



# A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations

---

Received: 5 June 2024

A list of authors and their affiliations appears at the end of the paper

---

Accepted: 9 January 2025

---

Published online: 7 March 2025

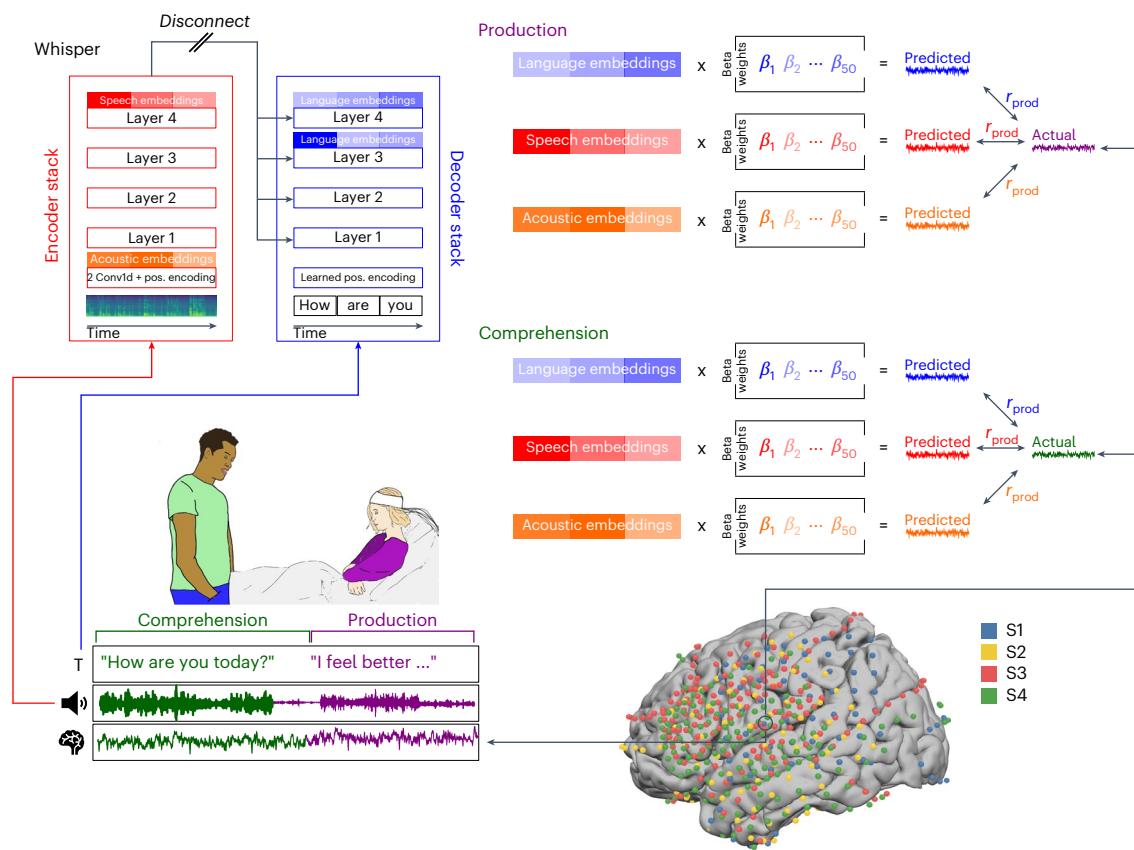
---

Check for updates

This study introduces a unified computational framework connecting acoustic, speech and word-level linguistic structures to study the neural basis of everyday conversations in the human brain. We used electrocorticography to record neural signals across 100 h of speech production and comprehension as participants engaged in open-ended real-life conversations. We extracted low-level acoustic, mid-level speech and contextual word embeddings from a multimodal speech-to-text model (Whisper). We developed encoding models that linearly map these embeddings onto brain activity during speech production and comprehension. Remarkably, this model accurately predicts neural activity at each level of the language processing hierarchy across hours of new conversations not used in training the model. The internal processing hierarchy in the model is aligned with the cortical hierarchy for speech and language processing, where sensory and motor regions better align with the model's speech embeddings, and higher-level language areas better align with the model's language embeddings. The Whisper model captures the temporal sequence of language-to-speech encoding before word articulation (speech production) and speech-to-language encoding post articulation (speech comprehension). The embeddings learned by this model outperform symbolic models in capturing neural activity supporting natural speech and language. These findings support a paradigm shift towards unified computational models that capture the entire processing hierarchy for speech comprehension and production in real-world conversations.

One of the ultimate goals of our collective research endeavour in human neuroscience is to model and understand how the brain supports dynamic, context-dependent behaviours in the real world. Perhaps the most distinctly human behaviour—and the focus of this paper—is our capacity for using language to communicate our thoughts to others during free, open-ended conversations. In daily conversations, language is highly complex, multidimensional and context dependent<sup>1–3</sup>. Traditionally, neurolinguistics has relied on an

incremental divide-and-conquer strategy, dividing language into distinct subfields, including phonetics, phonology, morphology, syntax, semantics and pragmatics. Psycholinguists aim to build a closed set of well-defined symbolic features and linguistic processes for each subfield. For example, classical psycholinguistic models use symbolic units, such as phonemes, to analyse speech (that is, processing spoken acoustic signals) and curated part-of-speech units, such as nouns, verbs, adjectives and adverbs, to analyse syntactic structures.



**Fig. 1 | An ecological, dense-sampling paradigm for modelling neural activity during real-world conversations.** We monitored continuous neural activity in 4 ECoG patients during their interactions with hospital staff, family and friends, providing a unique opportunity to investigate real-world social communication. Simultaneously recorded verbal interactions are transcribed and segmented into production (purple) and comprehension (green) components (bottom left). We used Whisper, a deep speech-to-text model, to process our speech recordings and transcripts, and extracted embeddings from different parts of the model: for each word, we extracted ‘acoustic embeddings’ from Whisper’s static encoder layer, ‘speech embeddings’ from Whisper’s top encoder layer (red), and ‘language

embeddings’ from Whisper’s decoder network (blue) (top left). The embeddings were reduced to 50 dimensions using PCA. We used linear regression to predict neural signals from the acoustic embeddings (orange), speech embeddings (red) and language embeddings (blue) across tens of thousands of words. We calculated the correlation between predicted and actual neural signals for left-out test words to evaluate encoding model performance. This process was repeated for each electrode and each lag, using a 25-ms sliding window ranging from  $-2,000$  to  $+2,000$  ms relative to word onset (top right). Bottom right: brain coverage across 4 participants comprising 654 left hemisphere electrodes.

Although interactions exist between these different levels of representations<sup>4–7</sup>, individual labs have traditionally focused on modelling each subfield in isolation using targeted experimental manipulations. The implicit aspiration behind this collective effort is to eventually integrate these fragmented studies into a comprehensive neurocomputational model of natural language processing<sup>8–10</sup>. After decades of research, however, there is increasing awareness of the gap between natural language processing and formal psycholinguistic theories<sup>11,12</sup>. Psycholinguistic models and theories often fail to account for the subtle, non-linear, context-dependent interactions within and across levels of linguistic analysis in real-world conversations<sup>13–15</sup>.

Deep learning provides a unified computational framework that can serve as an alternative approach to natural language processing in the human brain<sup>16,17</sup>. Recent breakthroughs in large language models (LLMs) have led to remarkable improvements in processing, summarizing and generating language for natural conversations<sup>18,19</sup>. Alongside remarkable advances in processing syntactic, semantic and pragmatic properties in written texts, deep learning has also come to excel in recognizing speech in acoustic recordings<sup>20</sup>. These multimodal, end-to-end models provide a theoretical advance over unimodal text-based models by offering a unified computational framework for modelling how continuous auditory input is transformed into speech and word-level linguistic dimensions during natural conversations (that is, acoustic-to-speech-to-language processing).

Notably, deep acoustic-to-speech-to-language models do not rely on symbolic representations of phonemes for speech recognition or parts of speech for language processing. The critical distinction between deep and symbolic models is the shift from discrete symbols to a multidimensional vectorial representation (that is, embedding space). This approach embeds all elements of speech and language into continuous vectors across a population of simple computing units (‘neurons’) by optimizing simple objectives such as predicting the next word in context or deciphering words from auditory stimuli. Combining speech and language embeddings into a unified multimodal model provides a numerical ‘code’ for linking across levels of linguistic representation, which are traditionally studied in isolation.

In this work, we leverage a multimodal acoustic-to-speech-to-language model called Whisper<sup>20</sup> that learns to transcribe acoustic recordings of natural conversations recorded in real-life contexts<sup>20</sup>. The Whisper architecture incorporates a multilayer encoder network and a multilayer decoder network (Fig. 1): the encoder maps continuous acoustic inputs into a high-dimensional embedding space, capturing speech features which are transferred into a word-level decoder, effectively mapping them into contextual word embeddings<sup>21–23</sup>. It is important to note that the model was designed and trained without using traditional linguistic elements (such as phonemes, parts of speech, syntactic rules and so on). Despite the absence of these symbolic units, the model can process natural language with a level

of accuracy comparable to that of a human<sup>20</sup>. Here, 'speech' refers to processing spoken signals, while 'language' refers to analysing conversations on the basis of word-level transcripts.

In this work, we report on the alignment between the internal representations of an acoustic-to-speech-to-language model and the human brain when processing real-life conversations. To study the neural basis of natural language processing in the real world, we developed a new dense-sampling electrocorticography (ECoG) paradigm to measure human neural activity at scale in unconstrained, real-world conversations. Unlike traditional ECoG studies, which typically rely on brief, controlled experiments, our dense-sampling paradigm enabled continuous, 24/7 recording of ECoG and speech data during each patient's days- to week-long stay in the epilepsy unit at NYU Langone Health. This ambitious effort resulted in a uniquely large ECoG dataset of natural conversations: 4 patients recorded during free conversations, yielding ~50 h (289,971 words) of neural recordings during speech comprehension and 50 h (230,238 words) during speech production in real-world settings. Modelling the 24/7 conversational data presents an unprecedented challenge, given that we have not imposed any experimental constraints on our participants and no two conversations are the same. Patients are free to say whatever they want, whenever they want; each conversation has its unique context and purpose.

To model and predict the underlying neural activity that supports our ability to produce or comprehend daily conversations, we opened the 'black box' of the acoustic-to-speech-to-language model (Whisper). We interrogated its internal representations—the embeddings—at each layer. We extracted three types of embedding from Whisper (Fig. 1) for every word the patients spoke or heard during their conversations. These embeddings include (1) acoustic embeddings derived from the auditory input layer of the speech encoder, (2) speech embeddings derived from the final layer of the speech encoder and (3) language embeddings derived from the final layers of the decoder. For each set of embeddings, we constructed electrode-wise encoding models to estimate a linear mapping from the embeddings to the neural activity for each word during speech production and comprehension (Fig. 1).

Our encoding models revealed a remarkable alignment between the human brain and the internal population code of the acoustic-to-speech-to-text model. We demonstrate that the embeddings provide surprisingly accurate predictions of human neural activity for each utterance and word across hundreds of thousands of words in our conversational dataset. Speech embeddings better captured cortical activity in lower-level speech perception and production areas, including the superior temporal cortex and precentral gyrus. On the other hand, linguistic embeddings were better aligned with higher-order language areas such as the inferior frontal gyrus and angular gyrus. Before each word onset during speech production, we observed a temporal sequence from language-to-speech encoding across cortical areas; during speech comprehension, we observed the reverse progression from speech-to-language encoding after word articulation. Our findings demonstrate that deep acoustic-to-speech-to-language models can provide a unified computational framework for the neural basis of language production and comprehension across large volumes of real-world data without sacrificing the diversity and richness of natural language.

## Results

We collected continuous 24/7 recordings of ECoG and speech signals from 4 patients as they spontaneously conversed with their family, friends, doctors and hospital staff during their entire days-long stay at the epilepsy unit (for patient demographics and clinical characteristics, see Supplementary Table 1). Across the 4 patients, we recorded neural signals from 676 intracranial electrodes (Fig. 1). Because only 1 of the 4 patients had 22 electrodes implanted in the right hemisphere, we focused on left hemisphere electrodes ( $n = 654$ ) in our analyses; 10 electrodes were excluded due to corrupted recordings, leaving

644 electrodes for analysis. We obtained extensive coverage of key language areas, including in the inferior frontal gyrus (IFG, also known as Broca's area;  $n = 75$ ) and superior temporal gyrus (STG;  $n = 45$ ), with a sparser sampling of the angular gyrus (AG;  $n = 35$ ). We built a pre-processing pipeline to identify the occurrence of speech, remove identifying information, transcribe each conversation and align each word with the concurrent ECoG signals. We then divided the data into two categories: comprehension (when patients were listening to speech) and production (when patients were producing speech). This unconstrained recording paradigm yielded neural activity from multiple electrodes per patient (104–255 electrodes) for dozens of hours (17–37 h), comprising tens of thousands of words (79,654–213,473 words). For details about linguistic features, see Supplementary Tables 1 and 2. For a comprehensive description of the speech collected, patient demographics and clinical characteristics, see Supplementary Tables 1 and 2 and Fig. 1.

In our dataset, each conversation is unique: patients freely express themselves without any intervention from experimenters. We input the audio recordings and the transcribed text into a multimodal, acoustic-to-speech-to-language model (Whisper)<sup>20</sup>. To leverage the multimodal architecture of Whisper, we separately extracted 'acoustic embeddings,' 'speech embeddings' and 'language embeddings' for each word in every conversation (Fig. 1 and Methods): acoustic embeddings were extracted from the acoustic input layer fed into the speech encoder. Speech embeddings were extracted from the top layer of the speech encoder, and language embeddings were extracted from the top layers of the decoder (Fig. 1). We conducted experiments to examine how speech input affects language embeddings in the Whisper model. We used two different methods to extract embeddings from the decoder. First, we disconnected the cross attention and separated it into a speech encoder stack and a language decoder stack. By providing the transcription to the decoder, we could extract language embeddings that were not influenced by the speech input. Second, we extracted language embeddings from the intact model, which receives both speech and textual inputs, to test how speech input modulates the language embeddings. It is important to note that while Whisper's encoder can provide direct input to its decoder, the activity in the decoder cannot influence the activity in the encoder.

### Acoustic-to-speech-to-language prediction of neural activity

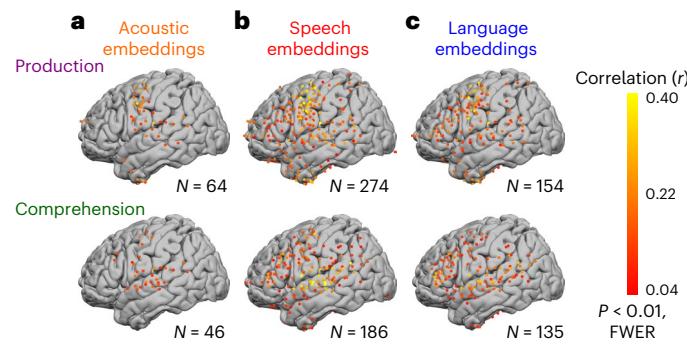
To assess whether the embeddings extracted from Whisper can capture neural activity during natural conversations, we constructed six sets of encoding models on the basis of acoustic embedding, speech embeddings and language embeddings during both speech production and speech comprehension (Fig. 1). We segmented the data from each patient into 10 temporally contiguous, non-overlapping folds for 10-fold leave-one-out cross-validation. The encoding models estimated a linear mapping between the Whisper embeddings and the neural activity for each word in the training set using 9 folds for training. Subsequently, we used the trained encoding models to predict the neural activity for each word at each electrode in the left-out unseen new conversations within the test fold (Fig. 1). This procedure was repeated 10 times to cover all folds. A separate encoding model was trained for each electrode at various time points, ranging from ~2,000 ms to +2,000 ms relative to the word onset (time 0). The performance of the encoding model was evaluated by calculating the correlation between the predicted and actual neural signals for the held-out conversations. All analyses were adjusted for multiple comparisons using a non-parametric procedure to control the family-wise error rate (FWER).

Whisper's acoustic, speech and language embeddings predicted neural activity with remarkable accuracy across conversations comprising hundreds of thousands of words during both speech production and comprehension for numerous electrodes in various regions of the cortical language network (Fig. 2). To minimize bias, we estimated the lag that yielded the maximum correlation in the training fold and used it to extract the matching correlation in the test fold (to determine the

colour of the electrode in Fig. 2). These brain regions include areas known to be involved in auditory speech processing (for example, superior temporal gyrus (STG)), language comprehension and production (for example, inferior frontal gyrus (IFG)), somatomotor (SM) planning and execution (for example, precentral and postcentral gyrus (preCG, postCG)), and high-level semantic cognition (for example, angular gyrus and temporal pole (AG, TP))<sup>24,25</sup>. Overall, acoustic embeddings yielded fewer significant electrodes than speech embeddings for production (64 vs 274, chi-square (1,  $N = 644$ ) = 175.21,  $P < 0.001$ , Bonferroni corrected,  $\varphi = 0.27$ ) and comprehension (46 vs 186, chi-square (1,  $N = 644$ ) = 101.58,  $P < 0.001$ , Bonferroni corrected,  $\varphi = 0.16$ ), and fewer significant electrodes than language embeddings (production: 64 vs 154, chi-square (1,  $N = 644$ ) = 43.73,  $P < 0.001$ , Bonferroni corrected,  $\varphi = 0.06$ ; comprehension: 46 vs 135, chi-square (1,  $N = 644$ ) = 49.55,  $P < 0.001$ , Bonferroni corrected  $\varphi = 0.08$ ). Speech embeddings yielded more significant electrodes than language embeddings for both production (274 vs 154, chi-square (1,  $N = 644$ ) = 49.55,  $P < 0.001$ , Bonferroni corrected,  $\varphi = 0.08$ ) and comprehension (186 vs 135, chi-square (1,  $N = 644$ ) = 10.37,  $P < 0.005$ , Bonferroni corrected,  $\varphi = 0.02$ ). Remarkably, the predicted signals were strongly correlated with the actual signals (Pearson correlations of up to 0.50) across hours of left-out speech segments. Moreover, prediction performance in the left-out testing segments was robust and did not meaningfully change even when we used only 25% of the data for training (Supplementary Fig. 2). We also extracted language embeddings from the decoder stack of layer 4 (instead of layer 3 which was used for Figs. 2–7) and a unimodal language model (GPT-2), and obtained similar encoding results (Supplementary Fig. 3). Because the speech encoder receives continuous speech recordings, we could also run encoding models for continuous acoustic and speech embeddings, encompassing all time points in each recording, including non-speech segments, irrespective of the spoken word boundaries (Supplementary Fig. 4a,b and Methods). Even when using continuous signals, we observed statistically higher encoding for the speech embeddings than for the acoustic embeddings in all electrodes (Supplementary Fig. 4c,d). This demonstrates that speech embeddings, which contain contextual speech information, model all cortical areas better than the simple acoustic embeddings derived from the model input layer.

**Selectivity and integration of speech and language information**  
In contrast to a modular view that assigns acoustic, speech and language processing to distinct circuits or brain areas, our analyses reveal that speech and language information are encoded in multiple brain areas. We utilized a variance partitioning approach to identify the proportion of the predicted signal in each electrode uniquely explained by the acoustic, speech and language embeddings. We fitted separate encoding models for speech and language embeddings and a joint encoding model by concatenating speech and language embeddings. The analysis measures the unique variance captured by each set of embeddings and the extent to which the information in one set is already embedded in another. A similar analysis was also done for acoustic and speech embeddings (Supplementary Fig. 5).

We observed different selectivity patterns for speech and language embeddings, each accounting for different portions of the variance across different cortical areas (Fig. 3). During spontaneous speech production (Fig. 3a), we observed organized hierarchical processing, where articulatory areas along the preCG and postCG, as well as STG, were better predicted by speech embeddings (red), while higher-level language areas such as IFG, pMTG and AG were better predicted by language embeddings (blue). A similar hierarchical organization was evident in speech comprehension (Fig. 3b): perceptual areas such as STG and somatomotor areas such as preCG and postCG showed a preference for speech embeddings, while higher-level language areas, including IFG and AG, displayed a preference for language embeddings. Our predictions had a high level of precision, with a

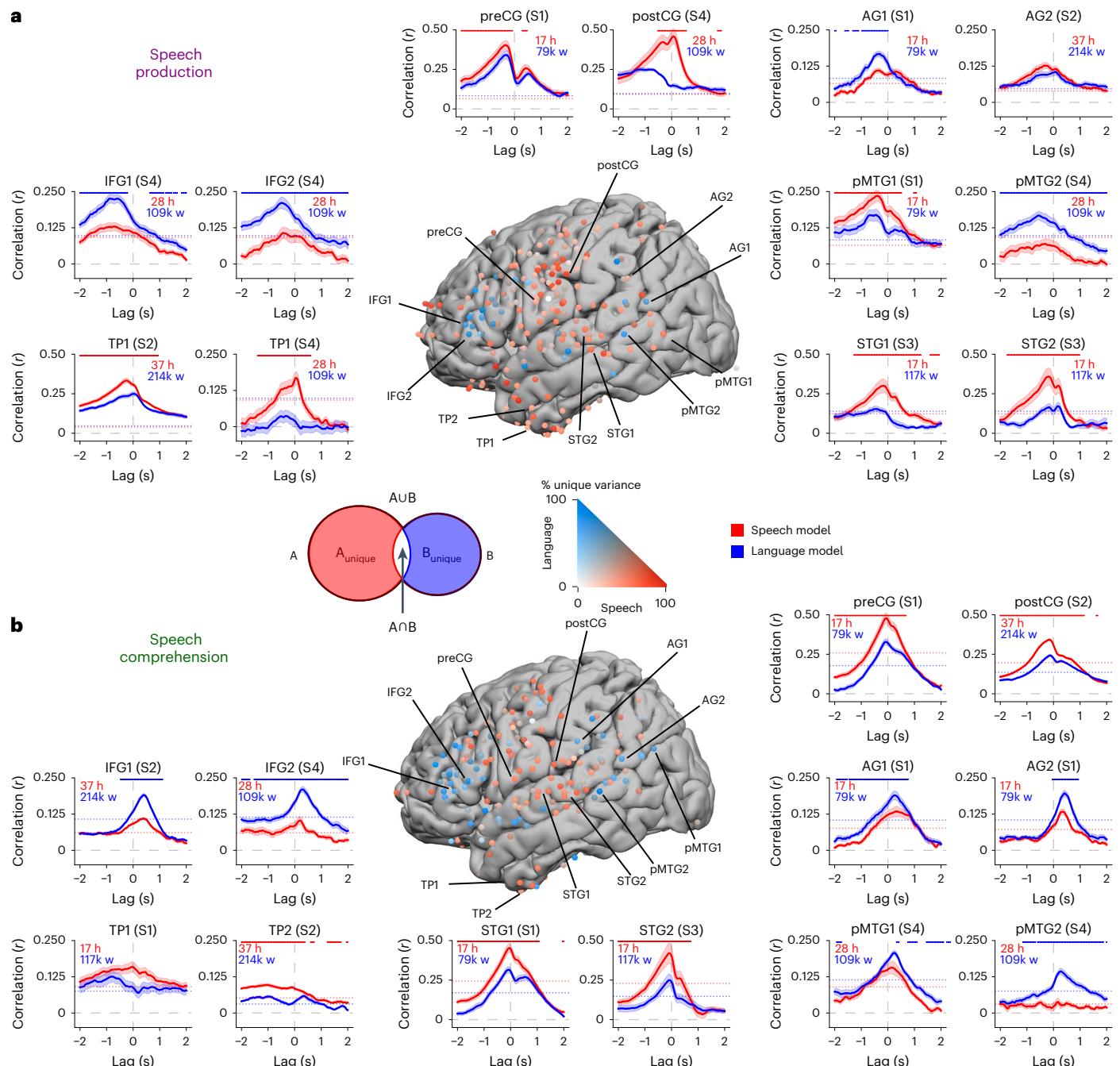


**Fig. 2 | Acoustic, speech and language encoding model performance during speech production and comprehension.** Encoding performance (correlation between model-predicted and actual neural activity) for each electrode for acoustic embeddings, speech embeddings and language embeddings during speech comprehension (~50 h, 289,971 words) and speech production (~50 h, 230,238 words). The plots illustrate the correlation values associated with the encoding for each electrode, with the colour indicating the highest correlation value across lags ( $P < 0.01$ , FWER). **a**, Encoding based on acoustic embeddings revealed significant electrodes in auditory and speech areas along the superior temporal gyrus (STG) and somatomotor areas (SM). During speech production, we observed enhanced encoding in SM, and during speech comprehension, we observed enhanced encoding in the STG. **b**, Encoding based on speech embeddings revealed significant electrodes in STG and SM, as well as the inferior frontal gyrus (IFG; Broca's area), temporal pole (TP), angular gyrus (AG) and posterior middle temporal gyrus (pMTG; Wernicke's area). **c**, Encoding based on language embeddings highlighted regions similar to speech embeddings (**b**) but notably fewer electrodes (with lower correlations) in STG and SM, and higher correlations in IFG.

correlation between predicted and actual neural responses ranging from 0.2 to 0.5 across electrodes and models (Fig. 3). This high predictive power was achieved for hundreds of thousands of words and tens of hours of speech from previously unseen, unique conversations not used to train the encoding model. Finally, we utilized a variance partitioning approach to identify the proportion of the predicted signal in each electrode uniquely explained by the acoustic versus speech embeddings. Our results indicate that the speech embeddings captured more variance than acoustic embeddings in most electrodes located along the superior temporal cortex, IFG and somatomotor cortex (Supplementary Fig. 5). Acoustic embeddings only captured additional variance in a few electrodes along the lateral fissure and ventral motor cortex.

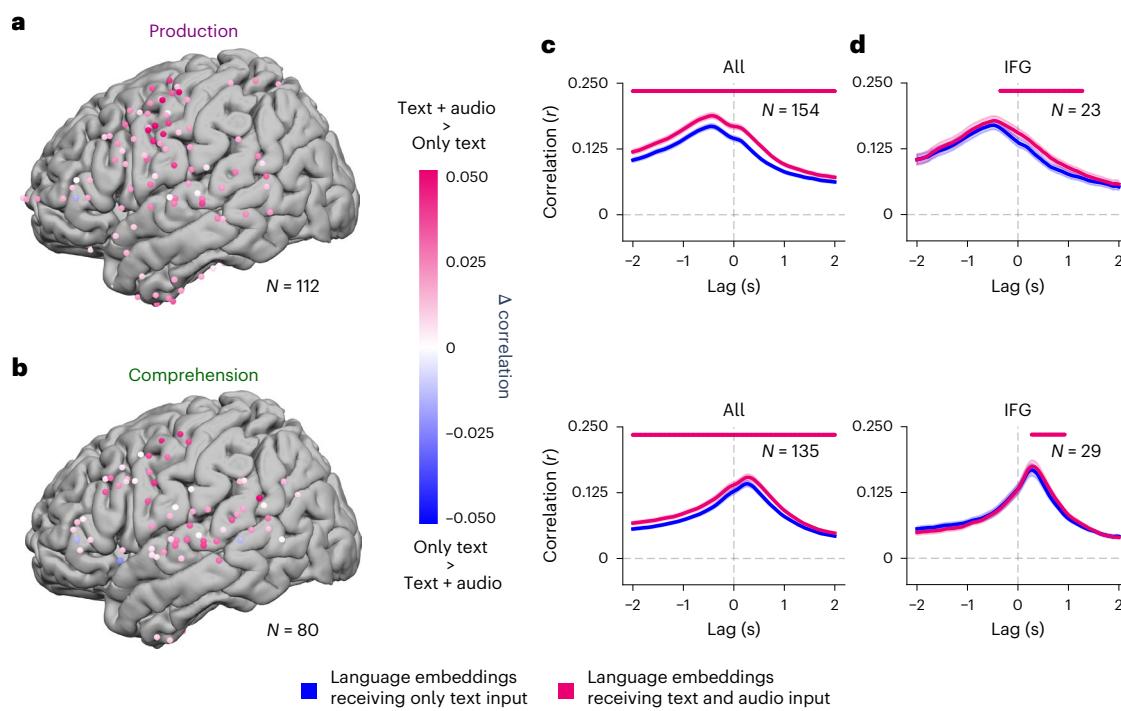
### Auditory speech signals inform language representations

Our multimodal model allowed us to study how speech information is combined with and influences language processing across different language areas. First, we treated Whisper's language decoder as a unimodal model and gave it text-only input. While providing Whisper with text-only input, we treated it as a regular unimodal language model (for example, GPT-2). Next, we utilized Whisper's multimodal capability by providing it with speech and text information. In other words, Whisper's encoder receives speech recordings, while Whisper's decoder receives the text transcription. This allows input from the speech embedding to influence the activity in the language decoder (as in the original architecture). In testing both sets of embeddings, we observed that encoding performance for language embeddings was significantly higher when the language decoder received speech information from the encoder, during both production (Fig. 4a) and comprehension (Fig. 4b). This pattern was consistent across most electrodes in STG and SM, as well as in IFG (Fig. 4c,d). These results demonstrate that speech information can modify the representation of linguistic information in Whisper. Furthermore, infusing speech information into the language embedding improves our ability to model neural responses in language



**Fig. 3 | Mixed selectivity for speech and language embeddings during speech production and comprehension.** **a.** Variance partitioning was used to identify the proportion of variance uniquely explained by either speech or language embeddings relative to the variance explained by the joint encoding model during speech production. Surrounding plots display encoding performance during speech production for selected individual electrodes across different brain areas and patients. Models were estimated separately for each lag (relative to word onset at 0 s) and evaluated by computing the correlation between predicted and actual neural activity. Data are presented as mean  $\pm$  s.e. across the 10 folds. The dotted horizontal line indicates the statistical threshold ( $q < 0.01$ , two-sided, FDR corrected). During production, the speech encoding model (red) achieved correlations of up to 0.5 when predicting neural responses to each word over hours of recordings in the STG, preCG and postCG. The language encoding model yielded significant predictions (correlations up

to 0.25) and outperformed the speech model in IFG and AG indicated by blue dots ( $q < 0.01$ , two-sided, FDR corrected). The variance partitioning approach revealed a mixed selectivity for speech and language embeddings during speech production. Language embeddings (blue) better explain IFG, while speech embeddings (red) better explain STG and SM. **b**, During comprehension, we observed a similar pattern of encoding performance. Language embeddings better explain IFG and AG, while speech embeddings better explain STG and SM indicated by red dots ( $q < 0.01$ , two-sided, FDR corrected). The variance partitioning analysis also revealed mixed selectivity for speech (red) and language (blue) embeddings during comprehension. Matching the flow of information during conversations, encoding models accurately predicted neural activity -500 ms before word onset during speech production and 300 ms after word onset during speech comprehension. Data are presented as mean  $\pm$  s.e. across the 10 folds.



**Fig. 4 | Enhanced encoding for language embeddings fused with auditory speech features.** Comparing electrode-wise encoding performance for language embeddings receiving only text input (that is, conversation transcripts) and language embeddings receiving audio and text inputs (that is, speech recordings and conversation transcripts). Language embeddings fused with auditory features outperform text-only language embeddings in predicting neural activity across multiple electrodes. **a**, During speech production, language embeddings fused with auditory features (pink) significantly improved encoding performance in SM electrodes ( $q < 0.01$ , FDR corrected). **b**, During speech comprehension, language embeddings fused with auditory features

(pink) significantly improved encoding performance in STG and SM electrodes ( $q < 0.01$ , FDR corrected). **c**, The advantage of the language embedding fused with auditory features (pink) persists across multiple time points at all significant electrodes. Data are presented as mean  $\pm$  s.e.m. across the electrodes. **d**, Even though the IFG is associated with linguistic processing, it can be seen that across multiple lags, the audio-fused language embeddings (pink) yield higher encoding performance during both production and comprehension. Pink markers indicate lags with a significant difference ( $q < 0.01$ , FDR corrected) between text-only and audio-fused language embeddings. Data are presented as mean  $\pm$  s.e. across electrodes.

areas. This suggests that language areas, similar to Whisper, encode the intricate relationship between speech and language representations in a multidimensional space.

#### Fine-scale temporal dynamics of speech processing

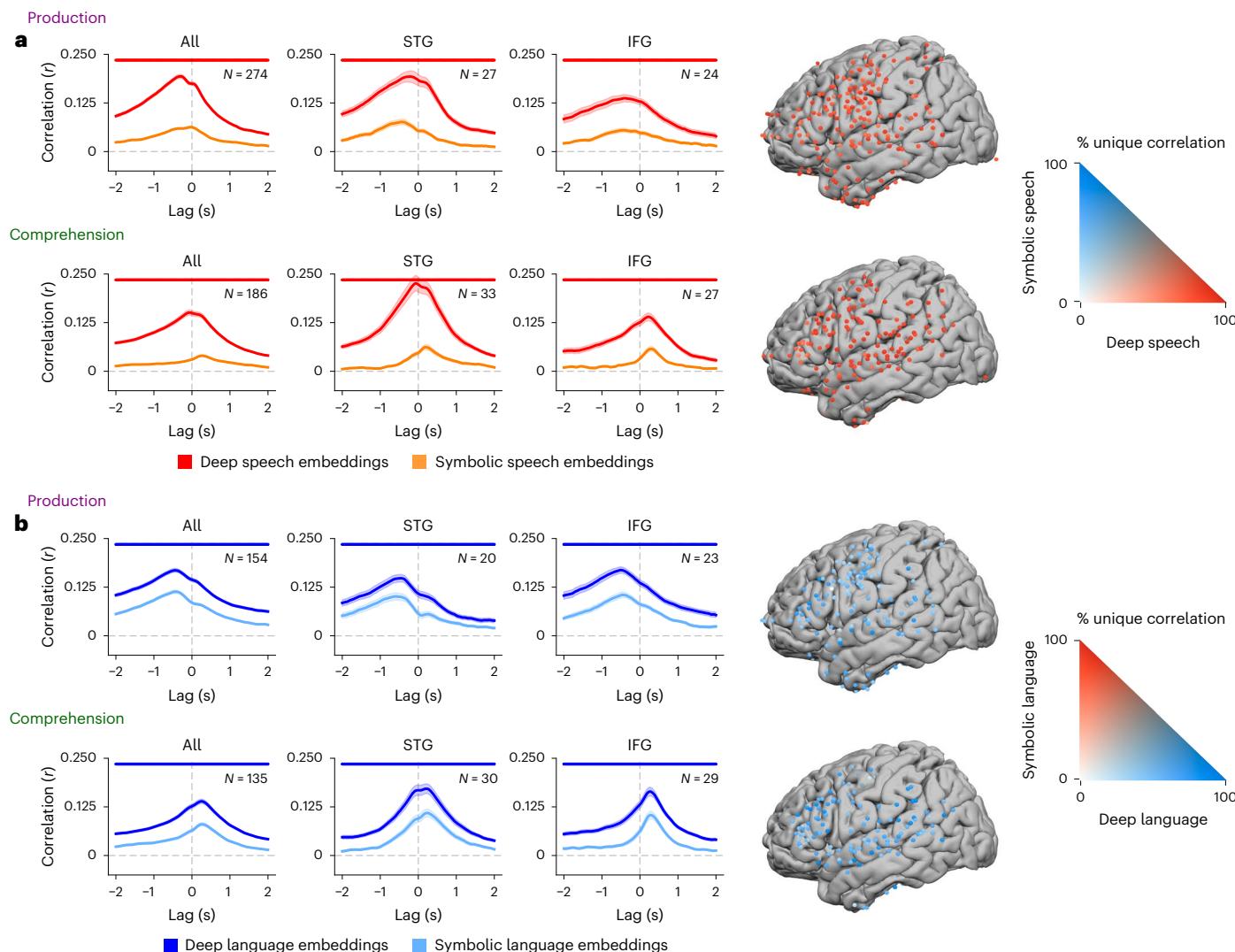
The high spatiotemporal resolution of our ECoG recordings allowed us to study the temporal dynamics of speech and language signals during real-life conversations. We calculated a separate encoding model for each embedding type over time, using 161 lags from  $-2,000$  to  $+2,000$  ms in 25-ms increments relative to word onset (lag 0). Our research showed different dynamic patterns for production vs comprehension across cortical areas. Our encoding models document a remarkable temporal specificity. Encoding performance peaks more than 300 ms before word onset during speech production (Fig. 3a) and more than 300 ms after word onset during speech comprehension (Fig. 3b). Although both the speech and language embeddings yield significant predictions in all regions of interest (ROIs), each embedding type captures different aspects of neural activity. A statistical contrast between models revealed that the speech embeddings better predict neural activity in early perceptual language areas along the STG and articulatory somatomotor areas. Conversely, language embeddings better predict neural activity in high-order language areas such as the IFG. In addition, while we observed biases of IFG towards language representation, and STG and SM towards speech representation, we could predict a substantial portion of the variance using either speech or language embeddings, suggesting a mixed representation in those ROIs. Supplementary Figs. 6 and 8 display the mean encoding results during production and comprehension in

three ROIs (SM, IFG and STG) per patient. Aggregated analysis across patients is presented in Supplementary Fig. 7.

In addition, we observed a different hierarchical selectivity during speech production and comprehension. Speech areas in STG and language areas in anterior and medial IFG yielded higher encoding performance during speech comprehension (Supplementary Fig. 8b, green), while posterior IFG and SM (preCG and postCG), as well as the TP, yielded higher encoding performance during speech production (Supplementary Fig. 8b, purple). Similar results were seen for language embeddings (Supplementary Fig. 8b). These results suggest a gradient from speech comprehension at the anterior part of IFG to speech production at the posterior IFG towards SM areas. We found that SM areas play a surprisingly notable role in real-life unconstrained conversations in terms of both speech and language features (Supplementary Fig. 8 shows results per cortical area).

Our ability to predict the neural responses of new conversations, which consisted of  $\sim 100$  h of audio recordings and 520,209 words, is a testament to the remarkable alignment between the neural activity and the internal population codes of the acoustic-to-speech-to-language model during our real-world conversations. The ability of the encoding model to generalize and predict minutes-long new conversations not seen during training is unrelated to the data size. A similar size effect was obtained even if only 50% or 25% of data were used, with only a slight decrease in power while using 10% of the data.

**Acoustic-to-speech-to-language model vs symbolic models**  
Deep acoustic-to-speech-to-language models provide an alternative, unified framework for modelling neural activity during



**Fig. 5 | Comparing speech and language embeddings to symbolic features.** **a**, We used a variance partitioning analysis to compare encoding models on the basis of speech embeddings (red; extracted from Whisper's encoder) and symbolic speech features (orange; phonemes, manner of articulation, place of articulation, speech or non-speech). Data are presented as mean  $\pm$  s.e.m. across electrodes. Red dots indicate lags with a significant difference ( $q < 0.01$ , FDR corrected) between deep speech embeddings and symbolic speech features. Encoding performance for deep speech embeddings is consistently higher than encoding performance for symbolic speech features across all significant electrodes, specifically in IFG and STG. **b**, We used a variance partitioning

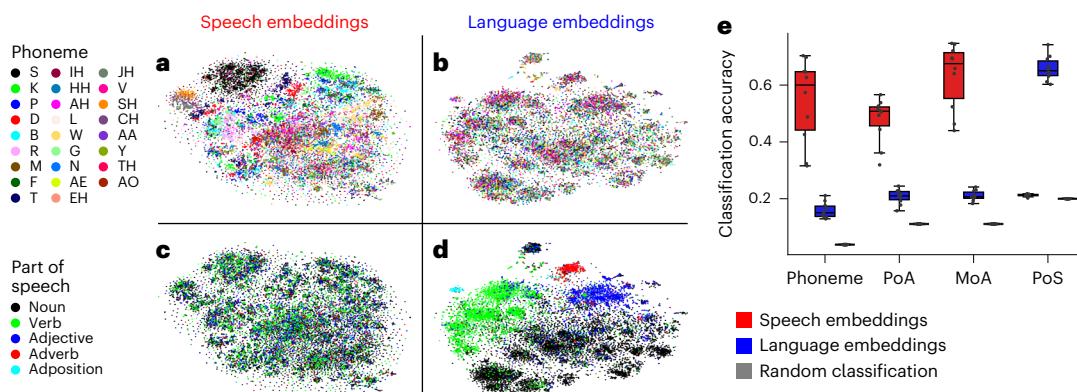
approach to compare encoding models on the basis of deep language embeddings (dark blue; extracted from Whisper's decoder) and symbolic language features (light blue; part of speech, dependency, prefix, suffix, stop word). Data are presented as mean  $\pm$  s.e.m. across electrodes. Blue dots indicate lags with a significant difference ( $q < 0.01$ , FDR corrected) between deep speech embeddings and symbolic language features. Encoding performance for deep language embeddings is consistently higher than encoding performance for symbolic language features across all significant electrodes, specifically in IFG and STG.

real-world conversations. Here we compare deep speech and language embeddings with symbolic speech and language models. We vectorize symbolic speech and linguistic features into binarized vectors. Vectorizing the symbolic models allows us to evaluate these symbolic models against the Whisper embeddings in the same regression-based encoding framework. We vectorize symbolic speech features (phonemes, voice, voiceless, place of articulation (PoA) and manner of articulation (MoA)) into a 60-dimensional binarized vector for each spoken word in the conversation. We also vectorize symbolic linguistic features (parts of speech (PoS), syntactic dependencies, prefixes, suffixes, stop words) into a 137-dimensional binarized vector for each word in the conversation (see Supplementary Table 3 for a comprehensive list of features).

Our findings indicate that speech and language embeddings extracted from the multimodal, deep acoustic-to-speech-to-language model outperform symbolic speech and language features (Fig. 5) in

predicting neural activity during natural conversations. This is evident in individual ROIs as well as across all electrodes. In addition, a variance partitioning analysis indicates that symbolic features account for very little unique variance beyond the deep multimodal embeddings.

Finally, we tested whether Whisper's speech and language embeddings implicitly learned classical psycholinguistic constructs. While phonemes and parts of speech do not function as fundamental computational (symbolic) units in the deep speech-to-text model, they nonetheless emerge as high-level descriptors of natural language. To visualize this, we used a nonlinear dimensionality reduction technique that maps high-dimensional data to a low-dimensional space (*t*-SNE) to project the multidimensional embeddings (3,840 dimensions for speech and 384 dimensions for language, sampled from each encoder layer and decoder layer) onto two-dimensional manifolds for visualization (Fig. 6a–d and Supplementary Fig. 9). Furthermore, we used a logistic classification procedure to classify phonemes with ~54% accuracy



**Fig. 6 | Representations of phonetic and lexical information in Whisper.**

**a–d.** Speech embeddings and language embeddings were visualized in a two-dimensional space using t-SNE. Each data point corresponds to the embedding for either an audio segment (speech embeddings from the encoder network) or a word token (language embeddings from the decoder network) for a unique word (averaged across all instances of a given word). Clustering according to phonetic categories is visible in speech embeddings (a) but far less prominent in language embeddings (b). Clustering according to lexical information

(part of speech) is visible in language embeddings (d) but not in speech embeddings (c). **e.** Classification of phonetic and lexical categories based on speech and language embeddings. We observed robust classification for phonetic information based on speech embeddings. We also observed robust classification for parts of speech based on language embeddings. The classification was performed using logistic regression, and the performance was measured on held-out data using a 10-fold cross-validation procedure.

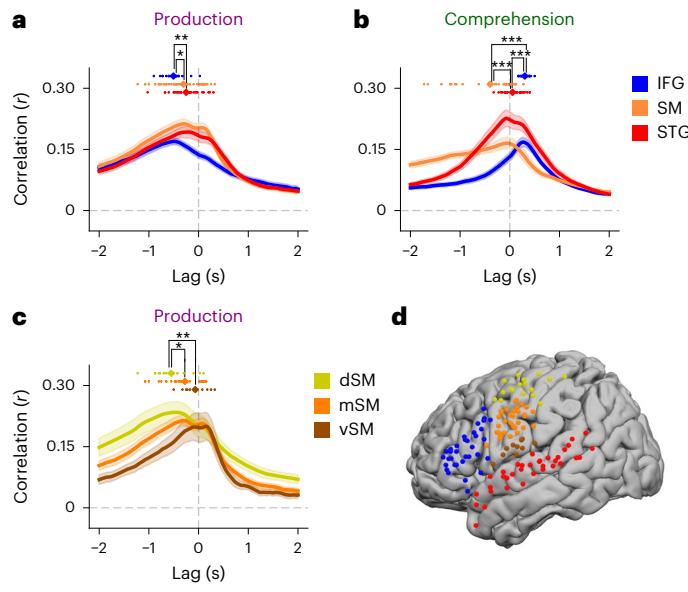
(chance level 4%,  $P < 0.001$ , determined using permutation test; see Methods) from Whisper's speech encoder embeddings (Fig. 6e). Similar clustering results were obtained for PoA and MoA (Supplementary Fig. 9). This indicates that high-level, symbolic descriptors of human speech emerge from speech embeddings learned using a simple objective function against training samples of real-world speech. Similarly, we successfully clustered and classified PoS (nouns, verbs, adjectives and so on) with ~67% accuracy (chance level 20%,  $P < 0.001$ ; see Methods for details) from the language embeddings (Fig. 6e). This suggests that language embeddings can capture high-level syntactic properties without relying on built-in symbolic processing or representational units. Note that Whisper was trained end-to-end to predict upcoming words given the audio as input; the encoder was not explicitly trained to recognize phonemes, and the decoder was not trained to recognize parts of speech. Our findings confirm that deep end-to-end multimodal models can capture language statistics without relying on predefined symbolic units, commonly considered the fundamental building blocks for natural language processing in psycholinguistics.

### Information flow during speech production and comprehension

Evaluating encoding models at each lag relative to word onset allows us to trace the temporal flow of information from STG (speech comprehension ROI) to IFG (language-related ROI) to SM (speech production ROI) during the production and comprehension of natural conversations (Fig. 7). In congruence with the flow of information during speech production, language encoding in IFG peaked first at ~500 ms before word onset ( $M = -505$  ms,  $s.d. = 201$  ms), whereas in SM (comprising preCG and postCG), the speech model encoding peaked significantly closer to speech onset ( $M = -200$  ms,  $s.e. = 7$  ms,  $t_{(78)} = 2.23$ ,  $P < 0.05$ ,  $CI(95\%) = [-180, -220]$  ms, Cohen's  $d = 0.03$ ; Fig. 7). A reverse dynamic was observed during speech comprehension (Fig. 7; see also ref. 26). During speech comprehension, speech areas along the STG peaked shortly after word onset ( $M = 54$  ms,  $s.d. = 186$  ms), while language model encoding in IFG peaked significantly later, ~300 ms after word onset ( $M = 247$  ms,  $s.e. = 4$  ms,  $t_{(60)} = -6.48$ ,  $P < 0.001$ ,  $CI(95\%) = [221, 273]$  ms, Cohen's  $d = 0.04$ ; Fig. 7). Finally, we found an unexpected temporal pattern of speech encoding during speech production: peak encoding performance proceeded from dorsal SM to middle SM, and finally to ventral SM before word articulation (Fig. 7).

Upon closer examination of the activity pattern, we observed two distinct peaks in the STG and somatosensory areas during speech production (Fig. 3a). The first peak appears ~300 ms before word onset. In contrast, the second peak occurs ~200 ms after word onset. Additional analyses indicate that the first peak is associated with motor planning, while the second peak is associated with the speaker processing their own voice (Supplementary Fig. 10).

To further dissociate neural activity before and after word onset during speech production and comprehension, we utilize the high precision of Whisper's encoder to extract speech embedding and construct encoding models for each 20-ms segment of speech (see Methods for details). This fine-grained analysis allows us to map the sequence of neural activity in unconstrained, real-world conversations with a temporal resolution of 20 ms. We observed that during speech comprehension, neural encoding begins to peak around word onset and gradually shifts over time (Fig. 8b,d). This indicates that the processing sequence in the speech encoder's top layer matches the sequence of neural activity in the human brain. Note that the embeddings at word onset carry some contextual information about the previous word and, thus, can fit responses about ~50 ms before word onset. We observed a different sequence of neural responses during speech production (Fig. 8a,c). Before word onset, neural encoding peaks across speech units occur with a fixed delay of about ~300 ms and do not shift over time. This suggests that during the planning phase, the brain already has information about the entire sequence of speech articulation for each word at approximately ~300 ms before speech articulation (Fig. 8a,c). After word onset, neural encoding peaks gradually shift over time in a similar pattern to speech comprehension (Fig. 8a,c). This finding indicates that the second post-word onset neural encoding peak is associated with neural mechanisms for processing self-generated speech as the speakers hear their own voice. To statistically evaluate the relationship between the encoder unit and the peak in encoding performance while considering patient variability, we constructed linear mixed models including a random intercept per patient. During comprehension, we observe a temporal shift in the encoding peak with increasing distance between the temporal segment covered by the encoder unit and word onset ( $\beta = 0.028$ ,  $P < 0.001$ ,  $CI(95\%) = [0.021, 0.035]$ ). During production, we observe a comparable shift in the encoding peak after word onset ( $\beta = 0.017$ ,  $P < 0.001$ ,  $CI(95\%) = [0.015, 0.019]$ ) but not before word onset ( $\beta = 0.001$ ,  $P = 0.59$ ,  $CI(95\%) = [-0.003, 0.005]$ ).



\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$

**Fig. 7 | Temporal dynamics of speech production and speech comprehension across different brain areas.** On the basis of tuning preferences for each ROI, we assessed temporal dynamics using the language model for IFG and the speech model for STG and SM. Coloured dots show the lag of the encoding peak for each electrode per ROI. Data are presented as mean  $\pm$  s.e. across electrodes. To determine significance, we performed independent-sample  $t$ -tests between encoding peaks;  $P$  values are one-sided. **a**, During speech production, encoding for language embeddings in IFG peaked significantly before speech embeddings in SM and STG. **b**, The reverse pattern was observed during speech comprehension: encoding performance for language embeddings encoding in IFG peaked significantly after speech encoding in SM and STG. **c**, For speech production, we observed a temporal pattern of encoding peaks shifting towards word onset within SM, proceeding from dorsal (dSM) to the middle (mSM) to ventral (vSM). **d**, Map showing the distribution of electrodes per ROI.

SM and STG articulation areas during speech production, suggesting a potential coupling between motor and perceptual processes<sup>30,31</sup>.

The unified, multimodal model provides a precise numerical code for how acoustic, speech and language features can be integrated across different levels of the cortical hierarchy. For example, acoustic information is preserved in speech embeddings (Fig. 3a), while speech and language embeddings capture different portions of the variance across areas (Fig. 3b). Allowing information to flow from the speech encoder into the language decoder, however, did improve the ability of the language embeddings to model neural activity across language areas (Fig. 4). This illustrates how the acoustic-to-speech-to-language model provides a holistic computational framework for how the brain integrates acoustic, speech and language information while processing natural conversations<sup>17,32</sup>. Overall, these results shed new light on the interaction between lower-level speech and higher-level semantic processing, where linguistic prediction can facilitate speech processing in auditory areas, and acoustic information can facilitate the processing of words in language areas<sup>33–37</sup>.

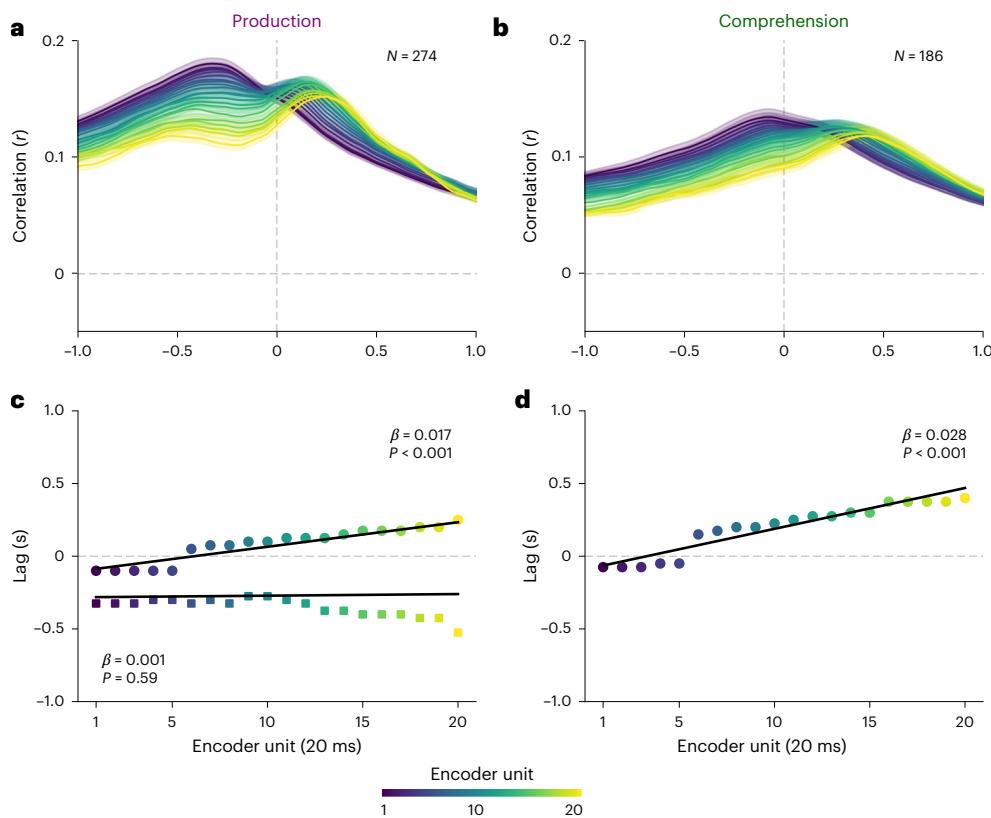
The acoustic-to-speech-to-language model processes natural speech with a temporal resolution of 20 ms. This gives us unprecedented precision in modelling how speech and language information are processed during real-life conversations. Regarding speech comprehension, the model revealed a sequence of speech-related activity at 20-ms resolution triggered around word onset. On the other hand, during speech production, the model revealed that information about the entire sequence of word articulation is already present 300 ms before word onset (Fig. 7). Interestingly, we also observed a secondary cascade of activity after word onset during speech production, which matches the activity wave during speech comprehension. These findings suggest that the same cortical areas that process incoming information from other speakers also process the speaker's own speech (Supplementary Fig. 10; see also ref. 38). In our investigation of sensorimotor areas, we observed a distinct dynamic of neural encoding following speech onset. Notably, these responses were accurately predicted only by a model trained during production, while models trained for comprehension yielded lower correlations (Supplementary Fig. 10). This divergence suggests unique neural representations of articulatory and speech features in the SM areas during speech comprehension and production. However, further research is required to test this hypothesis.

How should we interpret the relationship between the internal representations of the acoustic-to-speech-to-language model and the human brain when processing human speech? There are two potential options to consider. The first option is that our encoding model effectively learns the transformation between distinct codes for processing natural language. This is significant because it positions deep language models as a powerful computational tool to study and predict how the brain processes everyday conversations. They enable us to robustly predict the neural responses to speech and language information across multiple conversations and contexts on a scale that was not previously possible. This breakthrough was instrumental in modelling our unique, entirely unconstrained conversational dataset. The second interpretation is that deep language models and the human brain share computational principles for natural language processing<sup>39,40</sup>. This stronger theoretical claim challenges traditional rule-based symbolic linguistic models of language representation and processing<sup>41</sup>. Some arguments support the stronger theoretical claim. First, our encoding models established that a simple linear mapping between the internal neural activity in Whisper and the human brain yields remarkably high prediction performance. This suggests that the two internal representations may be more similar than initially anticipated. Second, deep speech and language embeddings dramatically outperform symbolic models for speech and language processing of our natural conversations (Fig. 5). Combined, our finding of a linear relationship between the internal activity in the acoustic-to-speech-to-language model and the internal

## Discussion

We analysed neural processes involved in natural speech production and comprehension using ECoG recordings collected over ~100 h of spontaneous open-ended conversations, comprising approximately half a million words. The unprecedented size of this dataset provides us with a detailed and uniquely comprehensive look at the richness of human conversations as they unfold in real-world contexts. We extracted internal acoustic, speech and language-related activity from embeddings at different layers of a unified acoustic-to-speech-to-language model (Whisper). Next, we built encoding models that learn a simple linear mapping between the model's internal embeddings and human brain activity—word by word during speech production and comprehension. Using the encoding model, we predicted, with remarkable precision, neural activity associated with acoustic, speech and language processing in speech-related and language-related areas for hours of new conversations not used in training the model.

Our encoding models revealed a distributed processing hierarchy in which sensory areas along the superior temporal gyrus and somatomotor areas along the precentral gyrus were better modelled by speech embeddings (red, Fig. 3). This result aligns with previous findings that used a unimodal speech model (Hubert) to encode speech information during passive listening to a closed set of sentences<sup>27</sup>. Higher-order language areas in the inferior frontal gyrus, as well as the posterior temporal and parietal cortex, were better modelled by language embeddings (blue, Fig. 3). This was true for speech production and comprehension. These results recapitulate the known hierarchy of natural language processing during free-flowing conversations<sup>28,29</sup>. Notably, we found strong alignment to speech embeddings in both



**Fig. 8 | Fine-grained temporal sequence of speech encoding during production and comprehension.** **a**, Encoding models for encoder units 1–20 time locked to word onset (corresponding to a temporal segment of 20–400 ms after word onset) during production. The encoding performance exhibits two peaks (one before and one after word onset). Data are presented as mean  $\pm$  s.e. across electrodes. **b**, Encoding models for encoder units 1–20 time locked to word onset during comprehension. The encoding performance peaks mainly

after word onset. Data are presented as mean  $\pm$  s.e. across electrodes. **c**, Coloured squares correspond to peaks encoding during production before word onset, and round dots after word onset. The model was found to be significant, the  $P$  value is two-sided. **d**, Coloured dots correspond to peaks encoding during comprehension after word onset. Data are presented as mean  $\pm$  s.e.m. across electrodes. The model was found to be significant, the  $P$  value is two-sided.

activity in the human brain during natural speech and language processing offers an alternative, unified computational framework for how the brain learns to process many aspects of natural speech.

Finally, although phonemes, place of articulation, manner of articulation and parts of speech are not considered fundamental computational units in the deep speech and language model, they emerge as high-level statistical descriptors of natural language embedded in the neural code of the model. This highlights the dual power of our unified acoustic-to-speech-to-language model to (1) account for how the brain processes language in real-life conversations collected in the wild across a diversity of real-life contexts<sup>12</sup> and (2) account for high-level phenomena documented by psycholinguistics over the years<sup>42</sup>.

In summary, the acoustic-to-speech-to-language model provides a new unified computational framework for studying the neural basis of natural language processing. This integrated framework signifies the beginning of a paradigm shift towards a new family of non-symbolic models based on statistical learning and high-dimensional embedding spaces. As these models improve at processing natural speech, their alignment with cognitive processes may also improve. For instance, new models are being developed to process speech-to-language-to-articulation without written text, referred to as audio-to-audio language models<sup>43</sup>. Such models allow for a more comprehensive analysis of linguistic phenomena, covering all levels of linguistic analysis, from acoustic and speech perception to language and motor articulation. Some models, such as GPT-4o, add a third visual modality to the speech and text multimodal model<sup>44</sup>,

while others incorporate embodied articulation systems that mimic human speech articulation systems<sup>45</sup>. The fast improvement of these models supports a shift to a unified linguistic paradigm that emphasizes the role of usage-based statistical learning in language acquisition as it is materialized in real-life contexts.

## Methods

### Ethics oversight

The study was approved by the NYU Grossman School of Medicine Institutional Review Board (approved protocol s14-02101) which operates under NYU Langone Health Human Research Protections and Princeton University's Review Board (approval protocol 4962). Studies were performed in accordance with the Department of Health and Human Services policies and regulations at 45 CFR 46. Before obtaining consent, all participants were confirmed to have the cognitive capacity to provide informed consent by a clinical staff member. Participants provided oral and written informed consent before beginning study procedures. They were informed that participation was strictly voluntary and would not impact their clinical care. Participants were informed that they were free to withdraw participation in the study at any time. All study procedures were conducted in accordance with the Declaration of Helsinki.

### Participants

Four patients (2 females, gender assigned on the basis of medical record; 24–53 years old) with treatment-resistant epilepsy undergoing intracranial monitoring with subdural grid and strip electrodes for

clinical purposes participated in the study. No statistical method was used to predetermine sample size. Three study participants consented to have a US Food and Drug Administration (FDA)-approved hybrid clinical–research grid implanted that includes additional electrodes in between the standard clinical contacts. The hybrid grid provides a higher spatial coverage without changing clinical acquisition or grid placement. Each participant provided informed consent following protocols approved by the New York University Grossman School of Medicine Institutional Review Board. Patients were informed that participation in the study was unrelated to their clinical care and that they could withdraw from the study without affecting their medical treatment.

### Preprocessing the speech recordings

We developed a semi-automated pipeline for preprocessing the dataset. The pipeline can be broken down into four steps:

1. De-identifying speech. All conversations in a patient's room were recorded using a high-quality microphone and stored locally. These audio recordings contain sensitive information about the patient's medical history and private life. To comply with the Health Insurance Portability and Accountability Act of 1996 (HIPAA)'s data privacy and security provisions for safeguarding medical information, any identifiable information (for example, names of people and places) was censored. Given the sensitivity of this phase, we employed a research specialist dedicated to the manual de-identification of recordings for each patient.
2. Transcribing speech. Although many speech-to-text transcription tools have been developed, extracting text from 24/7 noisy, multispeaker audio recordings is challenging. We used a human-in-the-loop annotation pipeline integrated with human transcribers from Amazon's Mechanical Turk to achieve the transcription quality necessary for our preliminary analyses.
3. Aligning text to speech. Text transcripts (that is, sequences of words) must be aligned with the audio recordings at the individual word level to provide an accurate time stamp for the production of each word. We used the Penn Forced Aligner<sup>46</sup>, which yields timestamps with 20-ms precision, to generate rough word onsets and offsets. We further improved this automated forced alignment by manually verifying and adjusting each word's onset and offset times.
4. Aligning speech to neural activity. To provide a precise mapping between neural activity and the conversational transcripts, we engineered one of the ECoG channels to record the microphones' output directly. The concurrent recordings of the audio and neural signals allowed us to align both signals with ~20 ms of precision.

### Preprocessing the ECoG recordings

The ECoG preprocessing pipeline mitigated artefacts due to movement, faulty electrodes, line noise, abnormal physiological signals (for example, epileptic discharges), eye blinks and cardiac activity<sup>47</sup>. We built a semi-automated analysis pipeline to identify and remove corrupted data segments (for example, due to epileptic seizures or loose wires) and mitigate other noise sources using fast Fourier transform (FFT), independent component analysis (ICA) and de-spiking methods<sup>48</sup>. We then bandpassed the neural signals using a broadband (75–200 Hz) filter and computed the power envelope, a proxy for each electrode's average local neural firing rate<sup>49</sup>. The signal was z-scored and smoothed with a 50-ms Hamming kernel. Three thousand samples were trimmed at each signal end to avoid edge effects. Signal preprocessing was performed using custom preprocessing scripts in MATLAB 2019a (MathWorks).

### Acoustic embedding extraction

To prepare audio recordings for subsequent processing by the speech model, we downsampled the audio recordings from 16 kHz. Since Whisper is trained on 30-s audio segments, audio recordings were fed to the model using a sliding window of 30 s. Whisper encoder's internal representations are not aligned to discrete word tokens (as in the decoder); instead, the encoder embeddings correspond to temporal segments of the original audio input. In our data, the median word duration is 189 ms (mean = 227 ms, s.d. = 158 ms), with the shortest word being 12 ms ('I') and the longest being 2,000 ms ('hysterical'). Other long words include 'mademoiselle' (1,850 ms), 'two-hundred-and-fifty-six' (1,995 ms) and 'narcolepsy' (1,996 ms). To temporally align the embeddings to word onsets, we defined the endpoint of each sliding window to the word's onset plus 200 ms so that the extracted 'word embedding' contained no information before word onset after the spectrogram and convolution layers. Inside Whisper, each 30-s audio segment was transformed into 1,500 encoder hidden state embeddings, where each hidden state represents a temporal segment of ~20 ms. We concatenated the last 10 hidden states to extract embeddings on the word level ( $d = 10 \times 384 = 3,840$ ), corresponding to 200 ms of the audio input. The acoustic embedding was extracted from the zeroth encoder layer (before any transformer blocks); therefore, no previous context was incorporated into the embedding.

### Speech embedding extraction

The speech embedding extraction process is the same as that for acoustic embedding extraction, where we aligned the temporal segments of audio input to word onsets. However, instead of the zeroth layer, we extracted embeddings from the fourth encoder layer since our classification analysis indicated that embeddings extracted from the fourth encoder layer have the most structured representation of phonetic categories compared with embeddings extracted from other encoder layers (Supplementary Fig. 9e).

### Speech embedding extraction with varied length

Since word duration is highly variable in conversational speech, we calculated the number of hidden states needed to capture the full word, from word onset to offset. For example, since each hidden state roughly represents a temporal segment of 20 ms, we would need 5 hidden states for a 100-ms word, 10 for a 200-ms word, and 20 for a 400-ms word. To temporally align the word embedding to the word onset, we defined the endpoint of each sliding window to the word's onset plus 20 ms times the number of hidden states needed to capture the word. This process created embedding vectors with different dimensions. Since the encoding model requires the same embedding size for all words, we used principal component analysis (PCA) for each word embedding to the same dimensionality of a single embedding unit ( $d = 384$ ). We re-ran the encoding models for comprehension and production using the speech-aligned embeddings and received results similar to those of the fixed-length speech embeddings. The results indicated that the original (fixed 200-ms length) speech encoding and the new, word length-based speech embeddings are almost identical (production,  $r_{(159)} = 0.99$ ,  $P < 0.001$ , CI(95%) = [0.996, 0.998]; comprehension,  $r_{(159)} = 0.99$ ,  $P < 0.001$ , CI(95%) = [0.996, 0.998]). This shows that our speech encoding results are robust when extracting speech embeddings on the basis of a fixed duration or over a dynamic range.

### Continuous acoustic and speech embedding extraction

Instead of using a sliding window of 30 s, audio recordings were fed to the model by non-overlapping 30-s segments. Because the 30-s audio segments were transformed into 1,500 encoder hidden state embeddings, each hidden state roughly represents a temporal segment of 20 ms. For each hidden state, we extracted its embeddings and calculated its onset and offset. Notably, instead of concatenating temporal hidden states to align with words, we treated each hidden

state as independent. Consequently, the embeddings represent a continuous audio stream rather than discrete words. We extracted embeddings from the zeroth (continuous acoustic) and fourth (continuous speech) encoder layers, corresponding to our previous acoustic and speech embedding extraction process. Due to the inherent continuous nature of the embeddings and the challenges in identifying clean boundaries between production and comprehension, we limited our selection to 30-s audio segments that are entirely either production or comprehension. We performed encoding on both continuous acoustic and continuous speech embeddings. When averaging neural signals, we used a 20-ms window at each lag (at 20-ms increments) to account for the finer temporal resolution of the continuous embeddings. We also replicated our results with the original 200-ms window at each lag (at 25-ms increments).

### Language embedding extraction

For each word, text transcripts corresponding to the 30-s context window were tokenized and given as contextual input to the decoder ( $M = 70$  words,  $s.d. = 28$  words in a 30-s window). We extracted the embedding corresponding to the last word in the sequence. We extracted embeddings from the third decoder layer in line with previous results, indicating that late-intermediate layers of language models show the best encoding performance for neural data.

### Electrode-wise encoding

We used linear regression to estimate encoding models for each electrode and lag relative to word onset to map the Whisper embeddings onto the neural activity. To construct the outcome variable, we averaged the neural signal across a 200-ms window at each lag (at 25-ms increments) for each electrode across all words (the results replicate for varying windows of 50 ms, 100 ms and 200 ms; Supplementary Fig. 5b). Using a 10-fold cross-validation procedure, we trained two sets of encoding models to predict the word-by-word neural signal magnitude on the basis of either speech or language embeddings. Within each training fold, we standardized the embeddings and used PCA to reduce the embeddings to 50 dimensions. We then estimated the regression weights using ordinary least-squares multiple linear regression from the training set and applied those weights to predict the neural responses for the test set. We calculated the Pearson correlation between the predicted and actual neural signals for each held-out test fold to assess model performance. The correlations were averaged across the 10 test folds. This procedure was repeated at 161 lags from  $-2,000$  to  $2,000$  ms in 25-ms increments relative to word onset; the exact predictor embeddings were used at each lag. To determine the maximum correlation across lags for each fold, we used the 9 training folds to estimate the lag that yielded the maximum correlation, then extracted the corresponding correlation for that specific lag from the test fold.

### Variance partitioning analysis

We employed a variance partitioning scheme to estimate the variance that different models uniquely explain. We built encoding models on the basis of two different embeddings A and B (for example, speech and language embeddings) and an additional combined encoding model where we concatenate the embeddings of A and B. Evaluating the encoding performance of the concatenated model gives us  $r^2_{A \cup B}$ . Using set arithmetic, we can derive the unique variance explained by embeddings A and B: we calculated the shared variance explained by both embeddings A and B as  $r^2_{A \cap B} = r^2_A + r^2_B - r^2_{A \cup B}$ . Now we can calculate the unique variance explained by embeddings A and B as  $r^2_{A_{\text{unique}}} = r^2_A - r^2_{A \cap B}$  and  $r^2_{B_{\text{unique}}} = r^2_B - r^2_{A \cap B}$ . We further calculated the percent variance uniquely explained by embeddings A and B as  $\%r^2_{A_{\text{unique}}} = r^2_{A_{\text{unique}}} / r^2_{A \cup B}$  and  $\%r^2_{B_{\text{unique}}} = r^2_{B_{\text{unique}}} / r^2_{A \cup B}$ . Our colour scheme reflects the relative variance explained. This way, we can identify which electrodes are better explained by  $r^2_{A_{\text{unique}}}$ ,  $r^2_{B_{\text{unique}}}$  or  $r^2_{A \cap B}$ . Since  $\%r^2_{A_{\text{unique}}}$ ,

$\%r^2_{B_{\text{unique}}}$  and  $\%r^2_{A \cap B}$  must add up to one, if both  $\%r^2_{A_{\text{unique}}}$  and  $\%r^2_{B_{\text{unique}}}$  are low (indicated by white),  $\%r^2_{A \cap B}$  is high, that is, the percent shared variance explained by both embeddings A and B is higher than the percent variance explained by either A or B alone.

### Electrode selection

To identify significant electrodes, we used a randomization procedure. At each iteration, we performed a random shift in the assigned embeddings to each predicted signal, thus disconnecting the relationship between the words and the brain signal while preserving the order between the different embeddings. The random shift was restricted to avoid rolling the assignment inside the context window. We then performed the entire encoding procedure for each electrode on the mismatching words. We repeated this process 1,000 times. After each iteration, the encoding model's score was calculated on the basis of the maximal value minus the minimal value across all 161 lags for each electrode. We then took the maximum value for each patient for each permutation across all electrodes. This resulted in a distribution of 1,000 maximum values for each patient, which was used to determine the significance of all electrodes. For each electrode, a  $P$  value was computed as the percentile of the original maximum–minimum values of the encoding model across all lags from the null distribution of 1,000 similarly calculated values. Performing a significance test using this randomization procedure evaluates the null hypothesis that there is no systematic relationship between the brain signal and the corresponding word embedding. This procedure yielded a family-wise error rate corrected  $P$  value for each electrode, correcting for the multiple lags<sup>50</sup>. Electrodes with  $P$  values less than 0.01 were considered significant.

### Differences in the overall magnitude of encoding performance

We used the same randomization procedure described in the electrode selection section to identify electrodes with significant differences in the magnitude of encoding performance for speech and language embeddings. We only statistically evaluated differences in model performance for electrodes with significant encoding performance for at least one model (see 'Electrode selection' above). For each permutation, we computed the difference in model performance by subtracting the two maximal encoding performance (correlation) values for each electrode across all 161 lags. This resulted in a distribution of 1,000 difference values between speech and language embeddings' encoding performance at each electrode. For each electrode, a  $P$  value was computed as the percentile of the non-permuted maximum difference values in encoding performance between speech and language embeddings across all lags from the null distribution of 1,000 difference values. We used false discovery rate (FDR) correction to correct for testing across multiple electrodes<sup>51</sup>. Electrodes with  $q$ -values less than 0.005 (significance of 0.01 standardized for the two-sided test) were considered to have significant differences in model performance. We used the same procedure to identify electrodes that showed a significant difference in the magnitude of encoding performance between speech production and comprehension.

### Differences in lag-by-lag encoding performance

To test for significant differences in electrode-wise encoding performance between the speech and language embeddings for each lag, we used a paired-sample permutation procedure: in each permutation, we randomly shuffled the labels of all observations for both models (we obtained a correlation coefficient for each fold during a 10-fold validation procedure, thus collecting 10 observations per electrode for each model). Then, we computed the difference in encoding performance between speech and language embeddings. We computed the exact null distribution of different values for the 10 observations ( $2^{10} = 1,024$  permutations). For each lag, a  $P$  value was computed as the percentile of the non-permuted difference relative to the null distribution of 1,024 difference values. To correct

for multiple lags, we used the FDR correction procedure<sup>51</sup>. Lags with  $q$ -values less than 0.005 (significance of 0.01 for the two-sided test) were considered significant.

We used a similar procedure to test for significant differences in electrode-wise encoding performance for the speech and language embeddings averaged across electrodes in different ROIs: we randomly shuffled the labels of all observations ( $10 \times n$ , where 10 is the number of folds and  $n$  corresponds to the number of electrodes in the ROI) and computed the difference in mean encoding performance between the speech embeddings and language embeddings. This process was repeated 10,000 times, resulting in a distribution of 10,000 difference values. For each lag, a  $P$  value was computed as the percentile of the non-permuted difference relative to the null distribution. FDR correction was applied to correct for multiple lags. Lags with  $q$ -values less than 0.005 (significance of 0.01 for the two-sided test) were considered statistically significant.

### Differences in the temporal lag of peak encoding performance

To test for significant differences in the temporal dynamics of encoding performance between ROIs, we performed independent-sample  $t$ -tests. First, we hypothesized that peak encoding in IFG electrodes would occur significantly earlier than in electrodes in somatomotor and auditory areas for production. To test this hypothesis, we performed an independent-sample  $t$ -test (one-sided) on the lags at peak encoding for electrodes in the given ROIs. Second, we hypothesized that for comprehension, peak encoding in electrodes in IFG would occur significantly later than peak encoding in electrodes in SM and STG. To test this hypothesis, we performed an independent-sample  $t$ -test (one-sided) on the lags at peak encoding for electrodes in the given ROIs. To test whether the peak encoding performance for electrodes in a given ROI occurred significantly before or after word onset, we performed one-sample  $t$ -tests (two-sided) on the lags at peak encoding for electrodes in the given ROI against lag 0 (word onset). We removed electrodes where the maximal lag exceeded three interquartile ranges above or below the median to reduce the influence of outliers.

### Implementing comprehension encoding model during production

We trained encoding models on speech comprehension data to further investigate the shared mechanisms between speech production and comprehension. We applied the beta weights of the best-performing lag to predict neural activity during production. Notably, the 10-fold cross-validation procedure was done on production and comprehension data together to avoid data leakage. We identified electrodes showing a double peak during speech production (at least one peak before and after word onset). We defined an encoding peak as a local maximum with a minimum correlation of 0.1 and a topographic prominence of at least 0.007. We implemented the peak-finding algorithm from the Scipy-signals package in Python (Scipy v.1.11.4.).

### Encoding models per speech unit

To test the temporal relationship between speech representation in Whisper's encoder and the brain, we constructed separate encoding models for 20 encoder hidden states (each receiving 20 ms of the original audio input in consecutive steps). All 20 encoder hidden states in the original audio input covered the range from word onset to 400 ms after word onset. Since there were more short words than long words in our dataset, the sample size decreases for later temporal segments (from 221,989 words for the encoding model corresponding to the first unit to 24,089 words for the encoding model corresponding to the 20th unit for production; and 276,812 words to 29,562 words for comprehension). We used a temporal smoothing window of 200 ms to average the neural signal. We replicated the results using a 20-ms smoothing window.

### Linear mixed model

We averaged encoding performance across all electrodes separately for each patient and computed each model's peak in encoding performance. For production, we computed two encoding peaks (before and after word onset), which align with our results showing a distinct double peak during production. The preprocessing procedure introduced a temporal uncertainty of 200 ms around word onset, where information from after leakage after word onset is bounded by -100 ms. Therefore, encoding peaks were defined as 'before word onset' when occurring between -2,000 ms and -100 ms before word onset and as 'after word onset' when occurring between -100 ms before and 2,000 ms after word onset. We computed the encoding peak between -2,000 ms and 2,000 ms around word onset for comprehension. To account for intersubject variability, we analysed time points of the neural encoding peaks with linear mixed models (LMMs), including a random intercept per patient using restricted maximum likelihood estimation. LMMs were implemented using the Statsmodels-regression package (Statsmodels v.0.14.1) in Python.

### Visualization of embedding space

We used t-SNE to project the high-dimensional embedding spaces down to two-dimensional manifolds to visualize the information structure represented in speech and language embeddings. This projection was computed separately for the speech embeddings (from the encoder network) and the language embeddings (from the decoder network). Each data point in the scatterplots (Fig. 6 and Supplementary Fig. 9) corresponds to a speech or language embedding for a unique word. We averaged the embeddings across instances throughout the transcript for each unique word ( $n = 13,347$ ) to get one embedding per word. We replicated the analysis using each word's first or random instances and obtained similar results. We then applied t-SNE to the averaged embeddings with perplexity = 50. To better understand the structure of this two-dimensional space, we coloured the data points (corresponding to word embeddings) according to several speech and language features: phonemes, place of articulation, manner of articulation and part of speech. Phonemes, PoA and MoA capture speech acoustic and articulatory features, whereas PoS captures lexical categories. We obtained phoneme classes from the Carnegie Mellon Pronouncing Dictionary<sup>52</sup>, which provides 39 classes (37 in our dataset). We further classified the phonemes on the basis of their place of articulation (total of 9 classes in our dataset) and manner of articulation (total of 9 classes in our dataset) according to the general American English consonants of the International Phonetic Alphabet. Because each word consists of multiple phonemes, we took the first phoneme for each word. We replicated the following visualizations and classification analyses for each word's second, third and fourth phonemes separately, and obtained similar results. To extract part of speech information, we used the part of the speech tagging process available in the NLTK Python package (total of 12 classes, 11 in our dataset). We removed classes with less than 100 occurrences (less than 1% of the data, resulting in 27 phoneme classes, 9 PoA classes, 9 MoA classes and 5 PoS classes).

### Classification of speech and linguistic features

To quantify the information encoded in the embeddings, we trained multinomial logistic regression classifiers (using the L2 penalty and default  $C = 1.0$  in sci-kit-learn) to predict phonetic (phonemes, PoA, MoA) and lexical categories (PoS) separately for both speech and language embeddings. We used a 10-fold cross-validation procedure with temporally contiguous training/test folds to train and evaluate classifier performance. On each fold of the cross-validation procedure, embeddings were standardized and reduced to 50 dimensions using PCA. To establish a baseline for comparing classifier accuracy, we trained dummy classifiers that learned to predict the most frequent class. Since the distribution of classes in our dataset was unbalanced, we used the balanced accuracy metric to evaluate classification performance<sup>53</sup>.

Balanced accuracy was calculated as the proportion of correct predictions per class averaged across all classes. This resulted in a value between 0 and 1, with higher values indicating better classification performance. For instance, a random classifier that always predicts the most frequent class will have a balanced accuracy of 1 divided by the number of classes at the chance level. The balanced accuracy metric assesses how well the classifier can differentiate between classes while minimizing misclassifications due to unbalanced data. The classification significance was computed using a non-parametric bootstrapping procedure where the labels of the classes tested (phoneme, PoA, MoA, PoA) were shuffled 1,000 times and the classification score was computed for each of the shuffle interactions. The actual score (non-shuffled) was higher than all the scores in the shuffled interactions.

## Reporting summary

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

## Data availability

The data contain patient–doctor conversations protected by HIPAA privacy regulations. Due to the size and complexity of our recordings, the data cannot be de-identified. Due to the sensitive nature of audio conversation data, we can only share data with researchers who directly contact the corresponding author and complete a signed data-sharing agreement with NYU Langone and onboard to our IRB. This process ensures that data sharing complies with HIPAA terms and our IRB terms, and that adequate resources are in place to prevent identifiable patient or audio data from leaving the Hospital’s ecosystem.

All data for reproducing the encoding results including encoding plots, error bars, thresholds and significance asterisks are available on GitHub at [https://github.com/hassonlab/247-plotting/blob/main/scripts/tfspaper\\_whisper.ipynb](https://github.com/hassonlab/247-plotting/blob/main/scripts/tfspaper_whisper.ipynb).

## Code availability

The codes for replicating the core analyses of this manuscript are available on GitHub at <https://github.com/hassonlab/247-pickling/tree/whisper-paper-1> (for embedding extraction) and at <https://github.com/hassonlab/247-encoding/tree/whisper-paper-1> (for encoding models).

## References

1. Hockett, C. F. *A Course in Modern Linguistics* (Macmillan College, 1960).
2. Crystal, D. *A Dictionary of Linguistics and Phonetics* (John Wiley & Sons, 2008).
3. Goldberg, A. E. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions* (Princeton Univ. Press, 2019).
4. Gunter, T. C., Stowe, L. A. & Mulder, G. When syntax meets semantics. *Psychophysiology* **34**, 660–676 (1997).
5. Hagoort, P. Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *J. Cogn. Neurosci.* **15**, 883–899 (2003).
6. Inkelas, S. *The Interplay of Morphology and Phonology* (Oxford Univ. Press, 2014).
7. Booij, G. in *Los Límites de la Morfología. Estudios Ofrecidos a Soledad Varela Ortega* (eds Fabrégas, A. et al.) 105–113 (Ed. Univ. Autónoma Madrid, 2012).
8. Friederici, A. D. The brain basis of language processing: from structure to function. *Physiol. Rev.* **91**, 1357–1392 (2011).
9. Jellinger, K. A. The heterogeneity of late-life depression and its pathobiology: a brain network dysfunction disorder. *J. Neural Transm.* **130**, 1057–1076 (2023).
10. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *Neuroimage* **30**, 1088–1096 (2006).
11. Gomez-Marin, A. & Ghazanfar, A. A. The life of behavior. *Neuron* **104**, 25–36 (2019).
12. Nastase, S. A., Goldstein, A. & Hasson, U. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* **222**, 117254 (2020).
13. Greenwald, A. G., Pratkanis, A. R., Leippe, M. R. & Baumgardner, M. H. Under what conditions does theory obstruct research progress? *Psychol. Rev.* **93**, 216–229 (1986).
14. Brewer, M. B. & Crano, W. D. in *Handbook of Research Methods in Social and Personality Psychology* (eds Reis, H. T. & Judd, C. M.) 3–16 (Cambridge Univ. Press, 2000).
15. Ibanez, A. The mind’s golden cage and cognition in the wild. *Trends Cogn. Sci.* **26**, 1031–1034 (2022).
16. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
17. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
18. OpenAI et al. GPT-4 technical report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2023).
19. Manyika, J., Hsiao, S., Manager, G. & Assistant, G. *An Overview of Bard: An Early Experiment With Generative AI* (Google, 2023).
20. Radford, A. et al. Robust speech recognition via large-scale weak supervision. Preprint at <https://doi.org/10.48550/arXiv.2212.04356> (2022).
21. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
22. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at <https://www.cs.ubc.ca/~amuhamed/LING530/papers/radford2018improving.pdf> (2018).
23. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
24. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
25. Humphreys, G. F., Lambon Ralph, M. A. & Simons, J. S. A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends Neurosci.* **44**, 452–463 (2021).
26. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
27. Li, Y. et al. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* **26**, 2213–2225 (2023).
28. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
29. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
30. Hickok, G., Houde, J. & Rong, F. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* **69**, 407–422 (2011).
31. Pickering, M. J. & Garrod, S. An integrated theory of language production and comprehension. *Behav. Brain Sci.* **36**, 329–347 (2013).
32. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
33. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
34. Gagnepain, P., Henson, R. N. & Davis, M. H. Temporal predictive codes for spoken words in auditory cortex. *Curr. Biol.* **22**, 615–621 (2012).
35. Obleser, J. & Kotz, S. A. Expectancy constraints in degraded speech modulate the language comprehension network. *Cereb. Cortex* **20**, 633–640 (2010).

36. Sohoglu, E., Peelle, J. E., Carlyon, R. P. & Davis, M. H. Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* **32**, 8443–8453 (2012).
37. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59 (2016).
38. Flinker, A. & Knight, R. T. Broca's area in comprehension and production, insights from intracranial studies in humans. *Curr. Opin. Behav. Sci.* **21**, 170–175 (2018).
39. Goldstein, A. et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).
40. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
41. Piantadosi, S. T. in *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett* (eds Gibson, E. & Poliak, M.) (Language Science Press, in the press).
42. Rajkumar, R. & White, M. Better surface realization through psycholinguistics. *Lang. Linguist. Compass* **8**, 428–448 (2014).
43. Nguyen, T. A. et al. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguist.* **11**, 250–266 (2023).
44. Schulman, J. et al. ImageBind: holistic AI learning across six modalities. *Meta AI Blog* <https://ai.meta.com/blog/imagebind-six-modalities-binding-ai/> (2023).
45. Beguš, G., Zhou, A., Wu, P. & Anumanchipalli, G. K. Articulation GAN: unsupervised modeling of articulatory learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5 (IEEE, 2023).
46. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878 (2008).
47. Sanchez, J. C., Gunduz, A., Carney, P. R. & Principe, J. C. Extraction and localization of mesoscopic motor control signals for human ECoG neuroprosthetics. *J. Neurosci. Methods* **167**, 63–81 (2008).
48. Honey, C. J., Thompson, C. R., Lerner, Y. & Hasson, U. Not lost in translation: neural responses shared across languages. *J. Neurosci.* **32**, 15277–15283 (2012).
49. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *J. Neurosci.* **29**, 13613–13620 (2009).
50. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
52. Carnegie Mellon Speech Group. *The CMU Pronouncing Dictionary*, Release 0.7a <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (Carnegie Mellon Univ., 1998).
53. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition* 3121–3124 (IEEE, 2010).

## Acknowledgements

We thank our funders: NIH NINDS R01NS109367, NIH NINDS R01NS115929, DP1HD091948 and R01DC022534.

## Author contributions

A.G. and U.H. conceptualized and designed the study, led the data analysis and wrote the manuscript. H.W., L.N., M.S. and T.S. analysed data. Z.Z., B.A. and H.G. preprocessed and analysed data. S.A.N. wrote the manuscript. A.S., A.R., G.C. and C.K. preprocessed and collected data. W.D. collected data. D.F. and P.D. provided clinical care. S.D. and O.D. assisted with patient recruitment and consent. A.H., M.B., Y.M. and A.F. conceptualized and designed the study.

## Competing interests

The authors declare no competing interests

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-025-02105-9>.

**Correspondence and requests for materials** should be addressed to Ariel Goldstein.

**Peer review information** *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

**Ariel Goldstein**  , **Haocheng Wang**  , **Leonard Niekerken**  , **Mariano Schain**  , **Zaid Zada**  , **Bobbi Aubrey**  , **Tom Sheffer**  , **Samuel A. Nastase**  , **Harshvardhan Gazula**  , **Aditi Singh**  , **Aditi Rao**  , **Gina Choe**  , **Catherine Kim**  , **Werner Doyle**  , **Daniel Friedman**  , **Sasha Devore**  , **Patricia Dugan**  , **Avinatan Hassidim**  , **Michael Brenner**  , **Yossi Matias**  , **Orrin Devinsky**  , **Adeen Flinker**   & **Uri Hasson**  

<sup>1</sup>Department of Cognitive and Brain Sciences and Business School, Hebrew University, Jerusalem, Israel. <sup>2</sup>Google Research, Mountain View, CA, USA.

<sup>3</sup>Department of Psychology and the Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. <sup>4</sup>Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands. <sup>5</sup>Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. <sup>6</sup>New York University School of Medicine, New York, NY, USA. <sup>7</sup>School of Engineering and Applied Science, Harvard University, Boston, MA, USA. <sup>8</sup>These authors contributed equally: Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A. Nastase.  e-mail: [ariel.y.goldstein@mail.huji.ac.il](mailto:ariel.y.goldstein@mail.huji.ac.il)

## 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	Recordings from grid, strip and depth electrode arrays were acquired using one of two amplifier types: NicoletOne C64 clinical amplifier (Natus Neurologics), band-pass filtered from 0.16–250Hz, and digitized at 512Hz; NeuroWorks Quantum Amplifier recorded at 2,048Hz, high-pass filtered at 0.01Hz and then resampled to 512Hz.
Data analysis	<p>Data were preprocessed using Matlab 2019b and The Fieldtrip toolbox (commit: 56769ab0s). The Python packages used: Scipy v1.11.4, Statsmodels v0.14.1, nltk 3.9.1. Codes for:</p> <p>For embedding extraction: <a href="https://github.com/hassonlab/247-pickling/tree/whisper-paper-1">https://github.com/hassonlab/247-pickling/tree/whisper-paper-1</a></p> <p>For encoding models: <a href="https://github.com/hassonlab/247-encoding/tree/whisper-paper-1">https://github.com/hassonlab/247-encoding/tree/whisper-paper-1</a></p> <p>For plotting: <a href="https://github.com/hassonlab/247-plotting/blob/main/scripts/tfpaper_whisper.ipyn">https://github.com/hassonlab/247-plotting/blob/main/scripts/tfpaper_whisper.ipyn</a></p>

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](#)

The current dataset contains sensitive patient-doctor conversations protected under HIPAA regulations. De-identifying such a large naturalistic dataset is a challenging task, and there is always a risk of researchers decoding sensitive information from neural signals. Therefore, we cannot make this dataset available to public data archives. We are developing a data-sharing protocol with selected research groups to promote collaboration.

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Four patients (2 females (gender assigned based on medical record); 24–48 years old)  
Sex or gender are not relevant variables to this study, and analysis was not done considering it.

Reporting on race, ethnicity, or other socially relevant groupings

Four patients (2 females (gender assigned based on medical record); 24–48 years old)

Population characteristics

All the patients suffer from drug immune epilepsy. We select patients with intact cognitive faculty as determined by neuropsychological tests that are done as part of the hospitalization process. Effects were not moderated by age.

Recruitment

Each participant provided informed consent following protocols approved by the New York University Grossman School of Medicine Institutional Review Board. Patients were informed that participation in the study was unrelated to their clinical care and that they could withdraw from the study without affecting their medical treatment. As the patients volunteer to the experiment there is a potential to self-selection based on this trait. However, the researcher could not think how it could impact the results.

Ethics oversight

The study was approved by the NYU Grossman School of Medicine Institutional Review Board (approved protocol s14-02101) which operates under NYU Langone Health Human Research Protections and Princeton University's Review Board (approval protocol 4962). Studies are performed in accordance with the Department of Health and Human Services policies and regulations at 45 CFR 46. Before obtaining consent, all participants were confirmed to have the cognitive capacity to provide informed consent by a clinical staff member. Participants provided oral and written informed consent before beginning study procedures. They were informed that participation was strictly voluntary, and would not impact their clinical care. Participants were informed that they were free to withdraw participation in the study at any time. All study procedures were conducted in accordance with the Declaration of Helsinki.

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](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

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

Study description

We used the recorded neural activity of four participants using dense intracranial arrays while they stayed in the hospital. This is a quantitative research as we correlate the brain signal recorded with the representation induced by deep ASR model (whisper)

Research sample

No statistical method was used to predetermine the sample size. We collected dozens of hours from each participants with dozens of thousands of words which are the unit for the analysis. We demonstrate that the results can replicate per-patient and for fractions of the sample size within patient (even as small as 25% from the obtained dataset).

Sampling strategy

The sample is 4 patients that gave their consent for the audio recording of their speech

Data collection	Recordings from grid, strip and depth electrode arrays were acquired using one of two amplifier types: NicoletOne C64 clinical amplifier (Natus Neurologics), band-pass filtered from 0.16–250Hz, and digitized at 512Hz; NeuroWorks Quantum Amplifier recorded at 2,048Hz, high-pass filtered at 0.01Hz and then resampled to 512Hz. The participants as well as the personal who interacted with the patients were not aware of the study hypothesis.
Timing	Each patient came to the hospital at 2019. The recording was continuous during their stay. The exact date of collection is an identifiable information and should not be shared: <a href="https://privacyruleandresearch.nih.gov/pr_08.asp">https://privacyruleandresearch.nih.gov/pr_08.asp</a>
Data exclusions	No data was excluded
Non-participation	No patients were removed.
Randomization	As analyses can be seen for each participant (separately and together) there is no between participant analysis nor allocation to conditions.

## 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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Plants

Seed stocks

NA

Novel plant genotypes

NA

Authentication

NA