MEGConformer: Conformer-Based MEG Decoder for Robust Speech and Phoneme Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present Conformer-based decoders for the LibriBrain 2025 PNPL competition, targeting two foundational MEG tasks: Speech Detection and Phoneme Classification. Our approach adapts a compact Conformer to raw 306-channel MEG signals, with a lightweight convolutional projection layer and task-specific heads. For Speech Detection, a MEG-oriented SpecAugment provided a first exploration of MEG-specific augmentation. For Phoneme Classification, we used inverse-square-root class weighting and a dynamic grouping loader to handle 100-sample averaged examples. In addition, a simple instance-level normalization proved critical to mitigate distribution shifts on the holdout split. Using the official Standard track splits and F1-macro for model selection, our best systems achieved 88.9% (Speech) and 65.8% (Phoneme) on the leaderboard, surpassing the competition baselines and ranking within the top-10 in both tasks. For further implementation details, the technical documentation, source code, and checkpoints are available at `link-hidden-for-review`.

1 Introduction

In the emerging fields of deep learning and neuroscience, decoding speech from non-invasive brain signals remains one of the central problems in neural signal processing and brain-computer interfaces (BCIs) [27, 11, 25]. The recently released LibriBrain dataset [26] and its associated PNPL 2025 Competition [18] provide a unique opportunity to address this challenge by enabling large-scale within-subject modeling of magnetoencephalography (MEG) recordings. LibriBrain contains over 50 hours of continuous MEG data from a single participant listening to the complete Sherlock Holmes corpus, accompanied by precise voice activity and phoneme-level alignments. The competition defines two foundational decoding tasks: Speech Detection, which distinguishes between speech and silence, and Phoneme Classification, which predicts the phoneme being perceived from averaged MEG signals. Both tasks include a Standard and an Extended track, where the Standard track restricts participants to using only the official training data. We participated exclusively in the Standard track.

Our approach explores an initial adaptation of state-of-the-art Automatic Speech Recognition (ASR) architectures to non-invasive neural signals. Specifically, we leverage the Conformer model [12], a convolution-augmented Transformer [33] that combines the global context modeling of self-attention [20] with the local feature extraction strength of convolutional networks [19]. Conformer-based architectures have been and continue to be the state-of-the-art in the ASR field [4, 15, 29, 14, 30]. We hypothesize that their ability to capture both temporal and spectral dependencies may serve as a good starting point for MEG-based speech tasks, which also exhibit structured spatiotemporal patterns over multiple timescales.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this work, we evaluate a Conformer model adapted to the LibriBrain competition tasks. The models process raw MEG windows to predict either speech activity or phoneme identity, sharing the same backbone while differing in input and output processing. To ensure fair comparison and reproducibility, all experiments follow the official data splits and evaluation metrics provided by the organizers [18].

Our main contributions are fourfold: (i) a unified Conformer architecture jointly optimized for both LibriBrain tasks; (ii) a robust yet straightforward input normalization method that substantially improves holdout generalization; (iii) an effective MEG-specific augmentation; (iv) and a dynamic grouping strategy to classify averaged samples in the Phoneme Task.

This study demonstrates that adapting modern ASR architectures to MEG decoding yields competitive and robust results in both voice activity detection and phoneme recognition, emphasizing the growing convergence between speech processing and neural decoding research.

2 Related work

Decoding speech from non-invasive neural signals has been explored with both EEG and MEG using supervised learning, from linear baselines to deep architectures. EEG studies reconstruct speech features or spectrograms with CNN/Transformer-style models and subject-independent training, showing steady gains with more data [1, 35]. On MEG, prior work demonstrated phrase and word-level decoding and trial-efficiency using wavelet denoising, CNNs, and transfer learning [6, 7, 5], as well as compact end-to-end networks adapted to sensor time series [31], and transformer-based neural encoding that links linguistic context to MEG responses [16]. Other studies have also started to explore phoneme-level decoding in perception and production modalities, comparing traditional and novel model architectures [8, 9]. More recently, LibriBrain established an unprecedented within-subject scale and standardized tasks for speech detection and phoneme classification [26, 18]. Beyond purely supervised setups, contrastive and self-supervised objectives, along with foundation-model guidance, have advanced non-invasive decoding across perception domains, improving retrieval and generation [10, 3, 2]. Recent work on non-invasive brain-to-text systems combines discriminative decoders with language-model rescoring and cross-dataset scaling to produce significant score improvements [13].

3 Methods

3.1 Model

We use a single Conformer encoder adapted to MEG for both tracks (Speech Detection and Phoneme Classification). The network takes raw sensor time series (306 channels) preprocessed with the official pipeline (bad-channel interpolation, head-position correction, signal-space separation, notch, and band-pass filtering), where windowed signal segments are downsampled to 250 Hz. For Speech, we use 2.5 s windows (625 samples), and for the Phoneme Task, 0.5 s windows (125 samples). A lightweight 1D convolutional projection adapts the 306 MEG channels to the Conformer input size (144), followed by a dropout of p=0.1, a stack of Conformer blocks, and a linear classifier.

Each Conformer block follows the standard macaron layout: feed-forward, multi-head self-attention, depthwise temporal convolution, and a second feed-forward, with residual connections. We keep the design compact with a hidden size of 144. For Speech, based on Conformer Small, we use 16 layers, 4 heads, a hidden layer dimension of FFNs of 576, and the layer's depthwise convolution layer with a kernel size of 31. For Phonemes, we created a custom-sized Conformer with 7 layers, 12 heads, FFN dimension of 2048, and a kernel size of 127, to better adapt the model to the smaller dataset.

3.2 Speech detection task specifics

The Speech model is a binary classifier with a single-logit head, binary cross-entropy loss, and loss-level label smoothing of 0.1 [24]. We found that simple, speech-style augmentation was sufficient for this task. Therefore, we develop and apply a light MEG-specific variant of SpecAugment [28], which we introduce as **MEGAugment**. It includes two operations: (i) *time masking*, which zeroes two random temporal spans (max width T=180 samples at 250 Hz) per window; and (ii) *bandstop*

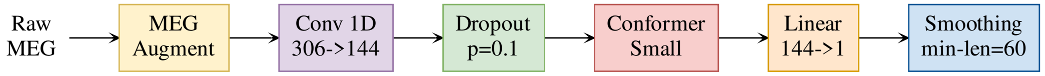

Figure 1: Conformer-based model architecture used for Speech Task.

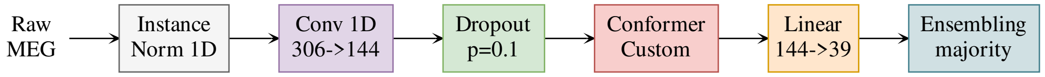

Figure 2: Conformer-based model architecture used for Phoneme Task.

masking, which randomly suppresses narrow frequency bands (theta, alpha, beta, gamma, high-gamma [22]) using fourth-order infinite impulse response (IIR) notches with probability p=0.4. Moreover, input windows slide with a 60-sample stride during training to increase diversity via overlapping.

The final layer outputs a probability of speech, which is smoothed during inference by removing speech sequences shorter than 60 samples ($240\,\mathrm{ms}$). The block diagram in Figure 1 summarizes the pipeline.

3.3 Phoneme classification task specifics

The Phoneme model uses a linear classification head that outputs 39 logits and is trained with cross-entropy. At inference, we select the class via `argmax`. To mitigate imbalance in the Phoneme Task, we use class weights w_c with the Inverse of Square root of Number of Samples (ISNS) rule following Equation 1, where n_c is the count of class c in the training set and C is the total number of classes.

$$w_c \propto \frac{1}{\sqrt{n_c}}, \qquad \frac{1}{C}\sum_c w_c = 1, \tag{1}$$

The architecture (detailed in Figure 2) includes an instance-level input normalization layer directly after the raw input. This layer computes per-sample, per-channel normalization across time with no running statistics. We find this choice essential for stable holdout performance: unlike validation and test, the holdout split exhibits different statistical characteristics, indicating distributional shift (see Appendix B for details). Instance normalization essentially removes amplitude and scale drift between windows and sessions. Consequently, it closes most of the holdout gap, even if it is sometimes slightly sub-optimal on the other splits. Because the competition holdout recordings are pre-averaged over 100 samples to improve SNR, we adapt our training setup accordingly. Particularly, we use a 100-sample dynamic grouping loader that reshuffles groups each epoch, allowing the model to see many independent averages of the same class while preserving temporal locality.

Lastly, we ensemble the five best model seeds to smooth predictions and select the final phonemes using majority voting.

4 Experimental setup

We follow the official LibriBrain data partitions for all experiments, using the provided train, validation, test, and holdout splits. In our experience, re-partitioning or rebalancing the data did not yield consistent improvements, so we retain the original configuration throughout.

Training uses the AdamW optimizer [21] with a learning rate of $1{\cdot}10^{-4}$, weight decay of $5{\cdot}10^{-2}$, batch size of 256, and early stopping based on validation F1-macro with a patience of 10. We monitor F1-macro on the validation set to select the final checkpoint. These hyperparameters are shared across both tasks.

Evaluation follows the competition protocol, using F1-macro as the primary metric on test and holdout splits.

All model seeds are trained on individual NVIDIA A100 and H100 GPUs, and each model configuration is trained with ten random seeds for better comparison.

Table 1: F1-macro scores in the holdout for both tasks (Standard track). These are preliminary results in the public leaderboard. Our model and scores are marked in bold.

Model	Speech Detection ↑	Phoneme Classification ↑
Naive Baseline	*45.30%*	*0.47%*
Baseline	*68.04%*	*60.39%*
Top-10	88.89%	60.96%
MEGConformer	**88.90%**	**65.83%**
Top-1	91.66%	73.82%

5 Results

Our Conformer-based models achieved competitive results in both tasks, as shown in Table 1. In Speech Detection, our best model reached an F1-macro of 88.9%, surpassing the official baseline (68.0%). In Phoneme Classification, our model reached 65.8% F1-macro and again exceeded the competition baseline (60.4%). These results confirm that Conformer architectures, initially developed for ASR, transfer productively to MEG-based decoding and perform competitively with traditional convolutional and transformer models.

Beyond leaderboard scores, several design choices contributed to these results. For Speech Detection, extending the input window from 0.5 to $2.5\,\mathrm{s}$ gave the largest gain (+10.8%), followed by adopting Conformer over the SEANet baseline (+9.0%) [32], and reducing the training stride to 60 (+2.8%); MEGAugment had a negligible net effect (+0.01%). In the Phoneme task, dynamic grouping (vs. fixed groups) improved performance by +13.3%, inverse-square-root class weighting outperformed no loss weighting by +7.6%, the custom Conformer added +0.5%, and ensembling of the best seeds contributed an additional +15.4% on the holdout set. Ultimately, instance-level normalization was essential for holdout generalization (over +200%), outperforming batch (+17.8%) and layer normalization (+88.2%). See Appendix A for further details and statistical analysis.

6 Discussion and conclusion

The Conformer architecture proves suitable to decode the spatiotemporal nature of MEG signals, combining convolutional blocks that capture local temporal structure with self-attention for longer-range dependencies. Using very similar models across both tasks, we achieved top-10 leaderboard performance, with minor input preprocessing differences between Speech and Phoneme decoding. Instance-level normalization is indispensable to mitigate distributional shifts between training and holdout splits in the Phoneme Task, substantially improving generalization. Furthermore, a simple SpecAugment-based [28] augmentation was effective for speech detection, but showed limited benefit for the more complex, imbalanced phoneme classification task. On the other hand, as LibriBrain is single-subject, cross-subject generalization remains an open challenge for future work.

Looking ahead, adapting speech-model architectures for MEG decoding could open the way toward end-to-end speech reconstruction using sequential objectives such as CTC or seq2seq heads, and even enable speech synthesis. Another promising direction is the use of linguistically grounded feature-based representations [17], which may help tackle data imbalance and improve interpretability by decomposing phoneme classification into binary articulatory features (see Appendix C for the challenges found). Finally, our scaling experiments (Appendix D) indicate that while speech decoding performance starts to saturate, phoneme classification continues to improve with additional data, suggesting further gains could be achieved as larger MEG datasets become available.

References

[1] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Scientific Reports*, 13(1):812, Jan 2023.

[2] H. Banville, Y. Benchetrit, S. d'Ascoli, J. Rapin, and J.-R. King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025.

[3] Y. Benchetrit, H. Banville, and J.-R. King. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] M. Burchi and V. Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE, 2021.

[5] D. Dash, P. Ferrari, D. Heitzman, and J. Wang. Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5531–5535, 2019.

[6] D. Dash, P. Ferrari, S. Malik, A. Montillo, J. A. Maldjian, and J. Wang. Determining the optimal number of MEG trials: A machine learning and speech decoding perspective. In S. Wang, V. Yamamoto, J. Su, Y. Yang, E. Jones, L. Iasemidis, and T. Mitchell, editors, *Brain Informatics*, pages 163–172, Cham, 2018. Springer International Publishing.

[7] D. Dash, P. Ferrari, S. Malik, and J. Wang. Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 489–493, 2018.

[8] X. de Zuazo, E. Navas, I. Saratxaga, M. Bourguignon, and N. Molinaro. Phone pair classification during speech production using MEG recordings. In *IberSPEECH 2024*, pages 76–80, 2024.

[9] X. de Zuazo, E. Navas, I. Saratxaga, M. Bourguignon, and N. Molinaro. Decoding phone pairs from MEG signals across speech modalities. *arXiv preprint arXiv:2505.15355*, 2025.

[10] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

[11] A. Dehgan, H. Abdelhedi, V. Hadid, I. Rish, and K. Jerbi. Artificial neural networks for magnetoencephalography: a review of an emerging field. *Journal of Neural Engineering*, 22(3):031001, jun 2025.

[12] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, pages 5036–5040, 2020.

[13] D. Jayalath, G. Landau, and O. P. Jones. Unlocking non-invasive brain-to-text. *arXiv preprint arXiv:2505.13446*, 2025.

[14] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe. E-branchformer: Branchformer with enhanced merging for speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 84–91. IEEE, 2023.

[15] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373, 2022.

[16] A. Klimovich-Gray, G. Di Liberto, L. Amoruso, A. Barrena, E. Agirre, and N. Molinaro. Increased top-down semantic processing in natural speech linked to better reading in dyslexia. *NeuroImage*, 273:120072, 2023.

[17] T. Kwon, S. Cho, G. Elvers, F. Mantegna, and O. Parker Jones. Language-inspired approaches to phoneme classification, 2025. Blog post.

[18] G. Landau, M. Özdogan, G. Elvers, F. Mantegna, P. Somaiya, D. Jayalath, L. Kurth, T. Kwon, B. Shillingford, G. Farquhar, M. Jiang, K. Jerbi, H. Abdelhedi, Y. Mantilla Ramos, C. Gulcehre, M. Woolrich, N. Voets, and O. Parker Jones. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *NeurIPS Competition Track*, 2025.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[20] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019.

[22] P. K. Mandal, A. Banerjee, M. Tripathi, and A. Sharma. A comprehensive review of magnetoencephalography (MEG) studies for brain functionality in healthy aging and alzheimer's disease (AD). *Frontiers in Computational Neuroscience*, 12:60, Aug 2018.

[23] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[24] R. Müller, S. Kornblith, and G. E. Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] S. A. Murad and N. Rahimi. Unveiling thoughts: A review of advancements in EEG brain signal decoding into text. *IEEE Transactions on Cognitive and Developmental Systems*, 17(1):61–76, 2025.

[26] M. Özdogan, G. Landau, G. Elvers, D. Jayalath, P. Somaiya, F. Mantegna, M. Woolrich, and O. P. Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *arXiv preprint arXiv:2506.02098*, 2025.

[27] J. T. Panachakel and A. G. Ramakrishnan. Decoding covert speech from EEG-a comprehensive review. *Frontiers in Neuroscience*, Volume 15 - 2021, 2021.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617, 2019.

[29] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR, 2022.

[30] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[31] M. Sarma, C. Bond, S. Nara, and H. Raza. MEGNet: A MEG-based deep learning model for cognitive and motor imagery classification. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2571–2578, 2023.

[32] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek. Seanet: A multi-modal speech enhancement network. In *Interspeech 2020*, pages 1126–1130, 2020.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[34] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[35] X. Xu, B. Wang, Y. Yan, H. Zhu, Z. Zhang, X. Wu, and J. Chen. ConvConcatNet: a deep convolutional neural network to reconstruct mel spectrogram from the EEG. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 113–114. IEEE, 2024.

Table 2: Speech task variant test scores over ten seeds before smoothing. Bold marks the best scores.

Variant	$\mathrm{F1}_{\text{macro}}$	$\mathrm{F1}_{\text{pos}}$	$\mathrm{Accuracy}_{\text{bal}}$	$\mathrm{AUROC}_{\text{macro}}$	Jaccard
Our model	**87.59 ± 0.70**	**94.72 ± 0.23**	85.72 ± 1.19	**96.15 ± 0.32**	**78.64 ± 1.02**
tmax=0.5	78.13 ± 1.32	90.62 ± 0.20	75.58 ± 1.74	90.59 ± 0.31	65.87 ± 1.47
SEANet	79.67 ± 0.50	90.97 ± 0.16	77.44 ± 0.59	90.58 ± 0.34	67.70 ± 0.61
No stride	85.18 ± 0.55	93.06 ± 0.14	83.59 ± 1.12	94.35 ± 0.20	75.01 ± 0.75
No augment	87.58 ± 0.55	94.20 ± 0.14	**85.82 ± 1.17**	96.08 ± 0.21	78.53 ± 0.79

Table 3: Phoneme task test scores over ten seeds before ensembling. Bold marks the best scores.

Variant	$\mathrm{F1}_{\text{macro}}$	$\mathrm{Accuracy}_{\text{bal}}$	$\mathrm{AUROC}_{\text{macro}}$
Our model	**44.29 ± 2.84**	47.75 ± 2.70	**96.67 ± 0.31**
Fixed groups	38.39 ± 2.83	40.43 ± 2.77	93.53 ± 1.66
No weights	40.92 ± 3.18	43.88 ± 3.53	95.87 ± 1.02
Conformer Small	44.09 ± 4.56	**48.49 ± 4.19**	96.33 ± 0.60

Appendix A Ablation details and statistical significance

For the ablation study, we start from our best-performing model and remove each improvement individually while keeping all other settings fixed. Each ablated variant is therefore identical to the full model except for one modification, allowing us to isolate the contribution of that component.

We report per-variant means and standard deviation over ten seeds and assess paired differences with the Wilcoxon signed-rank test [34] over the F1-macro scores in the test set. This non-parametric test does not assume normality and evaluates whether the median of paired score differences differs from zero, indicating a genuine effect rather than random variation. We will use a p-value threshold of 0.01 or lower to determine statistical significance.

Speech task ablation

As shown in Table 2, window extension from 0.5 to $2.5\,\mathrm{s}$ yields the largest gain over our default (+10.8% relative), and Conformer outperforms SEANet by +9.0%. Both effects are statistically significant: *tmax=0.5* vs. ours ($W = 0.0$, $p = 0.002$), and SEANet vs. Conformer ($W = 0.0$, $p = 0.002$). Reducing the training stride to 60 also significantly helps (+2.8%; $W = 0.0$, $p = 0.002$). Removing augmentation has a small effect (+0.01%) and is not significant ($W = 13.0$, $p = 0.160$). Although MEGAugment improvements are not significant in the final configuration, it did provide a significant gain of +1.8% in earlier model versions, motivating its inclusion ($W = 1.0$, $p = 0.004$).

SEANet adjustments. To match tmax $= 2.5\,\mathrm{s}$ (625 samples at $250\,\mathrm{Hz}$), we modify only the temporal downsampler, the third `conv1d` (k=50, s=25 in the original), by setting its stride to s=160. This keeps the intermediate length at L_{out}=4, so the subsequent k=4 stem still collapses to 1 (original: $L_{\text{out}} = \lfloor (125 - 50)/25 \rfloor + 1 = 4$; adjusted: $\lfloor (625 - 50)/160 \rfloor + 1 = 4$). We also align the classifier and loss with the binary task by replacing the final $1{\times}1$ `conv1d` from 512 to 2 with a single-logit 512 to 1 head and using BCE-with-logits plus label smoothing (0.1). Beyond architecture, we match the training and evaluation protocol used for our Conformer: sliding-window training with a stride of 60, MEGAugment, AdamW, early stopping on validation F1-macro, best-checkpoint selection, and validation and test scoring with a stride of 1.

Phoneme task ablation

As shown in Table 3, dynamic grouping (vs. fixed groups) yields the largest improvement (+13.3% relative; $W = 1.0$, $p = 0.004$). Inverse–square-root class weighting (our default) outperforms non-weighted loss by +7.6% (not significant with $W = 10.0$, $p = 0.084$). The custom Conformer provides a small lift over Conformer Small (+0.5%; $W = 16.0$, $p = 0.844$), consistent with similar capacity but with a better fit to the data characteristics. As noted in the main text, holdout ensembling adds +15.4% but is specific to the competition evaluation.

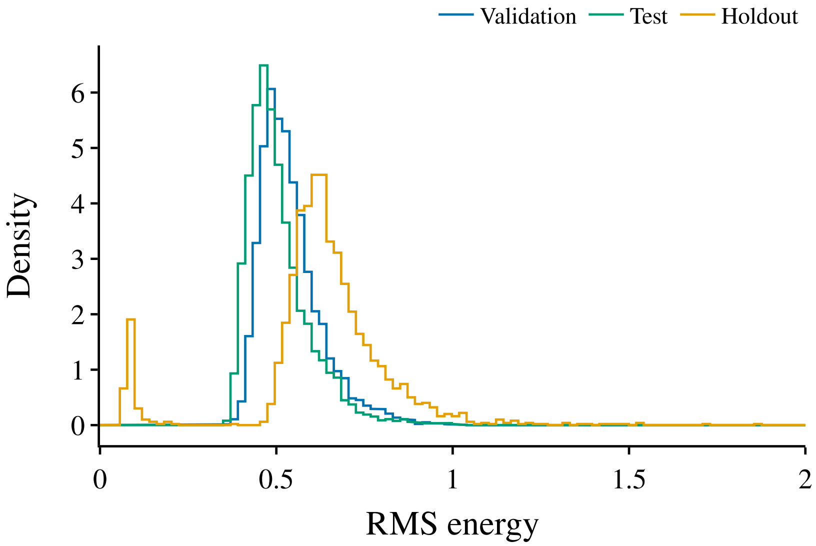

Figure 3: RMS energy distribution per split in Phoneme Task.

Appendix B RMS energy analysis across splits in Phoneme Task

To quantify the scale mismatch between partitions, we computed the per-window root-mean-square (RMS) energy across channels and time as in Equation 2, where x is the MEG window (C=306 channels, T=125 samples). For validation and test, windows were read with the standard competition preprocessing and standardization enabled. The holdout was read as stored (already standardized on disk). We then formed outline-only histograms using the same bin edges and density normalization across splits.

$$\mathrm{RMS}(x) = \sqrt{mean\big(x^2\big)} \quad \text{for } x \in \mathbb{R}^{C\times T}, \tag{2}$$

Figure 3 shows that the holdout distribution is bimodal (two peaks near 0.08 and 0.62), with markedly larger dispersion (mean 0.64, standard deviation 0.22), whereas validation and test are unimodal and tighter (validation 0.54 ± 0.09, test 0.51 ± 0.09). This shift motivated the instance-level normalization used in the Phoneme model to remove amplitude and scale differences window-by-window, thereby improving robustness on the holdout.

Appendix C Phonetic feature decoding challenges

To explore linguistically grounded representations for MEG decoding, we trained separate binary classifiers targeting core phonetic features related to the manner of articulation and voicing without sample averaging. Each model was adapted from the Phoneme Classification configuration described in Section 3.3, using the same Conformer backbone and training hyperparameters, but modified for binary output. Notably, we replaced the 39-class head with a single-logit output, used a binary cross-entropy loss with label smoothing and automatic positive-class weighting following Equation 1 to address class imbalance (based on the Speech Detection model in Section 3.2).

This approach follows the ideas proposed by the PNPL competition authors in their blog on language-inspired phoneme classification [17], where phonetic features are suggested to handle dataset imbalance and improve infrequent phoneme decoding.

Based on [23], we defined six phonetic features based on manner of articulation: Plosive, Fricative, Affricate, Nasal, Liquid, Glide, and a Voicing feature that distinguishes voiced from voiceless phonemes, including vowels and consonants. Table 4 reports preliminary decoding results using F1-macro, positive-class F1, balanced accuracy, and AUROC (mean ± std across ten seeds).

Table 4: Decoding performance for binary phonetic features in the non-averaged test set (in %).

Feature	Positives (%)	F1_{macro}	F1_{pos}	$\text{Accuracy}_{\text{bal}}$	$\text{AUROC}_{\text{macro}}$
Speech	*76.76*	*88.10 ± 0.52*	*94.42 ± 0.17*	*86.38 ± 0.97*	*96.38 ± 0.18*
Phoneme	–	*5.37 ± 0.21*	*4.00 ± 1.08*	*5.86 ± 0.22*	*64.35 ± 1.05*
Voicing	77.38	57.77 ± 0.56	83.23 ± 1.30	57.38 ± 0.69	63.88 ± 0.45
Plosive	18.82	55.58 ± 0.53	25.54 ± 1.86	55.31 ± 0.57	62.09 ± 1.22
Fricative	18.55	53.78 ± 0.78	23.85 ± 2.07	53.70 ± 0.71	57.71 ± 1.34
Affricate	0.97	49.69 ± 0.03	0.00 ± 0.00	49.98 ± 0.06	60.29 ± 1.92
Nasal	11.26	51.74 ± 0.54	12.98 ± 1.06	51.67 ± 0.50	55.97 ± 1.47
Liquid	7.51	50.80 ± 1.02	8.16 ± 3.03	51.02 ± 0.57	55.81 ± 1.20
Glide	3.62	52.16 ± 0.81	6.70 ± 1.65	51.83 ± 0.70	60.93 ± 2.25

Following Appendix A, statistical significance is evaluated with a one-sample Wilcoxon signed-rank test against chance (0.5 for F1-macro).

Among these features, Voicing showed the strongest decoding signal, approaching 58% F1-macro (significant with $W = 55.0$, $p = 0.001$). Despite its similarity in label balance to the Speech Task, scores are not on par, implying that decoding some fine-grained speech features may not be straightforward. Manner-based features such as Plosive and Fricative were moderately decodable (both significant with $W = 55.0$, $p = 0.001$), while Affricate, a composite and low-frequency category (0.97%), remained at chance (not significant with $W = 0.0$, $p = 1.0$). Attempts to address the affricate issue through ensembling of Plosive and Fricative models, transfer learning from these features, and larger effective batch sizes (via gradient accumulation) did not yield convergence improvements. This supports the view that decomposing phonemes into binary linguistic features alone is insufficient to overcome the data scarcity, as some linguistic features may also be rare enough not to be easily decodable, like affricates here.

Overall, these experiments suggest that feature-based representations are promising for improving interpretability and possibly scaling, but further research is needed to address low-frequency classes and explore multitask formulations that jointly learn shared articulatory subspaces.

Appendix D Data-size ablation

To examine the effect of dataset scale on decoding performance, we conducted data-size ablations for both tasks using progressively larger subsets of the LibriBrain training data. Figures 4 and 5 show the resulting F1-macro scores as a function of the total hours of MEG data used for training, with shaded bands indicating one standard deviation across random seeds.

For Speech Detection, performance rapidly improved with additional data and then began to saturate, suggesting that the task may already approach its ceiling with a single subject's data. In contrast, Phoneme Classification, even though with some diminishing returns signs, shows an upward trend with no apparent plateau, implying that larger-scale within-subject recordings could still yield substantial gains. These results align with prior observations [26] that decoding performance continues to scale with the amount of high-quality MEG data available for training.

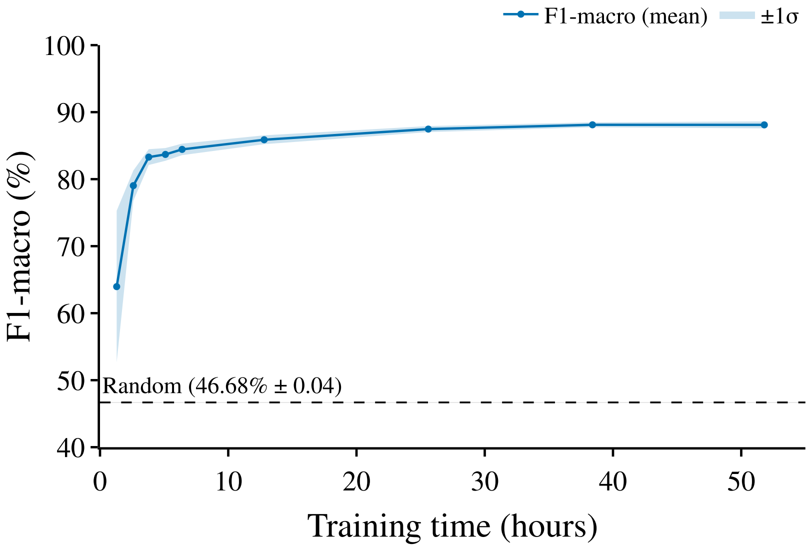

Figure 4: Data size ablation for Speech Task.

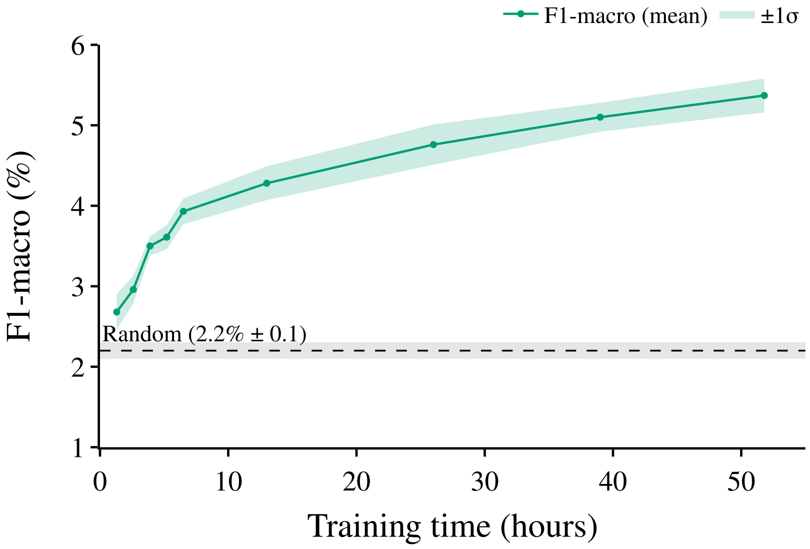

Figure 5: Data size ablation for Phoneme Task.