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

Dulhan Jayalath[1] **Gilad Landau**[1] **Brendan Shillingford**[2]
Mark Woolrich[3] **Oiwi Parker Jones**[1]
{[1]PNPL 🍍, [3]OHBA}, University of Oxford [2]Google DeepMind
{dulhan, oiwi}@robots.ox.ac.uk

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. To address this, we develop an initial set of neuroscience-inspired self-supervised objectives, together with a neural architecture, for representation learning from heterogeneous and unlabelled neural recordings. Experimental results show that representations learned with these objectives scale with data, generalise across subjects, datasets, and tasks, and outperform learning using only labelled data. In addition, we set new benchmarks for two foundational speech decoding tasks. Taken together, these methods now unlock the potential for training speech decoding models with orders of magnitude more existing data.

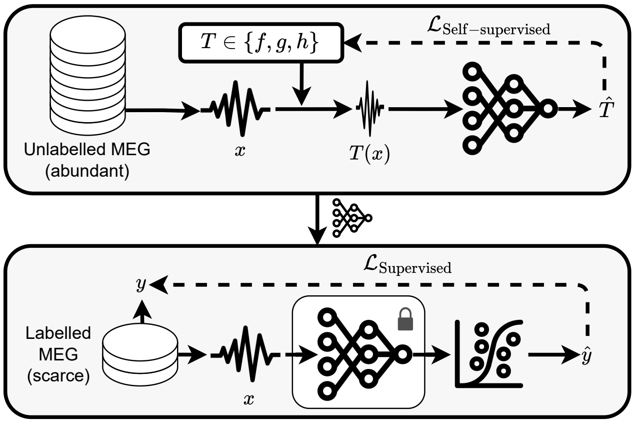
$T \in \{f, g, h\}$
$\mathcal{L}_{\text{Self-supervised}}$
Unlabelled MEG (abundant)
x
$T(x)$
$\hat{T}$
$\mathcal{L}_{\text{Supervised}}$
y
Labelled MEG (scarce)
x
$\hat{y}$

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

1^{st} Workshop on NeuroAI @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

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 [48]. 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 [27, 6, 40, 44].

In the domain of brain data, and of tasks like speech decoding, the bitter lesson has not yet been fully assimilated. State-of-the-art *brain-computer interfaces (BCIs)* have tried to scale up labelled datasets for individual subjects, using either invasive [38, 58] or non-invasive brain recordings [50], mapping these to transcripts of attempted or imagined speech. Yet, a number of obstacles to scale remain. With few exceptions at present (e.g. [15]), 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 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, as has begun to happen in neighbouring sub-fields (e.g. [25]), holds the promise of training models on collective, internet-scale data.

Given the scarcity of labelled data, *self-supervised learning (SSL)* appears promising as an avenue for domains where such data is rare or hard to obtain [3]. In the SSL paradigm, *pretext* tasks pre-train a model on unlabelled data by generating implicit training labels through transformations of the input data in order to help a downstream task. We develop a set of these tasks, informed by advances in neuroscience, for learning with unlabelled brain data (Figure 1) and design an architecture for processing continuous multi-sensor neuroimaging signals. In order to scale existing 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 different subjects and neuroimaging scanners. We evaluate the representations learned with our approach on heard speech datasets acquired with non-invasive *magnetoencephalography (MEG)*, setting the first baselines for speech detection and voicing classification on this data. The results not only demonstrate that scaling with unlabelled data works in speech decoding, but also shows that these representations can generalise across datasets, tasks, and even novel subjects for the first time. Our main contributions are:

- 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 **neural architecture** for learning these self-supervised objectives and training downstream speech decoding from brain data; and
- A comprehensive **experimental evaluation**, using more than twice 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 Method

We utilise a convolutional neural architecture adapted from neural audio codecs (Appendix A) for encoding heterogeneous brain signals into latent representations. By developing three pretext tasks with the objective of learning generalisable brain representations, we leverage this architecture for self-supervised learning from unlabelled data. 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.

Band prediction. In the literature, neural responses can be broadly segmented into functional frequency bands [20, 43, 34]. *Delta (δ)* waves (0.1–4 Hz) are commonly associated with the rhythmic structure of heard speech [33], *Theta (θ)* waves (4–8 Hz) reliably track [32] and phase-lock to the amplitude envelope of heard sentences [41], *Alpha (α)* waves (8–12 Hz) relate to attentional processes and the inhibition of irrelevant information, helping to focus on relevant speech signals [47], *Beta (β)* waves (12–30Hz) are implicated in top-down predictive coding [4] which affects lexical processing [57], *Gamma (γ)* waves (30–70 Hz) occur with higher cognitive functions (e.g. memory, learning, reasoning, and planning) [17, 5], and *High Gamma (γ^{high})* waves (>70 Hz) have been linked specifically to speech detection [24] and phonemic feature classification in the STG [36] as well as phonemic feature classification in the *ventral sensorimotor cortex (vSMC)* [8]. As High Gamma is a relatively wide band, we have split it into two sub-bands: *Lower High Gamma* ($\gamma^{\text{high}}_{\text{lower}}$*)* waves (70–100 Hz) and *Upper High Gamma* ($\gamma^{\text{high}}_{\text{upper}}$*)* waves (100–150 Hz). To learn representations that can distinguish between these, our band prediction task applies a band-stop filter for a randomly selected band ω to the sample x, passes the filtered sample $x^{\omega'}$ through the network backbone g and the corresponding predictor f_{band}, requiring the network to predict the frequency band that was rejected. This yields the loss

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

where B is a mini-batch of samples, $\omega \in \{\delta, \theta, \alpha, \beta, \gamma, \gamma^{\text{high}}_{\text{lower}}, \gamma^{\text{high}}_{\text{upper}}\}$, 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 [16, 52]. 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 [25] 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 ρ 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^{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. [24] 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 ρ 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.

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 a network backbone and apply a corresponding linear classifier to predict the transformation, summing the weighted losses such that our final pre-training loss is given by

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

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

Table 1: **Pre-training with pretext tasks leads to better representations for decoding speech.** In the *linear* case, we train a supervised linear classifier on the input MEG signals. In all other cases, we train a linear probe on top of our pre-trained backbone with its weights frozen. In the *no pre-training* baseline, the backbone is random. When *all* pretext tasks are used, their losses are weighted equally.

		Armeni			**Gwilliams**		
Experiment		ROC AUC	t	p	ROC AUC	t	p
Linear		0.559 ± 0.000	341	4e−6	0.527 ± 0.000	379	3e−6
Ours	No pre-training	0.519 ± 0.002	8	7e−3	0.498 ± 0.003	0	7e−1
	Amp$_{(\rho\,=\,0.2)}$	0.602 ± 0.001	114	4e−5	0.532 ± 0.005	6	1e−2
	Phase$_{(\rho\,=\,0.5)}$	0.603 ± 0.003	35	4e−4	0.535 ± 0.003	12	3e−3
	Band	0.616 ± 0.003	44	3e−4	0.542 ± 0.001	46	2e−4
	All tasks	$\mathbf{0.621} \pm 0.003$	36	4e−4	$\mathbf{0.543} \pm 0.003$	13	3e−3

3 Experiments

Datasets. We focus our evaluation on MEG data as the signal is rich, with better spatial resolution than EEG [30] and faster sampling rates than fMRI [23]. Unless specified otherwise, our experiments use Cam-CAN [46, 51] as an unlabelled representation learning dataset for pre-training. This is a study containing 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 participants listen to short stories or audiobooks. Armeni et al. [2] contains 3 subjects who listen to 10 hours of recordings each (30 hours total) while Gwilliams et al. [22] has 27 subjects, each recorded for 2 hours (54 hours total). Overall, we utilise over 200 hours of data. To the best of our knowledge, this is the largest volume of MEG data ever used for speech decoding.

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 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. Critically, these tasks remain unsolved within the domain of non-invasive speech decoding.

Training. 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, showing test ROC AUC at the best validation ROC AUC (early stopping). We quote uncertainty as the standard error of the mean over three seeds. We calculate the t-score and p-value from one-sample one-sided t-tests against chance.

Learning generalisable representations using pretext tasks. In Table 1, we show that all of our pretext tasks lead to statistically significant results, indicating that these pre-training objectives are useful for speech decoding. Interestingly, the combination of all pretext tasks leads to better generalisation than any task on its own. This may be because our pretext tasks capture complementary properties in time- and frequency-space. When comparing our pre-training approach to a similarly parameterised supervised method, pre-training comes out on top. Thus, it is apparent that pre-training with unlabelled data improves generalisation over using MEG signals as input directly.

Scaling speech decoding with unlabelled data. We examine how performance changes as we increase the amount of pre-training data with speech detection and voicing classification. Figure 2 shows that any unlabelled data is sufficient to pre-train a useful representation and adding unlabelled data improves generalisation. As we scaled up the pre-training dataset by increasing the number of subjects, our self-supervised method is an exception to the common consensus that pooling subjects worsens generalisation. As we pre-trained with a *different* dataset to those we fine-tuned on, our representation shows *cross-dataset generalisation*. This is notable as the Armeni et al. [2], Gwilliams et al. [22], and pre-training datasets all use different scanners entirely. To our surprise, scaling our pre-training data, which contained no language tasks whatsoever, improved speech task accuracy. Remarkably, this shows that non-linguistic tasks can be used to scale downstream performance.

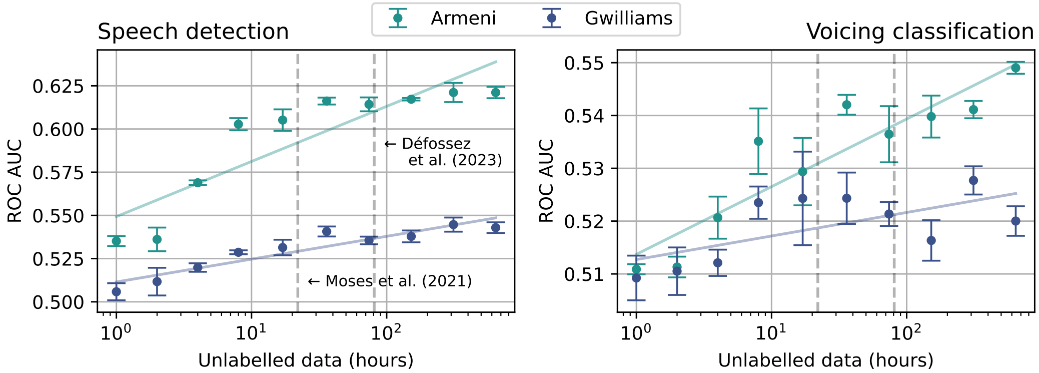

Figure 2: **Scaling unlabelled data improves generalisation.** We pre-train the model on increasing amounts of unlabelled data from Cam-CAN [46, 51]. The solid lines are the best linear fits to the data and the dashed lines show the amount of data used in prior surgical [38] and non-invasive [15] work.

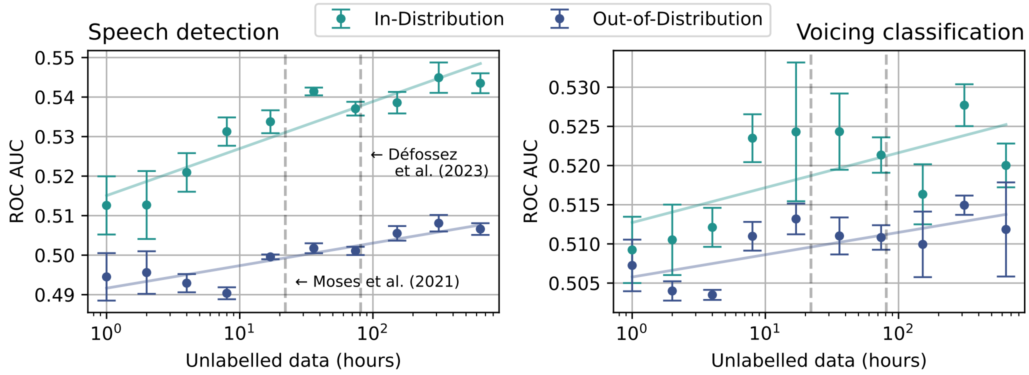

Figure 3: **Scaling unlabelled data improves novel subject generalisation.** We fine-tune on Gwilliams et al. [22]. When *in-distribution*, we evaluate on held-out sessions from subjects in the training set; when *out-of-distribution*, we evaluate on three held-out subjects. The solid lines are the best linear fits while the dashed lines show the amount of data used in prior surgical [38] and non-invasive [15] work.

Finally, it is notable that we have gone far beyond the data regime of prior surgical and non-surgical work and yet performance appears to continue to scale.

When generalising to novel subjects rather than only across subjects, models typically struggle since brain data is variable across participants [9]. However, when scaling up the pre-training data, Figure 3 reveals a positive trend in performance for *novel* subjects, suggesting that scale can alleviate these issues. As far as we are aware, this is also the first result to demonstrate novel subject generalisation in speech decoding from MEG.

3.1 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 voice 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 [38, 58] but not yet convincingly with non-invasive methods like MEG or EEG [26]. Speech detection has played an important role in the development of full brain-to-text in a surgical context [38] and we hope may play a similar role for non-invasive methods. Prior work has further used voice classification as a stand in for phoneme classification [21], and we have been able to improve on these results here. In future work, we would like to expand this to all English phonemes. Secondly, while we have been able to demonstrate the utility of a few pretext tasks here, 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. 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. 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 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.

4 Conclusion

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. Some studies have shown weak signals towards alleviating these issues. But until now, no one has developed a general solution. We provided a unified method that leverages unlabelled data using generic pretext tasks that shows that all of these problems can be solved. We verified this with experiments showing that our method scales with heterogeneous data and generalises across datasets, subjects, and tasks. We implore the research community to employ vast quantities of data and compute to realise this potential. If scale is all you need in speech decoding, then the bitter lesson may not be so bitter.

Acknowledgments and Disclosure of Funding

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 Gideoni 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 Özdogan for reviewing a draft of this work.

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).

References

[1] 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*, pages 37–45. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.13. URL `https://doi.org/10.1109/ICCV.2015.13`.

[2] Kristijan Armeni, Umut Güçlü, Marcel van Gerven, and Jan-Mathijs Schoffelen. 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`.

[3] 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`.

[4] Steven L Bressler and Craig G Richter. Interareal oscillatory synchronization in top-down neocortical processing. *Current Opinion in Neurobiology*, 31:62–66, 2015.

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

[6] 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*, pages 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL `https://doi.org/10.1109/ICCV48922.2021.00951`.

[7] Xupeng 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*, pages 1–14, April 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00824-8. URL `https://www.nature.com/articles/s42256-024-00824-8`.

[8] Connie Cheung, Liberty S Hamilton, Keith Johnson, and Edward F Chang. The auditory representation of speech sounds in human motor cortex. *eLife*, 5:e12577, 2016.

[9] 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`.

[10] 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`.

[11] 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`.

[12] 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`.

[13] 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*, pages 1422–1430. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.167. URL `https://doi.org/10.1109/ICCV.2015.167`.

[14] 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, editors, *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, December 10 - 16*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/1f2fd23309a5b2d2537d063b29ec1b52-Abstract-Conference.html`.

[15] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 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`.

[16] Pascal Fries. 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`.

[17] Pascal Fries. Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, 32:209–224, 2009.

[18] 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, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2962–2970, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/c59b469d724f7919b7d35514184fdc0f-Abstract.html`.

[19] 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=S1v4N2l0-`.

[20] Anne-Lise Giraud and David Poeppel. 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`.

[21] Laura Gwilliams, Jean-Rémi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, 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`.

[22] 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. *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`.

[23] Emma L. Hall, Siân E. Robson, Peter G. Morris, and Matthew J. Brookes. The relationship between MEG and fMRI. *NeuroImage*, 102:80–91, 2014. URL `https://doi.org/10.1016/j.neuroimage.2013.11.005`.

[24] Liberty S. Hamilton, Erik Edwards, and Edward F. Chang. 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`.

[25] Weibang Jiang, Liming Zhao, and Bao liang Lu. 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`.

[26] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are EEG-to-text models working? *arXiv*, 2024. doi: https://arxiv.org/abs/2405.06459.

[27] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, 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, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL `https://doi.org/10.1038/s41586-021-03819-2`.

[28] Peter Langland-Hassan and Agustín Vicente. *Inner Speech: New Voices*. Oxford University Press, 2018.

[29] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *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*, pages 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`.

[30] Fernando Lopes da Silva. 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`.

[31] 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`.

[32] Huan Luo and David Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010, 2007.

[33] Huan Luo, Zuxiang Liu, and David Poeppel. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLOS Biology*, 8(8):e1000445, 2010.

[34] Guangting Mai, James W. Minett, and William S. Y. Wang. 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`.

[35] 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. *Frontiers in Neuroengineering*, 7:14, 2014.

[36] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014. doi: DOI:10.1126/science.1245994.

[37] 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.

[38] 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`.

[39] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *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*, pages 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`.

[40] 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`.

[41] 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.

[42] 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, editors, *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*, pages 3942–3951. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11671. URL `https://doi.org/10.1609/aaai.v32i1.11671`.

[43] 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`.

[44] 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*, pages 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

[45] 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.

[46] 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.

[47] 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.

[48] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[49] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. SEANet: A multi-modal speech enhancement network. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1126–1130. ISCA, 2020. doi: 10.21437/INTERSPEECH.2020-1563. URL https://doi.org/10.21437/Interspeech.2020-1563.

[50] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 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.

[51] Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN Group, and Richard N. A. Henson. 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.

[52] Diego Vidaurre, Laurence T. Hunt, Andrew J. Quinn, Benjamin A. E. Hunt, Matthew J. Brookes, Anna C. Nobre, and Mark W. Woolrich. 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.

[53] Sarah K Wandelt, David A. Bjånes, Kelsie Pejsa, Brian Lee, Charles Y Liu, and Richard Andersen. 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.

[54] Bo Wang, Xiran Xu, Longxiang Zhang, Boda Xiao, Xihong Wu, and Jing Chen. Semantic reconstruction of continuous language from MEG signals. *CoRR*, abs/2309.07701, 2023. doi: 10.48550/ARXIV.2309.07701. URL https://doi.org/10.48550/arXiv.2309.07701.

[55] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. 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.

[56] Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, February 22 - March 1*, pages 5350–5358. AAAI Press, 2022. doi: 10.1609/AAAI.V36I5.20472. URL https://doi.org/10.1609/aaai.v36i5.20472.

[57] Sabine Weiss and Horst M. Mueller. “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.

[58] 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. Shenoy, and Jaimie M. Henderson. 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`.

[59] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 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`.

[60] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *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*, pages 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`.

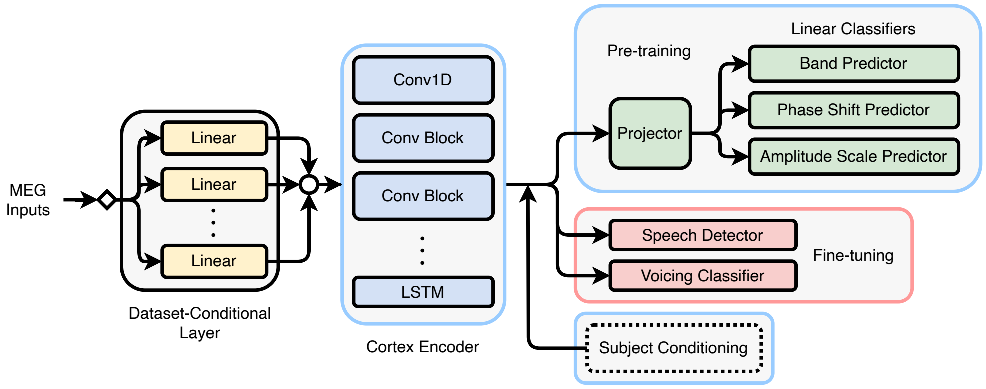

Figure 4: **Architecture overview.** Inputs are projected into a shared dimension, then encoded. In pre-training, all weights are trainable except for modules in light-red, while in fine-tuning, modules with light-blue borders are frozen and modules with light-red borders are unfrozen. Dashed borders indicate optional components.

A Network Architecture

We design a two-stage neural network architecture (Figure 4). The pre-training stage uses pretext tasks to train a representation with unlabelled brain data. Then, the fine-tuning stage uses this representation to learn the downstream task by training a linear probe with labelled data.

Normalisation. 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.

Backbone. The network takes as input the standardised sample windows. To combine heterogeneous datasets, which have different 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 [59, 12]. Specifically, our convolutional encoder adapts the implementation of the SEANet architecture [49] used in Défossez et al. [12]. 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 (where $\tau < t$), each of dimension d_{backbone} (i.e. an $\mathbb{R}^{d_{\text{backbone}}\times\tau}$ output).

Pre-training. Following the advice of Balestriero et al. [3, 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.

Fine-tuning. In this stage, we introduce classifiers for the downstream tasks and train with labelled data and supervised learning (Figure 4). To fine-tune, we train a classifier from scratch on top of the pre-trained representation, which remains frozen. Thus, we backpropagate only through the classifier. A new trainable dataset-specific linear layer can also 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 [10].

Subject conditioning. 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 [9]. In the speech literature, models include speaker conditioning to account for these differences [18]. We take a similar approach by introducing subject conditioning, represented as a subject embedding, to our model. With a SEANet-based architecture, Zeghidour et al. [59] find that conditioning is equally as effective at the encoder bottleneck as in other stages of the model. Hence, we place ours at the cortex encoder bottleneck for simplicity and use *feature-wise linear modulation (FiLM)* [42].

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 muscle movement artifacts, we apply a low-pass filter at 125Hz as well as a high-pass filter at 0.5Hz 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 multiples of 50Hz to account for line noise. Next, we downsample the signal to 250Hz, avoiding aliasing at frequencies up to our low-pass filter threshold. Finally, we detect bad sensor channels, those with significant noise and artifacts, using a variance threshold and replace them by interpolating the spatially nearest sensors.

B 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 [38, 37, 58, 7] and non-invasive studies [50]. 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, work by Tang et al. [50] 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.

Specific studies into the limitations of generalising models between subjects show that while performance decreases on average when subjects are pooled, there are exceptions. Csaky et al. [9] find that a subset of individuals perform better when evaluated with a group-level model than with individual models. Exploiting audio data in a multi-modal framework, Défossez et al. [15] show that decoding performance improves for a segment identification task as data from multiple subjects listening to connected speech are aggregated. Although they repeat the result within two MEG and two EEG datasets, Défossez et al. [15] do not show any improvements for pooling data across datasets. Moreover, they do not combine data from studies with different labels either; cf. [56, 14, 54]. Unfortunately, two of these papers [56, 14] included a bug in their evaluation code. As such, their methods may perform no better than a baseline that provides pure noise inputs to the model [26].

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 here on listening because it is easier to decode than imagined speech (e.g. [35]). There is also some evidence of a functional overlap between listening and imagined speech representations in the brain [53], though we acknowledge that the question of overlap has been contested [28]. Prior work has also investigated the two tasks that we focus on here [11, 38, 22]. The first of these, speech detection, formed the backbone to Moses et al. [38], where a speech detection model was trained and subsequently used to detect isolated words, which were in turn classified and checked against a language model to generate acceptable sentences. Hamilton et al. [24] further elaborated on the neural anatomy underlying speech detection, categorising neural responses in the *superior temporal gyrus (STG)* to sustained speech and speech onset. As for the second task, voicing classification, Gwilliams et al. [22] used this task as a proxy for phoneme classification, as pooling phonemes into unvoiced or voiced segments (e.g. /p t k f s/ vs /b d g v z/) improves data efficiency. We note that voicing classification and speech detection are related tasks as voicing is a subclass of speech. This makes them foundational for building hierarchical speech decoding pipelines similar to prior surgical decoding work [38, 58].

In the computer vision literature, there have been a plethora of methods that use self-supervised pretext tasks for representation learning [1, 13, 39, 29, 60, 19]. Until now, similar approaches have not translated to the brain decoding literature. However, prior work has used other methods to leverage unlabelled brain data. For example, Jiang et al. [25] succeeded in cross-dataset and cross-task generalisation, using a transformer with tokenised brain signals and a masked token prediction objective. Although this work combined unlabelled EEG datasets, it only achieved improvements on non-speech tasks. Wang et al. [55] used a similar approach, replacing tokens with contextualised embeddings of time-frequency input representations. They attained impressive speech detection results but with invasive neural recordings, which are comparatively rare and thus have much less potential to scale than non-invasive data.

C Experiment Details

We pre-train with non-overlapping sample windows from all subjects and sessions. We adjust the amount of unlabelled data used 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.

When fine-tuning with Armeni et al. [2], we hold out session 010 from all subjects during training and validation, using this for evaluation. Similarly, when fine-tuning with Gwilliams et al. [22], we hold out session 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. We always fine-tune to completion (usually around 300 epochs), taking the test metric at the best validation loss (early stopping).

In all experiments, 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.

D 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 2.

Table 2: **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 [31]
Learning rate	0.000066
Train ratio	0.8
Validation ratio	0.1
Test ratio	0.1

E 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.

F Licences For Datasets And Code

The Armeni et al. [2] dataset is distributed under CC-BY-4.0 while the Gwilliams et al. [22] dataset is distributed under the CC0 1.0 Universal licence. The Schoffelen et al. [45] dataset is distributed with a RU-DI-HD-1.0 licence from the Donders institute. The licence for the Cam-CAN [46, 51] dataset is unknown. The SEANet code adapted from Défossez et al. [12] is distributed under the MIT licence, and the OSL library, which we use for preprocessing, is under the BSD-3-Clause licence.

G Broader Impacts

Decoding speech from non-invasive brain recordings is likely to bring about significant positive societal impacts. Research in this field could enable paralysed patients to communicate freely and materially assist those with minor communication difficulty (e.g. stammering). As 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 technology could also present potential negative societal impacts. For one, reading inner speech creates new concerns over data controls as this information is likely to be highly sensitive and personal to individuals. Given access to this technology, there is also the risk that bad actors could extract sensitive information from target individuals without consent. Moreover, there are possible long horizon effects associated with speech decoding research. Broad adoption of this technology could lead to the gradual erosion of privacy over inner speech within society. In addition, asymmetric effects, where some individuals or organisations can read inner speech but others are unable to, could worsen societal inequality. Within the scope of this paper, we mitigate risks associated with inner speech by focusing on decoding heard speech where there is low potential for abuse. Nonetheless, we acknowledge that this is still a stepping stone towards solving inner speech decoding.