



HAL
open science

Decoding speech from non-invasive brain recordings

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

► **To cite this version:**

Alexandre Défossez, Charlotte Caucheteux, Jeremy Rapin, Ori Kabeli, Jean-Rémi King. Decoding speech from non-invasive brain recordings. 2022. hal-03808317

HAL Id: hal-03808317

<https://hal.science/hal-03808317>

Preprint submitted on 10 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Decoding speech from non-invasive brain recordings

Alexandre Défossez^{1,*}, Charlotte Caucheteux^{1,2}, Jérémy Rapin¹, Ori Kabeli¹, and Jean-Rémi King^{1,*}

¹Meta AI, ²Inria 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 \approx 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

21 potentially wearable setup [Boto et al., 2018]. However, these two devices produce notoriously noisy
 22 signals that vary greatly across sessions and across individuals [Schirrmester et al., 2017, King et al.,
 23 2018, Hämäläinen et al., 1993]. It is thus common to engineer pipelines that output hand-crafted
 24 features, which, in turn, can be learned by a decoder trained on a single participant [Lawhern et al.,
 25 2018, Lopopolo and van den Bosch, 2020, Chan et al., 2011, Nguyen et al., 2017].

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

30 Instead, we here propose to decode speech from non-invasive brain recordings by using (1) a single
 31 architecture trained across a large cohort of participants and (2) deep representations of speech learnt
 32 with self-supervised learning on a large quantity of speech data. For this, we introduce a convolutional
 33 neural network stacked onto a “Subject Layer” and trained with a contrastive objective to predict
 34 the representations of the audio waveform learnt by wav2vec 2.0 pretrained on 56k hours of speech
 35 [Baevski et al., 2020] (Figure 1). To validate our approach, we curate and integrate four public
 36 M/EEG datasets, encompassing the brain activity of 169 participants passively listening to sentences
 37 of short stories. With a sample of 3 seconds of M/EEG signals, our model identifies the matching
 38 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.

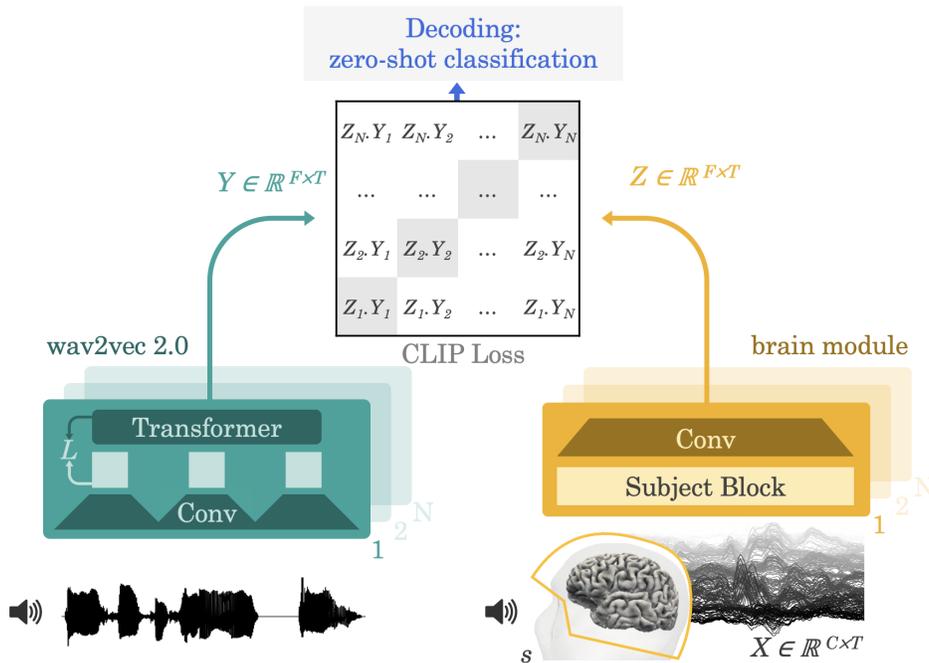


Figure 1: **Method** We aim to decode speech from the brain activity of healthy participants recorded with magnetoencephalography (MEG) or electroencephalography (EEG) while they listen to stories and/or sentences. For this, our model extracts the deep contextual representations of 3 s speech signals (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.

40 **2 Method**

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

43 **2.1 Problem formalization**

44 We aim to decode speech from a time series of high-dimensional brain signals recorded with
45 non-invasive magneto-encephalography (MEG) or electro-encephalography (EEG) while healthy
46 volunteers passively listened to spoken sentences in their native language. How spoken words are
47 represented in the brain is largely unknown [Hickok and Poeppel, 2007]. Thus, it is common to train
48 decoders in a supervised manner to predict a latent representation of speech known to be relevant to
49 the brain [Akbari et al., 2019, Angrick et al., 2019b,a, Krishna et al., 2020, Komeiji et al., 2022]. For
50 example, the Mel spectrogram is often targeted for neural decoding because it represents sounds
51 similarly to the cochlea [Mermelstein, 1976]. Let $X \in \mathbb{R}^{C \times T}$ be a segment of a brain recording
52 of a given subject while she listens to a speech segment of the same duration, with C the number
53 of M/EEG sensors and T the number of time steps. Let $Y \in \mathbb{R}^{F \times T}$ be the latent representation of
54 speech, using the same sample rate as X for simplicity, here the Mel spectrogram with F frequency
55 bands. Thus, supervised decoding consists of finding a decoding function: $f_{\text{reg}} : \mathbb{R}^{C \times T} \rightarrow \mathbb{R}^{F \times T}$
56 such that f_{reg} predicts Y given X . We denote by $\hat{Y} = f_{\text{reg}}(X)$ the representation of speech decoded
57 from the brain. When f_{reg} belongs to a parameterized family of models like deep neural networks, it
58 can be trained with a regression loss $L_{\text{reg}}(Y, \hat{Y})$ (e.g. the Mean Square Error),

$$\min_{f_{\text{reg}}} \sum_{X, Y} L_{\text{reg}}(Y, f_{\text{reg}}(X)). \quad (1)$$

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

63 **2.2 Model**

64 **2.2.1 Contrastive loss**

65 First, we reasoned that regression may be an ineffective loss because it departs from our objective:
66 decoding speech from brain activity. Consequently, we replaced it with a contrastive loss, namely,
67 the ‘‘CLIP’’ loss (originally for Contrastive Language-Image Pre-Training) by Radford et al. [2021],
68 which was originally designed to match latent representations in two modalities, text and images.
69 We implement the CLIP loss as follows: Let X be a brain recording segment and $Y \in \mathbb{R}^{F \times T}$
70 the latent representation of its corresponding sound (a.k.a ‘‘positive sample’’). We sample $N - 1$
71 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
72 want our model to predict the probabilities $\forall j \in \{1, \dots, N\}, p_j = \mathbb{P}[\bar{Y}_j = Y]$. We thus train a
73 model f_{clip} mapping the brain activity X to a latent representation $Z = f_{\text{clip}}(X) \in \mathbb{R}^{F \times T}$. The
74 estimated probability can then be approximated by the dot product of Z and the candidate speech
75 latent representations \bar{Y}_j , followed by a softmax:

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

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

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

79 Following [Radford et al., 2021], we use the other elements of the batch as negative samples at train
80 time. At test time, the negative samples correspond to all of the segments of the test but the positive
81 one.

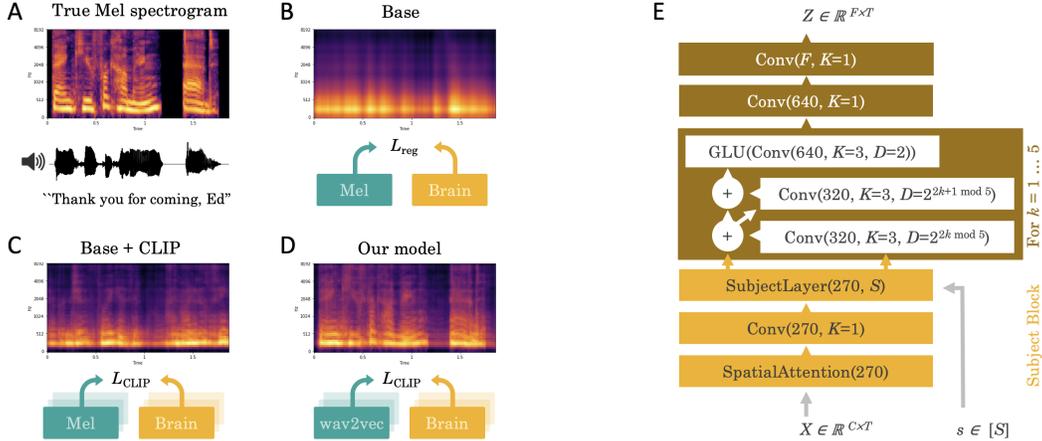


Figure 2: **Design choices.** **A.** Illustration of a 3 s speech sound segment (bottom) and its corresponding Mel spectrogram (top). **B.** Mel-spectrogram predicted with a direct regression loss L_{reg} of a brain decoder (orange). **C.** Replacing the regression loss with a CLIP loss [Radford et al.,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

82 2.2.2 Speech module

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

90 Wav2vec 2.0 is trained to transform the raw waveform with convolutional and transformer blocks to
 91 predict masked parts of its own latent representations. Baevski et al. [2020] showed that the resulting
 92 model can be efficiently fine-tuned to achieve state-of-the-art performance in speech recognition.
 93 Besides, this model effectively encodes a wide variety of linguistic features [Millet and Dunbar,
 94 2022, Adolfi et al., 2022]. Finally, recent work shows the existence of linear correspondence between
 95 the activations of the brain and those of wav2vec 2.0 [Millet et al., 2022, Vaidya et al., 2022].
 96 Consequently, we here test whether this model effectively helps the present decoding task. In practice,
 97 we use the wav2vec2-large-xlsr-53¹, which has been pre-trained on 56k hours of speech from
 98 53 different languages.

99 2.2.3 Brain module

100 Finally, for the brain module, we use a deep neural network f_{clip} , input with raw M/EEG times series
 101 X and a one-hot-encoding of the corresponding subject s , and outputs the latent brain representation
 102 Z , with the same sample rate as X . This architecture consists of (1) a spatial attention layer over the
 103 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.

| Dataset | Lang. | Type | # Chs. | # Subj. | Total duration | Train set | | Test set | |
|----------------|---------|------|--------|---------|----------------|------------|--------|------------|--------|
| | | | | | | # 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 |

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

107 **Spatial attention and subject layer.** The brain data is first remapped onto $D_1 = 270$ channels
 108 with a spatial attention layer based on the location of the sensors. The 3D sensor locations are first
 109 projected on a 2D plane obtained with the MNE-Python function `find_layout` [Gramfort et al.,
 110 2013], which uses a device-dependent surface designed to preserve the channel distances. Their 2D
 111 positions are finally normalized to $[0, 1]$. For each output channel, a function over $[0, 1]^2$ is learnt,
 112 parameterized in the Fourier space. The weights over the input sensors is then given by the softmax
 113 of the function evaluated at the sensor locations. Formally, each input channel i has a location (x_i, y_i)
 114 and each output channel j is attached a function a_j over $[0, 1]^2$ parameterized in the Fourier space as
 115 $z_j \in \mathbb{C}^{K \times K}$ with $K=32$ harmonics along each axis, $i.e.$

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

116 The output is given by a softmax attention based on the evaluation of a_j at each input position (x_i, y_i) :
 117

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

118 with SA the spatial attention. In practice, as a_j is periodic, we scale down (x, y) to keep a margin of
 119 0.1 on each side. We then apply a spatial dropout by sampling a location $(x_{\text{drop}}, y_{\text{drop}})$ and removing
 120 from the softmax each sensor that is within a distance of d_{drop} of the sampled location. We then add
 121 a 1x1 convolution (i.e. with a kernel size of 1) without activation and with the same number D_1 of
 122 output channels. Finally, to leverage inter-subject variability, we learn a matrix $M_s \in \mathbb{R}^{D_1, D_1}$ for
 123 each subject $s \in [S]$ and apply it after the spatial attention layer along the channel dimension. This
 124 is similar but more expressive than the subject embedding used by Chehab et al. [2021] for MEG
 125 encoding, and follows decade of research on subject alignment [Xu et al., 2012, Haxby et al., 2020].

126 **Residual dilated convolutions.** We then apply a stack of five blocks of three convolutional
 127 layers. For the k -th block, the first two convolutions are applied with residual skip connections
 128 (except for the very first one where the number of dimension potentially doesn't match), outputs
 129 $D_2 = 320$ channels and are followed by batch normalization [Ioffe and Szegedy, 2015] and a GELU
 130 activation [Hendrycks and Gimpel, 2016]. The two convolutions are also dilated to increase their
 131 receptive field, respectively by $2^{2k \bmod 5}$ and $2^{2k+1 \bmod 5}$ (with k zero indexed). The third layer in
 132 a block outputs $2D_2$ channels and uses a GLU activation [Dauphin et al., 2017] which halves the
 133 number of channels. All convolutions use a kernel size of 3 over the time axis, a stride of 1, and
 134 sufficient padding to keep the number of time steps constant across layers. The output of the model
 135 is obtained by applying two final 1x1 convolutions: first with $2D_2$ outputs, followed by a GELU,
 136 and finally with F channels as output, thus matching the dimensionality of speech representations.
 137 Given the expected delay between a stimulus and its corresponding brain responses, we further shift
 138 the input brain signal by 150 ms into the future to facilitate the alignment between Y and Z .

139 2.3 Datasets

140 We test our approach on four public datasets, two based on MEG recordings and two on EEG.
 141 All datasets and their corresponding studies were approved by the relevant ethics committee and
 142 are publicly available for fundamental research purposes. Informed consent was obtained from all
 143 human research participants. We provide an overview of the main characteristics of the datasets

144 on Table 1, including the number of train and test segments and vocabulary size over both splits.
145 For all datasets, healthy adult volunteers passively listened to speech sounds (accompanied with
146 some memory or comprehension questions to ensure participants were attentive), while their brain
147 activity was recorded with MEG or EEG. In Schoffelen et al. [2019], Dutch-speaking participants
148 listened to decontextualized Dutch sentences and word lists (Dutch sentences for which the words
149 are randomly shuffled). The study was approved by the local ethics committee (CMO – the local
150 “Committee on Research Involving Human Subjects” in the Arnhem-Nijmegen region). In Gwilliams
151 et al. [2022], English-speaking participants listened to four fictional stories from the Masc corpus
152 [Ide et al., 2010] in two identical sessions of one hour [Gwilliams et al., 2020]. The study was
153 approved by the Institution Review Board (IRB) ethics committee of New York University Abu
154 Dhabi. In Broderick et al. [2018], English-speaking participants listened to extracts of “The old
155 man and the sea”. The study was approved by the Ethics Committees of the School of Psychology
156 at Trinity College Dublin, and the Health Sciences Faculty at Trinity College Dublin. In Brennan
157 and Hale [2019], English-speaking participants listened to a chapter of “Alice in Wonderland”. See
158 Section A.1 in the Appendix for more details. The study was approved by the University of Michigan
159 Health Sciences and Behavioral Sciences Institutional Review Board (HUM00081060).

160 2.4 Preprocessing

161 M/EEG is generally considered to capture neural signals from relatively low frequency ranges
162 [Hämäläinen et al., 1993]. Consequently, we first resampled all brain recordings down to 120 Hz with
163 Torchaudio [Yang et al., 2021] and then split the data into *training*, *validation*, and *testing* splits with
164 a size roughly proportional to 70%, 20%, and 10%. We define a “sample” as a 3 s window of brain
165 recording with its associated speech representation. A “segment” is a *unique* 3 s window of speech
166 sound. As the same segment can be presented to multiple subjects (or even within the same subject in
167 Gwilliams et al. [2022]), the splits are defined so that one segment is always assigned to the same split
168 across repetitions. We ensure that there is no identical sentences across splits, and checked that each
169 sentence was pronounced by a unique speaker. Furthermore, we exclude all segments overlapping
170 over different splits. For clarity, we restrict the test segments to those that contain a word at a fixed
171 location (here 500 ms into the sample).

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

187 2.5 Training

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

192 2.6 Evaluation

193 **Segment-level evaluation.** In Figure 2, we estimate the Mel spectrogram from the model output.
194 Given a segment and its matching audio (here the sentence “Thank you for coming Ed”), we retrieve
195 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 | <i>Schoffelen2019</i> | <i>Gwilliams2022</i> | <i>Broderick2019</i> | <i>Brennan2019</i> | Mean |
|---------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-------------|
| Random model | 1.5 \pm 0.18 | 2.2 \pm 0.16 | 4.1 \pm 0.09 | 7.6 \pm 0.13 | 3.8 |
| Base model | 19.3 \pm 0.83 | 14.9 \pm 0.56 | 1.3 \pm 0.19 | 6.6 \pm 0.53 | 10.5 |
| + CLIP | 51.5 \pm 0.47 | 58.6 \pm 0.28 | 13.3 \pm 0.54 | 14.5 \pm 1.33 | 34.5 |
| + Deep Mel | 57.7 \pm 0.16 | 64.4 \pm 1.67 | 16.5 \pm 0.26 | 23.7 \pm 0.90 | 40.6 |
| + wav2vec 2.0 | 67.2 \pm 0.09 | 72.5 \pm 0.22 | 19.1 \pm 1.15 | 31.4 \pm 1.59 | 47.5 |

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

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

203 2.7 Code availability

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

205 3 Results

206 3.1 Accurately decoding speech from M/EEG recordings

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

214 3.2 Effect of contrastive loss, deep speech representations, and number of participants

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

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

224 Third, to test whether our model effectively leverage the inter-individual variability, we trained it
 225 on a variable number of subjects and computed its accuracy on the first 10% of subjects. As shown
 226 in Figure 3B, decoding performance increases as the model is trained with more subjects on the
 227 two MEG datasets. This ability to learn from multiple subjects is strengthened by another ablation
 228 experiment: training on all participants, but *without* the subject-specific layer, leads to a drop of 17%
 229 accuracy on average across the four datasets (Table 3). However, this last gain is relatively modest
 230 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 \pm 0.09 | 72.5 \pm 0.22 | 19.1 \pm 1.15 | 31.4 \pm 1.59 | 47.5 | 0.00 |
| \wo spatial attention dropout | 61.6 \pm 0.14 | 71.2 \pm 0.93 | 19.0 \pm 1.07 | 30.2 \pm 1.70 | 45.5 | -2.00 |
| \wo subj. embedding* | 59.5 \pm 0.24 | 72.0 \pm 0.77 | 20.2 \pm 1.24 | 30.2 \pm 0.77 | 45.4 | -2.08 |
| \wo GELU, \wo ReLU | 61.4 \pm 0.67 | 72.2 \pm 0.05 | 19.2 \pm 0.79 | 26.4 \pm 1.03 | 44.8 | -2.72 |
| \wo spatial attention | 60.0 \pm 1.32 | 69.5 \pm 0.44 | 17.9 \pm 0.34 | 26.0 \pm 0.61 | 43.3 | -4.18 |
| \wo final convs | 62.3 \pm 0.07 | 71.0 \pm 0.47 | 15.7 \pm 1.13 | 22.7 \pm 2.05 | 43.0 | -4.57 |
| \wo initial 1x1 conv. | 57.8 \pm 1.12 | 69.6 \pm 0.37 | 17.9 \pm 0.31 | 26.3 \pm 1.43 | 42.9 | -4.62 |
| \wo skip connections | 59.2 \pm 0.71 | 68.0 \pm 0.60 | 16.7 \pm 0.29 | 25.7 \pm 4.32 | 42.4 | -5.13 |
| \wo non-residual GLU conv. | 63.5 \pm 0.68 | 73.0 \pm 0.67 | 17.0 \pm 0.03 | 6.70 \pm 0.57 | 40.0 | -7.50 |
| \wo subject-specific layer | 38.3 \pm 0.77 | 49.2 \pm 0.23 | 11.8 \pm 0.14 | 21.5 \pm 0.59 | 30.2 | -17.30 |
| \wo clamping brain signal | 1.1 \pm 0.26 | 57.6 \pm 13.4 | 4.0 \pm 0.14 | 11.0 \pm 1.92 | 18.4 | -29.10 |

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

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

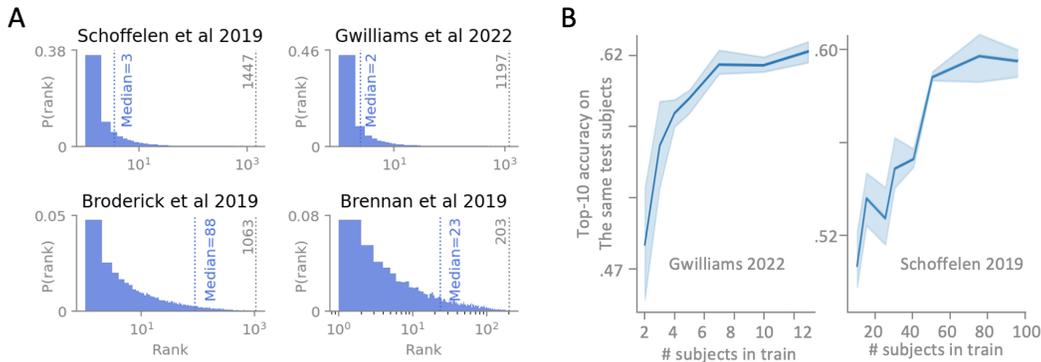


Figure 3: **Segment-level decoding.** **A.** Probability distribution of the decoded rank for each segment (lower is better) for each dataset. The gray dotted line indicates the number of segments in the test set. **B.** Top-10 accuracy obtained for the first 10% of subjects (y-axis) as a function of 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.

236 4 Discussion

237 Here, we aimed to decode the perception of natural speech from non-invasive brain recordings.
 238 Our results, based on the largest decoding study of M/EEG responses to speech to date, show that
 239 combining (1) a contrastive objective, (2) a convolutional architecture enhanced by a ‘‘Subject Layer’’,
 240 and (3) pretrained speech representations allows the decoding of new 3 s speech sounds with up to
 241 44% top-1 accuracy out of more than 1,500 possibilities.

242 Our approach contributes to the rapid transformation of hand-crafted pipelines into their end-to-end
 243 counterparts. In particular, this study shows how self-supervised and contrastive learning can improve
 244 both (1) the analyses of brain signals and (2) the definition of the linguistic features that should
 245 be used for decoding. In particular, previous models were typically trained on individual subjects
 246 to categorize a very small number of highly-repeated categories and/or hand-crafted features [Ali
 247 et al., 2022, Jayaram and Barachant, 2018, Lawhern et al., 2018]. For example, Sun and Qin [2016],
 248 Sree and Kavitha [2017], and Moindreau et al. [2018] all developed a decoder to classify 11, 5 and
 249 2 distinct imagined phonemes, respectively, from EEG signals. Similarly, Lopopolo and van den
 250 Bosch [2020], Chan et al. [2011], Nguyen et al. [2017] respectively developed a decoder to classify

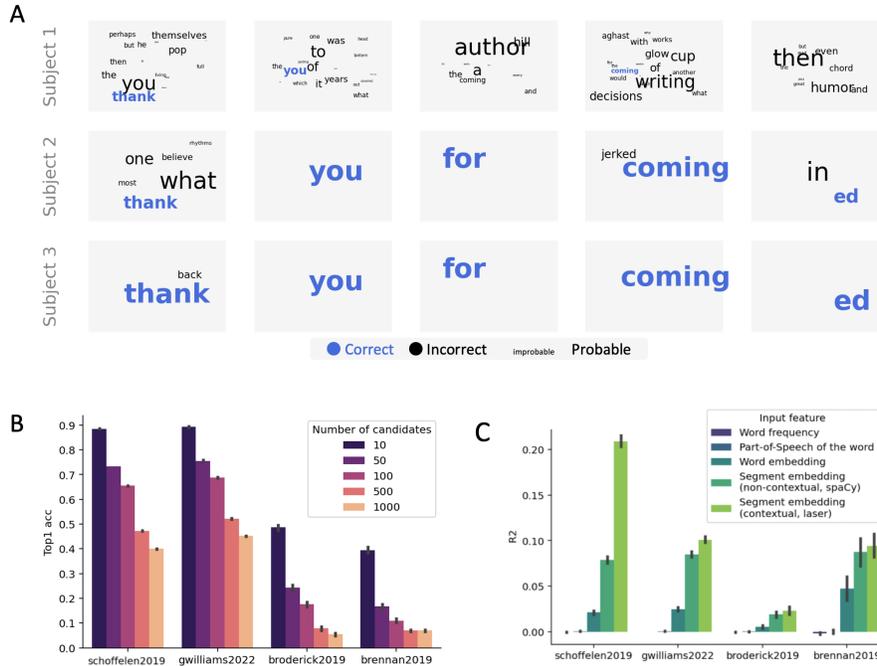


Figure 4: **A.** Single-word prediction for the first three subjects of Gwilliams et al. [2022] listening to the sentence “Thank you for coming, Ed”. Text color indicates whether the decoded word is accurate. Text size is proportional to the log-probability output by our model. **B.** Top-1 accuracy 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.

251 6 distinct part-of-speech (48% accuracy), 10 words (83% accuracy) and 3 words (70% accuracy),
 252 from MEG signals. Finally, both Dash et al. [2020] and Wang et al. [2017] trained a classifier to
 253 decode 5 distinct sentences from MEG activity (both around 94% accuracy). Given the combinatorics
 254 of language, such single-subject / limited-vocabulary approach necessarily limits the possibility to
 255 decode natural speech. By contrast, our model effectively achieves “zero-shot” decoding by matching
 256 a large number of brain recordings to the deep representations of their corresponding speech sounds.

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

268 The present non-invasive study is limited to speech *perception*. It thus differs from the recent
 269 achievements obtained in a small set of heavily-trained patients implanted for clinical purposes and
 270 tasked to produce language [Herff et al., 2015, Martin et al., 2016, Angrick et al., 2019b, Willett et al.,
 271 2021, Moses et al., 2021, Angrick et al., 2021, Kohler et al., 2021]. In particular, Willett et al. [2021]
 272 showed that a 1 s time window of neuronal activity in the motor cortex suffices to decode one of 26
 273 characters with 94.1% top-1 accuracy during a spelling task. Similarly, Moses et al. [2021] showed
 274 that 4 s of neuronal activity recorded in the sensory-motor cortices is sufficient to decode the intention
 275 to communicate one of 50 words with a median word error rate of 25.6%. We deliberately chose to

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

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

291 **References**

- 292 Federico Adolfi, Jeffrey S Bowers, and David Poeppel. Successes and critical failures of neural
293 networks in capturing human-like speech recognition. *arXiv preprint arXiv:2204.03740*, 2022.
- 294 Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani.
295 Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1):
296 1–12, 2019.
- 297 Omair Ali, Muhammad Saif-ur Rehman, Susanne Dyck, Tobias Glasmachers, Ioannis Iossifidis,
298 and Christian Klaes. Enhancing the decoding accuracy of eeg signals by the introduction of
299 anchored-stft and adversarial data augmentation method. *Scientific reports*, 12(1):1–19, 2022.
- 300 Miguel Angrick, Christian Herff, Garrett Johnson, Jerry Shih, Dean Krusienski, and Tanja Schultz.
301 Interpretation of convolutional neural networks for speech spectrogram regression from intracranial
302 recordings. *Neurocomputing*, 342:145–151, 2019a.
- 303 Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski,
304 and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural
305 networks. *Journal of neural engineering*, 16(3):036019, 2019b.
- 306 Miguel Angrick, Maarten C Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, Sophocles
307 Goulis, Jeremy Saal, Albert J Colon, Louis Wagner, Dean J Krusienski, et al. Real-time synthesis of
308 imagined speech processes from minimally invasive recordings of neural activity. *Communications
309 biology*, 4(1):1–10, 2021.
- 310 Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural
311 decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.
- 312 Alexei Baeviski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework
313 for self-supervised learning of speech representations. *Advances in Neural Information Processing
314 Systems*, 33:12449–12460, 2020.
- 315 Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea
316 Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed.
317 *Nature*, 398(6725):297–298, 1999.
- 318 Elena Boto, Niall Holmes, James Leggett, Gillian Roberts, Vishal Shah, Sofie S Meyer,
319 Leonardo Duque Muñoz, Karen J Mullinger, Tim M Tierney, Sven Bestmann, et al. Moving
320 magnetoencephalography towards real-world applications with a wearable system. *Nature*, 555
321 (7698):657–661, 2018.
- 322 Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions
323 during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.
- 324 Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C
325 Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of
326 natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.
- 327 Jonathan S Brumberg, Philip R Kennedy, and Frank H Guenther. Artificial speech synthesizer
328 control by brain-computer interface. In *Tenth Annual Conference of the International Speech
329 Communication Association*, 2009.
- 330 Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and
331 category-specific spatiotemporal representations from meg and eeg. *Neuroimage*, 54(4):3028–3039,
332 2011.
- 333 Omar Chehab, Alexandre Defossez, Jean-Christophe Loiseau, Alexandre Gramfort, and Jean-Rémi
334 King. Deep recurrent encoder: A scalable end-to-end network to model brain signals. *arXiv
335 preprint arXiv:2103.02339*, 2021.
- 336 Jan Claassen, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez,
337 Joshua U Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, et al. Detection of brain
338 activation in unresponsive patients with acute brain injury. *New England Journal of Medicine*, 380
339 (26):2497–2505, 2019.

- 340 Damian Cruse, Srivas Chennu, Camille Chatelle, Tristan A Bekinschtein, Davinia Fernández-Espejo,
341 John D Pickard, Steven Laureys, and Adrian M Owen. Bedside detection of awareness in the
342 vegetative state: a cohort study. *The Lancet*, 378(9809):2088–2094, 2011.
- 343 Debadatta Dash, Paul Ferrari, and Jun Wang. Decoding imagined and spoken phrases from non-
344 invasive neural (meg) signals. *Frontiers in neuroscience*, 14:290, 2020.
- 345 Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated
346 convolutional networks. In *Proceedings of the International Conference on Machine Learning*,
347 2017.
- 348 Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian
349 Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data
350 analysis with mne-python. *Frontiers in neuroscience*, page 267, 2013.
- 351 Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme
352 sequencing in real speech jointly encode order and invariant content. *bioRxiv*, 2020.
- 353 Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pyllkanen, David Poeppel, and Jean-Remi
354 King. Meg-masc: a high-quality magneto-encephalography dataset for evaluating natural speech
355 processing. *arXiv preprint arXiv:2208.11488*, 2022.
- 356 Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Mag-
357 netoencephalography—theory, instrumentation, and applications to noninvasive studies of the
358 working human brain. *Reviews of modern Physics*, 65(2):413, 1993.
- 359 James V Haxby, J Swaroop Guntupalli, Samuel A Nastase, and Ma Feilong. Hyperalignment:
360 Modeling shared information encoded in idiosyncratic cortical topographies. *Elife*, 9:e56601,
361 2020.
- 362 Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint*
363 *arXiv:1606.08415*, 2016.
- 364 Christian Herff, Dominic Heger, Adriana De Pestere, Dominic Telaar, Peter Brunner, Gerwin Schalk,
365 and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain.
366 *Frontiers in neuroscience*, 9:217, 2015.
- 367 Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews*
368 *neuroscience*, 8(5):393–402, 2007.
- 369 Nancy Ide, Collin F Baker, Christiane Fellbaum, and Rebecca J Passonneau. The manually annotated
370 sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010*
371 *conference short papers*, pages 68–73, 2010.
- 372 Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by
373 reducing internal covariate shift. Technical Report 1502.03167, arXiv, 2015.
- 374 Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis.
375 *Journal of neural engineering*, 15(6):066011, 2018.
- 376 Philip Kennedy, A Ganesh, and AJ Cervantes. Slow firing single units are essential for optimal
377 decoding of silent speech. 2022.
- 378 Jean-Rémi King, Frédéric Fugeras, Alexandre Gramfort, Aaron Schurger, Imen El Karoui, JD Sitt,
379 Benjamin Rohaut, C Wacongne, E Labyt, Tristan Bekinschtein, et al. Single-trial decoding of
380 auditory novelty responses facilitates the detection of residual consciousness. *Neuroimage*, 83:
381 726–738, 2013.
- 382 Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis
383 Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding neuronal dynamics:
384 Methodological framework to uncover the algorithms of cognition. 2018.
- 385 Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International*
386 *Conference on Learning Representations*, 12 2014.

- 387 Jonas Kohler, Maarten C Ottenhoff, Sophocles Goulis, Miguel Angrick, Albert J Colon, Louis Wagner,
388 Simon Tousseyn, Pieter L Kubben, and Christian Herff. Synthesizing speech from intracranial
389 depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*, 2021.
- 390 Shuji Komeiji, Kai Shigemi, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano,
391 Koichi Shinoda, and Toshihisa Tanaka. Transformer-based estimation of spoken sentences using
392 electrocorticography. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech
393 and Signal Processing (ICASSP)*, pages 1311–1315. IEEE, 2022.
- 394 Gautam Krishna, Co Tran, Yan Han, Mason Carnahan, and Ahmed H Tewfik. Speech synthesis
395 using eeg. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal
396 Processing (ICASSP)*, pages 1235–1238. IEEE, 2020.
- 397 Andrea Kübler, Boris Kotchoubey, Jochen Kaiser, Jonathan R Wolpaw, and Niels Birbaumer. Brain-
398 computer communication: Unlocking the locked in. *Psychological bulletin*, 127(3):358, 2001.
- 399 Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and
400 Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain-computer
401 interfaces. *Journal of neural engineering*, 15(5):056013, 2018.
- 402 Alessandro Lopopolo and Antal van den Bosch. Part-of-speech classification from magnetoen-
403 cephalography data using 1-dimensional convolutional neural network. 2020.
- 404 Stephanie Martin, Peter Brunner, Iñaki Iturrate, José del R Millán, Gerwin Schalk, Robert T Knight,
405 and Brian N Pasley. Word pair classification during imagined speech using direct brain recordings.
406 *Scientific reports*, 6(1):1–12, 2016.
- 407 Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern
408 recognition and artificial intelligence*, 116:374–388, 1976.
- 409 Juliette Millet and Ewan Dunbar. Do self-supervised speech models develop human-like perception
410 biases? *arXiv preprint arXiv:2205.15819*, 2022.
- 411 Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan
412 Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in
413 the brain with self-supervised learning. *arXiv preprint arXiv:2206.01685*, 2022.
- 414 Marc-Antoine Moинnereau, Thomas Brienne, Simon Brodeur, Jean Rouat, Kevin Whittingstall, and
415 Eric Plourde. Classification of auditory stimuli from eeg signals with a regulated recurrent neural
416 network reservoir. *arXiv preprint arXiv:1804.10322*, 2018.
- 417 David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F
418 Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthe-
419 sis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*,
420 385(3):217–227, 2021.
- 421 Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. Inferring imagined speech using
422 eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*,
423 15(1):016002, 2017.
- 424 Adrian M Owen, Martin R Coleman, Melanie Boly, Matthew H Davis, Steven Laureys, and John D
425 Pickard. Detecting awareness in the vegetative state. *science*, 313(5792):1402–1402, 2006.
- 426 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier
427 Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn:
428 Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- 429 Xiaomei Pei, Dennis L Barbour, Eric C Leuthardt, and Gerwin Schalk. Decoding vowels and
430 consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal
431 of neural engineering*, 8(4):046028, 2011.
- 432 Elmar GM Pels, Erik J Aarnoutse, Nick F Ramsey, and Mariska J Vansteensel. Estimated preva-
433 lence of the target population for brain-computer interface neurotechnology in the netherlands.
434 *Neurorehabilitation and neural repair*, 31(7):677–685, 2017.

- 435 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
436 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
437 models from natural language supervision. In *International Conference on Machine Learning*,
438 pages 8748–8763. PMLR, 2021.
- 439 Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin
440 Glasstetter, Katharina Eggenberger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and
441 Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization.
442 *Human brain mapping*, 38(11):5391–5420, 2017.
- 443 Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche H. L. Lam, Julia Uddén, Annika Hultén, and
444 Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing.
445 *Scientific Data*, 6(1):17, April 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0020-y.
- 446 R Anandha Sree and A Kavitha. Vowel classification from imagined speech using sub-band eeg
447 frequencies and deep belief networks. In *2017 fourth international conference on signal processing,
448 communication and networking (ICSCN)*, pages 1–4. IEEE, 2017.
- 449 Carol A Stanger and Michael F Cawley. Demographics of rehabilitation robotics users. *Technology
450 and Disability*, 5(2):125–137, 1996.
- 451 Sergey D Stavisky, Paymon Rezaii, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy,
452 and Jaimie M Henderson. Decoding speech from intracortical multielectrode arrays in dorsal
453 “arm/hand areas” of human motor cortex. In *2018 40th Annual International Conference of the
454 IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 93–97. IEEE, 2018.
- 455 Pengfei Sun and Jun Qin. Neural networks based eeg-speech models. *arXiv preprint
456 arXiv:1612.05369*, 2016.
- 457 Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively
458 explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*, 2022.
- 459 Jun Wang, Myungjong Kim, Angel W Hernandez-Mulero, Daragh Heitzman, and Paul Ferrari.
460 Towards decoding speech production from single-trial magnetoencephalography (meg) signals. In
461 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages
462 3036–3040. IEEE, 2017.
- 463 Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V
464 Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):
465 249–254, 2021.
- 466 Hao Xu, Alexander Lorbert, Peter J Ramadge, J Swaroop Guntupalli, and James V Haxby. Regularized
467 hyperalignment of multi-set fmri data. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*,
468 pages 229–232. IEEE, 2012.
- 469 Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen,
470 Ching-Feng Yeh, Christian Puhersch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z.
471 Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean
472 Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi.
473 TorchAudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*,
474 2021.
- 475 Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth
476 Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university
477 engineering department*, 3(175):12, 2002.

478 A Appendix

479 A.1 Datasets

480 The data from Schoffelen et al. [2019] was provided (in part) by the Donders Institute for Brain,
481 Cognition and Behaviour with a “RU-DI-HD-1.0” licence². The data for Gwilliams et al. [2022] is
482 available under CC0 1.0 Universal³. The data for Broderick et al. [2018] is available under the same
483 licence⁴. Finally, the data from Brennan and Hale [2019] is available under the CC BY 4.0 licence⁵
484 All audio files were provided by the authors of each dataset.

485 A.2 Extra Results

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

488 A.2.1 Effect of clamping

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

| | Clamping value | <i>Schoffelen2019</i> | <i>Gwilliams2022</i> | <i>Broderick2019</i> | <i>Brennan2019</i> | Mean |
|-----|-----------------------|-----------------------|----------------------|----------------------|--------------------|-------------|
| 501 | 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 |

502 A.3 Effect of the number of Mels

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

| | | # Mel bands | | | |
|-----|--------------------|--------------------|------|------|------|
| | Model value | 20 | 40 | 80 | 120 |
| 509 | 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 |

²https://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>

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

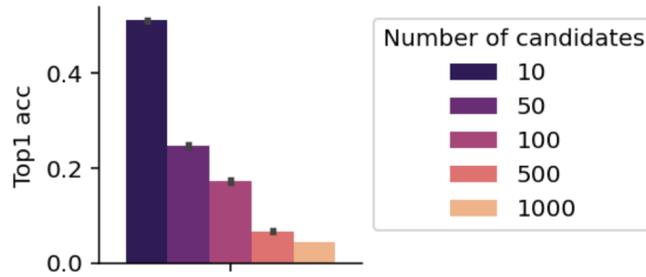


Figure A.1: Top-1 accuracy as a function of vocabulary size for word presented during random word lists in Schoffelen et al. [2019]. Error bar indicate SEM across participants.

510 A.4 Analyses of single-trial predictions

511 Does our model predict all words similarly? To address this question, we evaluate whether our
 512 model’s ability to decode individual words depends on their properties, namely their zipf frequency
 513 as provided by Wordfreq⁶, as well as their part-of-speech tag and their word embedding as provided
 514 by spaCy⁷. Similarly, we evaluate whether the decoding of the entire 3 s speech segment varies with
 515 its linguistic properties, as assessed by its average word embedding as well as its sentence embedding,
 516 as computed with Laser⁸. For this, we trained a regularized ridge regression with scikit-learn⁹’s
 517 default parameters to predict the softmax probability of the true word output by the decoder, given a
 518 feature. We then estimate the R^2 with a 5-split cross-validation: *i.e.* how well the feature predicts the
 519 probability of being selected by the decoder. The results, displayed in Figure 4-C, show that the word
 520 and segment embedding effectively explain the single-trial decoding accuracy. These results thus
 521 suggest that our decoder uses semantic and contextual information to make its predictions.

522 A.5 Decoding of isolated words

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

526 The results, displayed in Supplementary Figure A.1, show that our model reaches a top-1 accuracy
 527 of 25.0% with a vocabulary size of 50. While this performance is low, it is interesting to compare
 528 it to Moses et al. [2021] who report a top-1 accuracy of 39.5% with a model trained to decode the
 529 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>