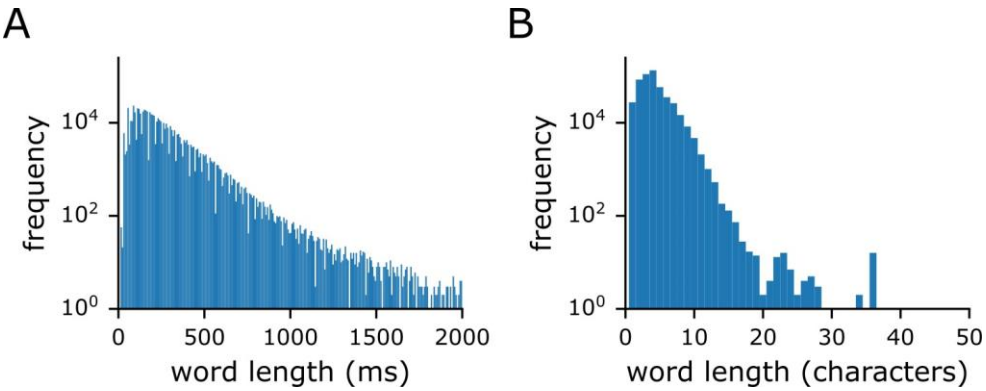
nature human behaviour

Supplementary information

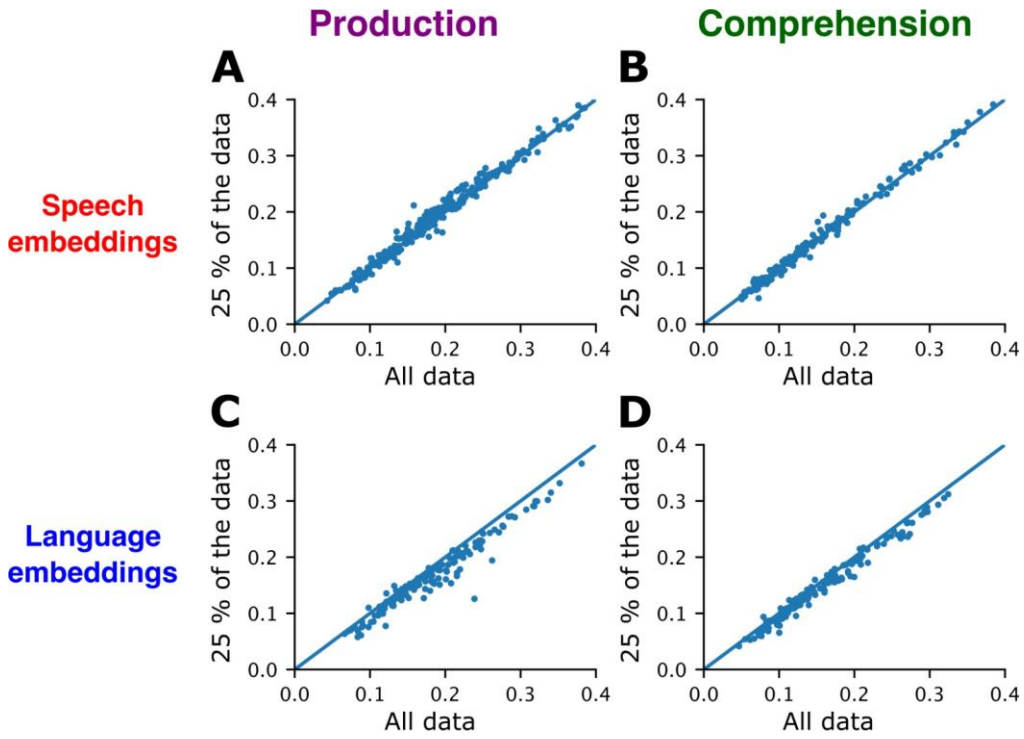
https://doi.org/10.1038/s41562-025-02105-9

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

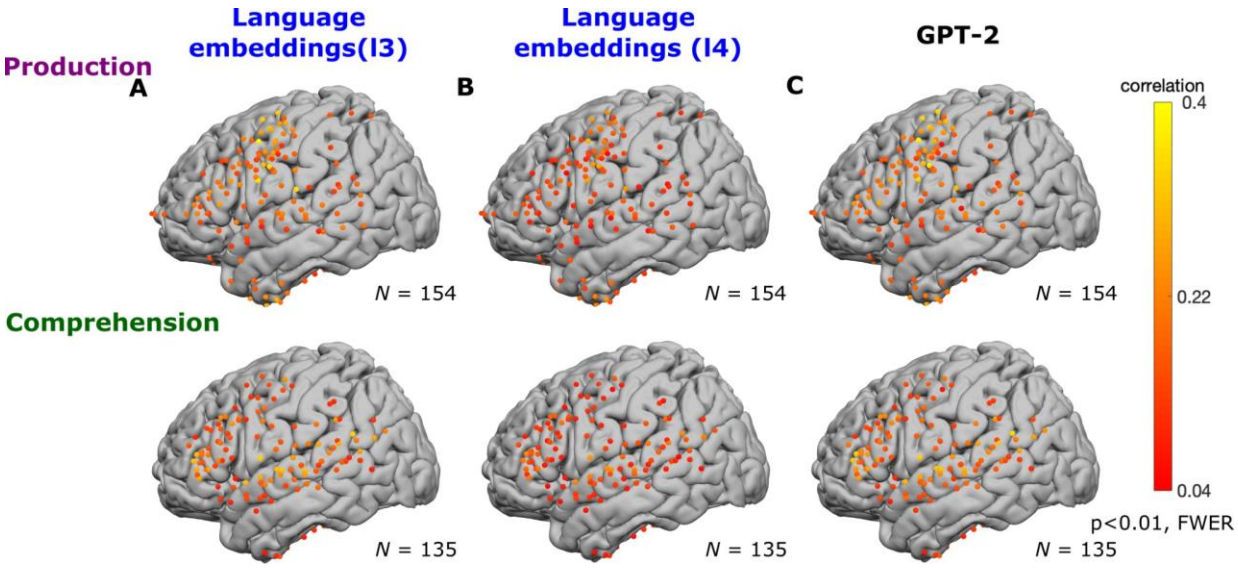
In the format provided by the authors and unedited



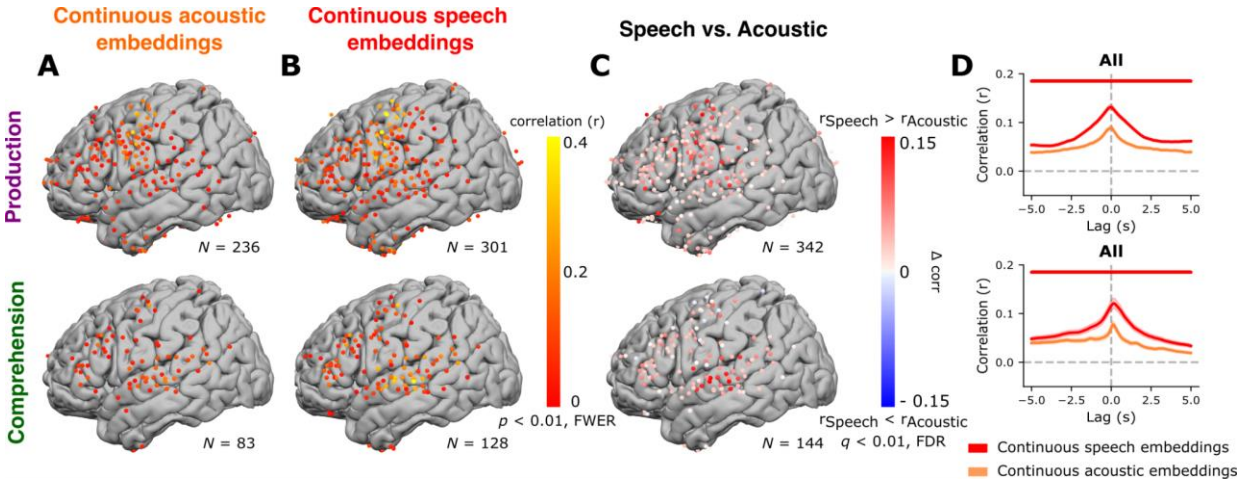
Supp. Figure 1. **Summary statistics of conversations (A)** Distribution of temporal word duration. **(B)** Distribution of number of characters in words.



Supp. Figure 2. Encoding performance for reduced sample size. Each point represents the encoding performance of a single electrode, with the highest correlation value across lags for the encoding model using complete data as the x-value and the highest correlation value across lags for the encoding model using 25% of the data as the y-value. The diagonal line indicates no deterioration of encoding performance when reducing the encoding data size to 25 % of the original sample size. We observe no systematic decrease in encoding performance for speech embeddings **(A, B)** and language embeddings **(C, D)** for speech production and comprehension.

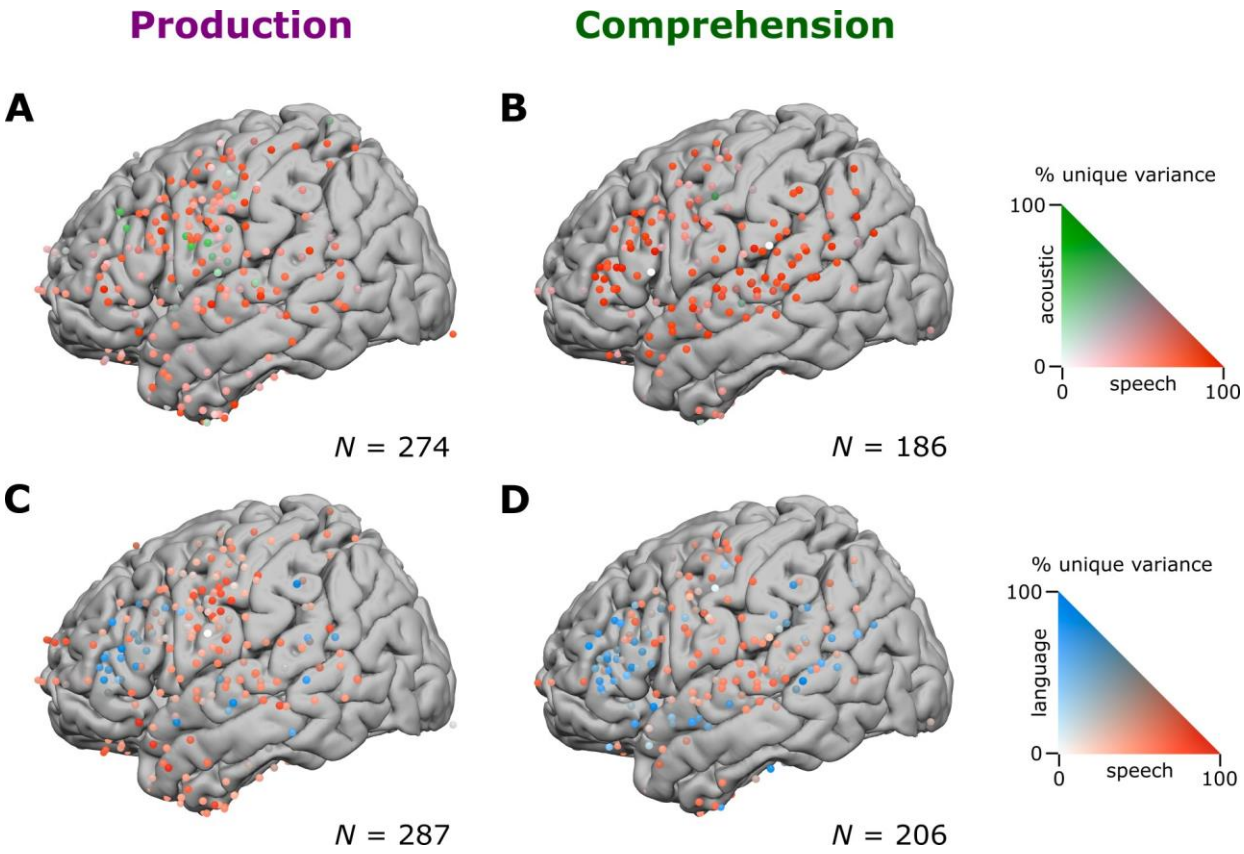


Supp. Figure 3. Comparing language embeddings across layers and models. We compared the encoding performance for language embeddings from Whisper decoder layers 3 (right) and 4 (center) to the performance of language embeddings from GPT-2 layer 8 (right). Encoding performance is shown for significant electrodes during speech production and comprehension, corrected for multiple comparisons. The color indicates the maximum correlation value across lags per electrode ($q < 0.01$, FWER).

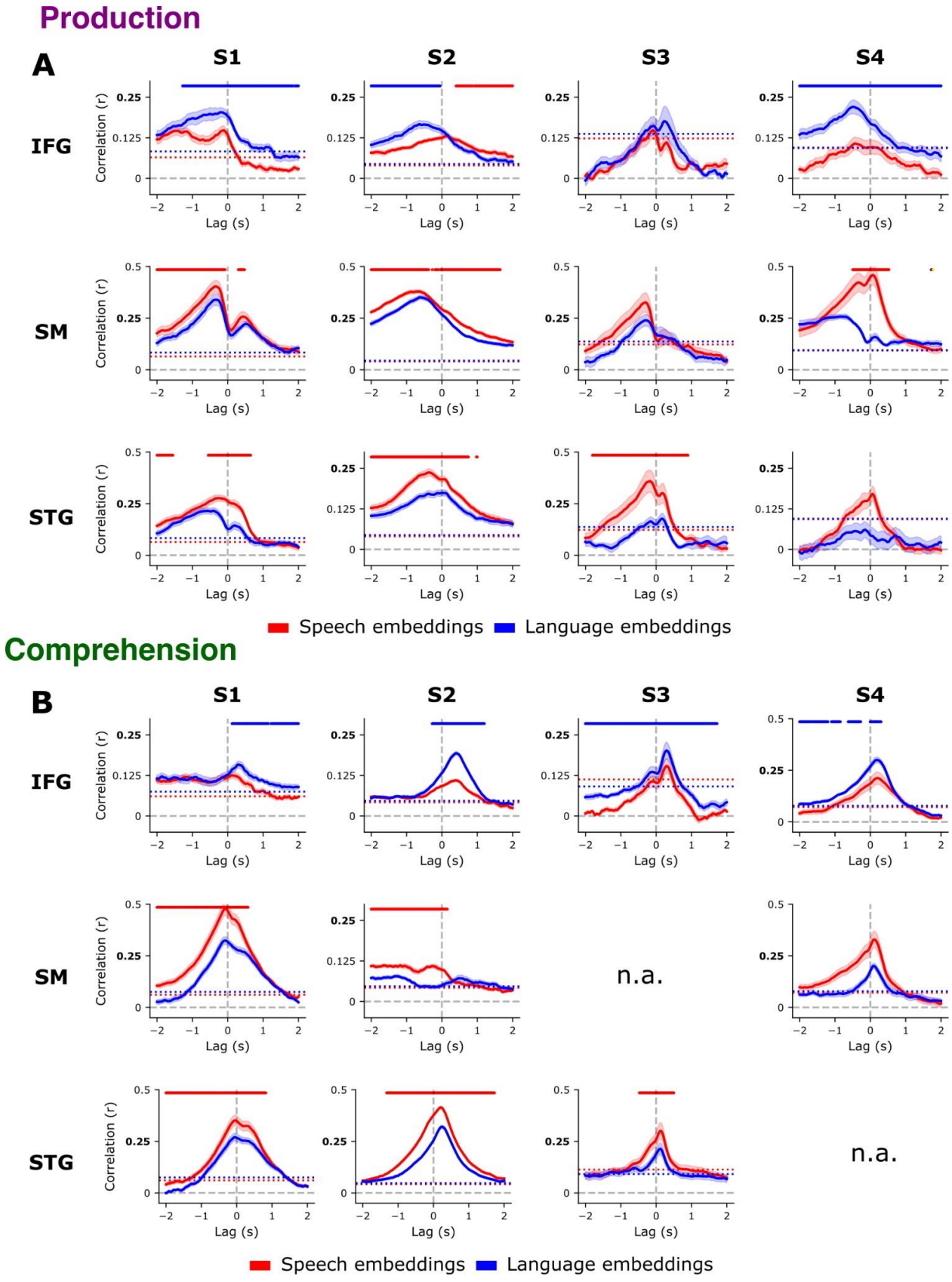


Supp. Figure 4. Continuous acoustic and speech encoding model performance during speech production and comprehension. (A) Electrode-wise encoding performance (correlation between predicted and actual neural activity) for continuous acoustic embeddings for comprehension and production ($p < 0.01$, FWER). The plots illustrate the correlation values associated with the encoding for each electrode, with the color indicating the highest correlation value across lags. **(B)** Electrode-wise encoding performance for continuous speech embeddings for comprehension and production ($p < 0.01$, FWER). **(C)** We computed the contrast between encoding performance for speech embeddings and acoustic embeddings, with electrodes colored in red showing better performance for speech embeddings and electrodes in blue showing better performance for acoustic embeddings ($q < 0.01$, FDR corrected). **(D)** We observe higher correlations at multiple time points for speech embeddings than acoustic embeddings. Data are presented as mean values across electrodes (*N* is

indicated in brackets). Red asterisks indicate lags with a significant difference ($q < 0.01$, FDR corrected) between continuous speech and acoustic embeddings. All analyses were done using permutation tests and used two-sided comparisons.

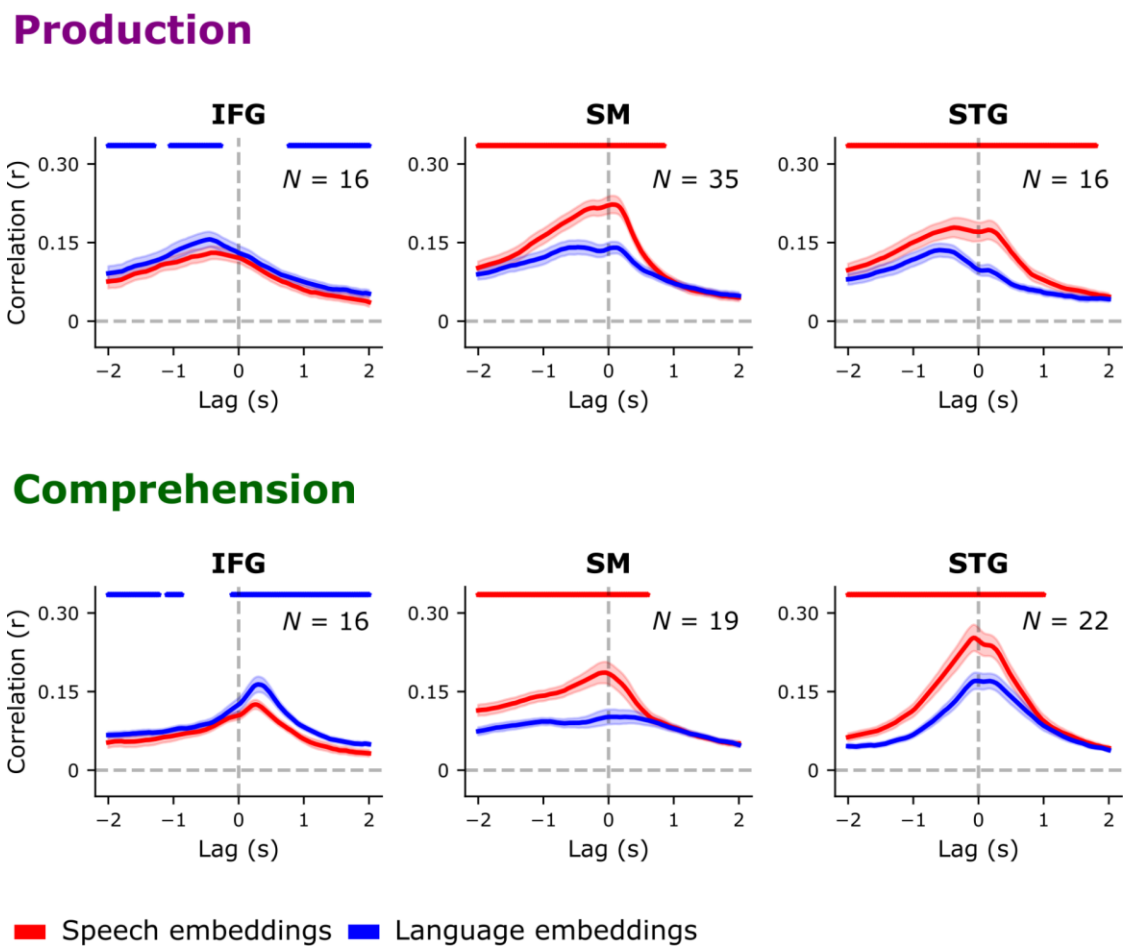


Supp. Figure 5. Unique variance explained by acoustic, speech, and language embeddings. We used a variance partitioning approach to identify the proportion of the variance uniquely explained by different embeddings (see Methods for details). We fitted separate encoding models for acoustic and speech embeddings and a joint encoding model by concatenating acoustic and speech embeddings. Next, we calculated the proportion of the variance uniquely explained by either acoustic or speech embeddings relative to the variance explained by the joint encoding model. Across all electrodes, speech embeddings explain more variance than acoustic embedding for speech production. **(A)** Speech embeddings (red) predict the neural activity during production, and the variance explained by acoustic embeddings is contained in the variance explained by speech embedding in most electrodes. **(B)** Speech embeddings (red) also predict neural activity during comprehension, and the variance explained by acoustic embeddings is contained in the variance explained by speech embedding in most of the electrodes. **(C)** The variance partitioning approach during production revealed a mixed selectivity for speech and language embeddings. Language embeddings (blue) better explain IFG, while speech embeddings (red) better explain STG and SM. **(D)**. Variance partitioning approach during comprehension also revealed a mixed selectivity for speech (red) and language (blue) embeddings. Language embeddings better explain IFG and AG, while speech embeddings better explain STG and SM.

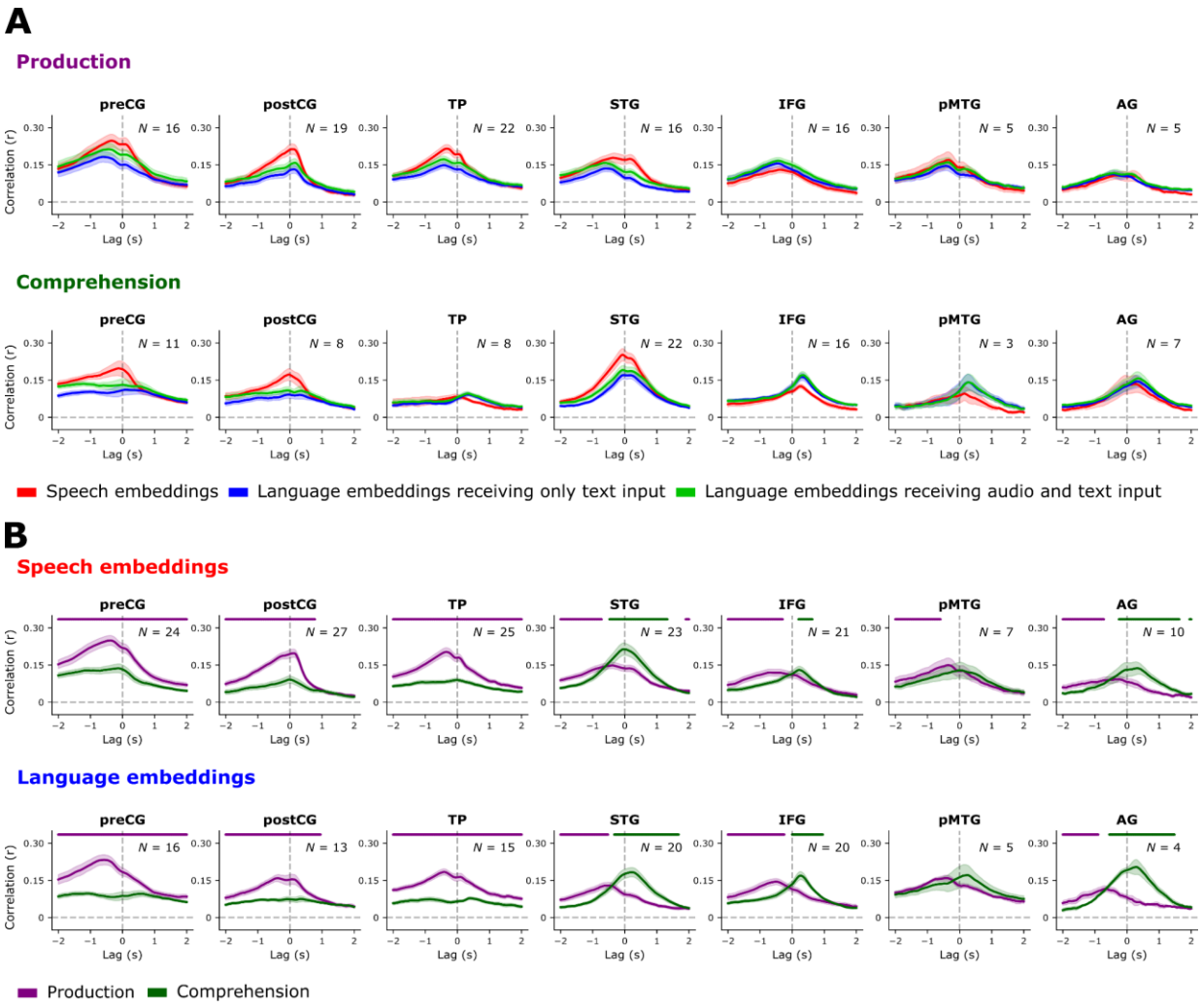


Supp. Figure 6. Mixed selectivity for speech and language embeddings during speech production and comprehension. (A) Single electrodes are encoded based on speech embeddings (red) versus language embeddings (blue) during speech production across different brain areas and subjects (S1-S4). Models were estimated separately for each lag (relative to word onset at 0 s) and

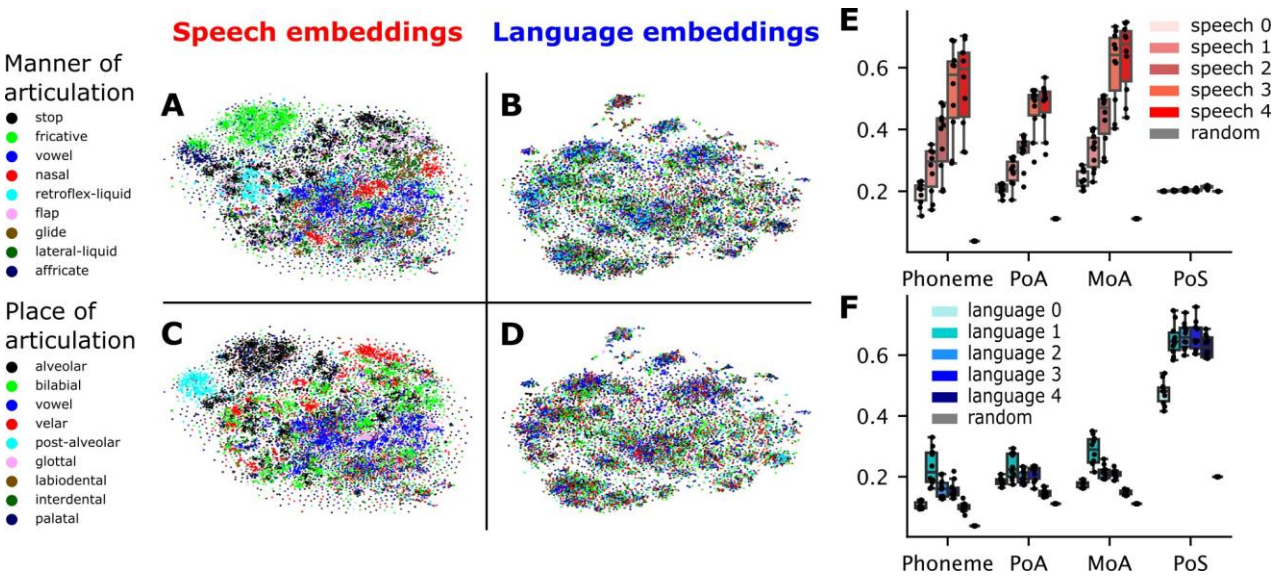
evaluated by computing the correlation between predicted and actual neural activity. Data are presented as mean values across folds (N = 10) +/- standard error. Red and blue asterisks indicate a significant difference between encoding based on speech and language embeddings (q < 0.01, FDR corrected). The dashed horizontal line indicates the statistical threshold (p < 0.01, FWER). Regions where we lack coverage are marked n.a for a subject. 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 superior temporal gyrus (STG), precentral gyrus (preCG), and postcentral gyrus (postCG). The language encoding model yielded significant accuracies (correlations up to 0.25) and outperformed the speech model in IFG and AG. **(B)** Encoding single electrodes based on speech embeddings (red) versus language embeddings (blue) during speech comprehension across different brain areas and patients. 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. Cases where data is missing in a specific ROI from a specific subject, are marked with an 'n.a'.



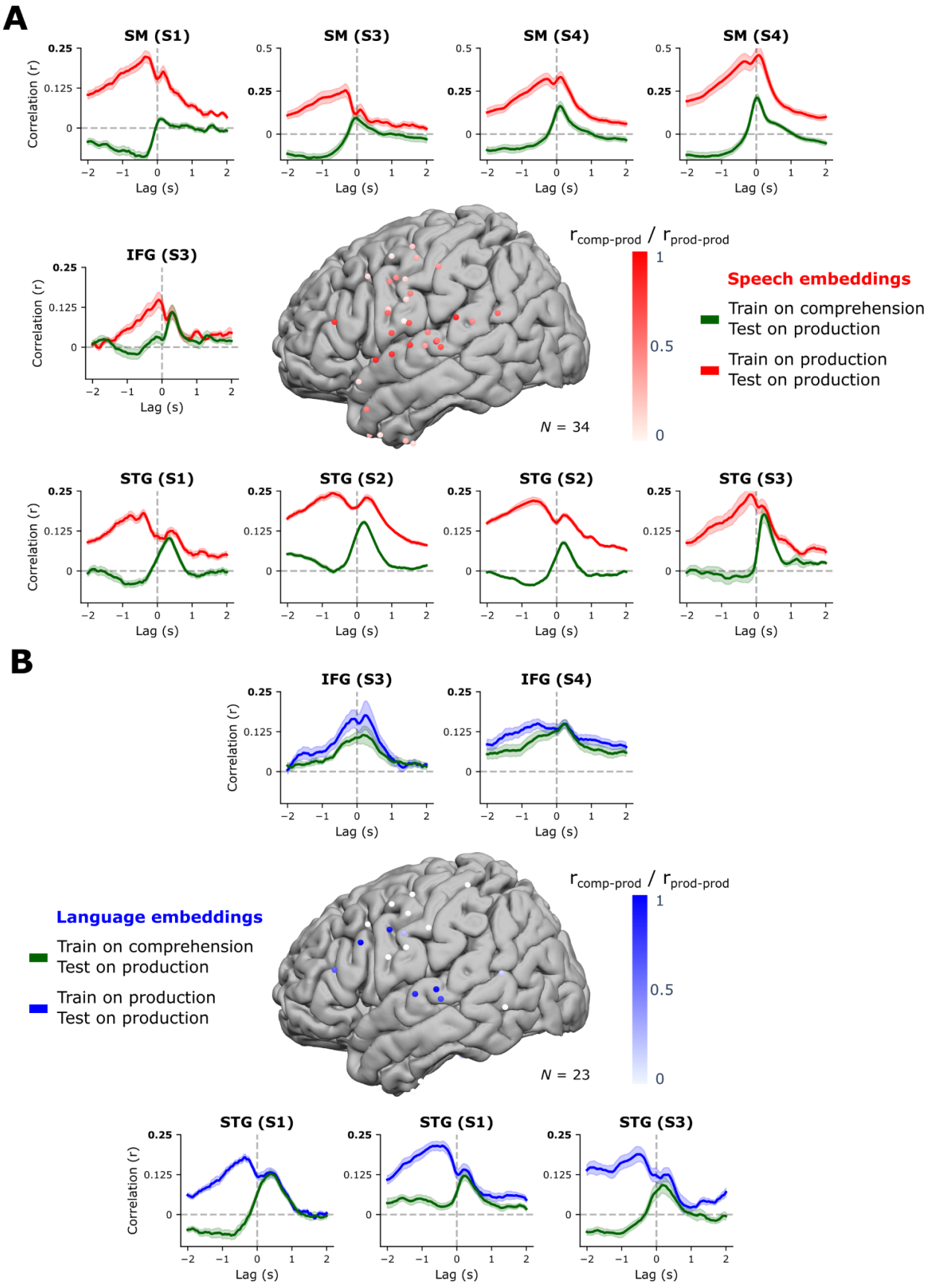
Supp. Figure 7. **Average speech and language encoding across ROIs.** Electrode-wise encoding performance values for speech (red) and language (blue) embeddings for speech production **(top row)** and comprehension **(bottom row)**. For each ROI, encoding performance was averaged across electrodes that showed a significant difference in maximum encoding performance (across lags) between speech and language embeddings (number of electrodes in parentheses) ($p < 0.01$, FWER). Data are presented as mean values across electrodes (N is indicated in brackets) +/- standard error. Blue and red asterisks indicate a lag-wise significant difference in ROI-wise encoding performance between speech and language embeddings ($q < 0.01$, FDR corrected). All analyses were done using permutation tests and used two-sided comparisons.



Supp. Figure 8. Comparing speech and language based encoding for comprehension and production. (A) Average ROI-level encoding for speech and language embeddings. Comparing electrode-wise encoding performance averaged per ROI using speech embeddings (red), language embeddings receiving only text input (blue), and language embeddings receiving text and audio input (green). **(B) Average ROI-level encoding for speech production and comprehension.** Electrode-wise encoding performance values were averaged per ROI for production (purple) and comprehension (green). For each ROI, encoding performance was averaged across electrodes, which showed a significant difference in maximum encoding performance (across lags) between production and comprehension (number of electrodes in parentheses). Data are presented as mean values across electrodes (*N* is indicated in brackets) +/- standard error. Purple and green asterisks indicate a lag-wise significant difference in ROI-wise encoding performance between production and comprehension ($q < 0.01$, FDR corrected). All analyses were done using permutation tests and used two-sided comparisons.



Supp. Figure 9. Representations of phonetic and lexical information in Whisper. (A-D) Visualization of speech embeddings and language embeddings in a two-dimensional space estimated using t-SNE. Each data point corresponds to the embedding for either an audio segment (speech model) or a word token (language model) for a unique word (averaged across all instances of one word). Clustering according to phonetic categories (manner of articulation, MoA; and place of articulation, PoA) is visible in speech embeddings (A, C) but not in language embeddings (B, D). (E, F) Classification of phonetic information (phoneme, MoA, PoA) and lexical information (PoS) based on embeddings taken from the different layers of the whisper encoder and decoder network. The last layer of speech embeddings (speech 4) shows the best classification accuracy for phonetic information and relatively low classification accuracy for lexical information (E). The third layer of language embeddings (language 3) shows the best classification accuracy for lexical information and relatively low classification accuracy for phonetic information (F).



Supp. Figure 10. Evidence for speech processing of the speaker's own voice during speech production. (A) We trained an encoding model using speech embeddings during speech

comprehension and then applied this model's weights to predict neural activity during speech production (green line). We compare the model performance with our original encoding model, which was trained and tested on speech production (red line). The speech comprehension model (green line) successfully predicts neural activity after word onset during speech production. Data are presented as mean values across folds (*N* = 10) +/- standard error. The brain map shows electrodes with a double peak during speech production (at least one peak before and after word onset). Electrodes are colored by the maximal encoding performance of the comprehension model (green line) relative to the maximal encoding performance of the production model (red line). We observe a high relative encoding performance in STG, SM, and IFG. **(B)** We trained an encoding model using language embeddings during speech comprehension and then applied this model's weights to predict neural activity during speech production (green line). We compare the model performance with our original encoding model, which was trained and tested on speech production (blue line). Data are presented as mean values across folds (*N* = 10) +/- standard error. Similarly, the language comprehension model (green line) successfully predicts neural activity after word onset during speech production.

In our investigation of sensorimotor (SM) areas, we observed distinct neural response patterns following speech onset. Notably, these responses were accurately predicted only by a model trained during production, and notably, lower correlations were induced by models trained for comprehension. This divergence suggests a unique neural representation of articulatory and speech features in the SM areas during speech comprehension and production. This differentiation raises essential questions about the underlying mechanisms in these regions—specifically, whether the observed responses are primarily driven by auditory feedback processes or by the execution of speech itself. These findings underscore the complexity of neural processing in speech-related tasks and highlight the necessity for further research to disentangle articulation-related processes from speech-comprehension-related processes. We noticed a pre-onset negative correlation in certain SM electrodes using the comprehension-based encoding model (green, Fig. S10). This may be linked to a documented suppression of speech areas resulting from an efferent copy of the speech articulation plans (Whitford et al., eLife, 2017). However, further research is required to test this hypothesis.

Participant	**1**	**2**	**3**	**4**
Age	53	26	48	24
Sex	F	M	F	M
Number of electrodes implanted	104	125	255	192
Hours of speech recorded	17	37	17	29
Number of words	79654	213473	117800	109282

recorded				
Number of comprehension words	47642	109967	71754	60608
Number of production words	32012	103506	46046	48674
Neuropsychological testing scores	VCI: 105 POI: 107 PSI: 120 WMI: 102	VCI: 89 POI: 75 PSI: 65 WMI: 74	VCI: 107 POI: 104 PSI: 111 WMI: 114	VCI: 145 POI: 96 PSI: 86 WMI: 95
Pathology/epilepsy type/seizure focus	posterior temporal lobe (neocortical) epilepsy. IEEG results were consistent with a seizure focus adjacent to or overlying the known posterior temporal cortical lesion	left anteromedial temporal lobe epilepsy	Right anteromedial temporal lobe epilepsy. ICEEG localized ictal onsets to the right temporal pole and right hippocampus	Focal epilepsy arising from the left hemisphere with a broad focus involving the left temporal neocortex (superior, middle, inferior temporal gyri) left frontal operculum, inferior postcentral gyrus, insula
Implant	Left grid, strips, and depth	Left grid, strips, and depth	Bilateral strips and depths, and a left grid	Left grid, strips, and depth

Supp. Table 1. Patient demographics and clinical characteristics.

Part of Speech	Prod Frequency	Comp Frequency
NOUN	64835	80018
VERB	53132	68717
PRON	32754	41426

ADP	23492	26079
ADV	19759	25711
DET	16884	22253
ADJ	16833	21799
CONJ	7698	8813
PRT	5799	8053
NUM	2425	2865
X	721	615
.	25	16

Supp. Table 2. Distribution of part of speech for all words in our dataset

Symbolic Speech Features	**Feature Categories / Dimensions**
Phonemes	39
Place of Articulation	9
Manner of Articulation	9
Voice or Voiceless	3
Sum	60
Symbolic Linguistic Features	**Feature Categories / Dimensions**
Part of Speech	17
Dependency	45
Prefix	30
Suffix	44

Stop Word	1
Sum	137

Supp. Table 3. Symbolic speech and linguistic features. Each feature is turned into one-hot embeddings with the dimension of the number of categories in the feature.