Decoding speech from non-invasive brain recordings

Alexandre Défossez, Charlotte Caucheteux, Jeremy Rapin, Ori Kabeli, Jean-Rémi King

Decoding speech from non-invasive brain recordings

Alexandre Défossez[1,*], Charlotte Caucheteux[1,2], Jérémy Rapin[1], Ori Kabeli[1], and Jean-Rémi King[1,*]

[1]Meta AI, [2]Inria Saclay, [*]{defossez;jeanremi}@fb.com

Abstract

Decoding speech from brain activity is a long-awaited goal in both healthcare and neuroscience. Invasive devices have recently led to major milestones in that regard: deep learning algorithms trained on intracranial recordings now start to decode elementary linguistic features (e.g. letters, words, spectrograms). However, extending this approach to *natural speech* and *non-invasive* brain recordings remains a major challenge. To address these issues, we introduce a contrastive-learning model trained to decode self-supervised representations of natural speech from the non-invasive recordings of a large cohort of individuals. To evaluate this approach, we curate and integrate four public datasets, encompassing 169 volunteers recorded with magneto- or electro-encephalography (M/EEG), while they listened to natural speech. The results show that our model can identify, from 3 seconds of MEG signals, the corresponding speech segment with up to 44% accuracy out of 1,594 distinct possibilities – a performance that allows the decoding of phrases absent from the training set. Model comparison and ablation analyses show that these results directly benefit from the use of (i) a contrastive objective, (ii) pretrained representations of speech and (iii) a common convolutional architecture simultaneously trained across multiple participants. Overall, these results delineate a promising path to assist patients with communication disorders, without putting them at risk for brain surgery.

1 Introduction

Every year, thousands of patients suffer from brain or spinal cord injuries and suddenly lose their ability to communicate [Stanger and Cawley, 1996, Pels et al., 2017, Kübler et al., 2001, Pels et al., 2017, Claassen et al., 2019, Owen et al., 2006, Cruse et al., 2011]. Brain Computer Interface (BCI) has been raising high expectations to detect [Owen et al., 2006, Claassen et al., 2019, Birbaumer et al., 1999, King et al., 2013] and restore language abilities in such patients [Brumberg et al., 2009, Herff et al., 2015, Stavisky et al., 2018, Willett et al., 2021, Moses et al., 2021, Kennedy et al., 2022]: Over the past decades, several teams used BCI to efficiently decode phonemes, speech sounds [Pei et al., 2011, Akbari et al., 2019], hand gestures [Stavisky et al., 2018, Willett et al., 2021] and articulatory movements [Anumanchipalli et al., 2019, Moses et al., 2021] from electrodes implanted in the cortex or over its surface. For instance, Willett et al. [2021] decoded 90 characters per minute (with a 94% accuracy, *i.e.* roughly ≈15-18 words per minute) from a spinal-cord injury patient recorded in the motor cortex during 10 hours of writing sessions. Similarly, Moses et al. [2021] decoded 15.2 words per minute (with 74.4% accuracy, and using a vocabulary of 50 words) in an anarthria patient implanted in the sensorimotor cortex and recorded over 48 sessions spanning over 22 hours.

However, such invasive recordings face major practical challenges: these high-quality signals require (1) brain surgery (2) long training sessions and (3) can be difficult to maintain chronically. Several laboratories have thus focused on decoding language from *non-invasive* recordings of brain activity like magneto- and electro-encephalography (M/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

Preprint. Under review.

potentially wearable setup [Boto et al., 2018]. However, these two devices produce notoriously noisy signals that vary greatly across sessions and across individuals [Schirrmeister et al., 2017, King et al., 2018, Hämäläinen et al., 1993]. 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 [Lawhern et al., 2018, Lopopolo and van den Bosch, 2020, Chan et al., 2011, Nguyen et al., 2017].

In sum, decoding language from brain activity is, to date, either limited to invasive recordings or to impractical tasks. Interestingly, both of these approaches followed a similar method: *i.e.* (1) training a model on a single patient and (2) aiming to decode a limited set of interpretable features (MEL spectrogram, letters, phonemes, small set of words).

Instead, we here 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 learnt with self-supervised learning on a large quantity of speech data. For this, we introduce a convolutional neural network stacked onto a "Subject Layer" and trained with a contrastive objective to predict the representations of the audio waveform learnt by wav2vec 2.0 pretrained on 56k hours of speech [Baevski et al., 2020] (Figure 1). To validate our approach, we curate and integrate four public M/EEG datasets, encompassing the brain activity of 169 participants passively listening to sentences of short stories. With a sample of 3 seconds of M/EEG signals, our model identifies the matching audio segment (*i.e.* zero-shot decoding) with up to 72.5% top-10 accuracy (out of 1,594 segments) for MEG and up to 19.1% top-10 accuracy (out of 2,604 segments) for EEG.

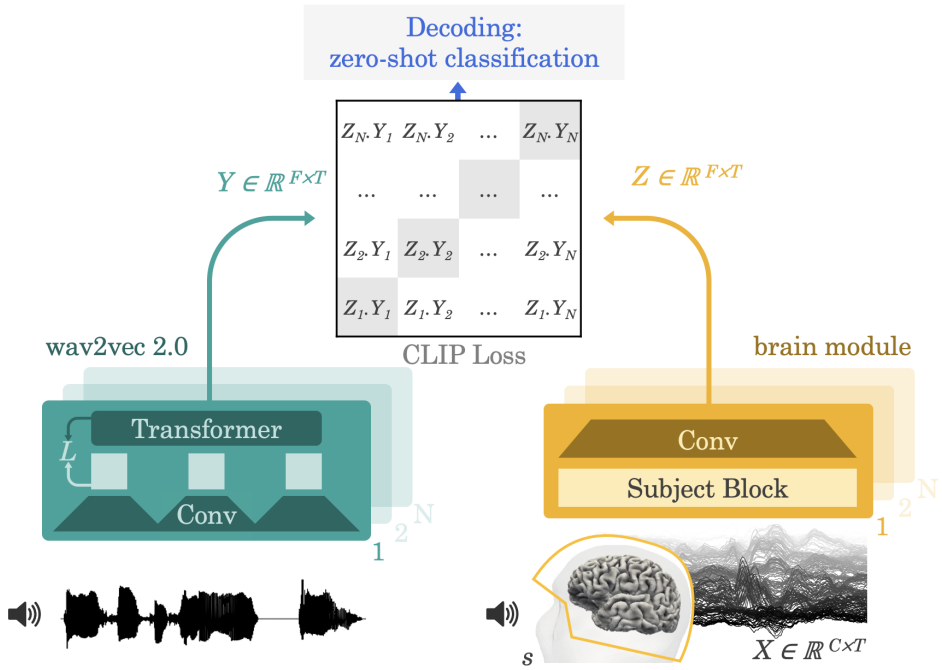

Figure 1: **Method** We aim to decode speech from the brain activity of healthy participants recorded with magnetoencephalography (MEG) or electroencephalography (EEG) while they listen to stories and/or sentences. For this, our model extracts the deep contextual representations of 3 s speech signals (Y) from a pretrained self-supervised model (wav2vec 2.0: Baevski et al. [2020]) and learns the representations Z of the brain activity on the corresponding 3 s window (X) that maximally align with these speech representations with a contrastive loss (CLIP: Radford et al. [2021]). 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.

2 Method

We first 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.

2.1 Problem formalization

We aim to decode speech from a time series of high-dimensional brain signals recorded with non-invasive magneto-encephalography (MEG) or electro-encephalography (EEG) while healthy volunteers passively listened to spoken sentences in their native language. How spoken words are represented in the brain is largely unknown [Hickok and Poeppel, 2007]. 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 [Akbari et al., 2019, Angrick et al., 2019b,a, Krishna et al., 2020, Komeiji et al., 2022]. For example, the Mel spectrogram is often targeted for neural decoding because it representats sounds similarly to the cochlea [Mermelstein, 1976]. Let $X \in \mathbb{R}^{C \times T}$ be a segment of a brain recording of a given subject while she listens to a speech segment of the same duration, with C the number of M/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. Thus, supervised decoding consists of finding a decoding function: $\boldsymbol{f}_{\text{reg}} : \mathbb{R}^{C \times T} \to \mathbb{R}^{F \times T}$ such that $\boldsymbol{f}_{\text{reg}}$ predicts Y given X. We denote by $\hat{Y} = \boldsymbol{f}_{\text{reg}}(X)$ the representation of speech decoded from the brain. When $\boldsymbol{f}_{\text{reg}}$ belongs to a parameterized family of models like deep neural networks, it can be trained with a regression loss $L_{\text{reg}}(Y, \hat{Y})$ (*e.g.* the Mean Square Error),

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

Empirically, we observed that this direct regression approach faces several challenges: decoding predictions appear to be dominated by a non-distinguishable broadband component when speech is present (Figure 2.A-B). This challenge motivates our three main contributions: the introduction of a contrastive loss, a pre-trained deep speech representation, and a dedicated brain decoder.

2.2 Model

2.2.1 Contrastive loss

First, we reasoned that regression may be an ineffective loss because it departs from our objective: decoding speech from brain activity. Consequently, we replaced it with a contrastive loss, namely, the "CLIP" loss (originally for Contrastive Language-Image Pre-Training) by Radford et al. [2021], which was originally designed to match latent representations in two modalities, text and images. 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 (a.k.a "*positive* sample"). We sample $N-1$ *negative* samples $\bar{Y}_{j \in \{1,\dots,N-1\}}$ over our dataset and we add the positive sample as $\bar{Y}_N = Y$. We want our model to predict the probabilities $\forall j \in \{1,\dots,N\}, p_j = \mathbb{P}\left[\bar{Y}_j = Y\right]$. We thus train a model $\boldsymbol{f}_{\text{clip}}$ mapping the brain activity X to a latent representation $Z = \boldsymbol{f}_{\text{clip}}(X) \in \mathbb{R}^{F \times T}$. 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:

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

with $\langle \cdot, \cdot \rangle$ the inner product over both dimensions of Z and $\hat{Y}$. We then train $\boldsymbol{f}_{\text{clip}}$ 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

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

Following [Radford et al., 2021], 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 but the positive one.

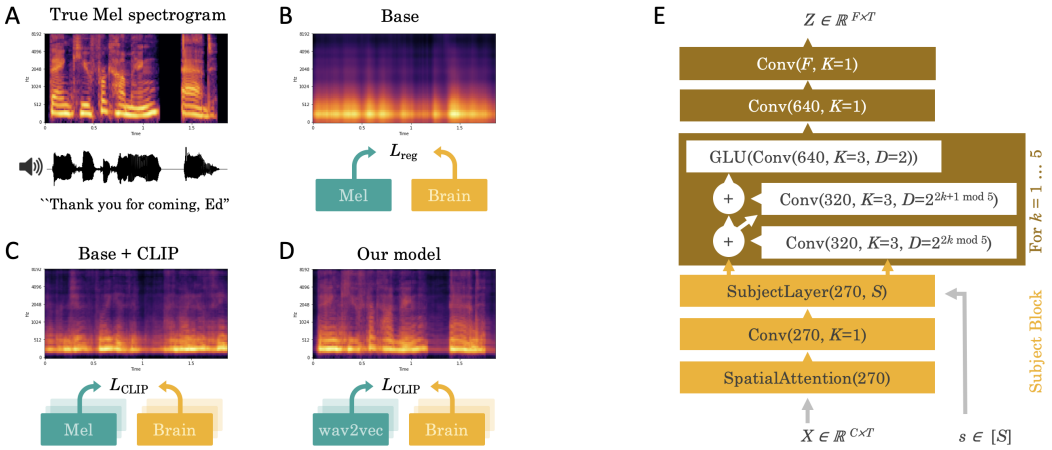

Figure 2: **Design choices. A.** Illustration of a 3 s speech sound segment (bottom) and its corresponding Mel spectrogram (top). **B.** Mel-spectrogram predicted with a direct regression loss L_{reg} of a brain decoder (orange). **C.** Replacing the regression loss with a CLIP loss [Radford et al., 2021] 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 [Baevski et al., 2020]. 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.

2.2.2 Speech module

Second, the Mel spectrogram is a low-level representation of speech and is thus unlikely to match the rich variety of cortical representations [Hickok and Poeppel, 2007]. Consequently, we replaced the Mel spectrograms Y with latent representations of speech, that are either learned end-to-end ("Deep Mel" model) or learned with an independent self-supervised speech model ("wav2vec 2.0", Baevski et al. [2020]) As detailed in the result section, the "Deep Mel" model uses an architecture similar to the brain module, but proved less efficient than its pretrained counterpart. We will thus focus the decoding results obtained with wav2vec 2.0.

Wav2vec 2.0 is trained to transform the raw waveform with convolutional and transformer blocks to predict masked parts of its own latent representations. Baevski et al. [2020] 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 [Millet and Dunbar, 2022, Adolfi et al., 2022]. Finally, recent work shows the existence of linear correspondence between the activations of the brain and those of wav2vec 2.0 [Millet et al., 2022, Vaidya et al., 2022]. Consequently, we here test whether this model effectively helps the present decoding task. In practice, we use the `wav2vec2-large-xlsr-53`[1], which has been pre-trained on 56k hours of speech from 53 different languages.

2.2.3 Brain module

Finally, for the brain module, we use a deep neural network $\boldsymbol{f}_{\text{clip}}$, input with raw M/EEG times series X and a one-hot-encoding of the corresponding subject 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 M/EEG sensors followed (2) by a subject-specific 1x1 convolution designed to leverage inter-subject

[1] https://github.com/pytorch/fairseq/blob/main/examples/wav2vec

Table 1: **Datasets**, noting chs. for channels and subj. for subjects.

						Train set		*Test set*	
Dataset	**Lang.**	**Type**	**# Chs.**	**# Subj.**	**Total duration**	**# Segments**	**Vocab.**	**# Segments**	**Vocab.**
Schoffelen2019	Dutch	MEG	273	96	80.7 h	5774	1755	1465	755
Gwilliams2022	English	MEG	208	21	49.2 h	6171	1870	1594	793
Broderick2019	English	EEG	128	19	18.8 h	7316	1393	2604	757
Brennan2019	English	EEG	60	33	6.7 h	1545	514	242	153

variability, which input to (3) a stack of convolutional blocks. An overview of the model is given in Figure 2. In the following, given a tensor U, we will note $U^{(i,\dots)}$ access to specific entries in the tensor.

Spatial attention and subject layer. The brain data is first remapped onto $D_1 = 270$ channels with a spatial attention layer based on the location of the sensors. The 3D sensor locations are first projected on a 2D plane obtained with the MNE-Python function `find_layout` [Gramfort et al., 2013], which uses a device-dependent surface designed to preserve the channel distances. Their 2D 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 is 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, *i.e.*

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

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

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

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_{drop} of the sampled location. We then add a 1x1 convolution (i.e. with a kernel size of 1) without activation and with the same number D_1 of output channels. Finally, to leverage inter-subject variability, we learn a matrix $M_s \in \mathbb{R}^{D_1,D_1}$ for each subject $s \in [S]$ and apply it after the spatial attention layer along the channel dimension. This is similar but more expressive than the subject embedding used by Chehab et al. [2021] for MEG encoding, and follows decade of research on subject alignment [Xu et al., 2012, Haxby et al., 2020].

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 [Ioffe and Szegedy, 2015] and a GELU activation [Hendrycks and Gimpel, 2016]. The two convolutions are also dilated to increase their receptive field, respectively by $2^{2k \bmod 5}$ and $2^{2k+1 \bmod 5}$ (with k zero indexed). The third layer in a block outputs $2D_2$ channels and uses a GLU activation [Dauphin et al., 2017] 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 1x1 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.

2.3 Datasets

We test our approach on four public datasets, two based on MEG recordings and two on EEG. 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 all human research participants. We provide an overview of the main characteristics of the datasets

on Table 1, including the number of train and test segments and vocabulary size over both splits. For all datasets, healthy adult volunteers passively listened to speech sounds (accompanied with some memory or comprehension questions to ensure participants were attentive), while their brain activity was recorded with MEG or EEG. In Schoffelen et al. [2019], 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 (CMO – the local "Committee on Research Involving Human Subjects" in the Arnhem-Nijmegen region). In Gwilliams et al. [2022], English-speaking participants listened to four fictional stories from the Masc corpus [Ide et al., 2010] in two identical sessions of one hour [Gwilliams et al., 2020]. The study was approved by the Institution Review Board (IRB) ethics committee of New York University Abu Dhabi. In Broderick et al. [2018], English-speaking participants listened to extracts of "The old man and the see". 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 [2019], English-speaking participants listened to a chapter of "Alice in Wonderlands". See Section A.1 in the Appendix for more details. The study was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (HUM00081060).

2.4 Preprocessing

M/EEG is generally considered to capture neural signals from relatively low frequency ranges [Hämäläinen et al., 1993]. Consequently, we first resampled all brain recordings down to 120 Hz with Torchaudio [Yang et al., 2021] and then split the data into *training*, *validation*, and *testing* splits with a size roughly proportional to 70%, 20%, and 10%. We define 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 subjects (or even within the same subject in Gwilliams et al. [2022]), the splits are defined so that one segment is always assigned to the same split across repetitions. We ensure that there is no identical sentences across splits, and checked that each sentence was pronounced by a unique speaker. Furthermore, we exclude all segments overlapping over different splits. For clarity, we restrict the test segments to those that contain a word at a fixed location (here 500 ms into the sample).

M/EEG data can suffer from large artifacts, e.g. eye movements, or variations in the electro-magnetic environment [Hämäläinen et al., 1993]. To limit their impact, we apply a "baseline correction" (*i.e.* we subtract to each input channel its average over the first 0.5 s) and a robust scaler with scikit-learn [Pedregosa et al., 2011]. We clamp values greater than 20 after normalization to minimize the impact of large outlier samples. For the Mel spectrogram, we use 120 Mel bands (see Section A.2 in the Appendix) [Young et al., 2002], with a normalized STFT with a frame size of 512 samples and hop length of 128 samples, using audio sampled at 16kHz. We apply log-compression, *i.e.* $\log(\epsilon + \text{mel})$, with $\epsilon{=}10^{-5}$. When using wav2vec 2.0, we average the activations of the last four layers of its transformer. We use standard normalization for both representations. To further assess the gains from using a self supervised representation, we also test a "Deep Mel" variant, where we train a deep transformation of the Mel, with the same architecture as the one applied to the brain recording, without the spatial attention and subject layer, and matching the output dimension of wav2vec 2.0. This transformation is trained along with the brain decoder using the contrastive objective (3). By definition, the Deep Mel model only sees the audio from the each of the studied datasets (unlike wav2vec 2.0).

2.5 Training

One training epoch is defined as 1,200 updates using Adam [Kingma and Ba, 2014] with a learning rate of $3{\cdot}10^{-4}$ and a batch size of 128. We stop training when no improvement is observed on the valid set for 10 epochs and keep the best model based on the valid loss. For the direct regression of the Mel spectrogram, we use the MSE loss. We use two V100 GPUs with 16GB of memory.

2.6 Evaluation

Segment-level evaluation. In Figure 2, we estimate the Mel spectrogram from the model output. Given a segment and its matching audio (here the sentence "Thank you for coming Ed"), we retrieve the predicted distribution over the $1,594$ segments given by (2). We use this distribution to average

Table 2: **Results.** 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 loss ('+ clip') and our model, *i.e.* trained to predict the features of wav2vec 2.0 with a contrastive loss ('+ wav2vec 2.0'). $\pm$ indicates the standard deviation across three random initializations of the model's weights.

Method	*Schoffelen2019*	*Gwilliams2022*	*Broderick2019*	*Brennan2019*	**Mean**
Random model	1.5 ± 0.18	2.2 ± 0.16	4.1 ± 0.09	7.6 ± 0.13	3.8
Base model	19.3 ± 0.83	14.9 ± 0.56	1.3 ± 0.19	6.6 ± 0.53	10.5
+ CLIP	51.5 ± 0.47	58.6 ± 0.28	13.3 ± 0.54	14.5 ± 1.33	34.5
+ Deep Mel	57.7 ± 0.16	64.4 ± 1.67	16.5 ± 0.26	23.7 ± 0.90	40.6
+ wav2vec 2.0	**67.2** ± 0.09	**72.5** ± 0.22	**19.1** ± 1.15	**31.4** ± 1.59	**47.5**

the Mel spectrogram of each candidate segment. Similarly, the top-10 segment accuracy indicates whether the true segment is in the top-10 most likely segments according to the same probabilities.

Word-level evaluation. We also evaluate the model at the word level (Figure 4). For each word of the test set, we select a 3 s segment starting with this word. We input the model with the corresponding brain recordings, and output the probability distribution over all test segments including the true segment. To obtain the distribution over the vocabulary, we group the candidate segments by their first word and sum the probabilities within each group.

2.7 Code availability

The code to reproduce the present study will be made publicly available upon publication.

3 Results

3.1 Accurately decoding speech from M/EEG recordings

Our model predicts the proper segment, out of more than 1,000 possible ones, with a top-10 accuracy of 72% and 67% for MEG datasets (top-1 accuracy of 44% and 36%) (Table 2). For more than half of samples, the true audio segment is ranked first or second in the decoders' predictions. For comparison, a model that predicts a uniform distribution over the vocabulary ('random model') only achieves a 2% top-10 accuracy on the same MEG datasets. Decoding performance for EEG datasets is lower: our model reaches 19% and 31% top-10 accuracy. While modest, these scores are four times higher than the random baseline.

3.2 Effect of contrastive loss, deep speech representations, and number of participants

Our ablation highlights the importance of: (1) the contrastive loss, (2) the use of deep speech representations [Baevski et al., 2020] and (3) the combination of a large number of participants. First, a model trained to predict the Mel spectrogram with a regression objective ('base model' in Table 2) achieves 10% top-10 accuracy on average across datasets – *i.e.* nearly five times lower than our model, when using the model output to rank the candidate segments by cosine similarity.

Second, predicting the Mel spectrogram with a contrastive loss leads to a 3X improvement over the base model, and gains another 16 points by using wav2vec 2.0 as the speech representation. We verified that the wav2vec 2.0's latent representations provide higher decoding performances than those learnt end-to-end with contrastive learning, as shown by the results of the Deep Mel model on Table 2.

Third, to test whether our model effectively leverage the inter-individual variability, we trained it on a variable number of subjects and computed its accuracy on the first 10% of subjects. As shown in Figure 3B, decoding performance increases as the model is trained with more subjects on the two MEG datasets. This ability to learn from multiple subjects is strengthened by another ablation experiment: training on all participants, but *without* the subject-specific layer, leads to a drop of 17% accuracy on average across the four datasets (Table 3). However, this last gain is relatively modest compared to the a subject embedding introduced recently [Chehab et al., 2021].

Table 3: **Ablations.** Top-10 segment-level accuracy (%) for our model and its ablated versions. Delta refers to the average decrease in accuracy of each ablated version compared to our model.

Arch. change	*Schoffelen2019*	*Gwilliams2022*	*Broderick2019*	*Brennan2019*	**Mean**	**Delta**
Our model	**67.2** ± 0.09	72.5 ± 0.22	19.1 ± 1.15	**31.4** ± 1.59	**47.5**	0.00
\wo spatial attention dropout	61.6 ± 0.14	71.2 ± 0.93	19.0 ± 1.07	30.2 ± 1.70	45.5	-2.00
\w subj. embedding*	59.5 ± 0.24	72.0 ± 0.77	**20.2** ± 1.24	30.2 ± 0.77	45.4	-2.08
\wo GELU, \w ReLU	61.4 ± 0.67	72.2 ± 0.05	19.2 ± 0.79	26.4 ± 1.03	44.8	-2.72
\wo spatial attention	60.0 ± 1.32	69.5 ± 0.44	17.9 ± 0.34	26.0 ± 0.61	43.3	-4.18
\wo final convs	62.3 ± 0.07	71.0 ± 0.47	15.7 ± 1.13	22.7 ± 2.05	43.0	-4.57
\wo initial 1x1 conv.	57.8 ± 1.12	69.6 ± 0.37	17.9 ± 0.31	26.3 ± 1.43	42.9	-4.62
\wo skip connections	59.2 ± 0.71	68.0 ± 0.60	16.7 ± 0.29	25.7 ± 4.32	42.4	-5.13
\wo non-residual GLU conv.	63.5 ± 0.68	**73.0** ± 0.67	17.0 ± 0.03	6.70 ± 0.57	40.0	-7.50
\wo subject-specific layer	38.3 ± 0.77	49.2 ± 0.23	11.8 ± 0.14	21.5 ± 0.59	30.2	-17.30
\wo clamping brain signal	1.1 ± 0.26	57.6 ± 13.4	4.0 ± 0.14	11.0 ± 1.92	18.4	-29.10

*: we used the subject embedding from [Chehab et al., 2021] instead of the subject layer.

Finally, other design choices modestly but significantly impact the performance of our model. Performance systematically decreases when removing skip connections, the spatial attention module, the initial or final convolutional layers (Table 3). We also show how essential clamping is to train the model, except for the [Gwilliams et al., 2022] dataset, which led to similar performances, although with a doubling of the training time. See Section A.2 in the Appendix for more ablations analyses.

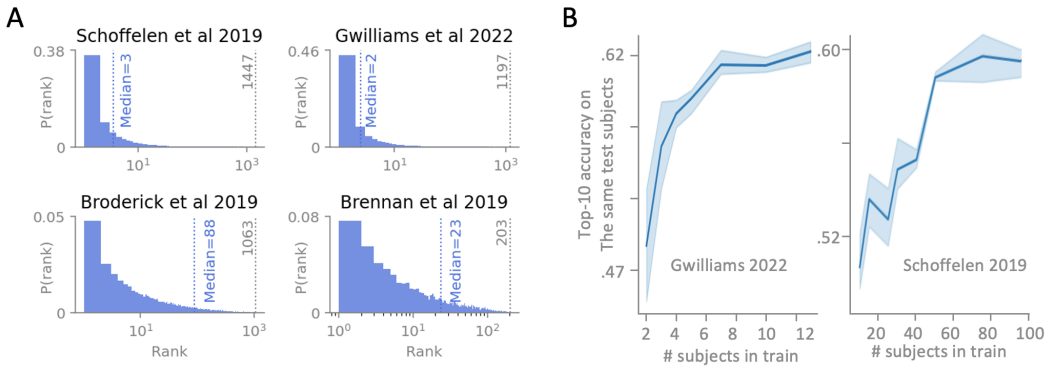

Figure 3: **Segment-level decoding. A.** Probability distribution of the decoded rank for each segment (lower is better) for each dataset. The gray dotted line indicates the number of segments in the test set. **B.** Top-10 accuracy obtained for the first 10% of subjects (y-axis) as a function of the number of subjects seen during training (x-axis). The line and confidence intervals represent the mean and standard error of the mean (SEM) across participants, respectively.

4 Discussion

Here, we aimed to decode the perception of natural speech from non-invasive brain recordings. Our results, based on the largest decoding study of M/EEG responses to speech to date, show that combining (1) a contrastive objective, (2) a convolutional architecture enhanced by a "Subject Layer", and (3) pretrained speech representations allows the decoding of new 3 s speech sounds with up to 44% top-1 accuracy out of more than 1,500 possibilities.

Our approach contributes to the rapid transformation of hand-crafted pipelines into their end-to-end counterparts. In particular, this study shows how self-supervised and contrastive learning can improve both (1) the analyses of brain signals and (2) the definition of the linguistic features that should be used for decoding. In particular, previous models were typically trained on individual subjects to categorize a very small number of highly-repeated categories and/or hand-crafted features [Ali et al., 2022, Jayaram and Barachant, 2018, Lawhern et al., 2018]. For example, Sun and Qin [2016], Sree and Kavitha [2017], and Moinnereau et al. [2018] all developed a decoder to classify 11, 5 and 2 distinct imagined phonemes, respectively, from EEG signals. Similarly, Lopopolo and van den Bosch [2020], Chan et al. [2011], Nguyen et al. [2017] respectively developed a decoder to classify

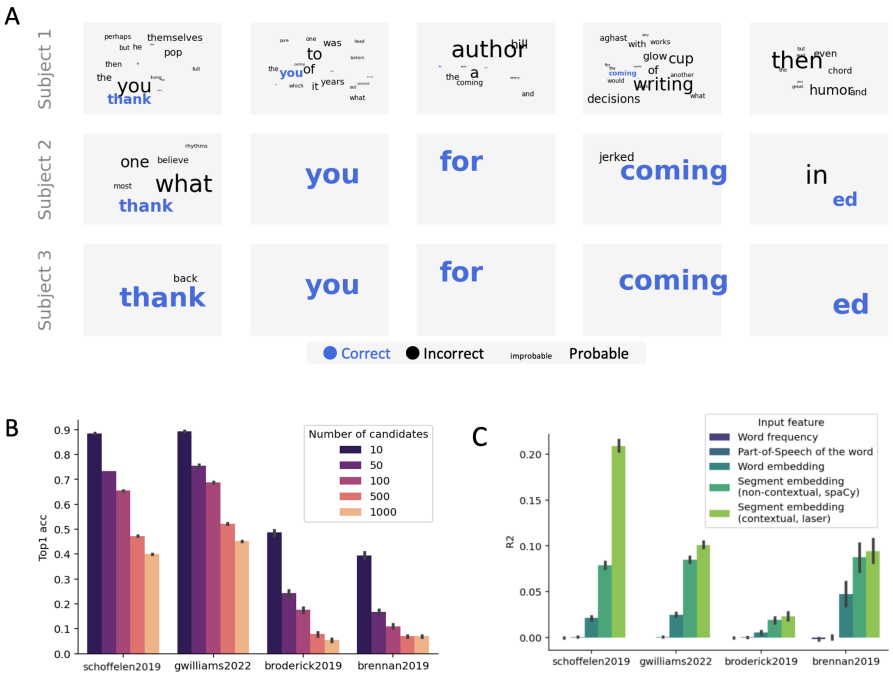

Figure 4: **A.** Single-word prediction for the first three subjects of Gwilliams et al. [2022] listening to the sentence "Thank you for coming, Ed". Text color indicates whether the decoded word is accurate. Text size is proportional to the log-probability output by our model. **B.** Top-1 accuracy at the word level (as explained in Section 2.6) as a function of the number of negatives during inference. **C.** The R^2 summarize how word frequency, part-of-speech tag, word embedding, and contextual embedding respectively predict the accuracy of single-word and single-segment decoding (Appendix A.4). Error bars are the SEM across participants.

6 distinct part-of-speech (48% accuracy), 10 words (83% accuracy) and 3 words (70% accuracy), from MEG signals. Finally, both Dash et al. [2020] and Wang et al. [2017] trained a classifier to decode 5 distinct sentences from MEG activity (both around 94% accuracy). Given the combinatorics of language, such single-subject / limited-vocabulary approach necessarily limits the possibility to decode natural speech. By contrast, our model effectively achieves "zero-shot" decoding by matching a large number of brain recordings to the deep representations of their corresponding speech sounds.

One remarkable phenomenon revealed by the present study is the difference of performance obtained between with EEG and MEG: while EEG is known to be less precise than MEG, we did not expect such a strong difference. This result thus holds great promises for the development of a safe and scalable system based on the analysis of magnetic – rather than electric – fields. It should be stressed that while the scientific community should remain vigilant that this approach will not be adapted to decode brain signals without the consent of the participants, this possibility appears unlikely at this stage: Unlike other biomarkers, such as fingerprints, DNA and facial features, electro-magnetic signals cannot be acquired unbeknownst to the participants. Furthermore, teeth clenching, eye blinks and other muscle movements are known to massively corrupt these signals, and thus presumably provide a simple way to counter downstream analyses. In any case, we believe that open science remains the best way to responsibly assess risks and benefits in this domain.

The present non-invasive study is limited to speech *perception*. It thus differs from the recent achievements obtained in a small set of heavily-trained patients implanted for clinical purposes and tasked to produce language [Herff et al., 2015, Martin et al., 2016, Angrick et al., 2019b, Willett et al., 2021, Moses et al., 2021, Angrick et al., 2021, Kohler et al., 2021]. In particular, Willett et al. [2021] showed that a 1 s time window of neuronal activity in the motor cortex suffices to decode one of 26 characters with 94.1% top-1 accuracy during a spelling task. Similarly, Moses et al. [2021] showed that 4 s of neuronal activity recorded in the sensory-motor cortices is sufficient to decode the intention to communicate one of 50 words with a median word error rate of 25.6%. We deliberately chose to

focus on speech perception here, because speech production generates muscle activity, which can be easily read with EEG and MEG. Consequently, decoding speech production could be trivial in healthy participant, without ensuring any kind of utility for patients with an inability to control facial muscles.

While our approach remains to be adapted to *language production*, the possibility of leveraging data from (i) multiple subjects and (ii) large natural language datasets, together with (iii) the multiplication of public neuroimaging datasets, makes us hopeful about possibility of decoding intended communication from non-invasive recordings of brain activity. This possibility could also be accelerated by the development of new MEG hardwares: the MEG used in the present study makes use of superconducting quantum interference device (SQUID) and necessitates to cool the sensors down to $\approx 4^{\circ}$K with a very large tank of liquid helium. However, several room-temperature sensors are now available, and already show signal-to-noise ratio comparable to SQUIDs [Boto et al., 2018]. Combined with A.I. systems, these new devices will thus likely contribute to improve the diagnosis, prognosis and restoration of language processing in non- or poorly-communicating patients without putting them at risks for brain surgery.

References

Federico Adolfi, Jeffrey S Bowers, and David Poeppel. Successes and critical failures of neural networks in capturing human-like speech recognition. *arXiv preprint arXiv:2204.03740*, 2022.

Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1): 1–12, 2019.

Omair Ali, Muhammad Saif-ur Rehman, Susanne Dyck, Tobias Glasmachers, Ioannis Iossifidis, and Christian Klaes. Enhancing the decoding accuracy of eeg signals by the introduction of anchored-stft and adversarial data augmentation method. *Scientific reports*, 12(1):1–19, 2022.

Miguel Angrick, Christian Herff, Garett Johnson, Jerry Shih, Dean Krusienski, and Tanja Schultz. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing*, 342:145–151, 2019a.

Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019b.

Miguel Angrick, Maarten C Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, Sophocles Goulis, Jeremy Saal, Albert J Colon, Louis Wagner, Dean J Krusienski, et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology*, 4(1):1–10, 2021.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

Elena Boto, Niall Holmes, James Leggett, Gillian Roberts, Vishal Shah, Sofie S Meyer, Leonardo Duque Muñoz, Karen J Mullinger, Tim M Tierney, Sven Bestmann, et al. Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, 555 (7698):657–661, 2018.

Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

Jonathan S Brumberg, Philip R Kennedy, and Frank H Guenther. Artificial speech synthesizer control by brain-computer interface. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and category-specific spatiotemporal representations from meg and eeg. *Neuroimage*, 54(4):3028–3039, 2011.

Omar Chehab, Alexandre Defossez, Jean-Christophe Loiseau, Alexandre Gramfort, and Jean-Remi King. Deep recurrent encoder: A scalable end-to-end network to model brain signals. *arXiv preprint arXiv:2103.02339*, 2021.

Jan Claassen, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez, Joshua U Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, et al. Detection of brain activation in unresponsive patients with acute brain injury. *New England Journal of Medicine*, 380 (26):2497–2505, 2019.

Damian Cruse, Srivas Chennu, Camille Chatelle, Tristan A Bekinschtein, Davinia Fernández-Espejo, John D Pickard, Steven Laureys, and Adrian M Owen. Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet*, 378(9809):2088–2094, 2011.

Debadatta Dash, Paul Ferrari, and Jun Wang. Decoding imagined and spoken phrases from non-invasive neural (meg) signals. *Frontiers in neuroscience*, 14:290, 2020.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, page 267, 2013.

Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*, 2020.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkanen, David Poeppel, and Jean-Remi King. Meg-masc: a high-quality magneto-encephalography dataset for evaluating natural speech processing. *arXiv preprint arXiv:2208.11488*, 2022.

Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

James V Haxby, J Swaroop Guntupalli, Samuel A Nastase, and Ma Feilong. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *Elife*, 9:e56601, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9:217, 2015.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Nancy Ide, Collin F Baker, Christiane Fellbaum, and Rebecca J Passonneau. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 conference short papers*, pages 68–73, 2010.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Technical Report 1502.03167, arXiv, 2015.

Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

Philip Kennedy, A Ganesh, and AJ Cervantes. Slow firing single units are essential for optimal decoding of silent speech. 2022.

Jean-Rémi King, Frédéric Faugeras, Alexandre Gramfort, Aaron Schurger, Imen El Karoui, JD Sitt, Benjamin Rohaut, C Wacongne, E Labyt, Tristan Bekinschtein, et al. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *Neuroimage*, 83: 726–738, 2013.

Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition. 2018.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Jonas Kohler, Maarten C Ottenhoff, Sophocles Goulis, Miguel Angrick, Albert J Colon, Louis Wagner, Simon Tousseyn, Pieter L Kubben, and Christian Herff. Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*, 2021.

Shuji Komeiji, Kai Shigemi, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, and Toshihisa Tanaka. Transformer-based estimation of spoken sentences using electrocorticography. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1311–1315. IEEE, 2022.

Gautam Krishna, Co Tran, Yan Han, Mason Carnahan, and Ahmed H Tewfik. Speech synthesis using eeg. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1235–1238. IEEE, 2020.

Andrea Kübler, Boris Kotchoubey, Jochen Kaiser, Jonathan R Wolpaw, and Niels Birbaumer. Brain–computer communication: Unlocking the locked in. *Psychological bulletin*, 127(3):358, 2001.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Alessandro Lopopolo and Antal van den Bosch. Part-of-speech classification from magnetoencephalography data using 1-dimensional convolutional neural network. 2020.

Stephanie Martin, Peter Brunner, Iñaki Iturrate, José del R Millán, Gerwin Schalk, Robert T Knight, and Brian N Pasley. Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, 6(1):1–12, 2016.

Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976.

Juliette Millet and Ewan Dunbar. Do self-supervised speech models develop human-like perception biases? *arXiv preprint arXiv:2205.15819*, 2022.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv preprint arXiv:2206.01685*, 2022.

Marc-Antoine Moinnereau, Thomas Brienne, Simon Brodeur, Jean Rouat, Kevin Whittingstall, and Eric Plourde. Classification of auditory stimuli from eeg signals with a regulated recurrent neural network reservoir. *arXiv preprint arXiv:1804.10322*, 2018.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*, 15(1):016002, 2017.

Adrian M Owen, Martin R Coleman, Melanie Boly, Matthew H Davis, Steven Laureys, and John D Pickard. Detecting awareness in the vegetative state. *science*, 313(5792):1402–1402, 2006.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Xiaomei Pei, Dennis L Barbour, Eric C Leuthardt, and Gerwin Schalk. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028, 2011.

Elmar GM Pels, Erik J Aarnoutse, Nick F Ramsey, and Mariska J Vansteensel. Estimated prevalence of the target population for brain-computer interface neurotechnology in the netherlands. *Neurorehabilitation and neural repair*, 31(7):677–685, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche H. L. Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6(1):17, April 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0020-y.

R Anandha Sree and A Kavitha. 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 (ICSCN)*, pages 1–4. IEEE, 2017.

Carol A Stanger and Michael F Cawley. Demographics of rehabilitation robotics users. *Technology and Disability*, 5(2):125–137, 1996.

Sergey D Stavisky, Paymon Rezaii, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. 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 (EMBC)*, pages 93–97. IEEE, 2018.

Pengfei Sun and Jun Qin. Neural networks based eeg-speech models. *arXiv preprint arXiv:1612.05369*, 2016.

Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.

Jun Wang, Myungjong Kim, Angel W Hernandez-Mulero, Daragh Heitzman, and Paul Ferrari. Towards decoding speech production from single-trial magnetoencephalography (meg) signals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3036–3040. IEEE, 2017.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, 2021.

Hao Xu, Alexander Lorbert, Peter J Ramadge, J Swaroop Guntupalli, and James V Haxby. Regularized hyperalignment of multi-set fmri data. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 229–232. IEEE, 2012.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3(175):12, 2002.

A Appendix

A.1 Datasets

The data from Schoffelen et al. [2019] was provided (in part) by the Donders Institute for Brain, Cognition and Behaviour with a "RU-DI-HD-1.0" licence[2]. The data for Gwilliams et al. [2022] is available under CC0 1.0 Universal[3]. The data for Broderick et al. [2018] is available under the same licence[4] Finally, the data from Brennan and Hale [2019] is available under the CC BY 4.0 licence[5] All audio files were provided by the authors of each dataset.

A.2 Extra Results

In this Section, we provide extra analysis with regard to the number of MEL band used, and the clamping value.

A.2.1 Effect of clamping

Clamping is essential due to the sensitivity of electro-magnetic recordings to perturbations. As explained in Section 2.4, we first use a quantile based robust scaler such that the range [-1, 1] maps to the [0.25, 0.75] quantile range. The scaling is computed independently for each recording. Thus it is expected most values for M/EEG recordings would have a scale of the order of 1. In the following table, we provide the top-10 accuracy for the Wav2Vec2.0 based model from Table 2. We observe that extending the clamping range from 20 to 100 doesn't allow the model to extract more information, which would be expected if large scale values are outliers without useful information on the underlying brain dynamics. On the other hand, when removing entirely clamping, we observe a collapse of the performance. This is expected, as extreme outliers will impact for instance the BatchNorm mean and standard deviation estimate, and one outlier can impact the entire batch. Outliers can also cause extreme gradients and throw off the optimization process. Interestingly, on Gwilliams2022, the drop is limited, potentially due to builtin preprocessing.

Clamping value	*Schoffelen2019*	*Gwilliams2022*	*Broderick2019*	*Brennan2019*	**Mean**
20	67.2 ± 0.09	72.5 ± 0.22	19.1 ± 1.15	31.4 ± 1.59	47.5
100	60.6 ± 2.38	72.4 ± 0.31	20.0 ± 0.54	31.5 ± 1.96	46.1
no clamping	1.1 ± 0.26	57.6 ± 13.39	4.0 ± 0.14	11.0 ± 1.92	18.4

A.3 Effect of the number of Mels

We now study the impact of the number of Mel bands. 120 bands is usually considered high enough for most practical use [Young et al., 2002], which we selected for the main evaluation in Table 2. We study the impact of the numer of Mel bands for different versions of the model. For clarity, we only provide the average top-10 accuracy overall datasets. We observe a small increase of the accuracy when using more Mel bands. Interestingly, when using the Deep Mel model, 20 bands is sufficient to achieve the best performance.

	# Mel bands			
Model value	20	40	80	120
Base model	9.5	10.0	10.3	10.5
+ CLIP	32.2	33.3	33.7	34.5
+ Deep Mel	40.7	40.6	40.3	40.6

[2] `https://data.donders.ru.nl/collections/di/dccn/DSC_3011220.01_297`

[3] `https://osf.io/rguwj/`

[4] `https://datadryad.org/stash/dataset/doi:10.5061/dryad.070jc`

[5] `https://deepblue.lib.umich.edu/data/concern/data_sets/bg257f92t`

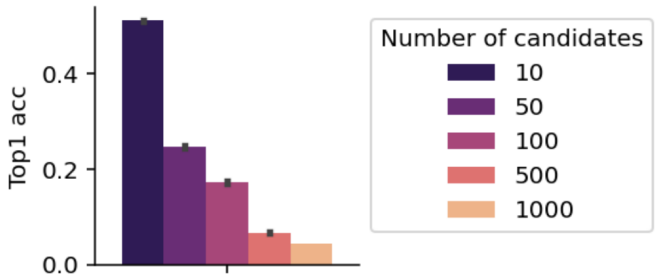

Figure A.1: Top-1 accuracy as a function of vocabulary size for word presented during random word lists in Schoffelen et al. [2019]. Error bar indicate SEM across participants.

A.4 Analyses of single-trial predictions

Does our model predict all words similarly? To address this question, we evaluate whether our model's ability to decode individual words depends on their properties, namely their zipf frequency as provided by Wordfreq [6], as well as their part-of-speech tag and their word embedding as provided by spaCy [7]. Similarly, we evaluate whether the decoding of the entire 3 s speech segment varies with its linguistic properties, as assessed by its average word embedding as well as its sentence embedding, as computed with Laser [8]. For this, we trained a regularized ridge regression with scikit-learn[9]'s default parameters to predict the softmax probability of the true word output by the decoder, given a feature. We then estimate the R^2 with a 5-split cross-validation: *i.e.* how well the feature predicts the probability of being selected by the decoder. The results, displayed in Figure 4-C, show that the word and segment embedding effectively explain the single-trial decoding accuracy. These results thus suggest that our decoder uses semantic and contextual information to make its predictions.

A.5 Decoding of isolated words

To what extent can our approach be used to decode words presented in isolation? To explore this issue, we evaluated our model using a subset from Schoffelen et al. [2019], where subjects are presented with random word lists. We use a segment ranging from -300 ms to +500 ms relative to word onset.

The results, displayed in Supplementary Figure A.1, show that our model reaches a top-1 accuracy of 25.0% with a vocabulary size of 50. While this performance is low, it is interesting to compare it to Moses et al. [2021] who report a top-1 accuracy of 39.5% with a model trained to decode the production of individual words, without the use of a language model, *i.e.* independently of context.

[6] https://pypi.org/project/wordfreq/
[7] https://spacy.io
[8] https://github.com/facebookresearch/laser
[9] https://scikit-learn.org