nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00714-5

Decoding speech perception from non-invasive brain recordings

Received: 20 September 2022

Accepted: 4 August 2023

Published online: 5 October 2023

Check for updates

Alexandre Défossez[1], Charlotte Caucheteux[1,2], Jérémy Rapin[1], Ori Kabeli[3] & Jean-Rémi King[1,4]

Decoding speech from brain activity is a long-awaited goal in both healthcare and neuroscience. Invasive devices have recently led to major milestones in this regard: deep-learning algorithms trained on intracranial recordings can now start to decode elementary linguistic features such as letters, words and audio-spectrograms. However, extending this approach to natural speech and non-invasive brain recordings remains a major challenge. Here we introduce a model trained with contrastive learning to decode self-supervised representations of perceived speech from the non-invasive recordings of a large cohort of healthy individuals. To evaluate this approach, we curate and integrate four public datasets, encompassing 175 volunteers recorded with magneto-encephalography or electro-encephalography while they listened to short stories and isolated sentences. The results show that our model can identify, from 3 seconds of magneto-encephalography signals, the corresponding speech segment with up to 41% accuracy out of more than 1,000 distinct possibilities on average across participants, and with up to 80% in the best participants—a performance that allows the decoding of words and phrases absent from the training set. The comparison of our model with a variety of baselines highlights the importance of a contrastive objective, pretrained representations of speech and a common convolutional architecture simultaneously trained across multiple participants. Finally, the analysis of the decoder's predictions suggests that they primarily depend on lexical and contextual semantic representations. Overall, this effective decoding of perceived speech from non-invasive recordings delineates a promising path to decode language from brain activity, without putting patients at risk of brain surgery.

Every year, traumatic brain injuries, strokes and neurodegenerative diseases cause thousands of patients lose their ability to speak or even communicate[1–6]. Brain–computer interfaces (BCIs) have raised high expectations for the detection[4,5,7,8] and restoration of communication abilities in such patients[9–14]. Over recent decades, several teams have used BCIs to efficiently decode phonemes, speech sounds[15,16], hand gestures[11,12] and articulatory movements[13,17] from electrodes implanted in the cortex or over its surface. For instance, Willett et al. [12] decoded 90 characters per minute (with a 94% accuracy, that is, roughly 15–18 words per minute) from a patient with a spinal-cord injury, recorded in the motor cortex during 10 hours of writing sessions. Similarly, Moses et al. [13] decoded 15.2 words per minute (with 74.4% accuracy, and using a vocabulary of 50 words) in a patient with anarthria and a BCI implanted in the sensori-motor cortex, recorded over 48 sessions

[1]Meta AI, Paris, France. [2]Inria Saclay, Saclay, France. [3]Meta AI, Tel Aviv, Israel. [4]LSP, Département d'Etudes Cognitives, École Normale Supérieure, PSL University, Paris, France. e-mail: defossez@meta.com; jeanremi@meta.com

spanning over 22 hours. Finally, Metzger et al.[18] recently showed that a patient with severe limb and vocal-tract paralysis and a BCI implanted in the sensori-motor cortex could efficiently spell words using a code word that represented each English letter (for example, 'alpha' for 'a'): this approach leads to a character error rate of 6.13% and a speed of 29.4 characters per minute, and hence starts to provide a viable communication channel for such patients.

However, such invasive recordings face a major practical challenge: these high-quality signals require brain surgery and can be difficult to maintain chronically. Several laboratories have thus focused on decoding language from non-invasive recordings of brain activity such as magneto-encephalography (MEG) and electro-encephalography (EEG). MEG and EEG are sensitive to macroscopic changes of electric and magnetic signals elicited in the cortex, and can be acquired with a safe and potentially wearable set-up[19]. However, these two devices produce notoriously noisy signals that vary greatly across sessions and across individuals[20–22]. It is thus common to engineer pipelines that output hand-crafted features, which, in turn, can be learned by a decoder trained on a single participant[23–28].

In sum, decoding language from brain activity is, so far, limited either to invasive recordings or to impractical tasks. Interestingly, both of these approaches tend to follow a similar method: that is, (1) training a model on a single participant and (2) aiming to decode a limited set of interpretable features (Mel spectrogram, letters, phonemes, small set of words).

Instead, here we propose to decode speech from non-invasive brain recordings by using (1) a single architecture trained across a large cohort of participants and (2) deep representations of speech learned with self-supervised learning on a large quantity of speech data. We focus the present work on speech perception in healthy volunteers rather than speech production in patients to design a deep-learning architecture that effectively addresses two core challenges: (1) the fact that non-invasive brain recording can be extremely noisy and variable across trials and across participants and (2) the fact that the nature and format of language representations in the brain remain largely unknown. For this, we introduce a 'brain module' and train it with contrastive learning to align its representations to those of a pretrained 'speech module', namely, wav2vec 2.0 (ref. 29) (Fig. 1). We train a single model for all participants, sharing most of the weights except for one participant-specific layer. Figure 1 provides a broad overview of our approach.

To validate our approach, we curate and integrate four public MEG and EEG datasets, encompassing the brain activity of 175 participants passively listening to sentences of short stories (see Table 1 for details). For each MEG and EEG recording, we evaluate our model on its ability to accurately identify the corresponding audio segment from a large set of more than 1,500 segments (that is, 'zero shot' decoding).

This study provides three main contributions for the development of a non-invasive method to decode speech from brain activity. First, it shows how pretrained speech models can leverage the decoding of speech in the brain, without exposing volunteers to a tedious repetition of every single word targeted by the decoder. Second, it shows how specific design choices—including contrastive learning and our multi-participant architecture—improve the processing of continuous EEG and MEG recordings. Finally, our results suggest that the speech decoder is primarily based on high-level and semantic representations of speech.

Results

Accurately decoding speech from MEG and EEG recordings

Our model predicts the correct segment, out of more than 1,000 possibilities, with a top-10 accuracy up to 70.7% on average across MEG participants (Table 2, top-1 accuracy up to 41.3%, and Extended Data Fig. 1). For more than half of the samples, the true audio segment is ranked first or second in the decoders' predictions. Interestingly, these performances can reach high top-1 accuracy in the best performing participants: for example, above 80% top-1 accuracy in the best participant of the Gwilliams 2022 dataset[30] (Fig. 2a). For comparison, a model that predicts a uniform distribution over the vocabulary ('random model') only achieves less than 1% top-10 accuracy on the same MEG datasets. Decoding performance for EEG datasets is substantially lower: our model reaches 17.7% and 25.7% top-10 accuracy for the two EEG datasets currently analysed. While modest, these scores are much higher than the random baseline.

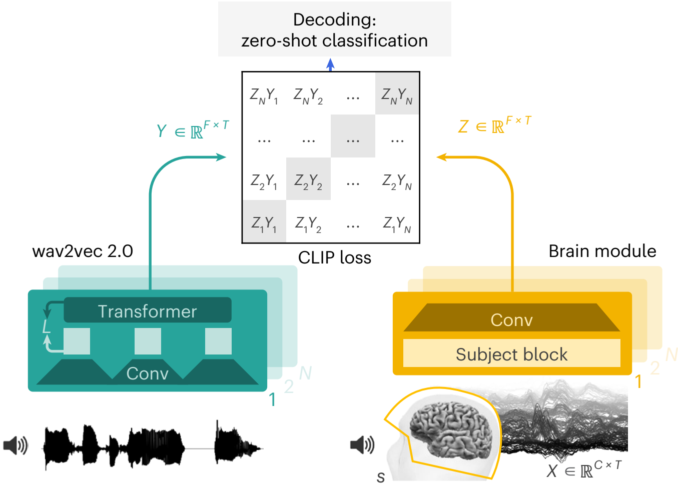


Fig. 1 | Model approach. We aim to decode speech from the brain activity of healthy participants recorded with MEG or EEG while they listen to stories and/or sentences. For this, our model extracts the deep contextual representations of 3 s speech signals (*Y* of *F* feature by *T* time samples) from a pretrained 'speech module' (wav2vec 2.0: ref. 29) and learns the representations (*Z*) of the brain activity on the corresponding 3 s window (*X* of *C* recording channels by *T* time samples) that maximally align with these speech representations with a contrastive loss (CLIP: ref. 44). The representation *Z* is given by a deep convolutional network. At evaluation, we input the model with left-out sentences and compute the probability of each 3 s speech segment given each brain representation. The resulting decoding can thus be 'zero shot' in that the audio snippets predicted by the model need not be present in the training set. This approach is thus more general than standard classification approaches where the decoder can only predict the categories learnt during training.

Is MEG really much better than EEG?

To investigate whether these performances depend on the total recording duration and/or the number of recording sensors, we train our model on a subset of the data that homogenizes recording time, the number of sensors and the number of participants. For this, we discard the dataset of Brennan and Hale[31], to avoid over-limiting the analysis dataset. Consequently, we match all datasets to the smallest number of channels of the three remaining datasets by keeping a random but fixed subset of channels (for example, 128). We keep only 19 participants per dataset, again aligning on the smallest for all three datasets. Finally, we keep the same average duration per participant for all three datasets, by dropping out some training segments (that is, the same segments are dropped for all participants or repetitions within one participant). All test segments are kept to maximize reliability. Overall, this subsampling diminishes decoding performance (for example, top 10: 30.3% for the Schoffelen dataset[32] and 31.7% for the Gwilliams dataset[33]), but MEG decoding remains much better than the EEG (Mann–Whitney across MEG and EEG participants: all $P < 10^{-6}$). Although these results should be confirmed by presenting the same stimuli to participants recorded with both EEG and MEG, they suggest that the difference in decoding performance observed between studies is mainly driven by the type of device.

Table 1 | Datasets

						Training set		Test set		
Dataset	**Language**	**Type**	**Sensors**	**Participants**	**Duration**	**Segments**	**Vocabulary**	**Segments**	**Vocabulary**	**Word overlap (%)**
Broderick 2019	English	EEG	128	19	19.2 h	2,645	1,418	1,842	764	67
Brennan and Hale 2019	English	EEG	60	33	6.7 h	1,211	513	190	148	60
Schoffelen 2019	Dutch	MEG	273	96	80.9 h	5,497	1,754	1,270	745	85
Gwilliams 2022	English	MEG	208	27	56.2 h	4,417	1,810	1,363	846	64

We study four datasets, two using MEG signals and two using EEG signals. We name each dataset by its author and year. For each dataset, we report the number of channels, the number of participants and the total duration in hours. Furthermore, we report the number of unique 3 s segments of words and vocabulary size over the training and test sets. 'Word overlap' indicates the percentage of the lexicon in the test set that is also present in the training set. We also have a validation set, roughly half the size of the test set used for early stopping. We define the training, validation and test split such that the same sentence for different participants is always in the same split.

'Speech module' evaluation

To evaluate our approach, we compare these decoding performances to those obtained with models that target different representations of speech (Table 2). While a model trained to predict the Mel spectrogram with a regression objective ('Base model' in Table 2) is systematically higher than chance, the use of a contrastive loss ('+ Contrastive') leads to decoding gains that range from 2% (for the Brennan and Hale dataset[31]) to 42.7% (for the Gwilliams dataset[33]). This gain is further supplemented by targeting the latent representations of the Mel spectrogram ('+ Deep Mel'). The latent representations of speech sounds, however, appear to be best identified with a pretrained speech module, that is, by using wav2vec 2.0, a model pretrained with self-supervised learning on speech sounds only, rather than by jointly learning speech and MEG and EEG representations (Table 2). Overall, these results show the importance, for decoding, of targeting deep representations of speech.

Table 2 | Results

Model	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Random model	5.3±0.1	0.5±0.1	0.7±0.1	0.8±0.1
Base model	6.0±0.9	1.0±0.3	12.4±1.2	20.6±1.8
+ Contrastive	8.0±4.8	9.7±1.0	55.1±0.7	55.1±0.9
+ Deep Mel	24.7±3.2	15.4±1.6	64.4±0.8	61.2±0.6
+ wav2vec 2.0	**25.7**±2.9	**17.7**±0.6	**70.7**±0.1	**67.5**±0.4

Top-10 segment-level accuracy (%) for a random baseline model that predicts a uniform distribution over the segments ('random'), a convolutional network trained to predict the Mel spectrograms with a regression loss ('base'), the same model trained with a contrastive CLIP loss ('+ Contrastive') and our model, which is trained to predict the features of wav2vec 2.0 with a contrastive loss ('+ wav2vec 2.0'). We also report the performance obtained with training, from scratch, a deep learning based speech representation using a contrastive loss ('+ Deep Mel'). Values are mean ±s.d. across three random initializations of the model's weights. The best accuracy across methods is indicated in bold.

'Brain module' evaluation

To evaluate the elements of the brain module, we performed a series of ablation experiments, and trained the corresponding models on the same data (Extended Data Fig. 2). Overall, these ablations show that several elements impact performance: performance systematically decreases when removing skip connections, the spatial attention module, and the initial or final convolutional layers of the brain module. These results also show the importance of clamping the MEG and EEG signals. Finally, additional experiments show that the present end-to-end architecture is robust to MEG and EEG artefacts, and requires little preprocessing of the MEG and EEG signals (Supplementary Sections A.3 and A.4).

Impact of the number of participants

To test whether our model effectively leverages the inter-individual variability, we trained it on a variable number of participants and computed its accuracy on the first 10% of participants. As shown in Fig. 2c, decoding performance steadily increases as the model is trained with more participants on the two MEG datasets. This result shows that our model effectively learns neural representations that are common across participants, while also accommodating participant-specific representations through the participant layer described in Methods.

Decoded representations best correlate with phrase embeddings

What type of representation does our model use to decode speech from brain signals? This interpretability question is notoriously difficult to address[22,34]. Figure 3 illustrates this issue: it shows the probability of each word given the MEG data of five representative participants listening to the phrase 'Thank you for coming, Ed'. Extended Data Fig. 3 shows additional predictions for five representative segments of the Gwilliams dataset [33]. In both cases, it can be difficult to judge whether the decoder's error tends to be related to the phonology or to the semantics of the actual sentence.

To address this issue, we analyse the single-word and single-segment predictions of our model with a linear model.

Specifically, we train a linear regression to predict the softmax probability of the true word estimated by the decoder, given different set of features, ranging from low-level representations (for example, phonemes) to high-level representations (for example, phrase embedding; see Methods for details). The results, shown in Fig. 4, show that the part-of-speech ($P < 0.004$), word embedding ($P < 10^{-8}$), bag-of-words embedding ($P < 10^{-23}$) and phrase embedding ($P < 10^{-23}$) significantly predict the single-trial decoding predictions. Overall, the higher level the representation, the more it accounts for the decoder's predictions. Given that phrase embeddings are known to capture semantic and syntactic representations[35–37], these results suggest that our decoder primarily relies on high-level and semantic representations of speech.

Methods

We formalize the general task of neural decoding and then describe and motivate the different components of our model, before describing the datasets, preprocessing, training and evaluation.

Problem formalization

We aim to decode speech from a time series of high-dimensional brain signals recorded with non-invasive MEG or EEG while healthy volunteers passively listened to spoken sentences in their native language. How spoken words are represented in the brain is largely unknown[37–39]. Thus, it is common to train decoders in a supervised manner to predict a latent representation of speech known to be relevant to the brain[16,34,40–42]. For example, the Mel spectrogram is often targeted for neural decoding because it represents sounds similar to the cochlea[43].

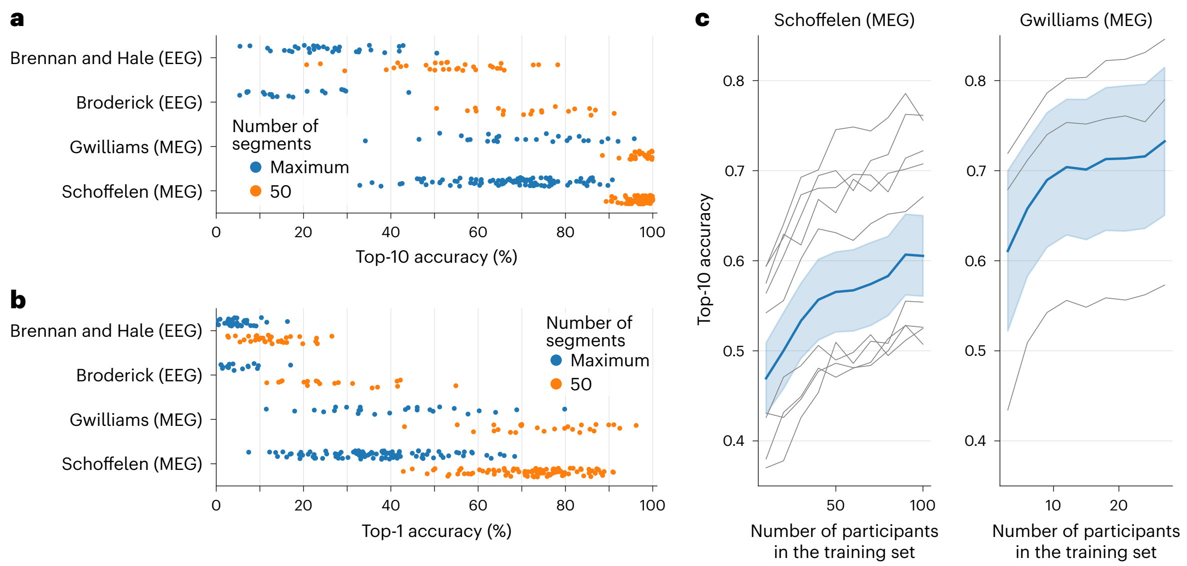


Fig. 2 | Decoding accuracy across subjects and datasets. a, Each dot represents the top-10 accuracy of a single participant, as estimated either with the full test set (blue) or with 50 possible segments (orange). **b**, The same as in **a**, but for top-1 accuracy. **c**, Top-10 accuracy as a function of the number of participants in the training set (blue line) as evaluated on the first 10% of the participants. The error bars indicate the s.e.m. across participants (grey lines).

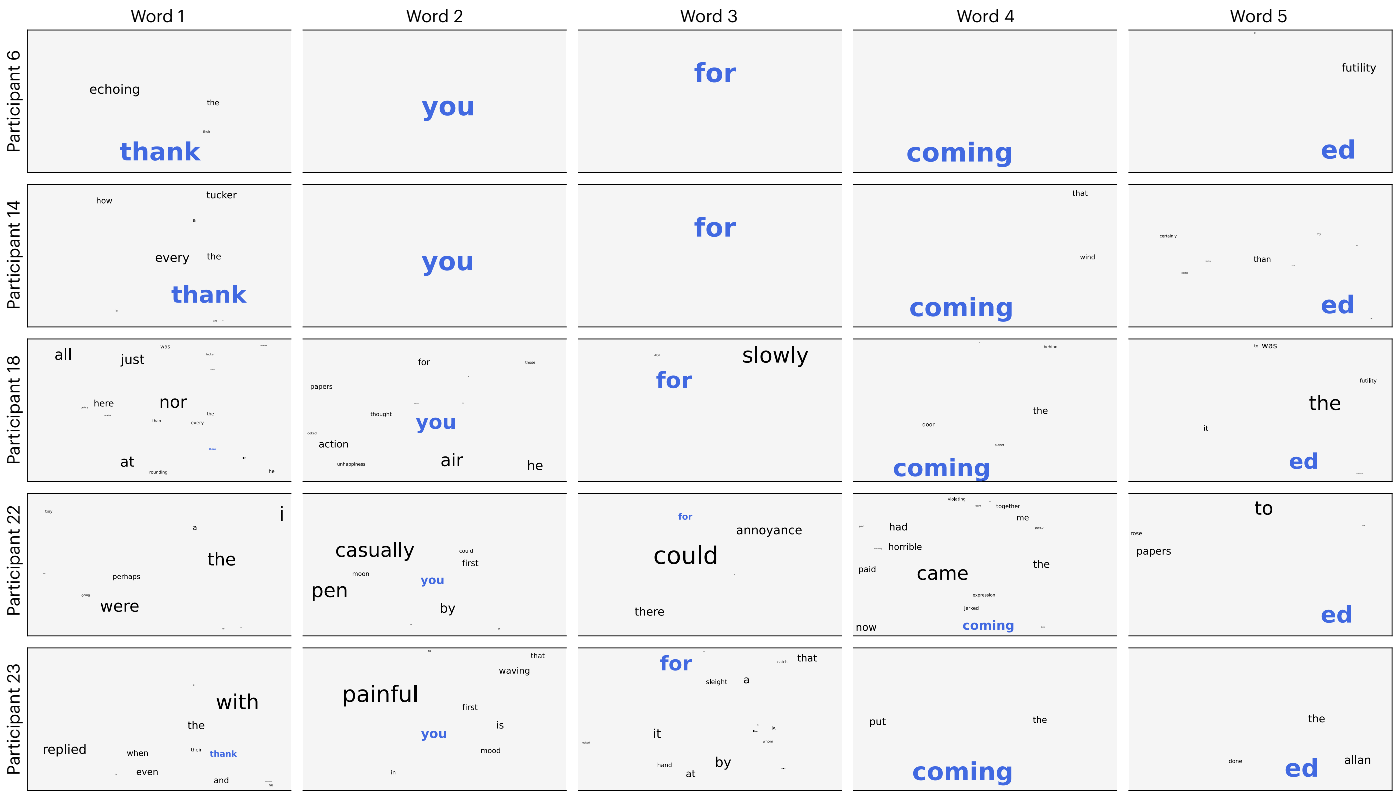


Fig. 3 | Word-level predictions. Word-level predictions for five representative participants (between the 20% (top) and the 80% percentiles (bottom) of the cohort) of the Gwilliams dataset[33] while they listened to the sentence 'Thank you for coming, Ed'. Blue words correspond to the correct word and black words correspond to negative candidates. Text size is proportional to the log-probability output by our model.

We formalize this problem as follows. Let $X \in \mathbb{R}^{C \times T}$ be a segment of a brain recording of a given participant while she listens to a speech segment of the same duration, with C the number of MEG or EEG sensors and T the number of time steps. Let $Y \in \mathbb{R}^{F \times T}$ be the latent representation of speech, using the same sample rate as X for simplicity, here the Mel spectrogram with F frequency bands. In this formalization, supervised decoding consists of finding a decoding function: $\mathbf{f}_{\text{reg}} : \mathbb{R}^{C \times T} \to \mathbb{R}^{F \times T}$ such that $\mathbf{f}_{\text{reg}}$ predicts Y given X. We denote by $\hat{Y} = \mathbf{f}_{\text{reg}}(X)$ the representation of speech decoded from the brain. When $\mathbf{f}_{\text{reg}}$ belongs to a parameterized family of models like deep neural

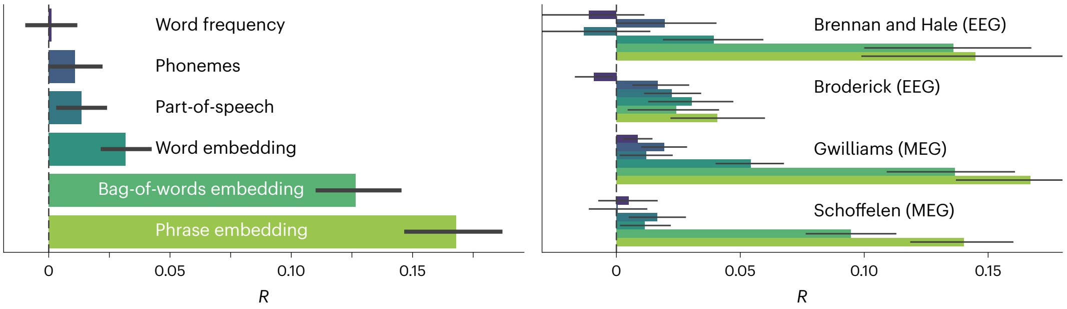


Fig. 4 | Decoding predictions mainly rely on high-level semantic features. The *R* values quantify the extent to which phonemes, word frequency, part-of-speech, word embedding and phrase embedding predict the probability of the predicted word to be correct. Error bars are the s.e.m. across participants (Table 1).

networks, it can be trained with a regression loss $L_{\mathrm{reg}}(Y, \hat{Y})$ (for example, the mean square error)

$$\min_{\mathbf{f}_{\mathrm{reg}}} \sum_{X,Y} L_{\mathrm{reg}}(Y, \mathbf{f}_{\mathrm{reg}}(X)). \tag{1}$$

This direct regression approach appears to be dominated by a non-distinguishable broadband component when speech is present (Extended Data Fig. 4a,b). This challenge motivates our three main contributions: the introduction of a contrastive loss, a pretrained deep speech representation and a dedicated brain decoder.

Model

Contrastive loss. We reasoned that regression may be an ineffective loss because it departs from our objective—that is, it requires maximally distinguishing different speech segments apart. Indeed, a regression objective stems from the principle that all of the dimensions of the Mel spectrogram are (1) equally important and (2) are scaled appropriately: the L2 objective inclines the model to predict low and high frequencies equally well, even if (1) some frequencies (for example, very low) may be irrelevant to speech and (2) some frequencies may vary in orders of magnitudes lowers than others. To relax this constraint, we opted for a contrastive objective and thus replaced the regression loss with the 'CLIP' loss (originally for Contrastive Language-Image Pre-Training) by ref. 44, which was originally designed to match latent representations in two modalities, text and images. Unlike the regression objective, this contrastive loss leads the model to find a combination of features that maximally discriminates samples in the batch. Consequently, the model is naturally inclined to focus on the informative dimensions of the Mel spectrograms and to scale them appropriately. We implement the CLIP loss as follows. Let X be a brain recording segment and $Y \in \mathbb{R}^{F\times T}$ the latent representation of its corresponding sound (also known as 'positive sample'). We sample $N-1$ negative samples $\bar{Y}_{j\in\{1,\ldots,N-1\}}$ over our dataset and we add the positive sample as $\bar{Y}_N = Y$. We want our model to predict the probabilities $\forall j \in \{1,\ldots,N\}, p_j = \mathbb{P}\left[\bar{Y}_j = Y\right]$. We thus train a model $\mathbf{f}_{\mathrm{clip}}$ mapping the brain activity X to a latent representation $Z = \mathbf{f}_{\mathrm{clip}}(X) \in \mathbb{R}^{F\times T}$. The estimated probability can then be approximated by the dot product of Z and the candidate speech latent representations Y_j, followed by a softmax:

$$\hat{p}_j = \frac{e^{\langle Z,\bar{Y}_j\rangle}}{\sum_{j'=1}^{N} e^{\langle Z,\bar{Y}_{j'}\rangle}}, \tag{2}$$

with $\langle\cdot,\cdot\rangle$ the inner product over both dimensions of Z and $\hat{Y}$. We then train $\mathbf{f}_{\mathrm{clip}}$ with a cross-entropy between p_j and $\hat{p}_j$. Note that for a large enough dataset, we can neglect the probability of sampling twice the same segment, so that we have $p_j = \mathbb{1}_{j=N}$, and the cross-entropy simplifies to

$$L_{\mathrm{CLIP}}(p,\hat{p}) = -\log(\hat{p}_N) = -\langle Z, Y\rangle + \log\left(\sum_{j'=1}^{N} e^{\langle Z,\bar{Y}_{j'}\rangle}\right). \tag{3}$$

Following ref. 44, we use the other elements of the batch as negative samples at train time. At test time, the negative samples correspond to all of the segments of the test set but the positive one.

Brain module. For the brain module, we introduce a deep neural network $\mathbf{f}_{\mathrm{clip}}$, input with raw MEG and EEG times series X and a one-hot encoding of the corresponding participant s, and outputs the latent brain representation Z, with the same sample rate as X. This architecture consists of (1) a spatial attention layer over the MEG and EEG sensors followed (2) by a participant-specific 1 × 1 convolution designed to leverage inter-individual variability, which input to (3) a stack of convolutional blocks. An overview of the model is given in the Extended Data Fig. 4e. In the following, given a tensor U, we note $U^{(i,\ldots)}$ access to specific entries in the tensor.

Spatial attention. The brain data are first remapped onto $D_1 = 270$ channels with a spatial attention layer based on the location of the sensors. The three-dimensional sensor locations are first projected on a two-dimensional plane obtained with the MNE-Python function find_layout[45], which uses a device-dependent surface designed to preserve the channel distances. Their two-dimensional positions are finally normalized to $[0, 1]$. For each output channel, a function over $[0, 1]^2$ is learnt, parameterized in the Fourier space. The weights over the input sensors are then given by the softmax of the function evaluated at the sensor locations. Formally, each input channel i has a location (x_i, y_i) and each output channel j is attached a function a_j over $[0, 1]^2$ parameterized in the Fourier space as $z_j \in \mathbb{C}^{K\times K}$ with $K = 32$ harmonics along each axis, that is

$$a_j(x,y) = \sum_{k=1}^{K}\sum_{l=1}^{K} \mathrm{Re}(z_j^{(k,l)})\cos\left(2\pi(kx+ly)\right) + \mathrm{Im}(z_j^{(k,l)})\sin\left(2\pi(kx+ly)\right). \tag{4}$$

The output is given by a softmax attention based on the evaluation of a_j at each input position (x_i, y_i):

$$\forall j \in \{1,\ldots,D_1\}, \mathrm{SA}(X)^{(j)} = \frac{1}{\sum_{i=1}^{C} e^{a_j(x_i,y_i)}}\left(\sum_{i=1}^{C} e^{a_j(x_i,y_i)} X^{(i)}\right) \tag{5}$$

with SA the spatial attention. In practice, as a_j is periodic, we scale down (x, y) to keep a margin of 0.1 on each side. We then apply a spatial dropout by sampling a location $(x_{\mathrm{drop}}, y_{\mathrm{drop}})$ and removing from the softmax each sensor that is within a distance of $d_{\mathrm{drop}} = 0.2$ of the sampled location. The initial motivation for spatial attention was to allow for a cross-dataset model to be defined in a way that would generalize across a diverse number location and set of sensors. Interestingly, we observed this layer to introduce an inductive bias that is beneficial to the prediction accuracy (Extended Data Fig. 2). See Extended Data Fig. 4 for a visualization of the learnt attention maps over each dataset. We then add a 1 × 1 convolution (that is, with a kernel size of 1) without activation and with the same number D_1 of output channels.

Participant layer. To leverage inter-individual variability, we learn a matrix $M_s \in \mathbb{R}^{D_1,D_1}$ for each participant $s \in [S]$ and apply it after the spatial attention layer along the channel dimension. This is similar to but more expressive than the participant embedding used by ref. 46 for MEG encoding, and follows decade of research on participant alignment[47,48].

Residual dilated convolutions. We then apply a stack of five blocks of three convolutional layers. For the *k*th block, the first two convolutions are applied with residual skip connections (except for the very first one where the number of dimension potentially doesn't match), outputs $D_2 = 320$ channels and are followed by batch normalization[49] and a GELU (Gaussian Error Linear Unit) activation[50]. The two convolutions are also dilated to increase their receptive field, by $2^{2k \bmod 5}$ and $2^{2k+1 \bmod 5}$ (with *k* zero indexed), respectively. The third layer in a block outputs $2D_2$ channels and uses a GLU (Gated Linear Unit) activation[51], which halves the number of channels. All convolutions use a kernel size of 3 over the time axis, a stride of 1 and sufficient padding to keep the number of time steps constant across layers. The output of the model is obtained by applying two final 1 × 1 convolutions: first with $2D_2$ outputs, followed by a GELU and finally with *F* channels as output, thus matching the dimensionality of speech representations. Given the expected delay between a stimulus and its corresponding brain responses, we further shift the input brain signal by 150 ms into the future to facilitate the alignment between *Y* and *Z*. The impact of this offset is considered in the Supplementary Section A.5.

Speech module. The Mel spectrogram is a low-level representation of speech inspired from the cochlea and is thus unlikely to match the rich variety of cortical representations[38]. Consequently, we replaced the Mel spectrograms with latent representations of speech. For this, we propose either to learn these representations end-to-end ('Deep Mel' model) or to rely on those learnt by an independent self-supervised speech model (wav2vec 2.0; ref. 29).

End-to-end speech representations with Deep Mel. The 'Deep Mel' module uses the same deep convolutional architecture to the brain module devoid of the participant block, and thus simultaneously learns to extract speech and MEG and EEG representations such that they are maximally aligned. By definition, and unlike wav2vec 2.0, Deep Mel sees only the audio used in the MEG and EEG datasets. As this end-to-end approach proved to be less efficient than its pretrained counterpart based on wav2vec 2.0, we will thereafter focus on the latter.

Pretrained speech representations with wav2vec 2.0. Wav2vec 2.0 is trained with audio data only to transform the raw waveform with convolutional and transformer blocks to predict masked parts of its own latent representations. A previous study[29] showed that the resulting model can be efficiently fine-tuned to achieve state-of-the-art performance in speech recognition. Besides, this model effectively encodes a wide variety of linguistic features[52,53]. In particular, recent studies have shown that the activations of wav2vec 2.0 linearly map onto those of the brain[54,55]. Consequently, we here test whether this model effectively helps the present decoding task. In practice, we use the wav2vec2-large-xlsr-53 (ref. 56), which has been pretrained on 56,000 hours of speech from 53 different languages.

Datasets

We test our approach on four public datasets, two based on MEG recordings and two based on EEG recordings. All datasets and their corresponding studies were approved by the relevant ethics committee and are publicly available for fundamental research purposes. Informed consent was obtained from each human research participant. We provide an overview of the main characteristics of the datasets in Table 1, including the number of training and test segments and vocabulary size over both splits. For all datasets, healthy adult volunteers passively listened to speech sounds (accompanied by some memory or comprehension questions to ensure participants were attentive), while their brain activity was recorded with MEG or EEG. In Schoffelen et al.[32], Dutch-speaking participants listened to decontextualized Dutch sentences and word lists (Dutch sentences for which the words are randomly shuffled). The study was approved by the local ethics committee (the local Committee on Research Involving Human Subjects in the Arnhem–Nijmegen region). The data are publicly and freely available after registration on the Donders Repository. In Gwilliams et al. [33], English-speaking participants listened to four fictional stories from the Masc corpus[57] in two identical sessions of 1 hour[30]. The study was approved by the institutional review board ethics committee of New York University Abu Dhabi. In Broderick et al. [58], English-speaking participants listened to extracts of *The Old Man and the Sea*. The study was approved by the ethics committees of the School of Psychology at Trinity College Dublin and the Health Sciences Faculty at Trinity College Dublin. In Brennan and Hale[31], English-speaking participants listened to a chapter of *Alice in Wonderland*. See Supplementary Section A.1 for more details. The study was approved by the University of Michigan Health Sciences and Behavioral Sciences institutional review board (HUM00081060).

Preprocessing

MEG and EEG are generally considered to capture neural signals from relatively low-frequency ranges[20]. Consequently, we first resampled all brain recordings down to 120 Hz with Torchaudio[59] and then split the data into training, validation and testing splits with a size roughly proportional to 70%, 20% and 10%, respectively. We defined a 'sample' as a 3 s window of brain recording with its associated speech representation. A 'segment' is a unique 3 s window of speech sound. As the same segment can be presented to multiple participants (or even within the same participant in ref. 33), the splits were defined so that one segment is always assigned to the same split across repetitions. We ensured that there were no identical sentences across splits. Furthermore, we excluded all segments overlapping over different splits. For clarity, we restricted the test segments to those that contain a word at a fixed location (here 500 ms before word onset).

MEG and EEG data can suffer from large artefacts, for example, eye movements or variations in the electro-magnetic environment[20]. To limit their impact, we applied a 'baseline correction' (that is, we subtracted from each input channel its average over the first 0.5 s) and a robust scaler with scikit-learn[60]. We normalized the data and clamp values greater than 20 s.d. to minimize the impact of large outlier samples. In Supplementary Section A.3, we study the effect of clamping and show that it is essential to ensure proper training. In Supplementary Section A.4, we further show that this approach is as effective as more complex data-cleaning procedures such as autoreject[61].

For the Mel spectrogram, we used 120 Mel bands (Supplementary Section A.6) (ref. 62), with a normalized STFT (short-time Fourier transform) with a frame size of 512 samples and hop length of 128 samples, using audio sampled at 16 kHz. We applied log-compression, that is, $\log(\epsilon + \mathrm{Mel})$, with $\epsilon = 10^{-5}$. When using wav2vec 2.0, we averaged the activations of the last four layers of its transformer. We used standard normalization for both representations.

Training

One training epoch is defined as 1,200 updates using Adam[63] with a learning rate of 3×10^{-4} and a batch size of 256. We stopped training when no improvement was observed on the valid set for ten epochs and kept the best model based on the valid loss. For the direct regression of the Mel spectrogram, we used the MSE (mean square error) loss. We used two V100 graphics processing units with 16 GB of memory. See Supplementary Section A.7 for an analysis of the impact of the training hyperparameters.

Evaluation

Mel reconstructions. In Extended Data Fig. 4, we illustrate some reconstructed Mel spectrograms using different models. With a regression loss, the generation of the Mel spectrogram is made directly. With a CLIP loss, we plot the weighted average across all test segments, where the weight corresponds to the probability of the segment to be true estimated with the CLIP loss. Specifically, given a segment and its matching audio (here the sentence 'Thank you for coming Ed'), we retrieve the predicted distribution over the 1,363 segments given by equation (2). We then use this distribution to average the Mel spectrogram of each candidate segment.

Segment-level evaluation. The top-10 segment accuracy indicates whether the true segment is in the top-10 most likely segments predicted by the decoder. We favour reporting this metric over the standard top-1 accuracy, given the large number of possible segments as the model may be able to decode useful information, without necessarily guessing the exact speech segment.

Word-level evaluation. To evaluate the model at the word level, we select a 3 s segment for each word of the test set (from −500 ms to 2.5 s). We input the model with the corresponding brain recordings, and output the probability distribution over all test segments. To obtain the distribution over the vocabulary, we group the candidate segments by the corresponding word (that is, starting at $t = 0$) and sum the probabilities of the same words spoken in different segments. Top-1 and top-10 word-level accuracy then quantify whether the true word is within the first or first 10 most likely predictions of the model, respectively.

Prediction analysis. To further inspect the predictions of the decoder, we quantify the extent to which they can be predicted from well-defined features $\tilde{Y} \in R^{n \times f_i}$. For this, we extract the phonetic features ($d = 60$) with Phonemizer[64], the 'zipf' frequency ($d = 1$) with Wordfreq[65], the part-of-speech tags ($d = 15$), the word embedding ($d = 300$) of each word with spaCy[66] as well as the phrase embedding of the corresponding 3 s speech segment ($d = 1, 024$) with Laser[67]. We refer to 'bag-of-words' as the sum of word embeddings over the segment. We then train a ridge regression with scikit-learn's default parameters[60] to predict the softmax probability of the true word output by the decoder, and estimate, with a five-split cross-validation, the correspondence between these two values with Pearson's R correlation. In sum, this analysis quantifies how well the feature predicts the probability of being selected by the decoder.

Statistics. Statistical comparison was performed on the test set. We used a Wilcoxon test across participants to compare different models on the same datasets. We used a Mann–Whitney test across participants to compare different datasets.

Discussion

Our model accurately identifies, from three seconds of non-invasive recordings of brain activity, the corresponding speech segment with up to 41% accuracy out of more than 1,000 distinct possibilities. This performance, sometimes reaching 80% in the best participants, allows the decoding of perceived words and phrases that are absent from the training set.

To decode speech perception from MEG and EEG, two major challenges must be overcome. First, these signals can be very noisy, making it difficult to extract useful information. Second, it is unclear which features of speech are, in fact, represented in the brain. Here we discuss how our 'brain' and 'speech' modules respectively address these two issues in the case of speech perception. Finally, we evaluate the performance of our model compared with previous works and outline the necessary steps to be taken before hoping to deploy this approach for the decoding of speech production in clinical settings.

Efficiently extracting brain signals

Non-invasive recordings are notoriously noisy: these signals present large variations across trials and participants and they are often contaminated by large artefacts[20–22]. Historically, solving this problem has involved complex preprocessing pipelines that include a variety of techniques—independent component analysis, outlier detection, artefact correction—that end with a linear model specific to each participant[68–70]. More recently, several deep-learning architectures have proved successful in solving simple classification tasks trained on single-participant recordings[71,72].

Building on these efforts, our end-to-end architecture requires minimal preprocessing of MEG and EEG signals and can be trained with a variety of participants, devices and stimuli. As decoding speech production can be challenged by the presence of muscular activity, we here evaluate this model on four public datasets where healthy participants listened to natural speech. Our analyses suggest that advanced MEG and EEG preprocessing does not provide a major advantage in the current decoding task and that a simple baseline correction followed by a robust scaler and clamping suffices. In addition, not only does our participant-specific layer improves decoding performance but also this performance increases with the number of participants present in the training set. These findings, combined with both the rise of publicly available datasets and the potential to learn informative features from unannotated data[73,74], suggest that this brain module may be a stepping stone for building a foundational model of brain recordings.

How is language represented in the brain?

Separating noise and signal in brain recordings is not, however, the only challenge. The nature of these representations in terms of their acoustic, phonetic, lexical and semantic properties remains poorly known. Consequently, determining the representations most suitable for decoding is an unresolved problem. To tackle this issue, previous studies have primarily used supervised models targeting well-defined features of language, such as individual letters, phonemes or frequency bands of the audio spectrogram[12,23,24,72,75–80]. Although this approach has demonstrated clear successes, it may impede the speed at which words are decoded from brain activity: for instance, spelling out a word letter by letter could be a slow and laborious process. As an alternative, others have proposed to learn to classify a small set of words[26,28,81–83]. This approach, however, is difficult to scale to a vocabulary size adequate for natural language. Finally, word semantics may be directly decoded from functional MRI signals[84–89]. However, the corresponding performances currently remain modest at the single-trial level.

Here we show how a neural network pretrained on a large corpus of speech sounds provides representations of language that are particularly valuable for brain decoding. Specifically, we leverage the recent discovery that these self-supervised speech models learn features that linearly relate to those of the brain[54,55] to build our speech module. By applying contrastive learning, we can effectively identify the most appropriate features for identifying new speech segments. Our analyses confirm that this approach outperforms (1) a supervised decoding of the Mel spectrogram as well as (2) 'Deep Mel', that is, latent representations of the Mel spectrogram optimized for decoding solely from the present MEG and EEG datasets. Finally, the inspection of the decoding predictions suggests that our model primarily captures the lexical and contextual features captured by modern word embeddings and language models. So far, however, what these high-level features represent and how these representations are structured and manipulated remain to be determined.

Comparison with previous works

Comparing the performance of our model to previous works is difficult because the variety of experimental paradigms is not compensated by a profusion of open datasets and reproducible code. Two elements may, nonetheless, substantiate such a comparison.

First, the size of vocabulary currently considered exceeds previous attempts, often by several orders of magnitude. For example, MEG and EEG studies typically used supervised decoders to discriminate a very small set of words[26,28,81–83] or sublexical classes (for example, phonemes, syllables, tones)[23,24,77–80]. For example, several studies[90–92] developed a decoder to classify 11, 5 and 2 distinct imagined phonemes from EEG signals, respectively. Similarly, several studies[25–27] developed an MEG decoder to classify 6 distinct part-of-speech (with 48% accuracy), 10 words (83% accuracy) and 3 words (70% accuracy), respectively. The limited vocabulary used in these non-invasive studies contrast with the present approach, which demonstrably accurately distinguishes several hundreds of words. Furthermore, the performances of our model are based on vocabularies that do not fully overlap with those used in the training set (Table 1). For example, for the Gwilliams dataset, the decoding performance reaches 40% in spite of the fact that nearly 36% of the words were never presented during training. Overall, such zero-shot decoding shows the versatility of this approach and opens the possibility to decode from even larger vocabulary.

Second, although our model's performance remains modest, it may not be too distant from the performance obtained with invasive recordings of brain activity. Indeed, decoding the perception of isolated words from a vocabulary of $n = 50$ words leads to a top-1 accuracy of 22.7% on average, but up to 42.9% in the best participants (Supplementary Section A.9). In comparison, Moses et al. [13] reported decoding produced words from intracranial recordings with a top-1 accuracy of 39.5% for isolated words out of $n = 50$ words. Similarly, still restricting the number of candidates to 50 and, this time, within the context of a sentence, our model decoding is above 72.5% top-1 accuracy on average across participants, and the best participants reach between 92.2% (ref. 32) and 95.9% (ref. 30; Fig. 2b), where Moses et al. [13] reached a top-1 accuracy of 74%. While comparing the decoding of perceived versus produced words should be considered with caution given their different brain bases, the performance of the current model thus leads us to be optimistic about its potential applicability in a speech production context.

Remaining steps to decode speech production in the clinics

Our non-invasive approach focuses on speech perception. To reach the performance obtained with clinical recordings[10,12,13,18,40,93–95], decoding intended communication will thus require addressing several challenges. Three specific challenges stand out.

First, the current model needs to be adapted to speech production. This can, in principle, be achieved by replacing the speech module with a neural network pretrained on production tasks such as handwriting or speech production.

Second, the current contrastive-learning objective can identify only the most likely word or speech segment from a predetermined set. The model thus needs to be supplemented with a generative module that can estimate the most likely phoneme, word or sentence given brain activity without requiring this set of candidates, similarly to what is being achieved with functional MRI[89,96].

Finally, our study reveals striking differences between EEG and MEG. While EEG is known to be less precise than MEG, we did not expect such an important difference in decoding performance. Adapting current MEG systems to the clinics will require substantial efforts, however: while new room-temperature sensors already show signal-to-noise ratio comparable to the superconducting quantum interference devices (SQUIDs) used in the present study, these systems are not commonly deployed in clinical settings, whose magnetic environment can be extremely noisy. Combined with artificial intelligence systems, these new devices could nevertheless contribute to improve the diagnosis, prognosis and restoration of language processing in non-communicating or poorly communicating patients without putting them at risk of brain surgery. In that regard, we hope that the release of a reproducible pipeline will contribute to the development of safe and scalable non-invasive methods for decoding intended communication.

Data availability

The data from Schoffelen et al.[32] were provided (in part) by the Donders Institute for Brain, Cognition and Behaviour with a 'RU-DI-HD-1.0' licence. The data for Gwilliams et al.[33] are available under the CC0 1.0 Universal licence. The data for Broderick et al.[58] are available under the same licence. Finally, the data from Brennan and Hale[31] are available under the CC BY 4.0 licence. All audio files were provided by the authors of each dataset.

Code availability

The complete source code for processing the datasets, training and evaluating the models and method presented here are available at github.com/facebookresearch/brainmagick. The code is provided under the CC-NC-BY 4.0 license.

References

1. Stanger, C. A. & Cawley, M. F. Demographics of rehabilitation robotics users. *Technol. Disabil.* **5**, 125–137 (1996).
2. Pels, E. G. M., Aarnoutse, E. J., Ramsey, N. F. & Vansteensel, M. J. Estimated prevalence of the target population for brain–computer interface neurotechnology in the netherlands. *Neurorehabil. Neural Repair* **31**, 677–685 (2017).
3. Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R. & Birbaumer, N. Brain–computer communication: unlocking the locked in. *Psychol. Bull.* **127**, 358 (2001).
4. Claassen, J. et al. Detection of brain activation in unresponsive patients with acute brain injury. *N. Engl. J. Med.* **380**, 2497–2505 (2019).
5. Owen, A. M. et al. Detecting awareness in the vegetative state. *Science* **313**, 1402–1402 (2006).
6. Cruse, D. et al. Bedside detection of awareness in the vegetative state: a cohort study. *Lancet* **378**, 2088–2094 (2011).
7. Birbaumer, N. et al. A spelling device for the paralysed. *Nature* **398**, 297–298 (1999).
8. King, J.-R. et al. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *Neuroimage* **83**, 726–738 (2013).
9. Brumberg, J. S., Kennedy, P. R. & Guenther, F. H. Artificial speech synthesizer control by brain–computer interface. In *Tenth Annual Conference of the International Speech Communication Association* (2009).
10. Herff, C. et al. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **9**, 217 (2015).
11. Stavisky, S. D. et al. Decoding speech from intracortical multielectrode arrays in dorsal 'arm/hand areas' of human motor cortex. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 93–97 (IEEE, 2018).
12. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
13. Moses, D. A. et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
14. Kennedy, P., Ganesh, A. & Cervantes, A. J. Slow firing single units are essential for optimal decoding of silent speech. *Front. Hum. Neurosci.* **16**, 874199 (2022).
15. Pei, X., Barbour, D. L., Leuthardt, E. C. & Schalk, G. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *J. Neural Eng.* **8**, 046028 (2011).
16. Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D. & Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 1–12 (2019).

17. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
18. Metzger, S. L. et al. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nat. Commun.* **13**, 6510 (2022).
19. Boto, E. et al. Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* **555**, 657–661 (2018).
20. Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. & Lounasmaa, O. V. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* **65**, 413 (1993).
21. Schirrmeister, R. T. et al. Deep learning with convolutional neural networks for eeg decoding and visualization. *Hum. Brain Mapp.* **38**, 5391–5420 (2017).
22. King, Jean-Rémi, et al. Encoding and decoding framework to uncover the algorithms of cognition. *Cogni. Neurosci.* **6**, 691–702 (2020).
23. Panachakel, J. T. & Ramakrishnan, A. G. Decoding covert speech from EEG—a comprehensive review. *Front. Neurosci.* **15**, 392 (2021).
24. Lawhern, V. J. et al. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **15**, 056013 (2018).
25. Lopopolo, A. & van den Bosch, A. Part-of-speech classification from magnetoencephalography data using 1-dimensional convolutional neural network. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/6gqj8 (2020).
26. Chan, A. M., Halgren, E., Marinkovic, K. & Cash, S. S. Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage* **54**, 3028–3039 (2011).
27. Nguyen, C. H., Karavas, G. K. & Artemiadis, P. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *J. Neural Eng.* **15**, 016002 (2017).
28. Murphy, A., Bohnet, B., McDonald, R. & Noppeney, U. Decoding part-of-speech from human eeg signals. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2201–2210 (2022).
29. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: a framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020).
30. Gwilliams, L., King, J. R., Marantz, A., & Poeppel, D. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nat. Commun* **13**, 6606 (2022).
31. Brennan, J. R. & Hale, J. T. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS ONE* **14**, e0207741 (2019).
32. Schoffelen, J.-M. et al. A 204-subject multimodal neuroimaging dataset to study language processing. *Sci. Data* **6**, 17 (2019).
33. Gwilliams, L. et al. MEG-MASC: a high-quality magneto-encephalography dataset for evaluating natural speech processing. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2208.11488 (2022).
34. Angrick, M. et al. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing* **342**, 145–151 (2019).
35. Hewitt, J. & Manning, C. D. A structural probe for finding syntax in word representations. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4129–4138 (2019).
36. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5.1**, 134 (2022).
37. Caucheteux, C., Gramfort, A. & King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.* **12**, 16327 (2022).
38. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
39. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
40. Angrick, M. et al. Speech synthesis from ECOG using densely connected 3D convolutional neural networks. *J. Neural Eng.* **16**, 036019 (2019).
41. Krishna, G., Tran, C., Han, Y., Carnahan, M. & Tewfik, A. H. Speech synthesis using EEG. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* 1235–1238 (IEEE, 2020).
42. Komeiji, S. et al. Transformer-based estimation of spoken sentences using electrocorticography. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* 1311–1315 (IEEE, 2022).
43. Mermelstein, P. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* **116**, 374–388 (1976).
44. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* 8748–8763 (PMLR (2021).
45. Gramfort, A. et al. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).
46. Chehab, O., Défossez, A., Jean-Christophe, L., Gramfort, A., & King, J. R. Deep recurrent encoder: an end-to-end network to model magnetoencephalography at scale. *Neurons Behav. Data Anal. Theory* https://doi.org/10.51628/001c.38668 (2022).
47. Xu, H., Lorbert, A., Ramadge, P. J., Guntupalli, J. S. & Haxby, J. V. Regularized hyperalignment of multi-set fMRI data. In *2012 IEEE Statistical Signal Processing Workshop (SSP)* 229–232 (IEEE, 2012).
48. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
49. Ioffe, S., & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* 448–456 (PMLR, 2015)
50. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at *arXiv* https://doi.org/10.48550/arXiv.1606.08415 (2016).
51. Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. Language modeling with gated convolutional networks. In *Proc. International Conference on Machine Learning* (2017), pp. 933–941
52. Millet, J. & Dunbar, E. uliette Millet and Ewan Dunbar. 2022. Do self-supervised speech models develop human-like perception biases?. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 7591–7605 (ACL, 2022).
53. Adolfi, F., Bowers, J. S., & Poeppel, D. Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Netw*. **162**, 199–211 (2023).
54. Millet, J. et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Adv. Neural Inf. Process.* **35**, 33428–33443 (2022).
55. Vaidya, A. R., Jain, S. & Huth, A. G. Self-supervised models of audio effectively explain human cortical responses to speech. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.14252 (2022).
56. Ott, M. et al. fairseq: a fast, extensible toolkit for sequence modeling. *GitHub* https://github.com/pytorch/fairseq/blob/main/examples/wav2vec (2019).
57. Ide, N., Baker, C. F., Fellbaum, C. & Passonneau, R. J. The manually annotated sub-corpus: a community resource for and by the people. In *Proc. ACL 2010 Conference Short Papers* 68–73 (2010).

58. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809 (2018).
59. Yang, Y. Y. et al. (2022, May). Torchaudio: Building blocks for audio and speech processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6982–6986 (IEEE, 2022).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F. & Gramfort, A. Autoreject: automated artifact rejection for MEG and EEG data. *NeuroImage* **159**, 417–429 (2017).
62. Young, S. et al. *The HTK Book* (Cambridge Univ. Engineering Department, 2002).
63. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1412.6980 (2014).
64. Bernard, M. & Titeux, H. Phonemizer: text to phones transcription for multiple languages in Python. *J. Open Source Softw.* **6**, 3958 (2021).
65. Speer, R. rspeer/wordfreq: v3.0. *Zenodo* https://doi.org/10.5281/zenodo.7199437 (2022).
66. Explosion AI. *spacy* https://spacy.io/ (2017).
67. Schwenk, H. & Douze, M. Learning joint multilingual sentence representations with neural machine translation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1704.04154 (2017).
68. Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
69. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* **8**, 679–685 (2005).
70. Nishimoto, S. et al. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
71. Roy, Y. et al. Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* **16**, 051001 (2019).
72. Dash, D. et al. Determining the optimal number of MEG trials: a machine learning and speech decoding perspective. In *Proc. Brain Informatics: International Conference 11* 163–172 (Springer, 2018).
73. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A. & Gramfort, A. Uncovering the structure of clinical EEG signals with self-supervised learning. *J. Neural Eng.* **18**, 046020 (2021).
74. Thomas, A., Ré, C., & Poldrack, R. Self-supervised learning of brain dynamics from broad neuroimaging data. *Adv. Neural Inf. Process* **35**, 21255–21269 (2022).
75. Miyawaki, Y. et al. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* **60**, 915–929 (2008).
76. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
77. Jayaram, V. & Barachant, A. Moabb: trustworthy algorithm benchmarking for bcis. *J. Neural Eng.* **15**, 066011 (2018).
78. Jahangiri, A., Chau, J. M., Achanccaray, D. R. & Sepulveda, F. Covert speech vs. motor imagery: a comparative study of class separability in identical environments. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2020–2023 (IEEE, 2018).
79. Orpella, J., Mantegna, F., Assaneo, F. & Poeppel, D. Speech imagery decoding as a window to speech planning and production. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.30.494046 (2022).
80. Ali, O. et al. Enhancing the decoding accuracy of EEG signals by the introduction of anchored-STFT and adversarial data augmentation method. *Sci. Rep.* **12**, 1–19 (2022).
81. Koizumi, K., Ueda, K. & Nakao, M. Development of a cognitive brain-machine interface based on a visual imagery method. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 1062–1065 (IEEE, 2018).
82. García-Salinas, J. S., Villaseñor-Pineda, L., Reyes-García, C. A. & Torres-García, A. A. Transfer learning in imagined speech EEG-based BCIs. *Biomed. Signal Process. Control* **50**, 151–157 (2019).
83. Dash, D., Ferrari, P., Heitzman, D. & Wang, J. Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 5531–5535 (IEEE, 2019).
84. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8**, 15037 (2017).
85. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.01244 (2019).
86. Affolter, N., Egressy, B., Pascual, D. & Wattenhofer, R. Brain2word: decoding brain activity for language generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.04765 (2020).
87. Pascual, D. et al. Improving brain decoding methods and evaluation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* 1476–1480 (IEEE, 2022).
88. Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J. & Binder, J. R. Decoding the information structure underlying the neural representation of concepts. *Proc. Natl Acad. Sci. USA* **119**, e2108091119 (2022).
89. Tang, J., LeBel, A., Jain, S., & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci* https://doi.org/10.1038/s41593-023-01304-9 (2023).
90. Sun, P. & Qin, J. Neural networks based EEG-speech models. Preprint at *arXIv* https://doi.org/10.48550/arXiv.1612.05369 (2016).
91. Sree, R. A. & Kavitha, A. Vowel classification from imagined speech using sub-band EEG frequencies and deep belief networks. In *2017 Fourth International Conference on Signal Processing, Communication and Networking* 1–4 (IEEE, 2017).
92. Moinnereau, M.-A. et al. Classification of auditory stimuli from EEG signals with a regulated recurrent neural network reservoir. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1804.10322 (2018).
93. Martin, S. et al. Word pair classification during imagined speech using direct brain recordings. *Sci. Rep.* **6**, 1–12 (2016).
94. Angrick, M. et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.* **4**, 1–10 (2021).
95. Kohler, J. et al. Synthesizing speech from intracranial depth electrodes using an encoder–decoder framework. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2111.01457 (2021).
96. Ozcelik, F. & VanRullen, R. Brain-diffuser: natural scene reconstruction from fMRI signals using generative latent diffusion. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.05334 (2023).

Acknowledgements

This work was funded in part by FrontCog grant ANR-17-EURE-0017 to J.R.K. for his work at PSL.

Author contributions

The project was led by A.D. and J.-R.K. J.-R.K. took care of the data curation. The training pipeline was built by J.R., O.K. and A.D. O.K. and A.D. handled model training and hyperparameter search. The speech module and evaluation pipeline was built by C.C. A.D., C.C., O.K. and J.-R.K. provided in depth data and results analysis. The present paper was written by A.D., C.C. and J.-R.K.

Competing interests

The authors declare no competing interests.

Additional information

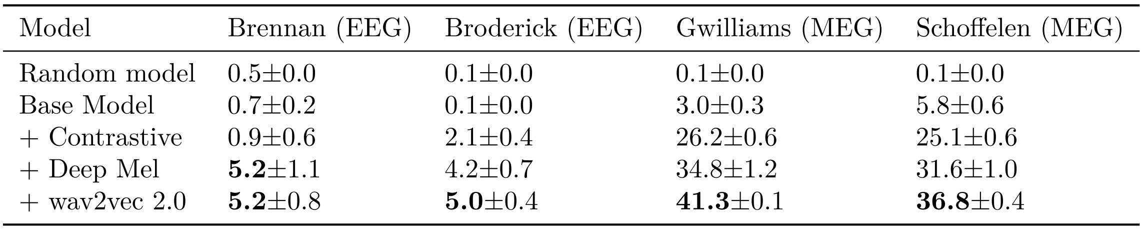
Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00714-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00714-5.

Correspondence and requests for materials should be addressed to Alexandre Défossez or Jean-Rémi King.

Peer review information *Nature Machine Intelligence* thanks Daniel Rubin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

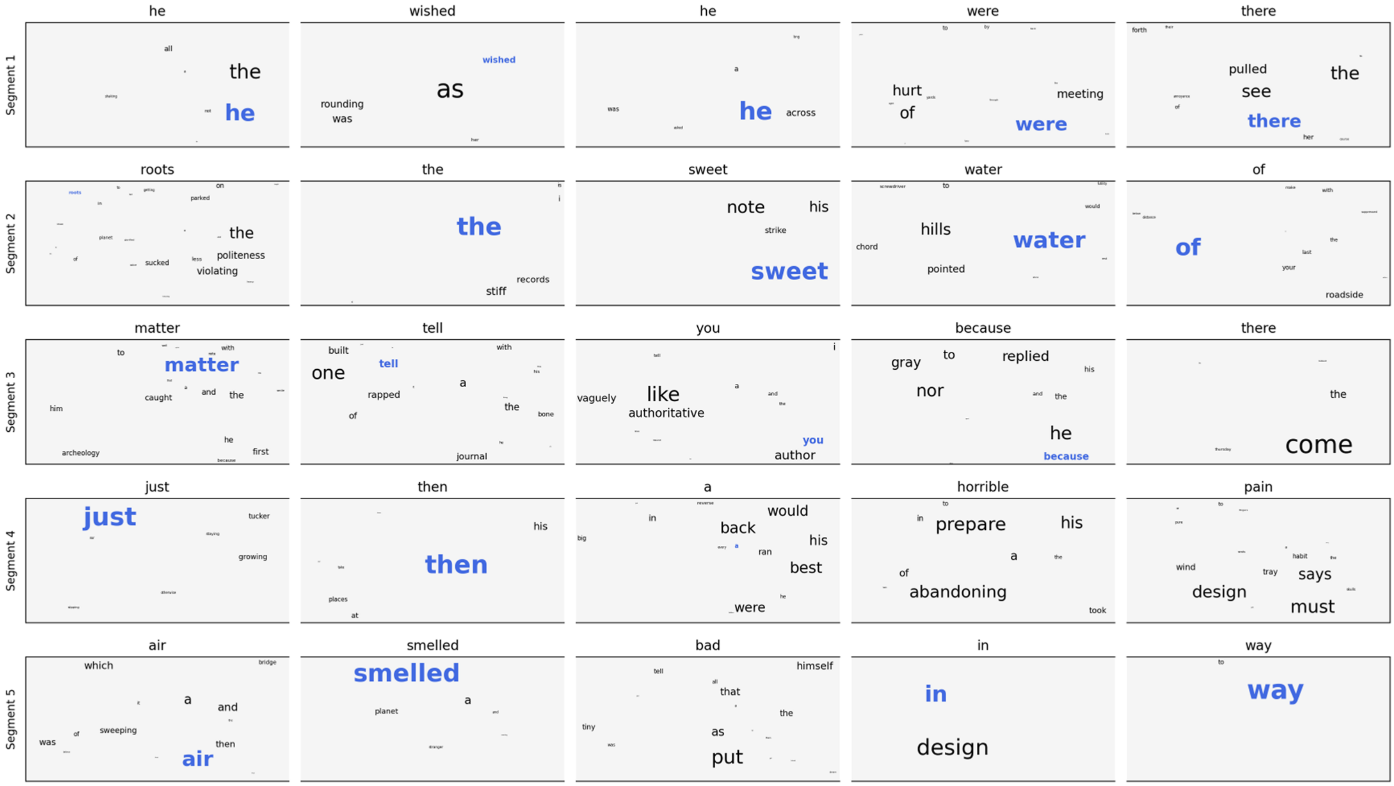


Model	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Random model	0.5±0.0	0.1±0.0	0.1±0.0	0.1±0.0
Base Model	0.7±0.2	0.1±0.0	3.0±0.3	5.8±0.6
+ Contrastive	0.9±0.6	2.1±0.4	26.2±0.6	25.1±0.6
+ Deep Mel	**5.2**±1.1	4.2±0.7	34.8±1.2	31.6±1.0
+ wav2vec 2.0	**5.2**±0.8	**5.0**±0.4	**41.3**±0.1	**36.8**±0.4

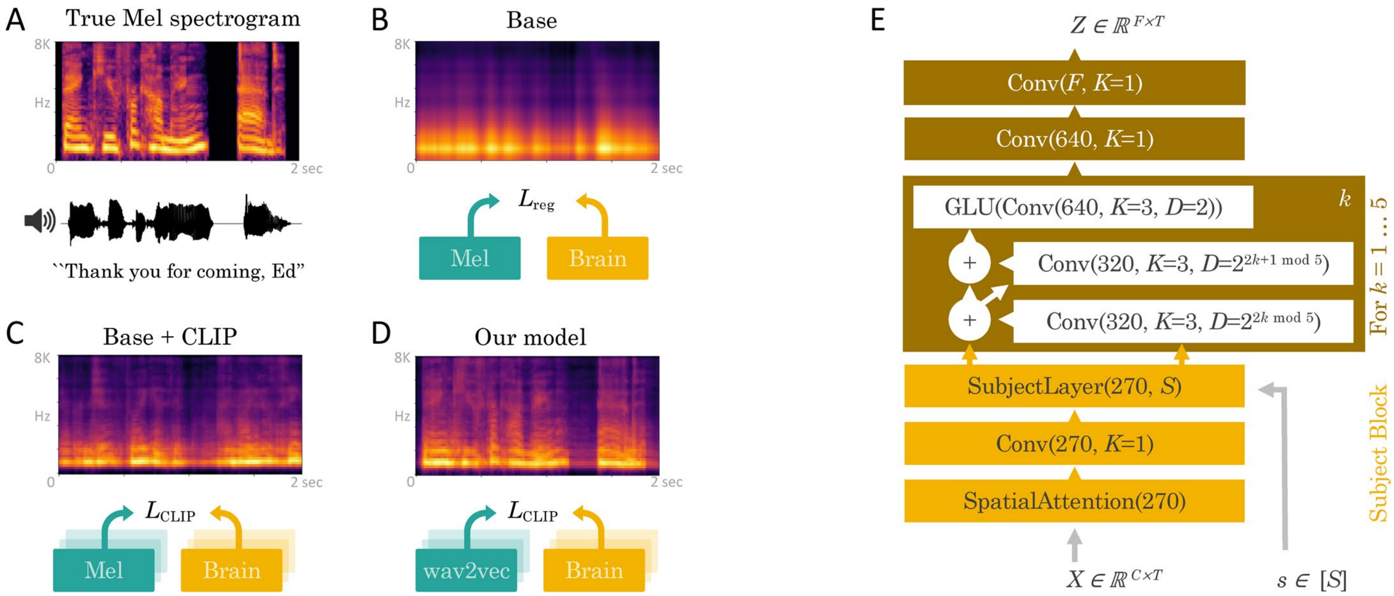
Extended Data Fig. 1 | Top-1 Accuracy. Segment-level top-1 accuracy related to Table 2.

Model	Broderick (EEG)	Brennan (EEG)	Schoffelen (MEG)	Gwilliams (MEG)	delta	p-val
Our model	**17.7** ± 0.6	25.7 ± 2.9	**67.5** ± 0.4	**70.7** ± 0.1		
- Spatial attention dropout	16.0 ± 1.7 *	**26.8** ± 0.7	67.5 ± 0.2	69.0 ± 0.2 *	0.4	0.009
- GELU + ReLU	16.4 ± 0.1	24.6 ± 2.1	65.8 ± 0.6 *	68.8 ± 1.3 *	1.6	$< 10^{-18}$
- Final convs	14.2 ± 1.1 *	19.0 ± 4.4 *	67.5 ± 0.3	68.9 ± 0.9 *	1.1	$< 10^{-10}$
- Non-residual GLU conv	8.4 ± 6.8 *	6.0 ± 0.2 *	67.0 ± 0.2	70.2 ± 0.2	1.6	$< 10^{-10}$
- Skip connections	13.9 ± 2.0 *	24.2 ± 2.7	65.4 ± 0.4 *	66.2 ± 0.3 *	2.4	$< 10^{-21}$
- Initial 1x1 conv	15.4 ± 0.6	22.1 ± 1.9 *	62.9 ± 0.9 *	67.7 ± 0.7 *	3.4	$< 10^{-26}$
- Spatial attention	15.4 ± 0.6 *	20.6 ± 2.2 *	65.9 ± 0.3 *	65.5 ± 0.4 *	2.5	$< 10^{-22}$
- Subj layer	8.1 ± 1.9 *	20.2 ± 1.3 *	42.4 ± 0.1 *	47.0 ± 1.3 *	14.4	$< 10^{-28}$
- Clamping	0.5 ± 0.0 *	14.1 ± 1.0 *	1.5 ± 0.3 *	23.6 ± 24.6 *	26.2	$< 10^{-29}$

Extended Data Fig. 2 | Ablations of the brain module. Segment-level top-10 accuracy (%) for our model and its ablated versions. Stars indicate significant gain ($p < 0.001$) across participants (dataset dependent, see Table 1). Confidence intervals are SEM over 3 runs.

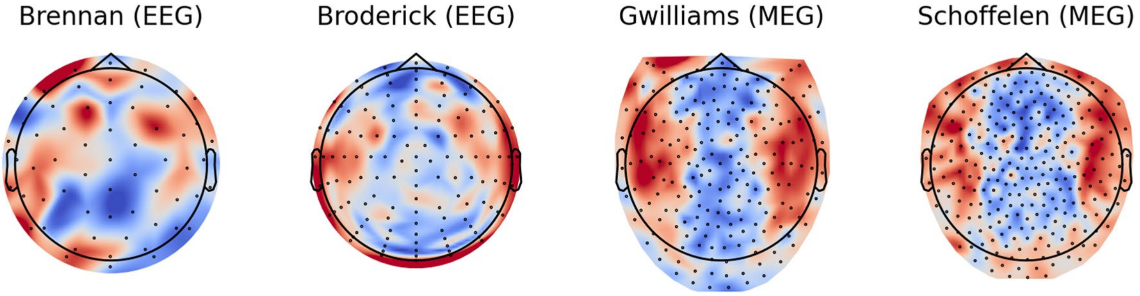


Extended Data Fig. 3 | Word-level predictions for different sentences. Similar illustration to Fig. 3 in the main paper but for five representative speech segments. The top and bottom segments are the easiest and hardest to decode, respectively. For each segment, we plot the predictions obtained for the subject with the median decoding scores across the cohort.



Extended Data Fig. 4 | Design choices and brain module. Design choices. A. Illustration of a 3 s speech sound segment (bottom) and its corresponding Mel spectrogram (top). **B.** Mel-spectrogram predicted with a direct regression loss of a brain decoder (orange). **C.** Replacing the regression loss with a contrastive loss improves reconstruction in the same subject, still using the mel-spectrogram as the speech representation. **D.** Now replacing the mel-spectrogram with wav2vec 2.0. The probabilities given by Eq. (2) are used to rebuild a mel-spectrogram. **E. Architecture of the brain module**. Architecture used to process the brain recordings. For each layer, we note first the number of output channels, while the number of time steps is constant throughout the layers. The model is composed of a spatial attention layer, then a 1x1 convolution without activation. A 'Subject Layer' is selected based on the subject index *s*, which consists in a 1x1 convolution learnt only for that subject with no activation. Then, we apply five convolutional blocks made of three convolutions. The first two use residual skip connection and increasing dilation, followed by a BatchNorm layer and a GELU activation. The third convolution is not residual, and uses a GLU activation (which halves the number of channels) and no normalization. Finally, we apply two 1x1 convolutions with a GELU in between.

https://doi.org/10.1038/s42256-023-00714-5



Extended Data Fig. 5 | Attention weights. Red color indicate that the M/EEG sensors is, on average, associated with a higher spatial attention weight. At the exception of the Brennan dataset, the topographies highlight channels typically activated during auditory stimulation.