRI data. To understand the relative contributions of the autoregressive context and the fMRI data to decoding performance, we evaluated decoders on perceived speech data in the absence of each component (Extended Data Fig. 9). We performed the standard decoding approach up to a cutoff point in the perceived speech test story. After the cutoff, we either reset the autoregressive context or removed the fMRI data. To reset the autoregressive context, we discarded all of the candidate sequences and re-initialized the beam with an empty sequence. We then performed the standard decoding approach for the remainder of the scan. To remove the fMRI data, we assigned random likelihoods (rather than encoding model likelihoods) to continuations for the remainder of the scan.

Isolated encoding model and language model scores

In practice, the decoder uses the previously predicted words to predict the next word. This use of autoregressive context causes errors to propagate between the encoding model and the language model, making it difficult to attribute errors to one component or the other. To isolate errors introduced by each component, we separately evaluated the decoder components on the perceived speech test story using the actual, rather than the predicted, stimulus words as context (Extended Data Fig. 10). At each word time *t*, we provided the encoding model and the language model with the actual stimulus word as well as 100 randomly sampled distractor words.

To evaluate how well the word at time *t* can be decoded using the encoding model, we used the encoding model to rank the actual stimulus word and the 100 distractor words based on the likelihood of the recorded responses. We computed an isolated encoding model score based on the number of distractor words ranked below the actual word. Because the encoding model scores are independent from errors

in the language model and the autoregressive context, they provide a ceiling for how well each word can be decoded from the fMRI data.

To evaluate how well the word at time *t* can be generated using the language model, we used the language model to rank the actual stimulus word and the 100 distractor words based on their probability given the previous stimulus words. We computed an isolated language model score based on the number of distractor words ranked below the actual word. Because the language model scores are independent from errors in the encoding model and the autoregressive context, they provide a ceiling for how well each word can be generated by the language model.

For both the isolated encoding model and the language model scores, 100 indicates perfect performance, and 50 indicates chance-level performance. The isolated encoding model and language scores were computed for each word. To compare against the full decoding scores from Fig. 1e, the word-level scores were averaged across 20-s windows of the stimulus.

Anatomical alignment

To test if decoders could be estimated without any training data from a target subject, volumetric[55] and surface-based[72] methods were used to anatomically align training data from separate source subjects into the volumetric space of the target subject.

For volumetric alignment, we used the *get_mnixfm* function in pycortex to compute a linear map from the volumetric space of each source subject to the MNI template space. This map was applied to recorded brain responses for each training story using the *transform_to_mni* function in pycortex. We then used the *transform_mni_to_subject* function in pycortex to map the responses in MNI152 space to the volumetric space of the target subject. We *z*-scored the response timecourse for each voxel in the volumetric space of the target subject.

For surface-based alignment, we used the *get_mri_surf2surf_matrix* function in pycortex to compute a map from the surface vertices of each source subject to the surface vertices of the target subject. This map was applied to the recorded brain responses for each training story. We then mapped the surface vertices of the target subject into the volumetric space of the target subject using the *line-nearest* scheme in pycortex. We *z*-scored the response timecourse for each voxel in the volumetric space of the target subject.

We used a bootstrap procedure to sample five sets of source subjects for the target subject. Each source subject independently produced aligned responses for the target subject. To estimate the encoding model and word rate model, we averaged the aligned responses across the source subjects. For the word rate model, we localized the speech network of the target subject by anatomically aligning the speech networks of the source subjects. To estimate the noise model $\sum$, we used aligned responses from a single, randomly sampled source subject to compute the bootstrap noise covariance matrix for each held-out training story. The cross-subject decoders were evaluated on actual responses recorded from the target subject.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data collected during the decoder resistance experiment are available upon reasonable request but were not publicly released due to concern that the data could be used to discover ways to bypass subject resistance. All other data are available at https://openneuro.org/datasets/ds003020 and https://openneuro.org/datasets/ds004510.

Code availability

Custom decoding code is available at https://github.com/HuthLab/semantic-decoding.

References

49. Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
50. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878 (2008).
51. Boersma, P. & Weenink, D. *Praat: doing phonetics by computer* (University of Amsterdam, 2014).
52. Casarosa, E. *La Luna* (Walt Disney Pictures; Pixar Animation Studios, 2011).
53. Sweetland, D. *Presto* (Walt Disney Pictures; Pixar Animation Studios, 2008).
54. Sohn, P. *Partly Cloudy* (Walt Disney Pictures; Pixar Animation Studios, 2009).
55. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
56. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
57. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23 (2015).
58. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
59. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
60. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* 8024–8035 (NeurIPS, 2019).
61. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020).
62. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. In *8th International Conference on Learning Representations* 1–16 (ICLR, 2020).
63. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (Association for Computational Linguistics, 2002).
64. Banerjee, S. & Lavie, A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72 (Association for Computational Linguistics, 2005).
65. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: evaluating text generation with BERT. In *8th International Conference on Learning Representations* 1–43 (ICLR, 2020).
66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
67. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
68. Pennington, J., Socher, R. & Manning, C. D. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* 1532–1543 (Association for Computational Linguistics, 2014).

69. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013).
70. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
71. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
72. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).

Acknowledgements

We thank J. Wang, X. X. Wei and L. Hamilton for comments on the manuscript and A. Arcot for writing answers to the behavioral comprehension questions. This work was supported by the National Institute on Deafness and Other Communication Disorders under award number 1R01DC020088-001 (A.G.H.), the Whitehall Foundation (A.G.H.), the Alfred P. Sloan Foundation (A.G.H.) and the Burroughs Wellcome Fund (A.G.H.).

Author contributions

Conceptualization: J.T. and A.G.H.; Methodology: J.T.; Software and resources: J.T. and S.J.; Investigation and data curation: J.T. and A.L.; Formal analysis and visualization: J.T.; Writing (original draft): J.T.; Writing (review and editing): J.T., A.L., S.J. and A.G.H.; Supervision: A.G.H.

Competing interests

A.G.H. and J.T. are inventors on a pending patent application (the applicant is The University of Texas System) that is directly relevant to the language decoding approach used in this work. All other authors declare no competing interests.

Additional information

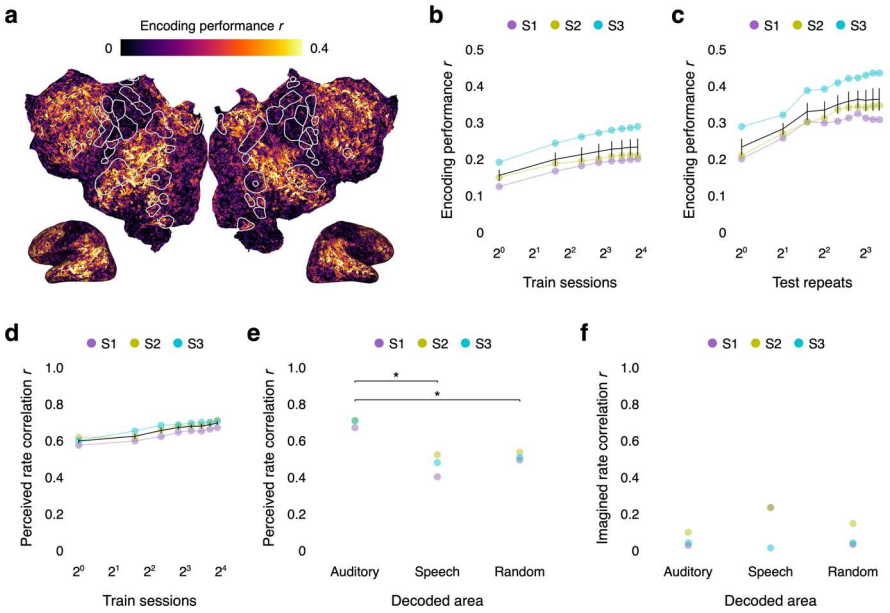
Extended data is available for this paper at https://doi.org/10.1038/s41593-023-01304-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-023-01304-9.

Correspondence and requests for materials should be addressed to Alexander G. Huth.

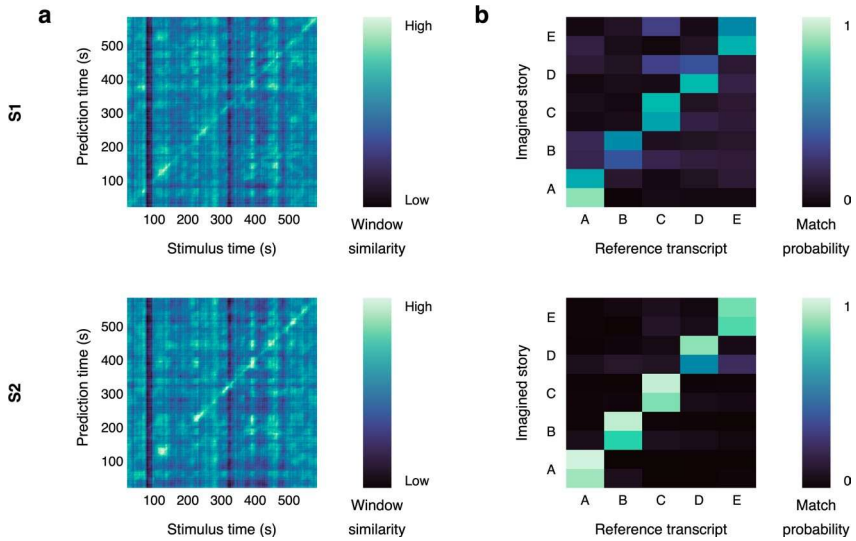
Peer review information *Nature Neuroscience* thanks Gregory Cogan, Stephen David and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41593-023-01304-9



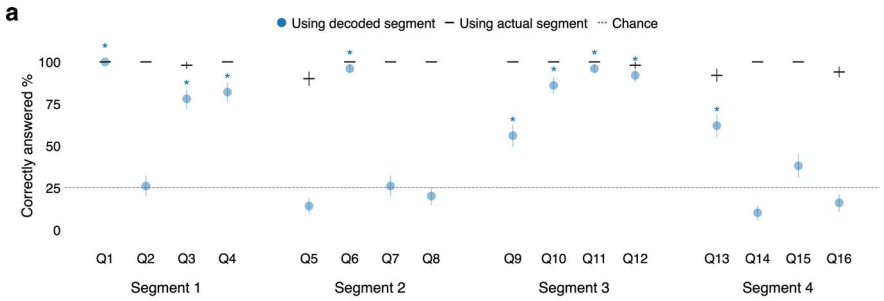
Extended Data Fig. 1 | Encoding model and word rate model performance. The two decoder components that interface with fMRI data are the encoding model and the word rate model. (**a**) Encoding models were evaluated by predicting brain responses to the perceived speech test story and computing the linear correlation between the predicted responses and the actual single-trial responses. Correlations for subject S3 were projected onto a cortical flatmap. The encoding model successfully predicted brain responses in most cortical regions outside of primary sensory and motor areas. (**b**) Encoding models were trained on different amounts of data. To summarize encoding model performance across cortex, correlations were averaged across the 10,000 voxels used for decoding. Encoding model performance increased with the amount of training data collected from each subject. (**c**) Encoding models were tested on brain responses that were averaged across different repeats of the perceived speech test story to artificially increase the signal-to-noise ratio (SNR). Encoding model performance increased with the number of averaged responses. (**d**) Word rate models were trained on different amounts of data. Word rate models were evaluated by predicting the word rate of a test story and computing the linear correlation between the predicted and the actual word rate vectors. Word rate model performance slightly increased with the amount of training data collected from each subject. (**e**) For brain responses to perceived speech, word rate models fit on auditory cortex significantly outperformed word rate models fit on frontal speech production areas or randomly sampled voxels (* indicates q(FDR) < 0.05 across $n = 3$ subjects, two-sided paired t-test). (**f**) For brain responses to imagined speech, there were no significant differences in performance for word rate models fit on different cortical regions. For all results, black lines indicate the mean across subjects and error bars indicate the standard error of the mean ($n = 3$).

https://doi.org/10.1038/s41593-023-01304-9



Extended Data Fig. 2 | Perceived and imagined speech identification performance. Language decoders were trained for subjects S1 and S2 on fMRI responses recorded while the subjects listened to narrative stories. (**a**) The decoders were evaluated on single-trial fMRI responses recorded while the subjects listened to the perceived speech test story. The color at (*i*,*j*) reflects the BERTScore similarity between the *i*th second of the decoder prediction and the *j*th second of the actual stimulus. Identification accuracy was significantly higher than expected by chance ($P < 0.05$, one-sided permutation test). Corresponding results for subject S3 are shown in Fig. 1f in the main text. (**b**) The decoders were evaluated on single-trial fMRI responses recorded while the subjects imagined telling five 1-minute test stories twice. Decoder predictions were compared to reference transcripts that were separately recorded from the same subjects. Each row corresponds to a scan, and the colors reflect the similarities between the decoder prediction and all five reference transcripts. For each scan, the decoder prediction was most similar to the reference transcript of the correct story (100% identification accuracy). Corresponding results for subject S3 are shown in Fig. 3a in the main text.

https://doi.org/10.1038/s41593-023-01304-9



b

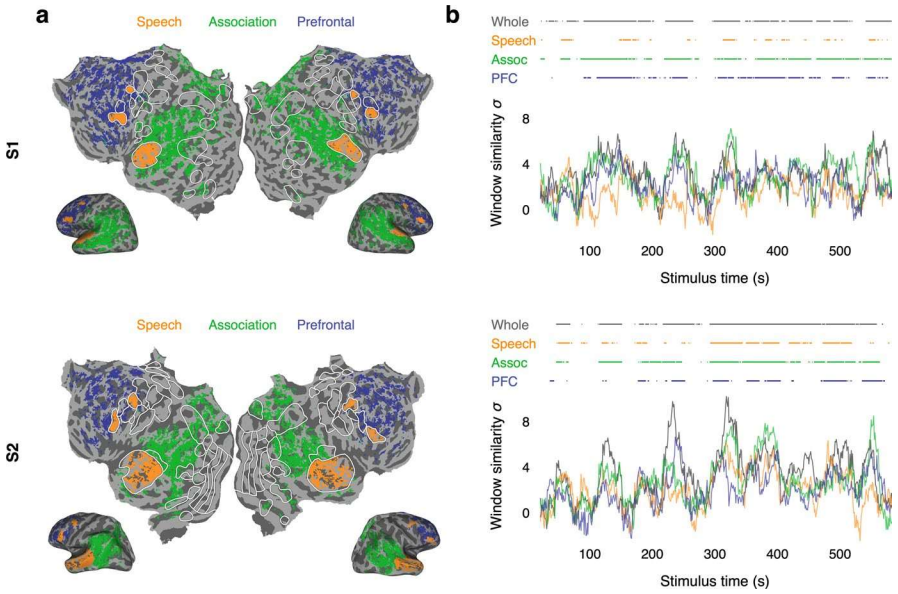
Decoded segment 1	Actual segment 1
we got in my car and i was crying i didn't have my purse i don't have any money to pay for gas i wasn't a very good friend to anyone that i had known since my dad was an alcoholic he was abusive to everyone and it was very hard on me as well as him to do anything that he said or did i would say a word to him in anger or a threatening way in his case it was always just an excuse to leave and that was why i did it and i think that when i finally did i ended up moving to an area with very few houses on the property so the neighbors wouldn't hear my car stop in the driveway and run out and tell me to leave and not come back i said ok and went home to pack a bag i told her about it and she was like wow you know how you get on with a boyfriend after a couple weeks i went home and it was a beautiful evening and the party was getting really packed we had a table set for six	i reached over and secretly undid my seatbelt and when his foot hit the brake at the red light i flung open the door and i ran i had no shoes on i was crying i had no wallet but i was ok because i had my cigarettes and i didn't want any part of freedom if i didn't have my cigarettes when you live with someone who has a temper a very bad temper a very very bad temper you learn to play around that you learn this time i'll play possum and next time i'll just be real nice or i'll say yes to everything or you make yourself scarce or you run and this was one of the times when you just run and as i was running i thought this was a great place to jump out because there were big lawns and there were cul de sacs and sometimes he would come after me and drive and yell stuff at me to get back in get back in and i was like no i'm out of here this is great and i went and hid behind a cabana and he left and i had my cigarettes and uh i started to walk in this beautiful neighborhood it was ten thirty at night and it was silent and lovely and there was no sound except for sprinklers

c

Q1) What kind of vehicle is the narrator in at the beginning of this passage?	Q2) Which word best describes the narrator's boyfriend?	Q3) What kind of place does the narrator find herself in?	Q4) When do the events of the story take place?
a) Boat	a) Gloomy	a) Restaurant	a) Morning
b) Train	**b) Angry**	b) Store	b) Noon
c) Car	c) Lazy	c) Hospital	c) Afternoon
d) Airplane	d) Friendly	**d) Neighborhood**	**d) Night**

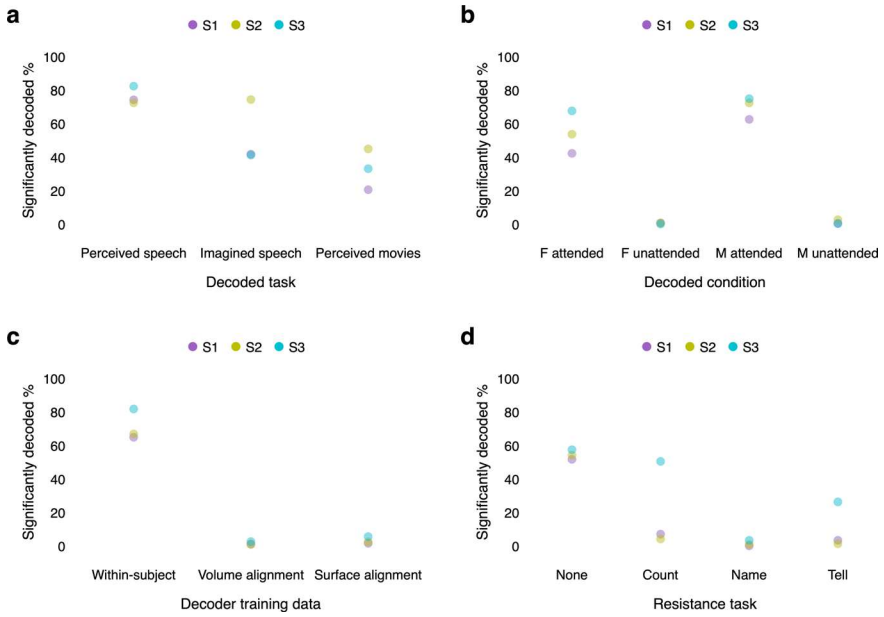
Extended Data Fig. 3 | Behavioral assessment of decoder predictions. Four 80 s segments were chosen from the perceived speech test story. For each segment, four multiple-choice questions were written based on the actual stimulus words without looking at the decoder predictions (Supplementary Table 7). 100 subjects were recruited for an online behavioral experiment and randomly assigned to experimental and control groups. For each segment, the experimental group subjects answered the questions after reading the decoded words from subject S3, while the control group subjects answered the questions after reading the actual stimulus words (see Methods). (**a**) Experimental group scores were significantly higher than expected by chance for 9 out of the 16 questions (* indicates q(FDR) < 0.05, two-sided binomial test). Error bars indicate the bootstrap standard error (n = 1,000 samples). (**b**) The decoded words and the actual stimulus words for a segment. (**c**) The multiple-choice questions cover different aspects of the stimulus story.

https://doi.org/10.1038/s41593-023-01304-9



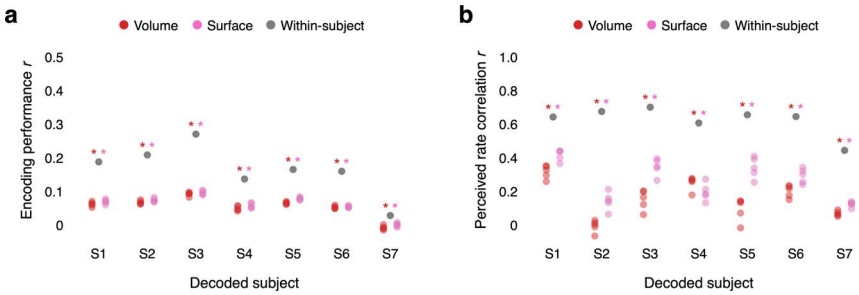
Extended Data Fig. 4 | Decoding across cortical regions. Cortical regions for subjects S1 and S2. (**a**) Brain data used for decoding (colored regions) were partitioned into the speech network, the parietal-temporal-occipital association region, and the prefrontal region (PFC). (**b**) Decoding performance time-course for the perceived speech test story from each region. Horizontal lines indicate when decoder predictions were significantly more similar to the actual stimulus words than expected by chance under the BERTScore metric (q(FDR) < 0.05, one-sided nonparametric test). Corresponding results for subject S3 are shown in Fig. 2a,c in the main text.

https://doi.org/10.1038/s41593-023-01304-9



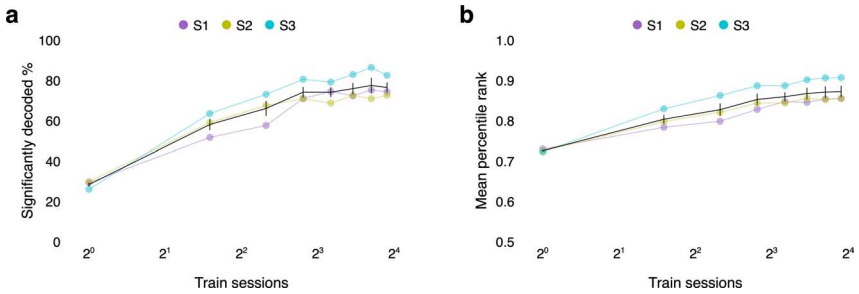
Extended Data Fig. 5 | Comparison of decoding performance across experiments. Decoder predictions from different experiments were compared based on the fraction of significantly decoded time-points under the BERTScore metric (q(FDR) < 0.05). The fraction of significantly decoded time-points was used because it does not depend on the length of the stimuli. (**a**) The decoder successfully recovered 72–82% of time-points during perceived speech, 41–74% of time-points during imagined speech, and 21–45% of time-points during perceived movies. (**b**) During a multi-speaker stimulus, the decoder successfully recovered 42–68% of time-points told by the female speaker when subjects attended to the female speaker, 0–1% of time-points told by the female speaker when subjects attended to the male speaker, 63–75% of time-points told by the male speaker when subjects attended to the male speaker, and 0–3% of time-points told by the male speaker when subjects attended to the female speaker. (**c**) During a perceived story, within-subject decoders successfully recovered 65–82% of time-points, volumetric cross-subject decoders successfully recovered 1–2% of time-points, and surface-based cross-subject decoders successfully recovered 1–5% of time-points. (**d**) During a perceived story, within-subject decoders successfully recovered 52–57% of time-points when subjects passively listened, 4–50% of time-points when subjects resisted by counting by sevens, 0–3% of time-points when subjects resisted by naming animals, and 1–26% of time-points when subjects resisted by imagining a different story.

https://doi.org/10.1038/s41593-023-01304-9



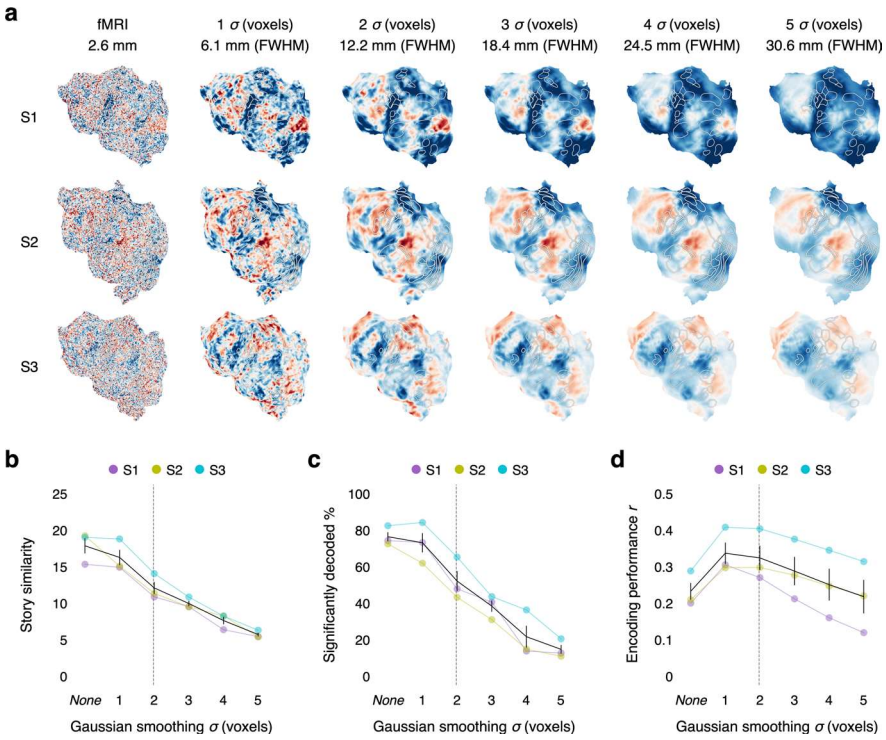
Extended Data Fig. 6 | Cross-subject encoding model and word rate model performance. For each subject, encoding models and word rate models were trained on anatomically aligned brain responses from 5 sets of other subjects (indicated by markers). The models were evaluated on within-subject single-trial responses to the perceived speech test story. (**a**) Cross-subject encoding models performed significantly worse than within-subject encoding models (* indicates q(FDR) < 0.05, two-sided t-test). (**b**) Cross-subject word rate models performed significantly worse than within-subject word rate models (* indicates q(FDR) < 0.05, two-sided t-test).

https://doi.org/10.1038/s41593-023-01304-9



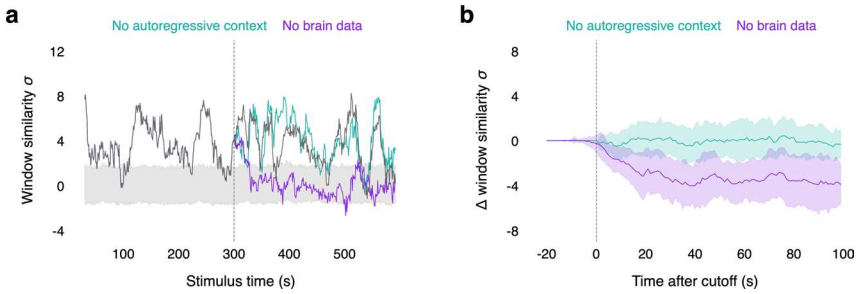
Extended Data Fig. 7 | Decoding performance as a function of training data. Decoders were trained on different amounts of data and evaluated on the perceived speech test story. (**a**) The fraction of significantly decoded time-points increased with the amount of training data collected from each subject but plateaued after 7 scanning sessions (7.5 h) and did not substantially increase up to 15 sessions (16 h). The substantial increase up to 7 scanning sessions suggests that decoders can recover certain semantic concepts after training on a small amount of data, but require much more training data to achieve consistently good performance across the test story. (**b**) The mean identification percentile rank increased with the amount of training data collected from each subject but plateaued after 7 scanning sessions (7.5 h) and did not substantially increase up to 15 sessions (16 h). For all results, black lines indicate the mean across subjects and error bars indicate the standard error of the mean ($n = 3$).

https://doi.org/10.1038/s41593-023-01304-9



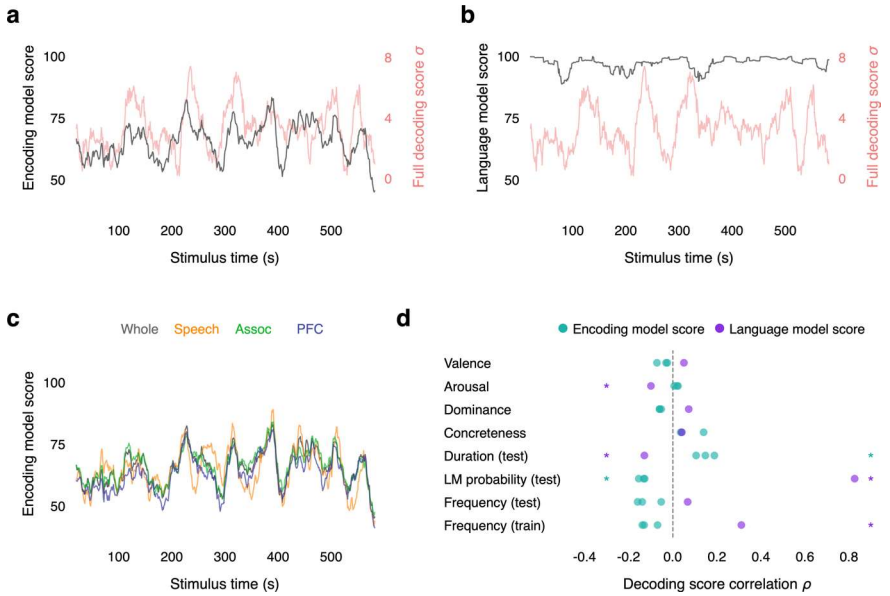
Extended Data Fig. 8 | Decoding performance at lower spatial resolutions. While fMRI provides high spatial resolution, current MRI scanners are too large and expensive for most practical decoder applications. Portable alternatives like functional near-infrared spectroscopy (fNIRS) measure the same hemodynamic activity as fMRI, albeit at a lower spatial resolution. To simulate how the decoder would perform at lower spatial resolutions, fMRI data were spatially smoothed using Gaussian kernels with standard deviations of 1, 2, 3, 4, and 5 voxels, corresponding to 6.1, 12.2, 18.4, 24.5, and 30.6 mm full width at half maximum (FWHM). The encoding model, noise model, and word rate model were estimated on spatially smoothed training data, and the decoder was evaluated on spatially smoothed responses to the perceived speech test story. (**a**) fMRI images for each subject were spatially smoothed using progressively larger Gaussian kernels. Blue voxels have above average activity and red voxels have below average activity. (**b**) Story similarity decreased as the data were spatially smoothed, but remained high at moderate levels of smoothing. (**c**) The fraction of significantly decoded time-points decreased as the data were spatially smoothed, but remained high at moderate levels of smoothing. (**d**) Encoding model prediction performance increased as the data were spatially smoothed, demonstrating that decoding performance and encoding model performance are not perfectly coupled. While spatial smoothing reduces information, making it harder to decode the stimulus, it also reduces noise, making it easier to predict the responses. For all results, black lines indicate the mean across subjects and error bars indicate the standard error of the mean ($n = 3$). Dashed gray lines indicate the estimated spatial resolution of current portable systems[43]. These results show that around 50% of the stimulus time-points could still be decoded at the estimated spatial resolution of current portable systems, and provide a benchmark for how much portable systems need to improve to reach different levels of decoding performance.

https://doi.org/10.1038/s41593-023-01304-9



Extended Data Fig. 9 | Decoder ablations. To decode new words, the decoder uses both the autoregressive context (that is the previously decoded words) and the fMRI data. To understand the relative contributions of the autoregressive context and the fMRI data, decoders were evaluated in the absence of each component. The standard decoding approach was performed up to a cutoff point in the perceived speech test story. After the cutoff, either the autoregressive context was reset or the fMRI data were removed. To reset the autoregressive context, all of the candidate sequences were discarded and the beam was re-initialized with an empty sequence. The standard decoding approach was then performed for the remainder of the scan. To remove the fMRI data, continuations were assigned random likelihoods rather than encoding model likelihoods for the remainder of the scan. (**a**) A cutoff point was defined 300 s into the stimulus for one subject. When the autoregressive context was reset, decoding performance fell but quickly rebounded. When the fMRI data were removed, decoding performance quickly fell to chance level. The gray shaded region indicates the 5th to 95th percentiles of the null distribution. (**b**) The ablations were repeated for cutoff points at every 50 s of the stimulus. The performance differences between the original decoder and the ablated decoders were averaged across cutoff points and subjects, yielding profiles of how decoding performance changes after each component is ablated. The blue and purple shaded regions indicate the standard error of the mean (n = 27 trials). These results demonstrate that the decoder continually relies on the encoding model and the fMRI data to achieve good performance, and does not require good initial context. In these figures, each time-point was scored based on the 20 s window ending at that time-point, whereas in all other figures, each time-point was scored based on the 20 s window centered around that time-point. This shifted indexing scheme emphasizes how decoding performance changes after a cutoff. Dashed gray lines indicate cutoff points.

https://doi.org/10.1038/s41593-023-01304-9



Extended Data Fig. 10 | Isolated encoding model and language model scores. The encoding model and the language model were separately evaluated on the perceived speech test story to isolate their contributions to the decoding error (see Methods). At each word time t, the encoding model and the language model were provided with the actual stimulus word and 100 random distractor words. The encoding model ranks the words by the likelihood of the fMRI responses, and the language model ranks the words by the probability given the previous stimulus words. Encoding model and language model scores were computed based on the number of distractor words ranked below the actual word (100 indicates perfect performance, 50 indicates chance level performance). To compare against the decoding scores from Fig. 1e, the word-level scores were averaged across 20 s windows of the stimulus. (**a**) Encoding model scores were significantly correlated with decoding scores (linear correlation $r = 0.22–0.58$, $P < 0.05$), suggesting that many of the poorly decoded time-points in Fig. 1e are inherently more difficult to decode using the encoding model. (**b**) Language model scores were not significantly correlated with decoding scores. (**c**) For each word, encoding model scores from 10 sets of distractors were compared to chance level. Most stimulus words with significant encoding model scores (q(FDR) < 0.05, two-sided t-test) for the whole brain also had significant encoding model scores for the speech network (80–87%), association region (88–92%), and prefrontal region (82–85%), suggesting that the results in Fig. 2c were not primarily due to the language model. Word-level encoding model scores were significantly correlated across each pair of regions (q(FDR) < 0.05, two-sided permutation test), suggesting that the results in Fig. 2d were not primarily due to the language model. (**d**) Word-level encoding model and language model scores were correlated against the word properties tested in Fig. 4d (* indicates q(FDR) < 0.05 for all subjects, two-sided permutation test). The encoding model and the language model were biased in opposite directions for several word properties. These effects may have balanced out in the full decoder, leading to the observed lack of correlation between the word properties and decoding scores (Fig. 4d).

nature portfolio

Corresponding author(s): Alexander G. Huth

Last updated by author(s): Mar 13, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL (5.0). Cortical surface meshes were generated from T1-weighted anatomical scans using FreeSurfer software (6.0). Flat maps were created using pycortex software (1.3.0). The Penn Phonetics Lab Forced Aligner (P2FA) (1.003) was used to automatically align stimulus audio with manual transcripts. Praat (6.2.01) was used to manually check and correct the alignments
Data analysis	All model fitting and analysis was performed using custom software written in Python (3.9.5), making heavy use of NumPy (1.20.2), SciPy (1.6.2), PyTorch (1.9.0), Transformers (4.6.1), and pycortex (1.3.0). Custom decoding code is available at https://github.com/HuthLab/semantic-decoding.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data collected during the decoder resistance experiment are available upon reasonable request, but were not publicly released due to concern that the data could be used to discover ways to bypass subject resistance. All other data are available at https://openneuro.org/datasets/ds003020 and https://openneuro.org/datasets/ds004510.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	fMRI data were collected from 3 female subjects and 4 male subjects. Behavioral data were collected from 50 female subjects, 49 male subjects, and 1 non-binary subject. Sex and gender were not considered in the study design because we do not expect language decoding performance or language comprehension performance to depend on sex or gender. Sex and gender were determined based on self-reporting.
Population characteristics	fMRI data were collected from 7 healthy human subjects (3 female, 4 male) between 23 and 36 years of age, with normal hearing, normal or corrected-to-normal vision, and native English language proficiency. Behavioral data were collected from 100 human subjects (50 female, 49 male, 1 non-binary) between 19 and 70 years of age, with native or fluent English language proficiency.
Recruitment	fMRI subjects were recruited from the lab. Lab members were asked if they would like to participate in the experiment. The subjects used in this experiment were the lab members that volunteered to get scanned for at least 6 sessions. Because there are no experimental manipulations in this study and subjects are only performing naturalistic tasks, we do not expect this to have an effect on our results in any way. Behavioral subjects were recruited through the Prolific online platform. The only inclusion criteria were age (at least 18 years) and native or fluent English language proficiency. While the subjects chose whether to participate in the experiment, we do not expect this self-selection to have an effect on our results in any way.
Ethics oversight	The experimental protocol was approved by the Institutional Review Board at the University of Texas at Austin. All subjects gave written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the fMRI experiment, no sample-size calculation was performed. Due to the large amount of data collected for each fMRI subject, the experimental approach does not rely on a large sample-size. The sample size of the fMRI experiment is comparable to those of previous decoding publications For the behavioral experiment, control group participants were expected to perform close to ceiling accuracy, so a sample size of 100 provides sufficient power to detect significance differences with test accuracies as high as 70% (G*Power, exact test of proportions with independent groups).
Data exclusions	No data were excluded from the analysis.
Replication	For the fMRI experiment, language decoding models were independently fit and evaluated for each subject, and the results of the study were consistent across subjects. This is effectively 3 separate and independent replications of the fMRI experiment. These sample sizes are similar to those reported in previous decoding publications. Exploratory decoding analyses were restricted to data collected while subject S3 listened to one test story. All exploratory analyses were qualitative. The analysis pipeline was frozen before we viewed results for the remaining

	subjects, stimuli, and experiments. For the behavioral experiment, multiple-choice answers were written by a separate researcher who had never seen the decoder predictions to remove researcher bias. Bootstrap re-sampling was used to test for reproducibility.
Randomization	For the fMRI experiment, subjects were not allocated into experimental groups. For the behavioral experiment, subjects were randomly allocated into experimental and control groups and blinded to group assignment.
Blinding	For the fMRI experiment, subjects were not allocated into experimental groups. For the behavioral experiment, investigators were blinded to group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	Naturalistic task design
Design specifications	Training data collection was broken up into 16 different scanning sessions, the first session involving the anatomical scan and localizers, and each successive session consisting of 5 or 6 spoken narrative stories from The Moth Radio Hour or Modern Love. Testing data was collected in 2 different scanning sessions for subjects S2 and S3, and 1 scanning session for subject S1.
Behavioral performance measures	The study does not involve behavioral performance.

Acquisition

Imaging type(s)	Functional
Field strength	3T
Sequence & imaging parameters	Gradient echo EPI sequence, field of view = 220mm, matrix size = 84x84, slice thickness = 2.6mm, flip angle = 71°
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0. FLIRT was used to align all data to a template that was made from the average across the first functional run in the first story session for each subject. These automatic alignments were manually checked for accuracy.
Normalization	Data were not normalized as we were looking for effects within individual subjects.
Normalization template	Data were not normalized as we were looking for effects within individual subjects.
Noise and artifact removal	Low frequency voxel response drift was identified using a 2nd order Savitzky-Golay filter with a 120 second window and then subtracted from the signal. To avoid onset artifacts and poor detrending performance near each end of the scan, responses were trimmed by removing 20 seconds (10 volumes) at the beginning and end of each scan.
Volume censoring	There were no volumes with sufficient motion to warrant censoring. All subjects wore customized headcases to prevent excessive movement.

Statistical modeling & inference

Model type and settings	Our study follows a Bayesian decoding framework. In the model estimation stage, predictive voxel-wise encoding models were separately estimated for each subject using brain recordings collected while that subject listened to narrative stories. In the language reconstruction stage, brain recordings were collected while the subject performed a range of naturalistic tasks. A language model generates candidate word sequences, and the subject's encoding model evaluates the candidates against the brain recordings.
Effect(s) tested	Decoding performance for each task was calculated by comparing decoder predictions to reference stimulus words. Paired t-tests were used to compare decoding performance across tasks.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	Anatomical ROIs were defined for each subject using Freesurfer. Functional ROIs were defined for each subject using an auditory cortex localizer and a motor localizer.
Statistic type for inference (See Eklund et al. 2016)	Decoding performance was calculated by comparing decoder predictions to reference stimulus words.
Correction	All tests were corrected for multiple comparisons when necessary using the false discovery rate (FDR).

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis	Voxel-wise encoding models were fit on a training dataset using L2-regularized linear regression to predict BOLD responses from semantic stimulus features. Semantic features of each stimulus word were extracted from a pre-trained GPT language model. Encoding models were used to decode novel BOLD responses in a test dataset. Prediction performance was computed using standard language similarity metrics on the decoder predictions and reference stimulus words.