Pretraining with Masked Autoencoding Improves Speech Decoding from ECoG

Abraar R. Samar[1] and Joseph G. Makin[1]

Abstract— **Accurately decoding speech from brain signals is a challenging problem for neuroscience and machine learning, with great potential for speech neuroprosthetics and communication aids. The problem is challenging because labeled neural data tend to be noisier and scarcer than their counterparts in mainstream machine learning, preventing simple deployment of state-of-the-art neural-network architectures. Here we show how some amount of data scarcity can be overcome through a combination of self-supervised pretraining, and data augmentation during supervised training. In particular, we train transformers to reconstruct partially masked, unlabeled ECoG data (masked autoencoding), and subsequently fine-tune these networks to map sequences of ECoG data—including time-shifted and masked augmentations—to text representations of the corresponding speech. Low-level features extracted from the model's initial layers are fused with high-level representations from its final layers, giving the classifier access to complementary information across feature hierarchies. We validate our methodology on a public ECoG dataset from four participants (who were being treated for unrelated conditions), achieving significant improvement in decoding accuracy compared to the previously best-performing RNN-based model on the same dataset. This study underscores the potential of integrating advanced preprocessing, self-supervised pretraining, data augmentation, and multi-layer feature fusion for decoding brain signals with high accuracy and efficiency.**

I. Introduction

In the last decade, decoding speech from neural recordings has moved from classification of isolated phonemes [1], [2] or sentences [3], to decoding of variable-length word sequences [4], [5] and speech audio [6], [7], and most recently to clinical trials with persons who have lost the ability to speak [8], [9], [10], [11]. Progress has been driven by better algorithms and, more recently, especially by better electrodes, with increased channel count and spatial resolution. Indeed, [11] have recently shown that word error rates (WERs) can be driven below 3% on essentially open vocabularies, at least in patients who retain some ability to vocalize, if speech motor cortex is implanted with four Utah arrays.

Nevertheless, it is not clear that algorithmic advances have been exhausted, and there are reasons to prefer this avenue of improvement to complicated surgeries with multiple electrode implants. One promising direction for algorithmic improvement begins with the observation that the large majority of data recorded from participants is not used to train the decoders. Training data are confined to special sessions in which the participant is prompted to produce, seriatim, specific sentences. Typically such sessions, which are boring and, for non-speaking participants, frustrating, yield less than an hour of spoken speech per day. For electrocorticography (ECoG), achieving tolerable WERs requires $\sim$15–20 hours of data, which can therefore take about two weeks to collect [9]. When decoding from single units like the Utah array, far fewer data are needed—about 200 sentences, or approximately one hour of data collection; but decoding can noticeably deteriorate over just one week, so the decoder must be regularly recalibrated [10].

On the other hand, with an implanted array, 24 hours of data can conceivably be recorded every day, even though the vast majority of this time lies outside of controlled sessions. Making use of these data to improve decoding, even with much lower return per minute of recording, would be pure profit, since at present they are entirely wasted.

Happily, the past half decade of machine-learning research has also seen the development and successful application of a suite of unsupervised and "self-supervised" techniques for learning representations from input data alone. For example, a neural network can be trained on sequence data to aggregate representations that are highly predictive of future elements of the sequence, or transformations thereof [12], [13], [14]. When applied to raw, unlabelled speech audio, the aggregated representations also make very good features for assigning phonemes or characters to the audio sequence, that is, for underwriting automatic speech recognition. Indeed, accuracies competitive with fully supervised methods can be achieved merely by "fine-tuning" the network—or even a simple linear, final layer—on a much smaller set of phoneme-labeled audio [12], [13], [14]. Such techniques have recently been shown to be effective on neural data [15].

In this study, we show that such techniques can indeed be successfully applied to neural data for speech decoding, or more precisely for word classification from ECoG. In particular, pretraining a neural network on unlabelled ECoG with a masked autoencoding task leads to better decoding on the downstream task of word classification, i.e. after supervised training. We also show that augmenting the *labeled* training data with masking and time shifting likewise improves decoder performance.

A number of recent studies have proposed to use unsupervised or self-supervised learning of neural data to improve performance on downstream tasks. Arguably the closest to our work are Brant (Brain Neural Transformer) [16] and BrainBERT [17]. Brant is a multi-modal transformer-based

*This work was supported by an NSF CAREER Award (2339781).

[1]Elmore Family School of Electrical and Computer Engineering, 465 Northwestern Ave, West Lafayette, IN 47907 U.S.A; asamar@purdue.edu, jgmakin@purdue.edu.

network that is pretrained on intracranial recordings (sEEG) with a masked autoencoding task [16]. The authors evaluate their model on signal forecasting and imputation, as well as seizure prediction (in data recorded from epilepsy patients). However, they do not consider speech, and it is not clear how to translate performance on the tasks they do consider to word classification. BrainBERT [17] is also a transformer-based network pretrained on a masked autoencoding task. Like Brant, BrainBert is applied to sEEG data. However, signals from different electrodes are treated as separate rather than joint data. This facilitates generalization to new subjects—the model essentially learns a generic model of sEEG signals—but prevents the model from exploiting all the information available at once from a single recording. Results are reported only for binary classification tasks (e.g., speech vs. non-speech).

Self-supervised pretraining has also been applied to EEG and MEG data [18], [19], [20], [21], [22], [23], [24]. The closest in spirit to our approach are the Large Brain Model (LaBraM) of [25] and the EEGformer of [26], both of which use a combination of masked reconstruction and vector quantization. However, speech decoding is (currently) beyond the reach of EEG and MEG, and accordingly these studies focus on different (and easier) downstream tasks. EEG and MEG also have significantly different properties from intracranial signals like ECoG and sEEG.

Crucially, all of these studies focus on downstream tasks that are significantly simpler than speech decoding. Furthermore, they were not designed for ECoG (although sEEG has similar signal characteristics). Rather than adapting these model to our task, then, we have constructed a new architecture that is more closely related to the ASR literature cited above [12], [13], [14]. We describe our approach below.

II. Methods

A. Data

We analyze a public dataset (provided upon request by the original authors [5]) in which participants undergoing treatment for epilepsy read aloud sentences while neural activity was recorded with an intracranial grid of electrodes (ECoG). Each of several recording "blocks" of sentences consisted of either a subset of 460 sentences/∼1800 unique words (MOCHA-TIMIT [27] subsets), or a set of 30 sentences/∼125 unique words ("picture descriptions"). Following [5], for each participant, we trained and tested speech decoders only on the first 50-sentence (∼150 unique words) subset of MOCHA-TIMIT (henceforth "MOCHA-1") or the picture descriptions—whichever set the participant had read more of (MOCHA-1 for participants **a** and **b**; picture descriptions for participants **c** and **d**). For brevity, we refer to these as "labeled" data. The remaining ("unlabeled") sentences were used without labels for self-supervised pretraining (see **Self-Supervised Training** below). Participants **a**, **c**, and **d** had roughly an hour of "unlabeled" data; participant **b** had roughly 30 minutes.

Simultaneous with speech, electrocorticograms were recorded from study participants from the cortical areas near the Sylvian fissure, including areas associated with speech production and perception (see the original study for details [5]). Electrode grids consisted of either 256 (participants **a**, **b**, **d**) or 128 (participant **c**) nominal electrodes per participant, with 4-mm spacing. The analog ECoG signals were digitized and recorded at roughly 3 kHz.

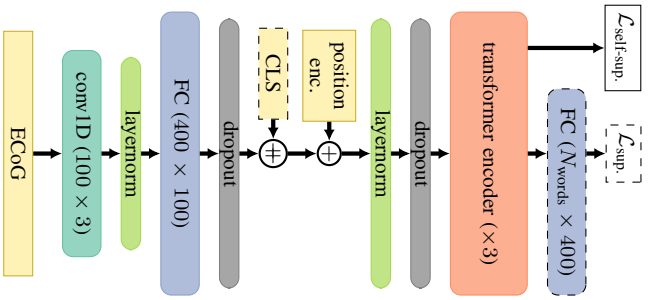

Fig. 1: Model architecture. Dashed-line elements appear only during supervised training. The symbol $+\!\!+$ indicates concatenation.

We closely followed the pre-processing scheme of the original study [5], which we omit here for brevity, with one difference: we removed channels not in speech-related areas. This mildly improved all results in preliminary experiments.

B. Network Architecture

We use a deep neural network to map each sequence of high-γ to the word that corresponds to it. The architecture, shown in Fig. 1, is inspired by recent networks for self-supervised learning [12], [13], [14], in which features are first extracted from the data (in this case, high-γ) by a convolutional network, and then autoencoded with a transformer encoder.

C. Self-Supervised Training

Our model is pretrained under a self-supervised learning strategy based on *masked autoencoding* [28]: random, contiguous segments of all input features are masked, and the encoder is tasked with reconstructing them. This approach enables the model to learn robust, high-level representations from unlabeled data. (We consider alternatives in the **Discussion**.) Rather than directly predicting the masked input features, we follow recent work in self-supervised training by imposing a *contrastive* loss [12], [13], [14]. That is, for every masked feature $\boldsymbol{q}_t$ and corresponding prediction $\boldsymbol{c}_t$, we minimize a loss of the form

$$L_{\mathrm{c}} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{c}_t, \boldsymbol{q}_t)}}{\sum_{m}^{M+1} e^{\mathrm{sim}(c_t, \boldsymbol{q}_m)}},$$

where $\boldsymbol{q}_m$ for all $m \neq t$ are "distractor" features sampled randomly from other parts of the feature sequence; and $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity. This forces the network to make its predictions different from other features, as well as similar to the target feature, and approximately maximizes the mutual information between $\boldsymbol{c}_t$ and $\boldsymbol{q}_t$ [12].

Following [14], the network is tasked with predicting a vector-quantized (VQ) version of the features rather than the features themselves. VQ reduces noise. But it also coerces

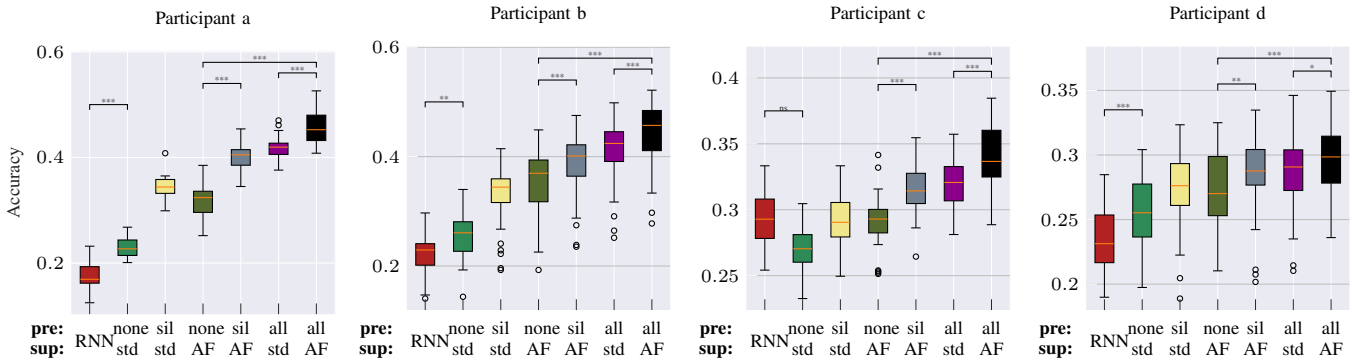

Fig. 2: Classification accuracies for all participants and all models. The box and whisker plots represent the distribution of accuracies across $n = 30$ networks trained separately for each model. The RNN [5] is the baseline model. All other models use the architecture in Fig. 1 and differ only in what training they received. Self-supervised pretraining (pre), none, only on ECoG from silent intervals (sil), or on both silent and speech intervals (all). Supervised training (sup): standard (std) or with the additional of data augmentation and feature fusion (AF). Asterisks show indicate significance under a one-sided Wilcoxon signed-rank test, after Holm-Bonferroni correction for multiple comparisons ($^{***}p < 0.0005$,$^{**}p < 0.005$,$^{*}p < 0.05$, ns. not significant).

the problem into a discrete space, which is more suitable for the downstream task, word classification, which must be made on the basis of discrete tokens—phonemes. We use the Gumbel softmax trick [29], [30] to obviate the derivative through this discretization. To encourage use of the entire VQ codebook, we also penalize its negative entropy (L_d), with the total self-supervised loss equal to $L = L_\mathrm{c} + \lambda L_\mathrm{d}$, with the hyperparameter $\lambda = 0.1$ as in [14].

D. Supervised Training

After self-supervised pretraining, we put the network to use for the task of interest, single-word classification. That is, the network must map a sequence of high-γ corresponding to a single word to that word's index in the vocabulary. In order to turn the network just described into a classifier, we make use of the so-called class token [31], [28]. A (learnable) class token was prepended to the high-γ sequence and then passed through the network (see Fig. 1). This allows the network to pass information useful for classification without re-purposing the features learned during pretraining. However, rather than classifying merely from the class token that is output by the final layer of the transformer encoder, we concatenated the tokens from the first and last layers, a form of multi-layer "feature fusion." The fused output is passed through a linear layer and thence the softmax function to compute the natural parameters for the multiway classification problem. Learning is carried out by stochastic gradient descent of the categorical cross entropy.

Performance of artificial neural networks is known to scale favorably with number of data [32], but labeled ECoG data are scarce. Accordingly, we augmented our supervised-learning dataset with randomly shifted and masked ECoG data. More precisely, we generated new training examples through some combination of

- shifting all ECoG channels left or right by up to 5 samples (25 ms);
- randomly masking (zeroing) 15 consecutive samples (75 ms of all channels);
- randomly masking a block of 15 channels for all time.

These augmentations were applied randomly to each batch of data, each with a probability of 50%.

E. Training Details

Each model was pretrained on high-γ with the self-supervised loss for about 275 epochs on a single Nvidia RTX A6000 GPU. For participants **b**, **c**, and **d**, supervised training was terminated with early stopping, with validation data consisting of a single block. For participant **a**, on the other hand, not enough supervised blocks existed for a three-way dataset split. Since participants **a** and **b** are similar, we used the early stopping epochs found for participant **b** for participant **a** as well. Both losses were optimized with stochastic gradient descent with AdaM optimization with a learning rate of 0.0001. The feature encoder consisted of 3 transformer encoder layers, each with a model dimension of 400 and 2 attention heads. ReLU activation functions were used in the feedforward network of the transformer, and the feedforward dimension was maintained at 400. A dropout rate of 0.5 was used throughout. For self-supervised training, the masking probability was set to 0.2 and the mask length set to 10 samples. The codebook had 2 groups each with 100 entries.

F. Baseline Model

We also considered a model analogous to that used in the original study of this data set [5]. That study focused on decoding sequences of text (sentences), and accordingly used an encoder-decoder framework. Since we are interested here only in word classification, we simply adopted the encoder from that study, a recurrent neural network with an initial convolutional layer, and classified the output with an attention-based classifier. This is very similar to the word classifier used in [8], which achieved high classification accuracies in a clinical trial for a person with anarthria.

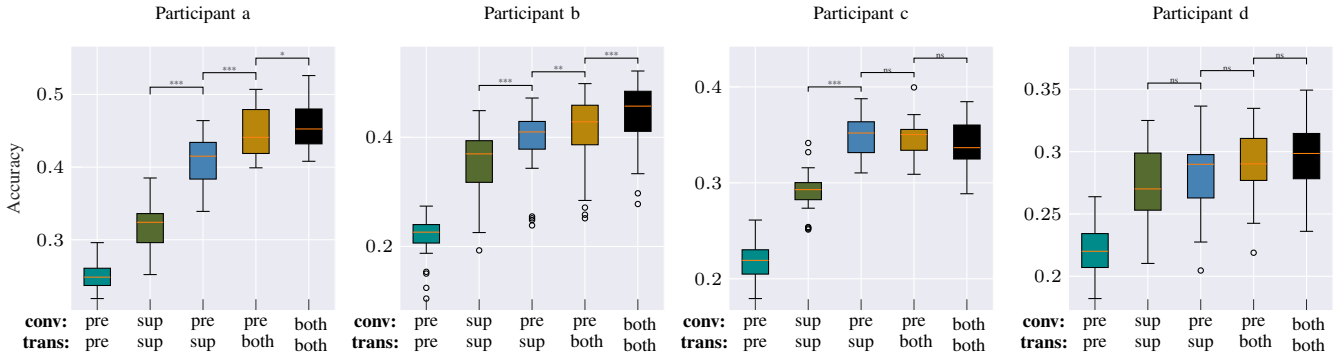

Fig. 3: Dissecting the effects of pretraining. Clasification accuracies for five different versions of the network, in which the convolutional and transformer components were trained with some combination of pre-training only (pre), supervised training only (sup), or both. Box plots show variation across 30 instances of each model version. Statistical comparisons made between versions are indicated with brackets (one-sided Wilcoxon signed rank test; $***p < 0.0005$, $**p < 0.005$, $*p < 0.05$). Chance is 0.04 (for **a** and **b**) or 0.16 (for **c** and **d**).

participant:		**a**	**b**	**c**	**d**
# speech-relevant channels:		204	240	122	159
pretraining	# blocks	16	8	16	19
	# seconds	4587	2293	4456	3866
	# words	∼6000	∼3200	∼5700	∼7000
sup. training	# blocks	3	10	28	45
	# seconds	420	1620	1380	1624
	# words	∼900	∼3550	∼4800	∼6650

TABLE I: Participant data details.

G. Statistical Testing

Data can vary widely across blocks, which were collected over the course of about a week for each participant. Therefore, for each participant and for each model just described, we report cross-validated performance on the supervised learning task. More precisely, we re-trained each model 30 times with randomly held-out validation and test blocks. However, by using the *same* random assignment of validation and test blocks across all models, we are able to apply pairwise statistical tests. To avoid the assumption of normality, we employ one-sided Wilcoxon signed rank tests throughout (except where otherwise noted).

III. Results

Our primary aim is to determine whether self-supervised pretraining on unlabeled ECoG data can improve performance on a subsequent ECoG word-classification task. The wide variation in number of training data available per participant (see Table I) complicates interpretation, so in our first experiment we limit each participant to just two recording blocks (see Section II-A) for supervised training, which is the maximum available from the participant with the fewest blocks (after reserving one block for testing).

1) Self-supervised pretraining improves performance: Fig. 2 shows the classification accuracy of seven models for each of the four participants, to wit, the baseline RNN (Section II-E), and the proposed architecture (Fig. 1) under six different training schemes. We note several comparisons, in order of increasing interest:

- Word classification always exceeds chance, even though training made use of only ∼400 labels in total. (For simplicity, we do not balance the datasets, so chance performance is ∼0.04 for **a** & **b** with a decoder that always emits "a" and ∼0.16, for **c** & **d** with a decoder that always emits "the.")
- Even without pretraining or data augmentation (none/std), the proposed architecture typically outperforms the RNN.
- Data augmentation and feature fusion during supervised training (AF) improve performance. Even for high-performing pretrained models (all/std), these additions always yield significant accuracy gains (all/AF).
- Self-supervised pretraining (sil or all) improves performance. In particular, the best model trained only on labeled data (none/AF) is always significantly outperformed by the best pretrained model (all/AF).
- Furthermore, this holds even if the latter is pretrained only on ECoG from periods of silence (sil/AF), which demonstrates the potential of the method to generalize to data gathered outside of controlled experimental blocks.

Effect sizes vary by participant. For all participants, the accuracy of the simplest purely-supervised model (none/std) is about 25%. For participants **a** and **b**, the improvement provided by the best model (all/AF; black) is about 20 percentage points, i.e. up to 45% accuracy; whereas for participants **c** and **d** it is only about 5 percentage points. The latter two participants read from a different dataset, but the more likely explanation is that they have fewer speech-relevant electrodes (Table I); we return to this in Section IV.

2) Dissecting the effects of pretraining: It is useful to know which components of the network (see Fig. 1) are most improved by pretraining. To that end, we compare the performance of models trained under several variants of supervised training. We find the following.

- If supervised learning is restricted to simply the linear readout and the class token (pre/pre), accuracy reaches only about 25%—on a par with supervised models trained without data augmentation (recall Fig. 2). This weak performance is unsurprising, since the transformer

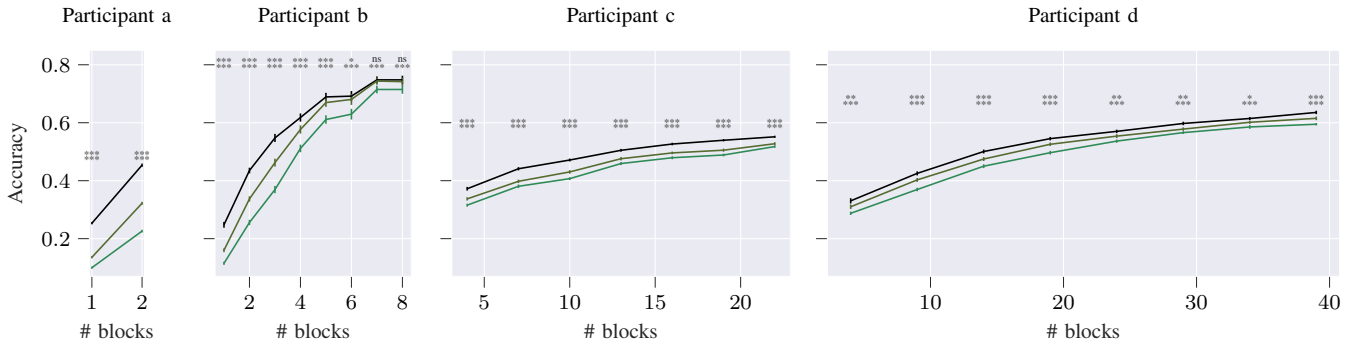

Fig. 4: Classification accuracies as a function of the number of *supervised* training blocks. For all participants and amounts of labeled data, pretraining (all/AF; black) significantly improves accuracy over models with (none/AF) or without (none/std; green) data augmentation and feature fusion (first and second rows of stars, respectively), except for participant **b**'s last two data sizes in all/AF vs. none/AF ($***p < 0.0005$,$**p < 0.005$,$*p < 0.05$, one-sided Wilcoxon signed-rank test, Holm-Bonferroni corrected for multiple comparisons). Error bars denote standard error of the mean over 30 independent networks trained from scratch.

cannot even learn what do to with the class token.

- For the convolutional layer, even simply *replacing* supervised training with self-supervised pretraining is generally beneficial (pre/sup vs. sup/sup). This can be seen by randomly re-initializing the transformer after pretraining, and keeping the convolutional layer frozen during supervised learning of the rest of the network (pre/sup)—although the effects is not statistically significant in participant **d**.
- Although it needs fine-tuning to learn what to do with the class token, the transformer does also appear to benefit from pretraining (pre/both vs. pre/sup). This can be inferred from the effect of resetting the transformer after pretraining (pre/sup), which damages performance (vs. pre/both, i.e. in both cases keeping the convolutional layer frozen). However, this effect is not visible in participants **c** and **d**, for whom pretraining overall has less effect.
- For participants **a** and **b**, allowing the convolutional layer to be fine-tuned (both/both vs. pre/both) yields further improvement.

3) The effect of more supervised training data: We have heretofore focused on the effect of pretraining when labeled data is scare, as this is where we anticipate it to be most useful. However, for some participants, we have additional blocks of labeled data that cover the same test-set vocabulary (see Section II-A). Here we ask whether pretraining continues to be beneficial as the number of supervised training data is scaled up.

Fig. 4 shows the accuracies achieved as a function of the number of supervised-training blocks (up to the maximum available), for the models trained with the three major schemes of this study: purely supervised (green); supervised with the addition of data augmentation and feature fusion (olive green); and trained with a combination of self-supervised pretraining and supervised fine tuning, including data augmentation & feature fusion (black). For all participants, the pretrained model is superior to both its randomly initialized (supervised-training-only) counterparts at all data sizes (except for the last two data sizes in participant **b** for black vs olive green). The effect size as a *percentage improvement* diminishes with number of supervised training blocks, as expected. But perhaps surprisingly, the effect size in *percentage points* is roughly constant for participants **a**, **b**, and **c**. This suggests that the effect of pretraining is a fixed increment of improvement, until ceiling effects kick in (as for participant **b**).

IV. Discussion

We have shown that self-supervised learning of unlabeled ECoG data improves subsequent performance on a supervised learning task, word classification from ECoG sequences. The results held for all four participants (Fig. 2).

For two participants (**c** and **d**), effect sizes were modest, on the order of 5%. For the other two (**a** and **b**), the effect sizes were much bigger, on the order of 20%. Participants **c** and **d** read from a different set of sentences, but the critical factor is much more likely to be the fact that far few channels happened to have been implanted in their speech-relevant areas (Table I and Fig. 5). (Recall that electrode placements were made for treatment of unrelated disorders in these patients, all of whom could speak normally.) This accords with the most recent advances in supervised speech decoding made by increasing electrode count [11].

A critical question now is how pretraining scales with more unlabeled data. In a standard recording setup, whether sub-chronic as in the treatment of epilepsy or in the context of a clinical trial, the number of unlabeled data available will typically be about 25–50 times the number of labeled data: Training sessions with prompted (and therefore labelable) speech are typically less than an hour per day, whereas recordings are (or could be) made for the entire 24 hour period of each day. Whether all 24 hours are equally useful is an important open question. Crucially, we found pretraining to be nearly as effective when limiting ECoG data to intervals during which the participant was not speaking (Fig. 2). On the other hand, the data occurred in the midst of a speaking task; it is possible that data collected during other activities would be less useful.

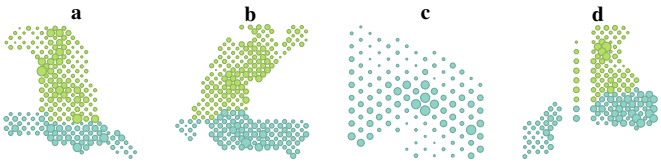

Fig. 5: Each channel's contribution to word classification. Contributions (computed according to the method in [5]) are proportional to marker area. Green markers denote vSMC; turquoise, STG.

In any case, the features generated by our self-supervised training on ECoG data are useful starting points for word classification. They are not, however, optimal for that task: if only the linear readout is updated during supervised training, classification is poor (Fig. 3). This contrasts with automatic speech recognition [14]. It is possible that this is likewise remediable with more unlabeled training data. But another possibility (which does not exclude the first) is that we have not yet found the optimal self-supervised pretext task for speech decoding. Here we have focused exclusively on masked autoencoding, but prediction and other self-supervised tasks (including more bespoke choices for the input mask) are important directions for future research.

References

[1] X. Pei, D. L. Barbour, and E. C. Leuthardt, "Decoding Vowels and Consonants in Spoken and Imagined Words Using Electrocorticographic Signals in Humans," *Journal of Neural Engineering*, vol. 8, no. 4, pp. 1–11, 2011.

[2] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, X. M. W. Slutzky, and M. W. Slutzky, "Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri," *The Journal of Neuroscience*, vol. 4653, no. 46, pp. 1206–18, 2018.

[3] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Communications*, vol. 10, no. 1, 2019.

[4] P. Sun, G. K. Anumanchipalli, and E. F. Chang, "Brain2Char: A deep architecture for decoding text from brain recordings," *Journal of Neural Engineering*, vol. 17, no. 6, Dec. 2020.

[5] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature Neuroscience*, vol. 23, no. 4, pp. 575–582, 2020.

[6] M. Angrick, C. Herff, E. M. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of Neural Engineering*, 2019.

[7] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[8] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, M. E. Dougherty, P. M. Liu, G. M. Abrams, A. Tu-Chan, K. Ganguly, and E. F. Chang, "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217–227, 2021.

[9] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, I. Zhuravleva, A. Tu-chan, K. Ganguly, G. K. Anumanchipalli, and E. F. Chang, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. August, pp. 1037–1046, 2023.

[10] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann, K. V. Shenoy, and J. M. Henderson, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. August, pp. 1031–1036, 2023.

[11] N. S. Card, M. Wairagkar, C. Iacobacci, X. Hou, T. Singer-Clark, F. R. Willett, E. M. Kunz, C. Fan, M. V. Nia, D. R. Deo, A. Srinivasan, E. Y. Choi, M. F. Glasser, L. R. Hochberg, J. M. Henderson, K. Shahlaie, S. D. Stavisky, and D. M. Brandman, "An Accurate and Rapidly Calibrating Speech Neuroprosthesis," *New England Journal of Medicine*, vol. 391, no. 7, pp. 609–618, Aug. 2024.

[12] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2018.

[13] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Interspeech*, 2019, pp. 3465–3469.

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, pp. 1–12.

[15] S. Schneider, J. H. Lee, and M. W. Mathis, "Learnable latent embeddings for joint behavioural and neural analysis," *Nature*, vol. 617, no. 7960, pp. 360–368, 2023.

[16] D. Zhang, Z. Yuan, Y. Yang, J. Chen, J. Wang, and Y. Li, "Brant: Foundation model for intracranial neural signal," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=DDkl9vaJyE

[17] C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "Brainbert: Self-supervised representation learning for intracranial recordings," 2023. [Online]. Available: https://arxiv.org/abs/2302.14367

[18] K. Yi, Y. Wang, K. Ren, and D. Li, "Learning topology-agnostic eeg representations with geometry-aware modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 875–53 891, 2023.

[19] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Machine Learning for Health*. PMLR, 2020, pp. 238–253.

[20] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for eeg-based sleep staging," *CoRR*, vol. abs/2109.07839, 2021. [Online]. Available: https://arxiv.org/abs/2109.07839

[21] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, p. 653659, 2021.

[22] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," *CoRR*, vol. abs/2007.04871, 2020. [Online]. Available: https://arxiv.org/abs/2007.04871

[23] S. Tang, J. A. Dunnmon, K. Saab, X. Zhang, Q. Huang, F. Dubost, D. L. Rubin, and C. Lee-Messer, "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," 2022. [Online]. Available: https://arxiv.org/abs/2104.08336

[24] D. Jayalath, G. Landau, B. Shillingford, M. Woolrich, and O. P. Jones, "The brain's bitter lesson: Scaling speech decoding with self-supervised learning," 2025. [Online]. Available: https://arxiv.org/abs/2406.04328

[25] W. Jiang, L. Zhao, and B. liang Lu, "Large brain model for learning generic representations with tremendous EEG data in BCI," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=QzTpTRVtrP

[26] Y. Chen, K. Ren, K. Song, Y. Wang, Y. Wang, D. Li, and L. Qiu, "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," *arXiv preprint arXiv:2401.10278*, 2024.

[27] A. Wrench, "MOCHA-TIMIT," 2019, online database. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[29] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with Gumbel-Softmax," in *International Conference on Learning Representations*, 2017, pp. 1–12.

[30] C. J. Maddison, A. Mnih, and Y. W. Teh, "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables," in *International Conference on Learning Representations*, 2017, pp. 1–20.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[32] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep Learning Scaling is Predictable, Empirically," Tech. Rep., 2017.