
The Brain’s Bitter Lesson: Scaling Speech Decoding With Self-Supervised Learning

Dulhan Jayalath¹ Gilad Landau¹ Brendan Shillingford² Mark Woolrich³ Oiwi Parker Jones¹

Abstract

The past few years have seen remarkable progress in the decoding of speech from brain activity, primarily driven by large single-subject datasets. However, due to individual variation, such as anatomy, and differences in task design and scanning hardware, leveraging data across subjects and datasets remains challenging. In turn, the field has not benefited from the growing number of open neural data repositories to exploit large-scale deep learning. To address this, we develop neuroscience-informed self-supervised objectives, together with an architecture, for learning from heterogeneous brain recordings. Scaling to nearly **400 hours** of MEG data and **900 subjects**, our approach shows generalisation across participants, datasets, tasks, and even to *novel* subjects. It achieves **improvements of 15-27%** over state-of-the-art models and **matches surgical decoding performance with non-invasive data**. These advances unlock the potential for scaling speech decoding models beyond the current frontier.

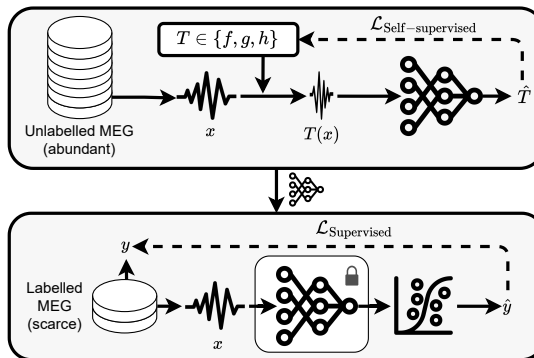



Figure 1: **Leveraging unlabelled data using pretext tasks for speech decoding.** We pre-train a neural network using tasks that generate implicit labels from abundant unlabelled MEG neuroimaging data, permitting learning from large heterogeneous datasets. The tasks apply a randomly selected neuroscientifically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

1. Introduction

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). The ability of deep learning to learn from ever-larger datasets has enabled seemingly arbitrary scaling with computation, leading to astounding advances across a diverse set of domains (Jumper et al., 2021; Caron et al., 2021; OpenAI, 2023; Radford et al., 2023).

In the domain of brain data, and of tasks like speech decoding, the bitter lesson has not yet been fully assimilated.

¹PNPL  ³OHBA, University of Oxford ²Google DeepMind. Correspondence to: <{dulhan, oiwi}@robots.ox.ac.uk>.

Recent work towards *brain-computer interfaces (BCIs)* have tried to scale up labelled datasets for individual subjects, using either invasive (Moses et al., 2021; Willett et al., 2023) or non-invasive brain recordings (Tang et al., 2023), mapping these to transcripts of attempted or imagined speech. Yet, a number of obstacles to scale remain, especially in *magnetoencephalography (MEG)* data. Current speech decoding models rarely train on multiple subjects, combine datasets, or utilise data from diverse tasks. Thus the size of training data has been limited to how much can be acquired for a 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 from different neuroimaging scanners differ, for example. But overcoming these limitations holds the promise of training models on collective, internet-scale data.

Decoding methods for MEG need to be highly data-efficient. While electroencephalography (EEG) data are abundant, MEG provides richer signals for decoding (Lopes da Silva, 2013; Hall et al., 2014) but are comparatively rare. Given the scarcity of speech-labelled MEG and the larger proportion of other 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). To data-efficiently learn from unlabelled MEG, we propose *pretext* training with neuroscience-informed input transformations that benefit downstream tasks. We use this for learning from unlabelled brain data (Figure 1) through an architecture for processing continuous multi-sensor neuroimaging signals. Our method provides a unified approach that enables leveraging data from other experiments that do not have the same labels (by treating them as unlabelled) and that come from different subjects and neuroimaging scanners. We evaluate representations learned with our approach on heard speech datasets acquired with MEG, setting the baselines for speech detection and voicing classification on this data.

Our main contributions are:

- A domain-specific **self-supervision method** and a **neural architecture** for representation learning from MEG that unlock scaling speech decoding over multiple subjects, multiple studies, and unlabelled data;
- Achieving **15-27% gains** over state-of-the-art self-supervised models, **matching surgical self-supervised decoding non-invasively**, and showing **novel subject generalisation** for the first time in MEG; and
- Demonstrating evidence for **scaling laws** arising from pre-training with unlabelled MEG recordings using multiple times the volume of data in prior work.

2. Related Work

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. Nevertheless, the method developed by Tang et al. (2023) is remarkable for showing an ability to generalise across labelled task data. They do not, however, use unlabelled data or show cross-subject generalisation.

Specific studies into the limitations of generalising models between subjects show that while performance decreases on average when subjects are pooled, there are exceptions

(e.g. Anumanchipalli et al. (2019) and Makin et al. (2019) in surgical settings and Csaky et al. (2022) non-invasively). Défossez et al. (2023) show cross-subject generalisation for a segment identification task from participants listening to connected speech. However, they do not demonstrate generalisation to novel subjects and retrain their model for new datasets rather than being able to generalise across datasets or pool them. Their method is also unable to incorporate data without corresponding audio labels and so does not scale with other kinds of tasks.

In general, speech decoding has centred on different kinds of speech: listening, imagining, speaking out loud, and, for paralysed patients, attempting to speak aloud. We focus on listening because it is easier to decode than imagined speech (e.g. Martin et al. (2014)). There is also evidence of functional overlap between listening and imagined speech representations in the brain (Wandelt et al., 2024), though the question of overlap has been contested (Langland-Hassan & Vicente, 2018). While some work on decoding text directly from heard speech tasks with MEG and EEG exist, it is unclear whether these methods perform any better than a baseline that provides pure noise inputs to the model (Jo et al., 2024). Non-invasive speech decoding remains a highly challenging and unsolved domain.

Self-supervised pretext tasks have been successful in computer vision (Agrawal et al., 2015; Doersch et al., 2015; Noroozi & Favaro, 2016; Larsson et al., 2016; Zhang et al., 2016; Gidaris et al., 2018) but rarely applied to brain decoding. There are, however, methods that leverage unlabelled brain data in other ways (Banville et al., 2019; Kostas et al., 2021; Le & Shlizerman, 2022; Zhang et al., 2023; Yi et al., 2023; Cai et al., 2023; Ye et al., 2023; Yuan et al., 2024; Chen et al., 2024b). Unfortunately, most of this literature is unable to scale and harmonize heterogeneous non-invasive data. The most notable of these works include Wang et al. (2023), who learn contextualised embeddings of time-frequency input representations through masked spectrogram in-filling. Their impressive speech detection results were achieved with invasive neural recordings, which are comparatively rare and thus have less potential to scale than non-invasive data. Another, BIOT (Yang et al., 2023), learns from generic heterogeneous bio-signals with a contrastive pre-training objective, rather than masking, and applies it to ECG/EEG data. Notably, none of these methods optimise their objectives for speech decoding or focus on MEG.

3. Method

We introduce a neural architecture to embed heterogeneous brain signals. Then, 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.

3.1. Network Architecture

Our neural network architecture has two stages (Figure 2): pre-training with pretext tasks on unlabelled data, and training a linear probe with labelled data for downstream tasks.

We divide recordings into windows of length w seconds or t samples. At train time, each batch of windows is standardised such that each sensor has zero mean and unit variance. The network takes as input the standardised sample windows. To combine heterogeneous datasets, which have varying numbers of sensors S , we apply a dataset-conditional linear layer to the sensor dimension, projecting the signal into a shared space with dimension d_{shared} . Then, to encode the signal, we construct a 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 adapts the SEANet architecture (Tagliasacchi et al., 2020) used in Défossez 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 channel dimension from 1 to match d_{shared} while also inflating the channel dimension of subsequent convolutions. We refer to the output dimension of embeddings from this backbone as d_{backbone} . Thus, the backbone takes as input a window in $\mathbb{R}^{S \times t}$, and encodes this into τ embeddings, each of dimension d_{backbone} (i.e. an $\mathbb{R}^{d_{\text{backbone}} \times \tau}$ output).

In the speech recognition literature, models include speaker conditioning to account for vocal and prosodic differences (Gibiansky et al., 2017). 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). We address this with a similar approach to the speech literature by introducing subject conditioning using *feature-wise linear modulation (FiLM)* (Perez et al., 2018). As Zeghidour et al. (2022) find that conditioning is as equally effective at the encoder bottleneck as in other stages of the model, we also condition at the bottleneck.

Following Balestriero et al. (2023, Section 3.2), we use a two-layer 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

Table 1: **Functional frequency bands in brain activity.**

Band	Hz	Association
Delta (δ)	.1-4	Rhythmic structure of heard speech (Luo et al., 2010)
Theta (θ)	4-8	Tracking (Luo & Poeppel, 2007) and phase-locking to the amplitude envelope of heard sentences (Peelle et al., 2012)
Alpha (α)	8-12	Attentional processes and the inhibition of irrelevant information (Strauß et al., 2015)
Beta (β)	12-30	Top-down predictive coding (Bressler & Richter, 2015) which affects lexical processing (Weiss & Mueller, 2012)
Gamma (γ)	30-70	Higher cognitive functions (e.g. memory, learning, reasoning, and planning) (Fries, 2009; Buzsáki & Wang, 2012)
High Gamma (γ^{high})	70+	Speech detection (Hamilton et al., 2018) and phonemic feature classification in the STG (Mesgarani et al., 2014) and the <i>ventral sensorimotor cortex (vSMC)</i> (Cheung et al., 2016)

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).

3.2. Pretext Tasks

To use our architecture for pre-training, we construct pretext objectives for unsupervised learning of generalisable speech decoding features. These objectives are inherently agnostic to the sensor count because they operate on properties that are independent of the specific sensor arrangement. This key design choice enables the tasks to work seamlessly across datasets with varying numbers of sensors—a critical requirement for combining heterogeneous brain data.

Band prediction. In the literature, neural responses can be segmented into functional frequency bands (Giraud & Poeppel, 2012; Piai et al., 2014; Mai et al., 2016) (Table 1). Sensitivity to these frequencies would bring about functional separability in the representation space. Thus, to learn such representations, we train the network to classify rejected bands. As High Gamma is a relatively wide band we split it into two sub-bands: *Lower High Gamma* ($\gamma_{\text{lower}}^{\text{high}}$) waves (70–100 Hz) and *Upper High Gamma* ($\gamma_{\text{upper}}^{\text{high}}$) waves (100–150 Hz). Our task applies a band-stop filter for a randomly selected band ω to the sample x , passes the filtered sample

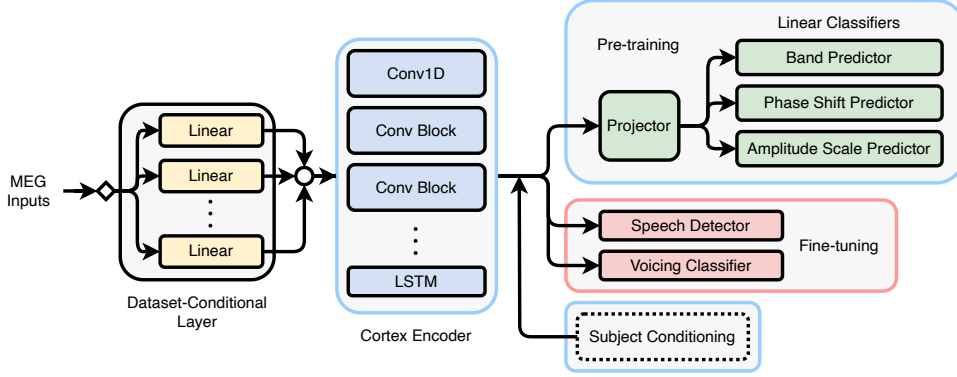


Figure 2: **Architecture overview.** Inputs are projected into a shared dimension by the dataset-conditional layer, then encoded. In pre-training, all weights are trainable except for modules in light-red, while in fine-tuning, modules with light-blue borders are frozen and modules with light-red borders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

$x^{\omega'}$ through the network backbone g and the corresponding linear predictor f_{band} , requiring the network to classify which frequency band ω was rejected. This yields the loss

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

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}_{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) and so 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 $\rho \in [0, 0.5]$ 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

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

where x^{ϕ} describes the signal with a phase shift ϕ 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 difference 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 $\rho \in [0, 0.5]$ 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

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

where x^A is the signal scaled with A .

Combined tasks. 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

$$\mathcal{L}_{\text{SSL}} = w_1 \mathcal{L}_{\text{band}} + w_2 \mathcal{L}_{\text{phase}} + w_3 \mathcal{L}_{\text{amplitude}}, \quad (4)$$

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

4. Experiments

We evaluate our self-supervised representations by measuring how they scale with unlabelled data and 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 train a linear probe 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 the class distribution. 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 up to five seeds.

4.1. Experimental setup

Datasets. In total, we use almost five times the volume of data in prior MEG work, totalling approximately 400 hours with nearly 900 subjects across pre-training and downstream training. Unless specified otherwise, we pre-train with Cam-CAN (Shafto et al., 2014; Taylor et al., 2017) as an unlabelled representation learning dataset. This is a study containing 641 subjects with resting and sensorimotor tasks, totalling approximately 160 hours of MEG recordings. When aggregating datasets, we also pre-train with MOUS (Schoffelen et al., 2019), which contains 204 subjects and another 160 hours from visual and auditory tasks. Downstream, we use labelled heard speech MEG datasets where participants listen to short stories or audiobooks. We mainly focus on Armeni et al. (2022) which contains 3 subjects who listen to 10 hours of recordings each (30 hours total). We also analyse Gwilliams et al. (2023) which has 27 subjects, each recorded for 2 hours (54 hours total). The latter dataset is particularly difficult to decode from as there is very little within-subject data and it did not enforce the use of head casts to immobilise participants. Nevertheless, given it has many more subjects, we use this dataset to study subject generalisation.

Preprocessing. Each recording is in $\mathbb{R}^{S \times T}$ where S is the number of sensors and T is the number of time points sampled by the scanner. To eliminate high-frequency artifacts, we apply a low-pass filter at 125Hz as well as a high-pass filter at 0.1Hz to remove slow-drift artifacts. Since the datasets were recorded in Europe, where the electric grid frequency is 50Hz, we apply a notch filter at 50Hz and its harmonics to account for line noise. Treating the low-pass filter threshold as the Nyquist frequency, we downsample the signal to twice that at 250Hz, avoiding aliasing within our band of interest. Finally, we detect sensor channels with significant noise and artifacts using a variance threshold and replace them by interpolating the spatially nearest sensors.

Downstream tasks. We evaluate our methods primarily

on *speech detection*. This task determines whether speech occurs in the auditory stimulus using the neural response. This is a fundamental task in understanding speech perception and is one of the few tasks that so far show statistically significant results in highly noisy MEG signals. It also has direct applications to BCIs as it can be used to segment words or sentences for decoding and activate a speech BCI when a patient wishes to communicate. Secondly, we also study *voicing classification* to demonstrate the versatility of our representations for general speech decoding tasks. Given data aligned at the onset of a phoneme, the task is to recognise whether the phoneme is *voiced* or *voiceless*, where voicing is a phonetic feature that categorises whether a speech sound is associated with vocal cord vibration. This task is also directly relevant to a speech BCI as it involves classifying phonemes which can be used to decode words.

4.2. Learning Generalisable Representations Using Pretext Tasks

Table 2 shows that our approach achieves two key feats: outperforming comparable state-of-the-art self-supervised methods by 15-27% (part C), and matching the performance of prior self-supervised methods with surgical data (11) while using only non-invasive data. Since non-invasive data has a much lower signal-to-noise ratio than surgical data, this is quite unprecedented. In the rest of this section, we analyse this table in detail.

In part B, we show the results of pre-training models with each pretext task independently, together, and without any pre-training at all. Using pretext tasks (3, 4, 5) outperforms no pre-training (2). Interestingly, the combination of all pretext tasks (5) leads to better generalisation than any task on its own (the improvement over (3) is statistically significant). We conjecture that this is because our pretext tasks capture complementary properties in time- and frequency-space, ensuring that the representation includes more salient features than any individual task could encode. Finally, we apply Gaussian filtering to the predictions (7), smoothing out anomalies in the predicted speech envelope.

Next, we turn to the baselines, starting with Table 2 part A. Our method outperforms random selection (0) and training a linear layer with the MEG signal directly (1). The latter even has substantially more trainable parameters because the input dimension is larger without an encoder. In part C, we compare our approach to two state-of-the-art self-supervised methods. In each experiment, we apply Gaussian filtering as in (7). Although better than random, BrainBERT (9) does not generalise as well as our method (11). BrainBERT employs a generic masked spectrogram in-filling pre-training objective. While it is intended for speech decoding tasks, the pre-training objective is not designed to specifically capture features salient to neural speech processing. Furthermore,

Table 2: **Our approach surpasses baselines in speech detection by up to 27% and matches surgical decoding.** For *linear*, we train a supervised linear classifier on the MEG signals. For ours and BrainBERT, we train a linear layer on top of a backbone pre-trained on CamCAN, with the rest of the model frozen. For BIOT, we use their pre-trained weights. In the *no pre-training* baseline, the backbone uses randomly initialised and frozen weights. In the surgical context, we quote the result from Wang et al. (2023, Table 2). With *all* pretext tasks, losses are weighted equally.

Part / ID	Model	ROC AUC
A	0 Random select	.500
	1 Linear	.539 \pm .002
B	2 Ours No pre-train.	.519 \pm .002
	3 + linear Amp($\rho = 0.2$)	.624 \pm .001
	4 Phase($\rho = 0.5$)	.615 \pm .001
	5 Band	.588 \pm .001
	6 All tasks	.630 \pm .000
	7 + smoothing	.700 \pm .002
C*	8 BIOT ¹ + linear	.615 \pm .002
	9 BrainBERT ² + linear	.556 \pm .007
	10 EEGPT ³ + linear	.602 \pm .006
	11 Ours (best) + linear	.705 \pm .003
12 BrainBERT ² + lin. (surgical)	.71 \pm .06	

¹Yang et al. (2023) ²Wang et al. (2023) ³Wang et al. (2024)

their method is less data-efficient because in-filling is a harder generative task compared to classification and while their methodology was developed for relatively high-fidelity intracranial recordings, the inherently lower signal-to-noise ratio of MEG presents an even greater challenge.

Our last baselines are BIOT (8) and EEGPT (10). We leverage publicly released weights pre-trained with thousands of hours of EEG. However, both still fall short of our pre-training method for reasons which we believe are similar to BrainBERT—their objectives do not leverage neuroscientific understanding of speech processing.

Finally, and quite remarkably, our best result matches the AUROC quoted in Wang et al. (2023, Table 2) who use *intracranial* data from heard speech (12). We achieved this score with *non-invasive* data which is typically substantially more difficult to decode due to the low signal-to-noise ratio.

4.3. Scaling Speech Decoding With Unlabelled Data

Figure 3 shows that performance scales predictably with unlabelled data volume, following distinct patterns for differ-

*Due to the large computational cost of processing embeddings for MEG data in BrainBERT, we restrict pre-training of experiments in part C to approximately 30 hours and use only subject 001 of the downstream dataset for training and evaluation.

ent tasks and datasets. For speech detection on Armeni et al. (2022), we observe logarithmic scaling in log-space (log-log scaling), suggesting diminishing but continued returns with increased data. For other tasks, ROC AUC improves log-linearly, indicating robust scaling potential. Importantly, even our smallest pre-training dataset beats chance performance, while our largest (160 hours) continues to show gains without plateauing. Notably, we have scaled far beyond the data regime of prior surgical and non-surgical work and yet performance has continued to scale. Thus, our self-supervision approach may remain useful as the volume of open data in the field continues to rapidly increase.

Our results also reveal several new and notable phenomena. We scaled up the pre-training dataset by increasing the number of subjects. Since this led to consistent and almost monotonic improvements in downstream accuracy, our method is an exception to the consensus that pooling subjects worsens generalisation. As we pre-trained our model with a *different* dataset to those we fine-tuned on, our representation shows *cross-dataset generalisation*. This is surprising as the Armeni et al. (2022), Gwilliams et al. (2023), and our pre-training dataset all use different scanners entirely. Performing well across these datasets indicates that, together, 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 data whatsoever yet still improved downstream accuracy on language tasks. Remarkably, this shows that unlabelled brain data collected from *any* task (including those that are not linguistic) can be used to improve speech decoding performance.

Since the results show improvements on both downstream tasks, this indicates that our pretext tasks are sufficiently generic to produce representations that work with multiple speech decoding tasks 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 in ROC AUC is steeper for Armeni et al. (2022). This is likely to be because this dataset has more within-subject data. The weaker results for Gwilliams et al. (2023) may be a consequence of shorter intra-subject recordings, greater subject variation, and the lack of head casts in data collection. These observations support the findings of 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

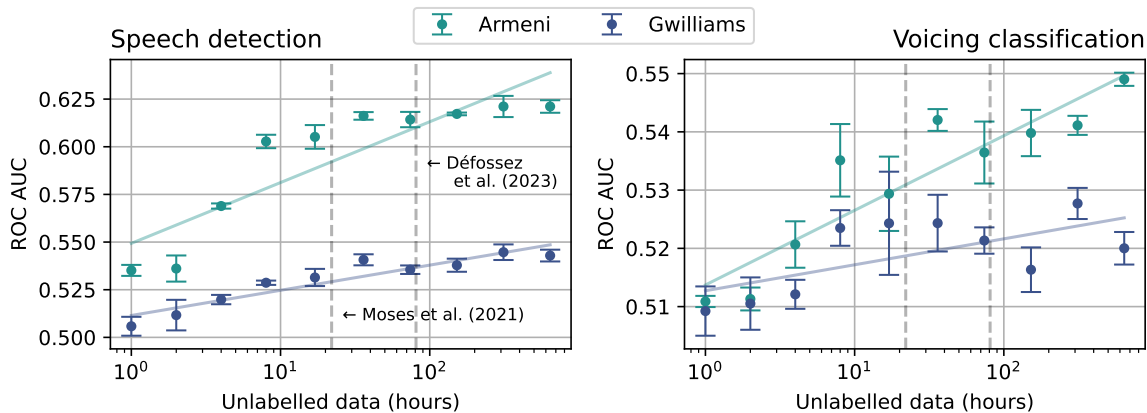


Figure 3: **Scaling unlabelled data improves generalisation.** We pre-train the model on increasing amounts of unlabelled data from Cam-CAN (Shafto et al., 2014; Taylor et al., 2017). The solid lines are linear fits and the dashed lines show the volume of data used in prior surgical (Moses et al., 2021) and non-invasive (Défossez et al., 2023) work. Unlike Table 2 (7) we do not apply Gaussian filtering to the predictions for simplicity. The best improvements are statistically significant.

difficult challenge. This is critical in order to widely deploy speech BCIs for new patients. In this experiment, we use Gwilliams et al. (2023) as our downstream dataset because of the large number of participants, holding 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. 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

Given the promising scaling results with single datasets, a natural question arises: can we achieve even better performance by combining multiple MEG datasets? This is particularly challenging since datasets often use different scanning hardware and experimental protocols. Thus, it has so far not been shown in MEG.

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.

The results in Table 3 show, for the first time, that combining datasets can improve performance on downstream speech decoding tasks. It leads to better performance compared to pre-training on either dataset alone. It is surprising that

Table 3: **Aggregating unlabelled datasets outperforms single studies in speech detection.** For the first time with MEG, we show that unlabelled pre-training data from multiple studies with different hardware profiles can be aggregated while gaining the benefits of scaling. Combining data leads to a significant ($p < 0.05$) improvement.

Pre-training Data	Hours	ROC AUC
CamCAN ^{1,2}	159	.630 \pm .0001
MOUS ³	160	.614 \pm .0004
CamCAN ^{1,2} + MOUS ³	319	.638\pm.0002

¹Shafto et al. (2014) ²Taylor et al. (2017)

³Schoffelen et al. (2019)

pre-training on Cam-CAN was better than pre-training on MOUS given that MOUS and the downstream dataset both used speech tasks and were acquired on the same MEG scanner. Cam-CAN, by contrast, did not use a speech task and was acquired on a different MEG scanner. We hypothesize that the better results for Cam-CAN are due to it being cleaner. During our experiments, we found that data quality, even among unlabelled data, can have a significant effect as artefacts in recordings disrupt learning.

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 and data availability. Increasing the number of datasets (e.g., by including EEG too) could further improve results. Just as increasing the number of subjects (rather than only within-subject data) improves novel subject generalisation, a larger

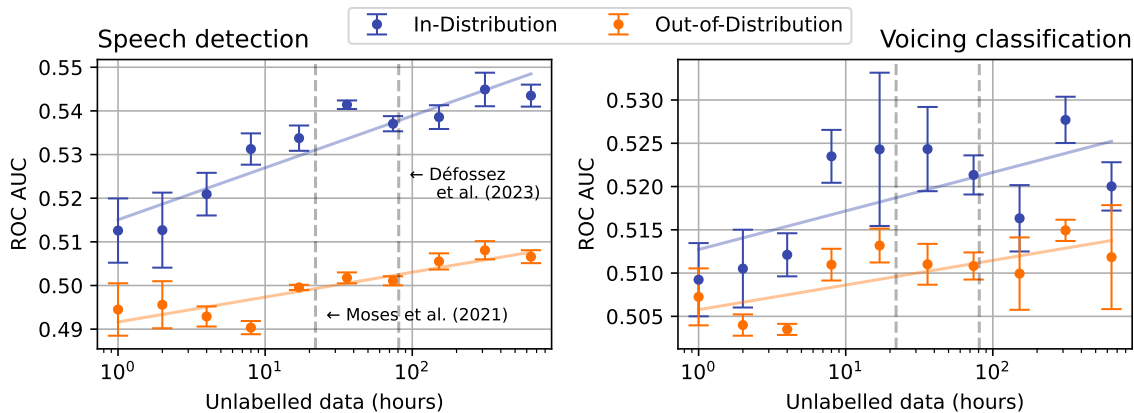


Figure 4: **Scaling unlabelled data improves novel subject generalisation.** We train a linear probe on [Gwilliams et al. \(2023\)](#). When *in-distribution*, we evaluate on held-out sessions; when *out-of-distribution*, we evaluate on held-out subjects. The lines represent the same as in Figure 3. The best improvements are statistically significant.

number of datasets may be key to scaling results when datasets are aggregated in pre-training.

4.6. Limitations

Although our results are significant in demonstrating a viable path forward to scale up speech BCIs, there remain a number of limitations to the present work. We focused here on two downstream tasks: speech detection and voicing classification. Ultimately, we would like to expand this work to predict full transcripts from brain recordings (i.e. *brain-to-text*). This has been achieved with surgical data ([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 of full brain-to-text in a surgical context ([Moses et al., 2021](#)) and we hope may play a similar role for non-invasive methods. In future work, we would also like to expand classification to all English phonemes as a step towards full transcript decoding.

Additionally, while we have been able to demonstrate the utility of a few pretext tasks, we do not claim to have exhausted the full set of useful tasks. Rather, we conjecture that more useful pretext tasks remain to be found and believe a useful avenue of research will be into other input representations for brain recordings. For example, this paper did not make use of spatial features when the geometry of a scanner’s sensor configuration is strongly correlated with the area of the brain from which the signal is derived. Another limitation is our emphasis on heard speech over other types of speech, such as attempted or imagined speech. We hypothesise that the same methods presented here will generalise to these other varieties of speech, though this has yet to be shown.

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 by scaling up the number of datasets used in training. In sharing this work now, we believe that the current proof of concept will be sufficiently impactful to the field as we continue to actively scale up the datasets that we can leverage.

5. Conclusion

Speech decoding from the brain has been limited by the field’s inability to scale up data to leverage deep learning. Prior methods have been unable to aggregate data across different datasets, labels, or subjects to scale up because of heterogeneity in recording hardware, experiment design, and participants. A handful of studies have shown weak signals towards alleviating these issues. But until now, no one has developed a general solution. We present a unified solution through data-efficient, self-supervised pretext tasks that overcome these fundamental scaling challenges. Our experiments demonstrate not just scaling with heterogeneous data, but generalisation across datasets, subjects, and tasks. They also show significant improvements of up to 27% compared to the prior state-of-the-art and even provide evidence of matching surgical decoding performance. Our method unlocks the potential of the bitter lesson, providing a general method to exploit more computation by using more data. We implore the research community to employ the vast 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.

Acknowledgements

We would like to thank Botos Csaba for many early insightful discussions which helped shaped the direction of this work. In alphabetical order, thanks also to Mats W.J. van Es for technical assistance with the OSL library, Yonatan Gideon for advice on data splits, Minqi Jiang for an encouraging conversation on scaling unsupervised representation learning, Brian Liu for technical contributions which did not reach the final paper, and Miran Özdoğan for reviewing a draft of this work. Finally, we thank the rest of the PNPL group for their continued and unwavering support.

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. <http://dx.doi.org/10.5281/zenodo.22558>.

DJ is supported by an AWS Studentship from the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (AIMS) (EP/S024050/1). GL is supported by an EPSRC Studentship. MW is supported by the Wellcome Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC, the Dementia Platform UK (RG94383/RG89702) and the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. OPJ is supported by the MRC (MR/X00757X/1), Royal Society (RG\R1\241267), NSF (2314493), NFRF (NFRFT-2022-00241), and SSHRC (895-2023-1022).

Impact Statement

This work uses publicly available datasets from human studies (Armeni et al., 2022; Gwilliams et al., 2023; Shafto et al., 2014; Taylor et al., 2017; Schoffelen et al., 2019), each with their own ethical approvals and documentation available in their respective publications.

Neural speech decoding research has transformative potential for healthcare and assistive technology. Advances could help paralysed patients communicate freely and assist those with communication difficulties. By developing non-invasive methods, the field opens up the possibility of broader access to these technologies without the risks of surgical implants.

However, we acknowledge potential societal risks as this technology matures:

- Privacy and Data Protection: Brain signals contain highly sensitive personal information, raising concerns about data security and individual privacy.
- Consent and Misuse: Advanced decoding capabilities

could enable unauthorized access to neural information, requiring robust safeguards against exploitation.

- Societal Impact: Widespread adoption could affect privacy norms around inner speech, while unequal access could exacerbate existing inequalities.

We focus specifically on decoding heard speech rather than inner speech, limiting potential misuse. Nevertheless, we recognize that advances in heard speech decoding contribute to the broader development of neural decoding technology. We encourage the research community to actively engage with these ethical considerations as the field progresses.

References

- Agrawal, P., Carreira, J., and Malik, J. 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>.
- Anumanchipalli, G. K., Chartier, J., and Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568:493 – 498, 2019.
- Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. A 10-hour within-participant 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>.
- Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A. G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., and Goldblum, M. 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>.
- Banville, H. J., Moffat, G., Albuquerque, I., Engemann, D., Hyvärinen, A., and Gramfort, A. Self-supervised representation learning from electroencephalography 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>.
- Bressler, S. L. and Richter, C. G. Interareal oscillatory synchronization in top-down neocortical processing. *Current Opinion in Neurobiology*, 31:62–66, 2015.

- Buzsáki, G. and Wang, X.-J. Mechanisms of gamma oscillations. *Annual Review of Neuroscience*, 35:203–225, 2012.
- Cai, D., Chen, J., Yang, Y., Liu, T., and Li, Y. Mbrain: A multi-channel self-supervised learning framework for brain signals. In Singh, A. K., Sun, Y., Akoglu, L., Gunopulos, D., Yan, X., Kumar, R., Ozcan, F., and Ye, J. (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>.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. 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>.
- Chen, X., Wang, R., Khalilian-Gourtani, A., Yu, L., Dugan, P., Friedman, D., Doyle, W., Devinsky, O., Wang, Y., and Flinker, A. 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>.
- Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., and Qiu, L. 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>.
- Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. The auditory representation of speech sounds in human motor cortex. *eLife*, 5:e12577, 2016.
- Csaky, R., van Es, M. W. J., Jones, O. P., and Woolrich, M. W. Group-level brain decoding with deep learning. *Human Brain Mapping*, 44:6105 – 6119, 2022. URL <https://doi.org/10.1002/hbm.26500>.
- Csaky, R., van Es, M. W., Jones, O. P., and Woolrich, M. 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>.
- Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. 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>.
- Doersch, C., Gupta, A., and Efros, A. A. 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>.
- Défossez, A., Caucheteux, C., Rapin, J., Kбели, O., and King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 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>.
- Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 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>.
- Fries, P. Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, 32:209–224, 2009.
- Gibiansky, A., Arik, S. Ö., Diamos, G. F., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. Deep voice 2: Multi-speaker neural text-to-speech. In *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.
- Gidaris, S., Singh, P., and Komodakis, N. 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->.
- Giraud, A.-L. and Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15(4): 511–517, April 2012. ISSN 1546-1726. doi: 10.1038/nn.3063. URL <https://www.nature.com/articles/nn.3063>.
- Gwilliams, L., Flick, G., Marantz, A., Pylkkänen, L., Poeppel, D., and King, J.-R. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 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>.
- Hall, E. L., Robson, S. E., Morris, P. G., and Brookes, M. J. The relationship between MEG and fMRI. *NeuroImage*, 102:80–91, 2014. URL <https://doi.org/10.1016/j.neuroimage.2013.11.005>.
- Hamilton, L. S., Edwards, E., and Chang, E. F. A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, 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>.
- Jiang, W., Zhao, L., and liang Lu, B. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=QzTpTRVtrP>.
- Jo, H., Yang, Y., Han, J., Duan, Y., Xiong, H., and Lee, W. H. Are EEG-to-text models working? *arXiv*, 2024. doi: <https://arxiv.org/abs/2405.06459>.
- Jumper, J. M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohli, S. A. A., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. URL <https://doi.org/10.1038/s41586-021-03819-2>.
- Kostas, D., Aroca-Ouellette, S. T., and Rudzicz, F. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15, 2021.
- Langland-Hassan, P. and Vicente, A. *Inner Speech: New Voices*. Oxford University Press, 2018.
- Larsson, G., Maire, M., and Shakhnarovich, G. Learning representations for automatic colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, 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.
- Le, T. and Shlizerman, E. STNDT: modeling neural population activity with spatiotemporal transformers. In *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.
- Lopes da Silva, F. EEG and MEG: Relevance to Neuroscience. *Neuron*, 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>.
- Loshchilov, I. and Hutter, F. 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>.
- Luo, H. and Poeppel, D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010, 2007.
- Luo, H., Liu, Z., and Poeppel, D. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLOS Biology*, 8(8):e1000445, 2010.
- Mai, G., Minett, J. W., and Wang, W. S. Y. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *NeuroImage*, 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>.
- Makin, J. G., Moses, D. A., and Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23:575–582, 2019. URL <https://api.semanticscholar.org/CorpusID:199639966>.
- Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., and Pasley, B. N. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7:14, 2014.
- Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014. doi: DOI: 10.1126/science.1245994.
- Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., and Chang, E. F. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023.

- Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., and Chang, E. F. 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>.
- Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (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>.
- Peelle, J. E., Gross, J., and Davis, M. H. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, 23(6): 1378–1387, 2012.
- Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. FiLM: Visual reasoning with a general conditioning layer. In McIlraith, S. A. and Weinberger, K. Q. (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>.
- Piai, V., Roelofs, A., and Maris, E. 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>.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 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>.
- Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., and Hagoort, P. 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>.
- Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J. S., Dalgleish, T., Henson, R. N. A., Brayne, C., and Matthews, F. E. 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.
- Strauß, A., Henry, M. J., Scharinger, M., and Obleser, J. Alpha phase determines successful lexical decision in noise. *Journal of Neuroscience*, 35(7):3256–3262, 2015.
- Sutton, R. The bitter lesson. *Incomplete Ideas (blog)*, 2019. URL <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- Tagliasacchi, M., Li, Y., Misiunas, K., and Roblek, D. SEANet: A multi-modal speech enhancement network. In Meng, H., Xu, B., and Zheng, T. F. (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 1126–1130. ISCA, 2020. doi: 10.21437/INTERSPEECH.2020-1563. URL <https://doi.org/10.21437/Interspeech.2020-1563>.
- Tang, J., LeBel, A., Jain, S., and Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5): 858–866, May 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL <https://www.nature.com/articles/s41593-023-01304-9>.
- Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., Tyler, L. K., Group, C.-C., and Henson, R. N. A. The Cambridge centre for ageing and neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262 – 269, 2017.
- Vidaurre, D., Hunt, L. T., Quinn, A. J., Hunt, B. A. E., Brookes, M. J., Nobre, A. C., and Woolrich, M. W. Spontaneous cortical activity transiently organises into 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/articles/s41467-018-05316-z>.

- Wandelt, S. K., Bjånes, D. A., Pejisa, K., Lee, B., Liu, C. Y., and Andersen, R. Representation of internal speech by single neurons in human supramarginal gyrus. *Nature human behaviour*, 2024. URL <https://doi.org/10.1038/s41562-024-01867-y>.
- Wang, C., Subramaniam, V., Yaari, A. U., Kreiman, G., Katz, B., Cases, I., and Barbu, A. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=xmcYx_reUn6.
- Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. EEGPT: pretrained transformer for universal and reliable representation of EEG signals. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- Weiss, S. and Mueller, H. M. “Too many betas do not spoil the broth”: the role of beta brain oscillations in language processing. *Frontiers in Psychology*, 3, 2012. doi: <https://doi.org/10.3389/fpsyg.2012.00201>.
- Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06377-x. URL <https://www.nature.com/articles/s41586-023-06377-x>.
- Yang, C., Westover, M. B., and Sun, J. BIOT: biosignal transformer for cross-data learning in the wild. In *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.
- Ye, J., Collinger, J. L., Wehbe, L., and Gaunt, R. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *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.
- Yi, K., Wang, Y., Ren, K., and Li, D. Learning topology-agnostic EEG representations with geometry-aware modeling. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (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.
- Yuan, Z., Shen, F., Li, M., Yu, Y., Tan, C., and Yang, Y. BrainWave: A brain signal foundation model for clinical applications. 2024. URL <https://api.semanticscholar.org/CorpusID:267740511>.
- Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: 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>.
- Zhang, D., Yuan, Z., Yang, Y., Chen, J., Wang, J., and Li, Y. Brant: Foundation model for intracranial neural signal. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (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.
- Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (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. URL https://doi.org/10.1007/978-3-319-46487-9_40.

A. Experiment Details

Pre-training data. 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.

Labelled training data. When training with Armeni et al. (2022), we hold out session 009 for validation and 010 for testing. Similarly, when fine-tuning with Gwilliams et al. (2023), we hold out task 1 from subjects 23, 24, 25, 26, and 27, using these sessions for evaluation only. As there is limited within-subject data in the latter dataset, we did not hold out a session from all subjects as before. For our novel subject experiments, we hold out subjects 1, 2, and 3 entirely 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 sessions for each task are repeats of the task. This means that while the recording itself is unseen, in this dataset, it is possible that held-out sessions use stimuli that may be shared.

Adapting models for more sensors. As BrainBERT is designed for single-electrode representations, for a fair comparison, to take into account all sensors, we ensured our linear classifier is applied to a concatenated embedding over all sensors. This is a large vector and leads to very computationally expensive training and is the reason we had to reduce the pre-training and downstream data in part C of Table 2. We similarly concatenate sets of 19 sensors when evaluating EEGPT.

Statistical testing. Significance was determined using one-sided t -tests with $p < 0.05$ as the threshold for significance.

B. Hyperparameters

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 4.

Table 4: **Experimental hyperparameters.**

Hyperparameter	Value
Window length (s)	0.5
ρ (phase)	0.5
ρ (amplitude)	0.2
$\{w_1, w_2, w_3\}$	$\{1.0, 1.0, 1.0\}$
d_{shared}	512
d_{backbone}	512
SEANet convolution channels	(512, 512, 512, 512)
SEANet downsampling ratios	(5, 5, 1)
FiLM conditioning dimension	16
Subject embedding dimension	16
Pre-training epochs	200
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Learning rate	0.000066
Train ratio	0.8
Validation ratio	0.1
Test ratio	0.1

Why ρ ? The choice of the proportion of sensors to apply transformations to, $\rho = 0.5$ for phase shift prediction and $\rho = 0.2$ for amplitude prediction, were determined through a hyperparameter search. It is important to note that $\rho \geq .5$ leads to the same effect as $1 - \rho$ for the complementary amplitude or phase shift. We conjecture that a smaller ρ is optimal for amplitude scale prediction since this leads to representations that are 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 the STG, which is associated with speech onset (Hamilton et al., 2018). In contrast, a larger ρ 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 the sensors in a MEG scanner should encode information about this feature.

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 Cam-CAN ([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.