# 000 THE BRAIN'S BITTER LESSON: SCALING SPEECH DECODING WITH SELF-SUPERVISED LEARNING

Anonymous authors

001

002 003 004

010 011

012

013

014

015

016

017

018

019

021

023

025

026

027

028

029

031 032

034

040

041

042 043

044

Paper under double-blind review

### ABSTRACT

The past few years have produced a series of spectacular advances in the decoding of speech from brain activity. The engine of these advances has been the acquisition of labelled data, with increasingly large datasets acquired from single subjects. However, participants exhibit individual differences, such as anatomy, and datasets use varied scanners and task designs. As a result, prior work has struggled to leverage data from multiple subjects, multiple datasets, multiple tasks, and unlabelled datasets. In turn, the field has not benefited from the rapidly growing number of open neural data repositories to exploit large-scale data and deep learning. This gap exists for all neural data, but especially for magnetoencephalography (MEG), where the scale of individual datasets has not yet caught up with other modalities. To address this, we develop a set of neuroscience-inspired selfsupervised objectives, together with a neural architecture, for representation learning from heterogeneous and unlabelled neural recordings. Experimental results with MEG show that representations learned with these objectives scale with data, generalise across subjects, datasets, and tasks, outperform using the raw input representation, and even surpass comparable self-supervised approaches. In addition, we set new benchmarks for two foundational speech decoding tasks. Collectively, these methods now unlock the potential for training speech decoding models with orders of magnitude more existing data.





Figure 1: Leveraging unlabelled data using pretext tasks for speech decoding. We pre-train a 048 neural network using tasks that generate implicit labels from abundant unlabelled MEG neuroimaging data, permitting learning from large heterogeneous datasets. The tasks apply a randomly selected neuroscientifically relevant transformation T to the data and the network predicts the transformation. 051 We then train a linear probe on top of the pre-trained model, which remains frozen, with labelled 052 data, achieving superior generalisation (cf. raw inputs) owing to the strength of the representation.

# 1 INTRODUCTION

055

In his *Bitter Lesson*, Richard Sutton argues that a major conclusion of 70 years of AI research is that general methods exploiting large-scale computation will outperform model-based approaches as the availability of compute increases (Sutton, 2019). In line with this, the generality of deep learning, via statistical learning from ever bigger datasets, has allowed the field to leverage computation in a way that appears to scale arbitrarily, leading to astounding advances across a diverse set of domains (Jumper et al., 2021; Caron et al., 2021; OpenAI, 2023; Radford et al., 2023).

062 In the domain of brain data, and of tasks like speech decoding, the bitter lesson has not yet been fully 063 assimilated. State-of-the-art brain-computer interfaces (BCIs) have tried to scale up labelled datasets 064 for individual subjects, using either invasive (Moses et al., 2021; Willett et al., 2023) or non-invasive 065 brain recordings (Tang et al., 2023), mapping these to transcripts of attempted or imagined speech. 066 Yet, a number of obstacles to scale remain. With few exceptions at present, e.g. Défossez et al. 067 (2023), speech decoding models tend not to train on data from more than one subject. Moreover, they do not combine data from multiple datasets and in general do not utilise unlabelled data, or data 068 from diverse tasks. Thus the size of training data has been limited to how much can be acquired for a 069 single subject, and data from other subjects, or from the growing number of public data repositories, has not been leveraged. There are many reasons for these limitations; individual brains and data 071 from different neuroimaging scanners differ, for example. But overcoming these limitations, as has 072 begun to happen in neighbouring sub-fields, such as Jiang et al. (2024), holds the promise of training 073 models on collective, internet-scale data. 074

While neuroimaging modalities such as electroencephalography (EEG) are more abundant, MEG 075 may be a better modality for decoding as it provides a richer signal (Lopes da Silva, 2013; Hall 076 et al., 2014). Given the scarcity of speech-labelled MEG data and the relative abundance of other 077 MEG data, self-supervised learning (SSL) appears promising as it is an avenue for domains where labels are rare or hard to obtain (Balestriero et al., 2023). But the scale of public MEG data, while 079 large, is still not at the volume of breakthroughs in self-supervised image and natural language processing, let alone EEG. Thus, SSL methods for MEG need to be highly data-efficient. Pretext tasks 081 are one such method in which domain-specific self-supervised tasks are used to pre-train a model on unlabelled data by generating implicit training labels through transformations of the input in order to 083 help a downstream task. We develop a set of these tasks, informed by advances in neuroscience, for 084 learning with unlabelled brain data (Figure 1) and design an architecture for processing continuous multi-sensor neuroimaging signals which we train using our pretext tasks. In order to scale exist-085 ing non-invasive datasets, we provide a unified method that allows us to leverage data from other experiments that do not have the same labels (by treating them as unlabelled) and that come from 087 different subjects and neuroimaging scanners. We evaluate the representations learned with our ap-088 proach on heard speech datasets acquired with non-invasive MEG, setting the baselines for speech detection and voicing classification on this data. The results not only demonstrate that scaling with 090 unlabelled data works in speech decoding, but also shows that these representations can generalise 091 across datasets, tasks, and even novel subjects for the first time. Our main contributions are: 092

- A set of domain-specific **self-supervised pretext tasks** for representation learning that can scale speech decoding over multiple subjects, multiple studies, and unlabelled data;
- A data-efficient **neural architecture** for learning these self-supervised objectives and training downstream speech decoding from brain data; and
- A comprehensive **experimental evaluation**, using multiple times the volume of data in prior work, that verifies the above claims and additionally provides evidence for the existence of **scaling laws** when pre-training models with unlabelled MEG recordings.
- 2 RELATED WORK
- 102 103

093

094

095

096

098

099

100 101

2

Prior work in speech decoding has focused almost entirely on supervised learning with decoding
models that typically do not generalise across participants or experiments. This is true both in
recent state-of-the-art invasive studies (Moses et al., 2021; Metzger et al., 2023; Willett et al., 2023;
Chen et al., 2024a) and non-invasive studies (Tang et al., 2023). These prior works have scaled
up the experimental data collected within individual subjects, but are unable to leverage data from

other subjects and experiments. Focusing on semantic rather than phonetic decoding, the method developed by Tang et al. (2023) is remarkable for showing an ability to generalise across labelled task data when listening to speech, imagining speech, or even watching videos. They do not, however, leverage unlabelled data and are unable to show generalisation between subjects.

112 Specific studies into the limitations of generalising models between subjects show that while perfor-113 mance decreases on average when subjects are pooled, there are exceptions (e.g. Anumanchipalli 114 et al. (2019) and Makin et al. (2019) in surgical settings and Csaky et al. (2022) non-invasively). 115 Exploiting audio data in a multi-modal framework, Défossez et al. (2023) show that decoding per-116 formance improves for a segment identification task as data from multiple subjects listening to con-117 nected speech are aggregated. However, they do not demonstrate the ability to generalise to novel 118 subjects and must retrain their model for new datasets. Moreover, although they repeat the result within two MEG and two EEG datasets, Défossez et al. (2023) do not show any improvements for 119 pooling data across datasets. Their method is also unable to incorporate data without corresponding 120 audio labels and so they do not combine data from studies with other kinds of labels either; cf. Wang 121 & Ji (2022); Duan et al. (2023); Wang et al. (2023a). Unfortunately, the first two of these papers in-122 cluded a bug in their evaluation code. As such, their methods may perform no better than a baseline 123 that provides pure noise inputs to the model (Jo et al., 2024). 124

In general, speech decoding has centred on different kinds of speech: listening, imagining, speak-125 ing out loud, and, for paralysed patients, attempting to speak aloud. We focus here on listening 126 because it is easier to decode than imagined speech (e.g. Martin et al. (2014)). There is also some 127 evidence of a functional overlap between listening and imagined speech representations in the brain 128 (Wandelt et al., 2024), though we acknowledge that the question of overlap has been contested 129 (Langland-Hassan & Vicente, 2018). Prior work has also investigated the two tasks that we focus 130 on here (Dash et al., 2020; Moses et al., 2021; Gwilliams et al., 2023). The first of these, speech 131 detection, formed the backbone to Moses et al. (2021), where a speech detection model was trained 132 and subsequently used to detect isolated words, which were in turn classified and checked against a 133 language model to generate acceptable sentences. Hamilton et al. (2018) further elaborated on the 134 neural anatomy underlying speech detection, categorising neural responses in the superior temporal 135 gyrus (STG) to sustained speech and speech onset. As for the second task, voicing classification, Gwilliams et al. (2023) used this task as a proxy for phoneme classification, as pooling phonemes 136 into unvoiced or voiced segments (e.g. /p t k f s/ vs /b d g v z/) improves data efficiency. We note 137 that voicing classification and speech detection are related tasks as voicing is a subclass of speech. 138 This makes them foundational for building hierarchical speech decoding pipelines similar to prior 139 surgical decoding work (Moses et al., 2021; Willett et al., 2023). 140

141 In the computer vision literature, there have been a plethora of methods that use self-supervised 142 pretext tasks for representation learning (Agrawal et al., 2015; Doersch et al., 2015; Noroozi & Favaro, 2016; Larsson et al., 2016; Zhang et al., 2016; Gidaris et al., 2018). Until now, similar 143 approaches have not translated to the brain decoding literature with few exceptions (e.g. Cai et al. 144 (2023)). However, prior work has used other methods to leverage unlabelled brain data (Banville 145 et al., 2019; Kostas et al., 2021; Le & Shlizerman, 2022; Zhang et al., 2023; Yi et al., 2023; Ye et al., 146 2023; Yuan et al., 2024; Chen et al., 2024b). For example, Jiang et al. (2024) succeeded in cross-147 dataset and cross-task generalisation, using a transformer with tokenised brain signals and a masked 148 token prediction objective. Although this work combined unlabelled datasets, their results studied 149 simpler non-speech tasks with EEG. Wang et al. (2023b) used a similar approach, replacing tokens 150 with contextualised embeddings of time-frequency input representations. Their impressive speech 151 detection results were achieved with invasive neural recordings, which are comparatively rare and 152 thus have much less potential to scale than non-invasive data. Perhaps the closest work to ours in terms of unlocking scaling with neural data is BIOT (Yang et al., 2023). This is a self-supervised 153 architecture for encoding bio-signals that is similarly capable of training with different datasets, 154 labels, and varied numbers of sensors. Like the previous works, the approach tokenises signals for 155 a transformer architecture, but instead of a masked loss it uses a contrastive pre-training objective. 156 While theoretically supporting MEG, Yang et al. (2023) evaluate BIOT on simple ECG/EEG tasks 157 rather than address the comparatively complex challenge of speech decoding with MEG data. 158

- 159
- 160
- 161



Figure 2: Architecture overview. Inputs are projected into a shared dimension by the datasetconditional layer, then encoded. In pre-training, all weights are trainable except for modules in
light-red, while in fine-tuning, modules with light-blue borders are frozen and modules with lightred borders are unfrozen. Dashed borders indicate optional components.

### 3 Method

179 180 181

182

183

177 178

> To encode continuous neuroimaging data, we introduce a neural architecture to embed heterogeneous brain signals. We leverage this architecture for self-supervised learning from unlabelled MEG data using a set of pretext tasks designed to generate generalisable brain representations for speech decoding. With this approach, we hope to replicate similar successes in computer vision (Gidaris et al., 2018; Chen et al., 2020).

185 186 187

188

### 3.1 NETWORK ARCHITECTURE

Our two-stage neural network architecture (Figure 2) uses pretext tasks in pre-training to learn a representation with unlabelled brain data. Then, the fine-tuning stage uses this representation to learn the downstream task by training with labelled data.

We divide recordings into windows of length w seconds or t samples. At train time, each batch of 193 windows is standardised such that each sensor has zero mean and unit variance. The network takes 194 as input the standardised sample windows. To combine heterogeneous datasets, which have different 195 numbers of sensors S, we apply a dataset-conditional linear layer to the sensor dimension, projecting 196 the signal into a shared space with dimension  $d_{\text{shared}}$ . Then, to encode the signal, we construct a 197 wave-to-wave convolutional encoder architecture, the *cortex encoder*, inspired by work in neural audio codecs (Zeghidour et al., 2022; Défossez et al., 2022). Specifically, our convolutional encoder 199 adapts the implementation of the SEANet architecture (Tagliasacchi et al., 2020) used in Défossez 200 et al. (2022) which we describe here and as part of Figure 2. As these codecs typically operate on mono audio signals in  $\mathbb{R}^{1 \times t}$ , while our signals are in  $\mathbb{R}^{d_{\text{shared}} \times t}$ , we increase the convolutional 201 channel dimension from 1 to match  $d_{\rm shared}$  while also inflating the channel dimension of subsequent 202 convolutions. We refer to the output dimension of embeddings from this backbone as  $d_{\text{backbone}}$ . 203 Thus, the backbone takes as input a window in  $\mathbb{R}^{S \times t}$ , and encodes this into  $\tau$  embeddings (where  $\tau < t$ ), each of dimension  $d_{\text{backbone}}$  (i.e. an  $\mathbb{R}^{d_{\text{backbone}} \times \tau}$  output). 204 205

Just as speakers have different voices, neural responses between subjects have different characteristics. Consequently, individual variation leads to models that do not generalise well across subjects (Csaky et al., 2022). In the speech literature, models include speaker conditioning to account for these differences (Gibiansky et al., 2017). We take a similar approach by introducing subject conditioning. Zeghidour et al. (2022) find that conditioning is equally effective at the encoder bottleneck as in other stages of the model. Hence, we place ours at the cortex encoder bottleneck for simplicity. We use *feature-wise linear modulation (FiLM)* (Perez et al., 2018) as our conditioning method.

Following the advice of Balestriero et al. (2023, Section 3.2), we use a two-layer feedforward projector to alleviate misalignment between our pretext and downstream tasks in the representation. After the projector, linear classifiers make predictions for each of the pretext tasks. When fine-tuning, we train a linear decoder, for a downstream task, on top of the pre-trained representation, which remains frozen. Thus, we backpropagate only through the classifier. A trainable dataset-specific linear layer can be introduced for a novel dataset.

For speech detection, our classifier makes a prediction for each individual embedding. For voicing classification, where there is only one label for each sample window, the embeddings are flattened into a tensor in  $\mathbb{R}^{d_{\text{backbone}} \times \tau}$  representing the entire window. This is the input to the voicing classifier and is referred to as full epoch decoding in neuroimaging literature (Csaky et al., 2023).

223 224 3.2 PRETEXT TASKS

Our pretext tasks are unsupervised feature learning tasks that aim to learn generalisable speech decoding features. Since different datasets use varied numbers of sensors, we construct these tasks with labels that are agnostic to the number of sensors in the signal.

228 **Band prediction.** In the literature, neural responses can be segmented into functional frequency 229 bands (Giraud & Poeppel, 2012; Piai et al., 2014; Mai et al., 2016). Delta ( $\delta$ ) waves (0.1–4 Hz) are 230 commonly associated with the rhythmic structure of heard speech (Luo et al., 2010), Theta ( $\theta$ ) waves 231 (4–8 Hz) reliably track (Luo & Poeppel, 2007) and phase-lock to the amplitude envelope of heard 232 sentences (Peelle et al., 2012), Alpha ( $\alpha$ ) waves (8–12 Hz) relate to attentional processes and the 233 inhibition of irrelevant information, helping to focus on relevant speech signals (Strauß et al., 2015), 234 Beta ( $\beta$ ) waves (12–30Hz) are implicated in top-down predictive coding (Bressler & Richter, 2015) which affects lexical processing (Weiss & Mueller, 2012), Gamma ( $\gamma$ ) waves (30–70 Hz) occur with 235 higher cognitive functions (e.g. memory, learning, reasoning, and planning) (Fries, 2009; Buzsáki 236 & Wang, 2012), and High Gamma ( $\gamma^{\text{high}}$ ) waves (>70 Hz) have been linked specifically to speech 237 detection (Hamilton et al., 2018) and phonemic feature classification in the STG (Mesgarani et al., 238 2014) as well as phonemic feature classification in the ventral sensorimotor cortex (vSMC) (Cheung 239 et al., 2016). As High Gamma is a relatively wide band, we have split it into two sub-bands: Lower 240 *High Gamma* ( $\gamma_{\text{lower}}^{\text{high}}$ ) waves (70–100 Hz) and *Upper High Gamma* ( $\gamma_{\text{upper}}^{\text{high}}$ ) waves (100–150 Hz). 241

To learn representations that can distinguish between these, our band prediction task applies a bandstop filter for a randomly selected band  $\omega$  to the sample x, passes the filtered sample  $x^{\omega'}$  through the network backbone g and the corresponding linear predictor  $f_{\text{band}}$ , requiring the network to classify which frequency band  $\omega$  was rejected. This yields the loss

246 247

248 249

263

264

 $\mathcal{L}_{\text{band}} = \sum_{x \in B} \mathcal{L}_{\text{CE}}(f_{\text{band}}(g(x^{\omega'})), \omega), \tag{1}$ 

where *B* is a mini-batch of samples,  $\omega \in \{\delta, \theta, \alpha, \beta, \gamma, \gamma_{\text{lower}}^{\text{high}}, \gamma_{\text{upper}}^{\text{high}}\}$ , and  $\mathcal{L}_{\text{CE}}$  is the cross-entropy loss as this is a multi-class classification task.

Phase shift prediction. Phase coupling between networks of neuron populations is necessary for
 coordinating brain activity (Fries, 2005; Vidaurre et al., 2018). Thus, since phase often synchronises
 between communicating brain areas, phase coupling between spatially distant sensors is likely to be
 a useful feature. Supporting this insight, recent work (Jiang et al., 2024) also finds phase to be an
 essential component of the signal.

To learn representations that encode phase differences between brain areas, this task applies a discrete uniform random phase shift  $\phi \in \{0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8}, \frac{3\pi}{4}, \frac{7\pi}{8}\}$  to a uniformly randomly selected proportion  $\rho$  of the sensors. Applying this shift to random sensors is critical since sensors are placed in different positions, capturing different regions of the brain. Uniform random selection ensures differences between any two regions of the brain are represented. The objective of this task is to predict the phase shift. This leads to a similar loss

$$\mathcal{L}_{\text{phase}} = \sum_{x \in B} \mathcal{L}_{\text{CE}}(f_{\text{phase}}(g(x^{\phi})), \phi), \tag{2}$$

where  $x^{\phi}$  describes the signal with a phase shift  $\phi$  applied to a proportion of the sensors. We use a discrete number of possible phase shifts, treating it as a multi-class task rather than a regression task, to ease the difficulty of the problem as MEG scanners typically have a large number of sensors.

**Amplitude scale prediction.** MEG and EEG signals use an array of sensors at different spatial locations, capturing different signal sources more intensely. Representing the relative amplitude dif-

ference between sensors could be important for differentiating between neural responses originating
 from distinct parts of the brain. Within speech, Hamilton et al. (2018) find that localised regions of
 the STG respond to sustained speech and speech onsets. Differentiating between neural responses
 from this region and others may be essential for decoding speech perception.

Thus, this pretext task focuses on learning representations that encode relative sensor amplitude differences. Similar to the phase shift task, we select a random proportion of the sensors  $\rho$  and apply a discrete random amplitude scaling coefficient  $A \in [-2, 2]$ , discretised into 16 scaling factors, to the signal. The objective is to predict the scaling factor, leading to the loss

$$\mathcal{L}_{\text{amplitude}} = \sum_{x \in B} \mathcal{L}_{\text{CE}}(f_{\text{amplitude}}(g(x^A)), A), \tag{3}$$

where  $x^A$  is the signal scaled with A.

279 280 281

282

283

284

285

286

291 292

293 294

295

296

297

298

305

306

These pretext tasks capture complementary time- and frequency-domain properties of the signal. Hence, during pre-training, we combine them, creating an augmented version of the input for *every* pretext task by applying the matching transformation. We feed the augmented inputs through the network backbone and apply the corresponding classifier to predict the transformation, summing the weighted losses such that our final pre-training loss is given by

$$\mathcal{L}_{\rm SSL} = w_1 \mathcal{L}_{\rm band} + w_2 \mathcal{L}_{\rm phase} + w_3 \mathcal{L}_{\rm amplitude},\tag{4}$$

where  $w_i$  is a constant coefficient for each self-supervised loss.

## 4 EXPERIMENTS

In this section, we evaluate the representations learned with our pretext tasks by measuring their ability to scale downstream performance with unlabelled data. This includes understanding how well they can generalise across datasets, subjects, and tasks. We focus our evaluation on MEG data as the signal is rich, with better spatial resolution than EEG (Lopes da Silva, 2013) and faster sampling rates than fMRI (Hall et al., 2014).

We pre-train all models to completion and then fine-tune on labelled data for each task. In all tables and figures, we quote the *receiver operating characteristic area under the curve (ROC AUC)* where chance is always 0.5 regardless of class balance. We show the test ROC AUC at the best validation ROC AUC (early stopping) and quote uncertainty as the standard error of the mean over three seeds. Additionally, we state the *t*-score and *p*-value from single-sample one-sided *t*-tests against chance.

### 4.1 EXPERIMENTAL SETUP

307 Datasets. Unless specified otherwise, our experiments use Cam-CAN (Shafto et al., 2014; Taylor 308 et al., 2017) as an unlabelled representation learning dataset for pre-training. This is a study con-309 taining 641 subjects with resting and sensorimotor tasks, totalling approximately 160 hours of MEG recordings. For our downstream tasks, we use two labelled heard speech MEG datasets where par-310 ticipants listen to short stories or audiobooks. Armeni et al. (2022) contains 3 subjects who listen 311 to 10 hours of recordings each (30 hours total) while Gwilliams et al. (2023) has 27 subjects, each 312 recorded for 2 hours (54 hours total). Overall, we utilise over 200 hours of data. To the best of our 313 knowledge, this is the largest volume of MEG data ever used for speech decoding. 314

**Preprocessing.** Each recording is in  $\mathbb{R}^{S \times T}$  where S is the number of sensors and T is the number 315 of time points sampled by the scanner. To eliminate high-frequency muscle movement artifacts, we 316 apply a low-pass filter at 125Hz as well as a high-pass filter at 0.5Hz to remove slow-drift artifacts. 317 Since the datasets were recorded in Europe, where the electric grid frequency is 50Hz, we apply a 318 notch filter at multiples of 50Hz to account for line noise. Treating the low-pass filter threshold as the 319 Nyquist frequency, we downsample the signal to twice that at 250Hz, avoiding aliasing within our 320 band of interest. Finally, we detect bad sensor channels, those with significant noise and artifacts, 321 using a variance threshold and replace them by interpolating the spatially nearest sensors. 322

**Downstream tasks.** We evaluate our methods with two fundamental speech decoding tasks of increasing difficulty. The first, *speech detection*, determines whether speech occurs in the auditory

Table 1: **Pre-training with pretext tasks leads to better representations for speech detection.** In the *linear*-only case, we train a supervised linear classifier on the input MEG signals. For BIOT, we train a linear layer on top of a backbone pre-trained on CamCAN, with the rest of the model frozen. Similarly, for *ours*, we train a linear probe on top of our pre-trained backbone with its weights frozen. In the *no pre-training* baseline, the backbone uses randomly initialised and subsequently unmodified weights. When *all* pretext tasks are used, their losses are weighted equally.

	Armeni				Gwilliams			
Experim	ent	ROC AUC	t	p	ROC AUC	t	p	
Linear		$0.559 \pm 2e{-4}$	341	4e-6	$0.527\pm7\mathrm{e}{-5}$	379	3e-6	
BIOT + 1	inear	$0.500 \pm 4e{-4}$	0	6e-1	$0.499 \pm 2\mathrm{e}{-4}$	-3	1e+0	
Ours	No pre-training	$0.519 \pm 0.002$	8	7e-3	$0.498 \pm 0.003$	0	7e-1	
+ linear	$\operatorname{Amp}_{(\rho = 0.2)}$	$0.602\pm0.001$	114	4e-5	$0.532 \pm 0.005$	6	1e-2	
	$Phase_{(\rho = 0.5)}$	$0.603 \pm 0.003$	35	4e-4	$0.535 \pm 0.003$	12	3e-3	
	Band	$0.616\pm0.003$	44	3e-4	$0.542 \pm 0.001$	46	2e-4	
	All tasks	$0.621 \pm 0.003$	36	4e-4	$0.543 \pm 0.003$	13	3e-3	

stimulus using the neural response. The second task is *voicing classification*. Given data aligned at the occurrence of a phoneme, the task is to recognise whether the phoneme is *voiced* or *voiceless*, where voicing is a binary phonetic feature that categorises whether a speech sound is associated with vocal cord vibration. We select these tasks as they are simpler than phoneme recognition, but are foundational because they must be solved to decode speech accurately into natural language.

343

344

345

## 4.2 LEARNING GENERALISABLE REPRESENTATIONS USING PRETEXT TASKS

350 Our first experiment investigates whether our self-supervised objectives produce generalisable rep-351 resentations. In Table 1, we show the results of pre-training models with each pretext task inde-352 pendently as well as together. Here, all of our pretext tasks lead to result that are statistically 353 significant, and outperform a baseline fine-tuned without pre-training. This provides initial evidence that our tasks are helpful in speech decoding. Interestingly, the combination of all pretext tasks leads 354 to better generalisation than any task on its own. As we hypothesised earlier, this may be because 355 our pretext tasks capture complementary properties in time- and frequency-space, enforcing that our 356 representation includes more salient features for speech decoding than any individual task. 357

358 Now, we turn to the other baselines. Our approach significantly outperforms the equivalent with a raw MEG input instead of a pre-trained representation (the *linear* experiment). Here, the baseline 359 has substantially more trainable parameters because the input dimension is far larger without an 360 encoder. Even with this bias favouring the experiment with the raw input, using our representation 361 still performs better. We also compare our approach to BIOT (Yang et al., 2023) which is a similar 362 state-of-the-art self-supervised method. When BIOT is pre-trained using exactly the same data, the 363 fine-tuned probe fails to generalise entirely after exhaustive hyperparameter tuning. We put this 364 down to three critical reasons. Firstly, BIOT was designed around considerably lower-dimensional signals. Their EEG evaluation used an order of magnitude fewer sensors than our MEG data. With 366 MEG, their transformer approach requires many more channel embeddings, leading to difficulty 367 learning the complex interactions between sensors. Secondly, our self-supervised objective extracts 368 speech decoding features which is essential for solving speech decoding tasks. BIOT performs well on simple EEG tasks in Yang et al. (2023)'s evaluation, but non-invasive speech decoding is 369 significantly more challenging. Together, these obstacles suggest a vast amount of data is required 370 to learn their objective with MEG. Indeed, given that they pre-train with over 50 thousand hours of 371 EEG data in their evaluation, their objective appears too general to efficiently learn a representation 372 for speech decoding from the limited amount of MEG pre-training data (160 hours) available to us. 373 This highlights the importance of data-efficiency in SSL methods for MEG. 374

Among the individual pretext tasks, band prediction leads the rest. Perhaps this is because, by learning to discriminate between meaningful bands, the representation easily identifies phase-locking to speech onset in theta waves (Peelle et al., 2012). Further investigation is necessary here. The choice of the proportion of sensors to apply transformations to,  $\rho = 0.5$  for phase shift prediction and



Figure 3: Scaling unlabelled data improves generalisation. We pre-train the model on increasing amounts of unlabelled data from Cam-CAN (Shafto et al., 2014; Taylor et al., 2017). The solid lines are the best linear fits to the data and the dashed lines show the amount of data used in prior surgical (Moses et al., 2021) and non-invasive (Défossez et al., 2023) work.

 $\rho = 0.2$  for amplitude prediction, were determined through a hyperparameter search. We conjecture that a smaller  $\rho$  is optimal for amplitude scale prediction since this leads to representations that are 396 especially strong at discriminating amplitude differences among small groups of sensors. Perhaps this makes it easier to distinguish between neural responses from distinct parts of the brain such as 398 the STG, which is associated with speech onset (Hamilton et al., 2018). In contrast, a larger  $\rho$  for phase shift prediction could lead to representations that better discriminate neural synchrony information which is distributed across the brain rather than localised. As a result, a large proportion of 401 the sensors in a MEG scanner should encode information about this feature.

402 403 404

405

389

390

391

392 393 394

397

399

400

#### 4.3 SCALING SPEECH DECODING WITH UNLABELLED DATA

406 Here, we analyse generalisation as we increase the volume of unlabelled data, analysing scaling 407 performance on downstream tasks. As before, we pre-train with the combined pretext tasks. Figure 3 shows ROC AUC as we increase the amount of unlabelled data in pre-training up to approximately 408 160 hours. For both tasks, pre-training with any amount of data is sufficient to beat chance and 409 there is a clear improvement in accuracy as the amount of unlabelled data increases. For speech 410 detection on Armeni et al. (2022), scaling appears logarithmic in log-space; for all others, ROC 411 AUC improves log-linearly within the data regime we study. In any case, adding unlabelled data has 412 improved generalisation. Notably, we have scaled far beyond the data regime of prior surgical and 413 non-surgical work and yet performance has continued to scale. Thus, our self-supervision approach 414 may remain useful as the volume of open data in the field continues to rapidly increase. 415

Our results also reveal several new and notable phenomena. Firstly, we scaled up the pre-training 416 dataset by increasing the number of subjects. Since this led to consistent and almost monotonic 417 improvements in downstream accuracy, our method is an exception to the common consensus that 418 pooling subjects worsens generalisation. Secondly, as we pre-trained our model with a *different* 419 dataset to those we fine-tuned on, our representation shows cross-dataset generalisation. This is 420 particularly surprising as the Armeni et al. (2022), Gwilliams et al. (2023), and our pre-training 421 dataset all use different scanners entirely. Performing well across these datasets indicates that, to-422 gether, our architecture and pretext tasks successfully generate representations that are generalisable across heterogeneous scanners. Finally, we note that our pre-training dataset contained no language 423 data whatsoever yet still improved downstream accuracy on language tasks. Remarkably, this shows 424 that unlabelled brain data collected from any task (including those that are not linguistic) can be 425 used to improve speech decoding performance. 426

427 Since the results show improvements on both downstream tasks, this indicates that our pretext tasks 428 are sufficiently generic to produce representations that work with multiple speech decoding tasks 429 while still generalising well on each task individually. This is generally a challenging trade-off to manage. However, we notice that in both tasks, the base accuracy is higher and the improvement 430 in ROC AUC is steeper for Armeni et al. (2022). This is likely to be because this dataset has more 431 within-subject data. The weaker results for Gwilliams et al. (2023) may be a consequence of the



Figure 4: Scaling unlabelled data improves novel subject generalisation. We fine-tune on Gwilliams et al. (2023). When *in-distribution*, we evaluate on held-out sessions from subjects in the training set; when *out-of-distribution*, we evaluate on three held-out subjects. The solid lines are the best linear fits while the dashed lines show the amount of data used in prior surgical (Moses et al., 2021) and non-invasive (Défossez et al., 2023) work.

larger number of subjects with shorter intra-subject recordings and greater subject variation. These observations support the findings of other recent work such as Csaky et al. (2022).

4.4 SCALING UNLABELLED DATA IMPROVES GENERALISATION TO NOVEL SUBJECTS

In neuroimaging, brain data is generally highly variable across participants, leading to difficulty transferring models to novel subjects (Csaky et al., 2022). Whilst we have shown generalisation *across* subjects, here, we investigate whether we can generalise to *novel* subjects—an even more difficult challenge. This is critical in order to widely deploy speech BCIs for new patients. In this experiment, we fine-tune only on Gwilliams et al. (2023) and hold out three subjects with which we evaluate novel subject generalisation.

Figure 4 shows that scaling up the amount of unlabelled data used in pre-training not only improves accuracy on subjects previously seen, but also demonstrates a positive log-linear trend in performance for novel subjects. This indicates that scaling our method is an encouraging direction for resolving the challenges of subject variance faced by prior work. Moreover, as far as we are aware, this is the first result to demonstrate *novel* subject generalisation in speech decoding from MEG.

4.5 Aggregating Unlabelled MEG Datasets

To scale up unlabelled data further than individual studies, we must be able to combine many existing datasets. As a preliminary investigation, we combine two of the largest public MEG datasets: MOUS (Schoffelen et al., 2019) and Cam-CAN (Shafto et al., 2014; Taylor et al., 2017). In this section, we investigate how pre-training with these combined datasets affects downstream performance using the same experimental setup as Figure 3.

Table 2: **Combining unlabelled datasets shows signs of outperforming single studies.** We examine performance on the speech detection task. We see a small improvement when the datasets are combined on Gwilliams et al. (2023), but not Armeni et al. (2022).

481	Pre-training		Armo	eni		Gwillia	ms	
482	dataset	Hours	ROC AUC	t	p	ROC AUC	t	p
483	Cam-CAN	159	$0.621 \pm 0.003$	36	4e-4	$0.543 \pm 0.003$	13	3e-3
484	MOUS	160	$0.605\pm0.000$	261	7e-6	$0.543 \pm 0.004$	9	5e-3
485	Cam-CAN + MOUS	319	$0.611 \pm 0.003$	40	3e-4	$0.546 \pm 0.002$	20	1e-3

486 The results in Table 2 show, for the first time, that combining datasets can improve performance 487 on downstream speech decoding tasks. It leads to better performance on Gwilliams et al. (2023) 488 compared to pre-training on either dataset alone. Interestingly, this was not the case for Armeni 489 et al. (2022) where pre-training on Cam-CAN alone performed best. Combined pre-training did, 490 however, outperform training only on MOUS. It is surprising that pre-training on Cam-CAN was better than pre-training on MOUS when evaluating on Armeni et al. (2022) given that MOUS and 491 Armeni et al. (2022) both used speech tasks and were acquired on the same MEG scanner. Cam-492 CAN, by contrast, did not use a speech task and was acquired on a different MEG scanner. We 493 hypothesise that the better results for Cam-CAN are due to it being a cleaner dataset. During our 494 experiments, we found that data quality, even among unlabelled data, can have a significant affect 495 as artefacts in recordings disrupt learning. 496

While the combination of the two datasets includes far more hours of data than any prior work on deep learning with MEG, further work needs to be done to aggregate more datasets. Here, we were limited by compute budget. Increasing the number of datasets could enable the network to eventually always improve over the best singular dataset. Just as increasing the number of subjects (rather than only within-subject data) improves novel subject generalisation, a larger number of datasets may be key to scaling results when datasets are aggregated in pre-training.

503 504 4.6 LIMITATIONS

505 Although our results are significant in demonstrating a viable path forward to scale up speech BCIs, 506 there remain a number of limitations to the present work. We focused here on two downstream tasks: 507 speech detection and voice classification. Ultimately, we would like to expand this work to predict 508 full transcripts from brain recordings (i.e. brain-to-text). This has been achieved with surgical data 509 (Moses et al., 2021; Willett et al., 2023) but not yet convincingly with non-invasive methods like MEG or EEG (Jo et al., 2024). Speech detection has played an important role in the development 510 of full brain-to-text in a surgical context (Moses et al., 2021) and we hope may play a similar role 511 for non-invasive methods. Prior work has further used voice classification as a stand in for phoneme 512 classification (Gwilliams et al., 2022), and we have been able to improve on these results here. In 513 future work, we would like to expand this to all English phonemes. Secondly, while we have been 514 able to demonstrate the utility of a few pretext tasks, we do not claim to have exhausted the full set 515 of useful tasks. Rather, we conjecture that more useful pretext tasks remain to be found and believe a 516 useful avenue of research will be into other input representations for brain recordings. For example, 517 this paper did not make use of spatial features. Another limitation is our emphasis on heard speech 518 over other types of speech, such as attempted or imagined speech. We hypothesise that the same 519 methods presented here will generalise to these other varieties of speech, though this has yet to be 520 shown. But, perhaps the biggest limitation of the present work is that, while it surpasses the amount of data used in other studies, it remains to be seen how much speech decoding tasks can be improved 521 by scaling up the number of datasets used in training. In sharing this work now, we believe that the 522 current proof of concept will be sufficiently impactful to the field as we continue to actively scale up 523 the datasets that we can leverage. 524

525 526

527

# 5 CONCLUSION

528 Ultimately, solving speech decoding could transform the lives of patients with severe communication difficulties. This promise has not yet materialised because the field has been blocked by its 529 inability to scale up data to leverage deep learning. Prior methods have been unable to aggregate 530 data across different datasets, labels, or subjects to scale up because of heterogeneity in recording 531 hardware, experiment design, and participants. A handful of studies have shown weak signals to-532 wards alleviating these issues. But until now, no one has developed a general solution. We provided 533 a unified method that leverages unlabelled recordings data-efficiently using generic pretext tasks that 534 shows that all of these problems can be solved. We verified this with experiments showing that our 535 method not only scales with heterogeneous data but even generalises across datasets, subjects, and 536 tasks. Our method unlocks the potential of the bitter lesson, providing a general method to exploit 537 more computation by using more data. We implore the research community to employ the vast 538 quantities of data and compute available to realise this potential. If scale is all you need in speech decoding, then the bitter lesson may not be so bitter.

#### 540 ETHICS STATEMENT 541

In this work, we use data from studies that involve human subjects (Armeni et al., 2022; Gwilliams
et al., 2023; Shafto et al., 2014; Taylor et al., 2017; Schoffelen et al., 2019). These datasets are
public, cited, and have their own ethical approvals. The documentation for these is available with
the publications for the respective datasets.

546 While there are clear positive impacts, we acknowledge that insights from neural speech decoding 547 research may not all be beneficial. Research in this field could enable paralysed patients to commu-548 nicate freely and materially assist those with minor communication difficulty (e.g. stammering). As 549 the technology matures, it could also enable new ways of communicating with others and interacting with devices without the risks of invasive surgical implants. Nevertheless, the maturity of this 550 technology could also present potential negative societal impacts. For one, reading inner speech cre-551 ates new concerns over data controls as this information is likely to be highly sensitive and personal 552 to individuals. Given access to this technology, there is also the risk that bad actors could extract 553 sensitive information from target individuals without consent. Moreover, there are possible long 554 horizon effects associated with speech decoding research. Broad adoption of this technology could 555 lead to the gradual erosion of privacy over inner speech within society. In addition, asymmetric ef-556 fects, where some individuals or organisations can read inner speech but others are unable to, could 557 worsen societal inequality. Within the scope of this paper, we mitigate risks associated with inner 558 speech by focusing on decoding heard speech where there is low potential for abuse. Nonetheless, 559 we acknowledge that this is still a stepping stone towards solving inner speech decoding.

560 561

562

568

570

571

572

573

574

575

576 577

578

579

580

### REPRODUCIBILITY STATEMENT

In the supplementary materials, we have provided an anonymised code repository with instructions
for reproducing our main experiments. We also include details on experiment design and setup
(Section 4.1 and Appendix A), hyperparameters (Appendix B), and compute (Appendix C). While
we attempt to be exhaustive with these details, any information not found directly in the main body
or appendices can be located in the supplementary materials.

### 569 **REFERENCES**

- Pulkit Agrawal, João Carreira, and Jitendra Malik. Learning to see by moving. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 37–45. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.13. URL https: //doi.org/10.1109/ICCV.2015.13.
- Gopala Krishna Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568:493 498, 2019.
- Kristijan Armeni, Umut Güçlü, Marcel van Gerven, and Jan-Mathijs Schoffelen. A 10-hour withinparticipant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, 9(1):278, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01382-7. URL https://www.nature.com/articles/s41597-022-01382-7.
- Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Grégoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *CoRR*, abs/2304.12210, 2023. doi: 10.48550/ARXIV.2304.12210. URL https://doi.org/10. 48550/arXiv.2304.12210.
- Hubert J. Banville, Graeme Moffat, Isabela Albuquerque, Denis-Alexander Engemann, Aapo Hyvärinen, and Alexandre Gramfort. Self-supervised representation learning from electroen-cephalography signals. In 29th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2019, Pittsburgh, PA, USA, October 13-16, 2019, pp. 1–6. IEEE, 2019. doi: 10. 1109/MLSP.2019.8918693. URL https://doi.org/10.1109/MLSP.2019.8918693.
- 593 Steven L Bressler and Craig G Richter. Interareal oscillatory synchronization in top-down neocortical processing. *Current Opinion in Neurobiology*, 31:62–66, 2015.

György Buzsáki and Xiao-Jing Wang. Mechanisms of gamma oscillations. *Annual Review of Neuroscience*, 35:203–225, 2012.

- Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. Mbrain: A multi-channel self-supervised learning framework for brain signals. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 130–141. ACM, 2023. doi: 10.1145/3580305. 3599426. URL https://doi.org/10.1145/3580305.3599426.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and
   Armand Joulin. Emerging properties in self-supervised vision transformers. In 2021 IEEE/CVF
   *International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10- 17, 2021*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https:
   //doi.org/10.1109/ICCV48922.2021.00951.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for
   contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event,* volume 119 of *Proceedings of Machine Learning Research,* pp. 1597–1607. PMLR, 2020. URL http://proceedings.
   mlr.press/v119/chen20j.html.
- Kupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, pp. 1–14, April 2024a. ISSN 2522-5839. doi: 10.1038/s42256-024-00824-8. URL https://www.nature.com/articles/s42256-024-00824-8.
- Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale EEG foundation model. *CoRR*, abs/2401.10278, 2024b. doi: 10.48550/ARXIV.2401.10278. URL https://doi.org/10.48550/arXiv.2401.10278.
- 623 Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang. The auditory represen-624 tation of speech sounds in human motor cortex. *eLife*, 5:e12577, 2016.
- Richard Csaky, Mats W. J. van Es, Oiwi Parker Jones, and Mark W. Woolrich. Group-level brain decoding with deep learning. *Human Brain Mapping*, 44:6105 6119, 2022. URL https://doi.org/10.1002/hbm.26500.

628

- Richard Csaky, Mats W.J. van Es, Oiwi Parker Jones, and Mark Woolrich. Interpretable many-class decoding for MEG. *NeuroImage*, 282:120396, November 2023. ISSN 10538119. doi: 10.1016/j.neuroimage.2023.120396. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811923005475.
- Debadatta Dash, Paul Ferrari, Satwik Dutta, and Jun Wang. NeuroVAD: Real-Time Voice Activity
   Detection from Non-Invasive Neuromagnetic Signals. Sensors, 20(8):2248, January 2020. ISSN
   1424-8220. doi: 10.3390/s20082248. URL https://www.mdpi.com/1424-8220/20/
   8/2248.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *CoRR*, abs/2210.13438, 2022. doi: 10.48550/ARXIV.2210.13438. URL https://doi.org/10.48550/arXiv.2210.13438.
- Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning
   by context prediction. In 2015 IEEE International Conference on Computer Vision, ICCV 2015,
   Santiago, Chile, December 7-13, 2015, pp. 1422–1430. IEEE Computer Society, 2015. doi:
   10.1109/ICCV.2015.167. URL https://doi.org/10.1109/ICCV.2015.167.
- Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, December

648 649 650	10 - 16, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ 1f2fd23309a5b2d2537d063b29ec1b52-Abstract-Conference.html.
651 652 653 654	Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. De- coding speech perception from non-invasive brain recordings. <i>Nature Machine Intelligence</i> , 5 (10):1097–1107, October 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00714-5. URL https://www.nature.com/articles/s42256-023-00714-5.
655 656 657 658 659	Pascal Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. <i>Trends in Cognitive Sciences</i> , 9(10):474–480, October 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.08.011. URL https://www.sciencedirect.com/science/ article/pii/S1364661305002421.
660 661 662	Pascal Fries. Neuronal gamma-band synchronization as a fundamental process in cortical computa- tion. <i>Annual Review of Neuroscience</i> , 32:209–224, 2009.
663 664 665 666 667 668 669	<ul> <li>Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 2962–2970, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/c59b469d724f7919b7d35514184fdc0f-Abstract.html.</li> </ul>
670 671 672 673	Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N210
674 675 676 677 678	Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging com- putational principles and operations. <i>Nature Neuroscience</i> , 15(4):511–517, April 2012. ISSN 1546-1726. doi: 10.1038/nn.3063. URL https://www.nature.com/articles/nn. 3063.
679 680 681 682	Laura Gwilliams, Jean-Rémi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. <i>Nature Communications</i> , 13(1): 6606, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34326-1. URL https://www.nature.com/articles/s41467-022-34326-1.
683 684 685 686 687 688	Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. <i>Scientific Data</i> , 10(1):862, December 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02752-5. URL https://www.nature.com/articles/s41597-023-02752-5.
689 690 691	Emma L. Hall, Siân E. Robson, Peter G. Morris, and Matthew J. Brookes. The relationship between MEG and fMRI. <i>NeuroImage</i> , 102:80–91, 2014. URL https://doi.org/10.1016/j.neuroimage.2013.11.005.
692 693 694 695 696	Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. A Spatial Map of Onset and Sus- tained Responses to Speech in the Human Superior Temporal Gyrus. <i>Current Biology</i> , 28(12): 1860–1871.e4, June 2018. ISSN 09609822. doi: 10.1016/j.cub.2018.04.033. URL https: //linkinghub.elsevier.com/retrieve/pii/S0960982218304615.
697 698 699	Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic represen- tations with tremendous EEG data in BCI. In <i>The Twelfth International Conference on Learning</i> <i>Representations</i> , 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.
700	Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working? <i>arXiv</i> , 2024. doi: https://arxiv.org/abs/2405.06459.

702 703 704 705 706 707 708	John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ron- neberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridg- land, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera- Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Se- bastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Push- meet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. <i>Na</i> - tum 50(582) 580, 2021 UBL https://doi.org/10.1020/cr41506.001102010
709 710 711 712 713	Demetres Kostas, Stephane T Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. <i>Fron-</i> <i>tiers in Human Neuroscience</i> , 15, 2021.
714 715	Peter Langland-Hassan and Agustín Vicente. <i>Inner Speech: New Voices</i> . Oxford University Press, 2018.
716 717 718 719 720 721 722	Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for auto- matic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), <i>Computer</i> <i>Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14,</i> <i>2016, Proceedings, Part IV</i> , volume 9908 of <i>Lecture Notes in Computer Science</i> , pp. 577–593. Springer, 2016. doi: 10.1007/978-3-319-46493-0\_35. URL https://doi.org/10.1007/ 978-3-319-46493-0_35.
723 724 725 726 727 728	Trung Le and Eli Shlizerman. STNDT: modeling neural population activity with spatiotemporal transformers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/72163dlc3c1726flc29157d06e9e93c1-Abstract-Conference.html.
729 730 731	Fernando Lopes da Silva. EEG and MEG: Relevance to Neuroscience. <i>Neuron</i> , 80(5):1112–1128, December 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.10.017. URL https://www.sciencedirect.com/science/article/pii/S0896627313009203.
732 733 734 735	Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.
736 737	Huan Luo and David Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. <i>Neuron</i> , 54(6):1001–1010, 2007.
739 740 741	Huan Luo, Zuxiang Liu, and David Poeppel. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. <i>PLOS Biology</i> , 8(8):e1000445, 2010.
742 743 744 745 746	Guangting Mai, James W. Minett, and William S. Y. Wang. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. <i>NeuroImage</i> , 133:516–528, June 2016. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2016.02.064. URL https://www.sciencedirect.com/science/article/pii/S1053811916001737.
747 748 749	Joseph G. Makin, David A. Moses, and Edward F. Chang. Machine translation of cortical activity to text with an encoder-decoder framework. <i>Nature Neuroscience</i> , 23:575 – 582, 2019. URL https://api.semanticscholar.org/CorpusID:199639966.
750 751 752 753	Stéphanie Martin, Peter Brunner, Chris Holdgraf, Hans-Jochen Heinze, Nathan E Crone, Jochem Rieger, Gerwin Schalk, Robert T Knight, and Brian N Pasley. Decoding spectrotemporal features of overt and covert speech from the human cortex. <i>Frontiers in Neuroengineering</i> , 7:14, 2014.
754 755	Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. <i>Science</i> , 343(6174):1006–1010, 2014. doi: DOI:10.1126/ science.1245994.

- Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023.
- 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, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, 385(3):217–227, July 2021. ISSN 0028-4793. doi: 10.1056/NEJMoa2027540. URL https://doi.org/10. 1056/NEJMoa2027540.
- Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jig-saw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), Computer Vision ECCV 2016 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI, volume 9910 of Lecture Notes in Computer Science, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4\\_5. URL https://doi.org/10.1007/978-3-319-46466-4\\_5.
- OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774.
   URL https://doi.org/10.48550/arXiv.2303.08774.
- Jonathan E. Peelle, Joachim Gross, and Matthew H. Davis. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, 23(6):1378–1387, 2012.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11671. URL https://doi.org/10.1609/aaai.v32i1.11671.
- Vitória Piai, Ardi Roelofs, and Eric Maris. Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, 53:146–156, January 2014. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2013.11.014. URL https://www.sciencedirect.com/science/article/pii/S0028393213004119.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.
- 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. URL
   https://www.nature.com/articles/s41597-019-0020-y.
- Meredith A. Shafto, Lorraine K. Tyler, Marie Dixon, Jason R. Taylor, James Benedict Rowe, Rhodri
   Cusack, Andrew J. Calder, William D. Marslen-Wilson, John S. Duncan, T. Dalgleish, Richard
   N. A. Henson, Carol Brayne, and Fiona E. Matthews. The Cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination
   of healthy cognitive ageing. *BMC Neurology*, 14, 2014.
- Antje Strauß, Molly J Henry, Mathias Scharinger, and Jonas Obleser. Alpha phase determines successful lexical decision in noise. *Journal of Neuroscience*, 35(7):3256–3262, 2015.
- **Richard Sutton.** The bitter lesson. *Incomplete Ideas (blog)*, 2019. URL http://www. incompleteideas.net/IncIdeas/BitterLesson.html.

- 810 Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. SEANet: A multi-811 modal speech enhancement network. In Helen Meng, Bo Xu, and Thomas Fang Zheng 812 (eds.), Interspeech 2020, 21st Annual Conference of the International Speech Communica-813 tion Association, Virtual Event, Shanghai, China, 25-29 October 2020, pp. 1126–1130. ISCA, 814 2020. doi: 10.21437/INTERSPEECH.2020-1563. URL https://doi.org/10.21437/ Interspeech.2020-1563. 815
- 816 Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of con-817 tinuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, May 818 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature. 819 com/articles/s41593-023-01304-9. 820
- Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, 821 Lorraine K. Tyler, Cam-CAN Group, and Richard N. A. Henson. The Cambridge centre for 822 ageing and neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and 823 cognitive data from a cross-sectional adult lifespan sample. Neuroimage, 144:262 – 269, 2017. 824
- 825 Diego Vidaurre, Laurence T. Hunt, Andrew J. Quinn, Benjamin A. E. Hunt, Matthew J. Brookes, 826 Anna C. Nobre, and Mark W. Woolrich. Spontaneous cortical activity transiently organises into 827 frequency specific phase-coupling networks. Nature Communications, 9(1):2987, July 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05316-z. URL https://www.nature.com/ 828 articles/s41467-018-05316-z. 829
- 830 Sarah K Wandelt, David A. Bjånes, Kelsie Pejsa, Brian Lee, Charles Y Liu, and Richard Andersen. 831 Representation of internal speech by single neurons in human supramarginal gyrus. *Nature human* 832 behaviour, 2024. URL https://doi.org/10.1038/s41562-024-01867-y. 833
- Bo Wang, Xiran Xu, Longxiang Zhang, Boda Xiao, Xihong Wu, and Jing Chen. Semantic recon-834 struction of continuous language from MEG signals. CoRR, abs/2309.07701, 2023a. doi: 10. 835 48550/ARXIV.2309.07701. URL https://doi.org/10.48550/arXiv.2309.07701. 836
- 837 Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Igna-838 cio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial 839 recordings. In The Eleventh International Conference on Learning Representations, ICLR 2023, 840 Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023b. URL https://openreview. 841 net/pdf?id=xmcYx\_reUn6.
- 842 Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and 843 zero-shot sentiment classification. In Thirty-Sixth AAAI Conference on Artificial Intelligence, 844 AAAI 2022, Virtual Event, February 22 - March 1, pp. 5350-5358. AAAI Press, 2022. doi: 845 10.1609/AAAI.V36I5.20472. URL https://doi.org/10.1609/aaai.v36i5.20472. 846
- Sabine Weiss and Horst M. Mueller. "Too many betas do not spoil the broth": the role of beta brain 847 oscillations in language processing. Frontiers in Psychology, 3, 2012. doi: https://doi.org/10. 848 3389/fpsyg.2012.00201. 849
- 850 Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. 852 Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. Nature, 620 853 (7976):1031-1036, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06377-x. URL 854 https://www.nature.com/articles/s41586-023-06377-x.

- 855 Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: biosignal transformer for cross-data 856 learning in the wild. In Advances in Neural Information Processing Systems 36: Annual Confer-857 ence on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 858 December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper\_files/paper/ 859 2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference. html. 861
- Joel Ye, Jennifer L. Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 862 Multi-context pretraining for neural spiking activity. In Alice Oh, Tristan Nau-2: 863 mann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances

 in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper\_files/paper/2023/hash/
 fe51de4e7baf52e743b679e3bdba7905-Abstract-Conference.html.

- 868 Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG 869 representations with geometry-aware modeling. In Alice Oh, Tristan Naumann, Amir 870 Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neu-871 ral Information Processing Systems 36: Annual Conference on Neural Information Pro-872 cessing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 873 2023, 2023. URL http://papers.nips.cc/paper files/paper/2023/hash/ 874 a8c893712cb7858e49631fb03c941f8d-Abstract-Conference.html.
  - Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. BrainWave: A brain signal foundation model for clinical applications. 2024. URL https://api. semanticscholar.org/CorpusID:267740511.
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Sound-stream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.
- Baoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li.
  Brant: Foundation model for intracranial neural signal. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 16, 2023, 2023. URL http://papers.nips.cc/paper\_files/paper/2023/hash/535915d26859036410b0533804cee788-Abstract-Conference.html.
- Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), Computer Vision ECCV 2016 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III, volume 9907 of Lecture Notes in Computer Science, pp. 649–666. Springer, 2016. doi: 10.1007/978-3-319-46487-9\_40.
  WRL https://doi.org/10.1007/978-3-319-46487-9\_40.
- 896 897

898

875

876

877

878

# A EXPERIMENT DETAILS

We pre-train with non-overlapping sample windows from all subjects and sessions. We adjust the amount of unlabelled data used from Cam-CAN by increasing the number of subjects in the sequence
1, 2, 4, 8, 17, 36, 74, 152, 312, and 641, successively randomly selecting more subjects to include. Each seed uses a different set of subjects to reduce negative effects from outlier subjects.

903 When fine-tuning with Armeni et al. (2022), we hold out session 010 from all subjects during train-904 ing and validation, using this for evaluation. Similarly, when fine-tuning with Gwilliams et al. 905 (2023), we hold out session 1 from subjects 23, 24, 25, 26, and 27, using these sessions for evalua-906 tion only. As there is limited within-subject data in the latter dataset, we did not hold out a session 907 from all subjects as before. For our novel subject experiments, we hold out subjects 1, 2, and 3 en-908 tirely and use the data for these subjects during evaluation. In Gwilliams et al. (2023), we note that they use four different tasks for each subject and their order is randomized between subjects. Both 909 sessions for each task are repeats of the task. This means that while the recording itself is unseen, in 910 this dataset, it is possible that heldout sessions use tasks that may have been seen in the training set. 911

In all experiments, we always fine-tune to completion (usually around 300 epochs), taking the test metric at the best validation loss (early stopping). We use three randomly selected seeds for each pre-training and corresponding fine-tuning run. For speech detection, since our encoder reduces the temporal dimension from 125 samples (the number of samples in a 0.5 second window with a sample rate of 250Hz) down to 5 embeddings, we downsample our speech detection labels to match using PyTorch's torch.nn.functional.interpolate. Therefore, each speech detection label represents a 0.1 second period of time.

#### 918 B HYPERPARAMETERS 919

925

926

970 971

We conducted a search over hyperparameters of interest to optimise our self-supervised objectives
and neural architecture. While these ablations indicated a theoretically ideal architectural configuration, in practice, we altered our final experimental architecture due to instabilities during training
when data was scaled up. Our final architecture hyperparameters achieve a balance between the best
values from our hyperparameter search and stable training. These values are detailed in Table 3.

928	Hyperparameter	Value
929		0.5
930	window length (s)	0.5
931	$\rho$ (phase)	0.5
932	$\rho$ (amplitude)	0.2
933	$\{w_1, w_2, w_3\}$	{1.0, 1.0, 1.0}
934	$d_{ m shared}$	512
935	dbackhone	512
936	SEANet convolution channels	(512, 512, 512, 512)
937	SEANet downsampling ratios	(5, 5, 1)
938	Fil M conditioning dimension	16
39	Subject embedding dimension	16
940		
)41	Pre-training epochs	200
942	Uptimizer Leorning rate	Adam w (Losnchilov & Hutter, 2019)
943	Train ratio	0.000000
44	Validation ratio	0.0
45	Test ratio	0.1
46		0.1
47		
48		
49		
50		
51		
52		
53		
54		
55		
56		
57		
58		
59		
60		
61		
62		
63		
64		
65		
66		
67		
68		
69		

#### Table 3: Experimental hyperparameters.

#### 972 C COMPUTE RESOURCES

All experiments were run on individual NVIDIA V100 and A100 GPUs with up to 40GiB of GPU memory on a system with up to 1TiB of RAM. Each pre-training run with the maximum amount of pre-training data took approximately 200 hours (8.3 days). Fine-tuning following pre-training took up to another 12 hours. We estimate that we used approximately 3000 hours of compute for the final experimental runs, including hyperparameter searches. In total, over the course of developing this work from idea to final paper, we used around 10,000 hours of GPU compute.

# D LICENCES FOR DATASETS AND CODE

The Armeni et al. (2022) dataset is distributed under CC-BY-4.0 while the Gwilliams et al. (2023)
dataset is distributed under the CC0 1.0 Universal licence. The Schoffelen et al. (2019) dataset
is distributed with a RU-DI-HD-1.0 licence from the Donders institute. The licence for the CamCAN (Shafto et al., 2014; Taylor et al., 2017) dataset is unknown. The SEANet code adapted from
Défossez et al. (2022) is distributed under the MIT licence, and the OSL library, which we use for
preprocessing, is under the BSD-3-Clause licence.