MEGState: Phoneme Decoding from Magnetoencephalography Signals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decoding linguistically meaningful representations from non-invasive neural recordings remains a central challenge in neural speech decoding. Among available neuroimaging modalities, magnetoencephalography (MEG) provides a safe and repeatable means of mapping speech-related cortical dynamics, yet its low signal-to-noise ratio and high temporal dimensionality continue to hinder robust decoding. In this work, we introduce MEGState, a novel architecture for phoneme decoding from MEG signals that captures fine-grained cortical responses evoked by auditory stimuli. Extensive experiments on the LibriBrain dataset demonstrate that MEGState consistently surpasses baseline model across multiple evaluation metrics. These findings highlight the potential of MEG-based phoneme decoding as a scalable pathway toward non-invasive brain–computer interfaces for speech.

1 Introduction

Decoding speech representations from brain activity holds considerable promise for restoring communication in individuals with paralysis or severe speech impairments [Moses et al., 2021]. Recent advances in invasive brain–computer interfaces [Nagashima et al., 2025, Suzuki et al., 2025] have enabled continuous speech reconstruction from intracranial recordings, achieving word error rates below 5% for vocabularies exceeding 100,000 words [Card et al., 2024]. However, their reliance on neurosurgical implantation limits scalability and clinical feasibility. In contrast, non-invasive approaches such as magnetoencephalography (MEG) offer a safe and repeatable alternative for probing speech-related neural activity. Nevertheless, decoding linguistically meaningful information from MEG remains challenging due to its low signal-to-noise ratio, high temporal resolution, and sparse neural representations [Yang et al., 2024b,a].

In this work, we introduce MEGState, a novel architecture designed for phoneme classification from MEG signals. The model integrates two complementary components: (i) a Multi-Resolution Convolution module that captures fine-grained temporal dynamics of phoneme-evoked cortical responses, and (ii) a Sensor-wise SSM that captures long-range temporal dependencies across individual sensors. This design effectively mitigates the challenges of MEG's sparsity and high sampling rate, allowing the model to capture both local and global neural dynamics. Comprehensive experiments on the LibriBrain dataset demonstrate that MEGState consistently surpasses baseline method across multiple evaluation metrics.

2 Method

2.1 Preliminaries

State space models. Recent progress in state space models (SSMs) [Gu et al., 2022, Dao and Gu, 2024] shows that they can outperform prevailing architectures, most notably Transformers, on a wide range of sequence modeling tasks. Grounded in control theory [Kalman, 1960], SSMs provide a

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

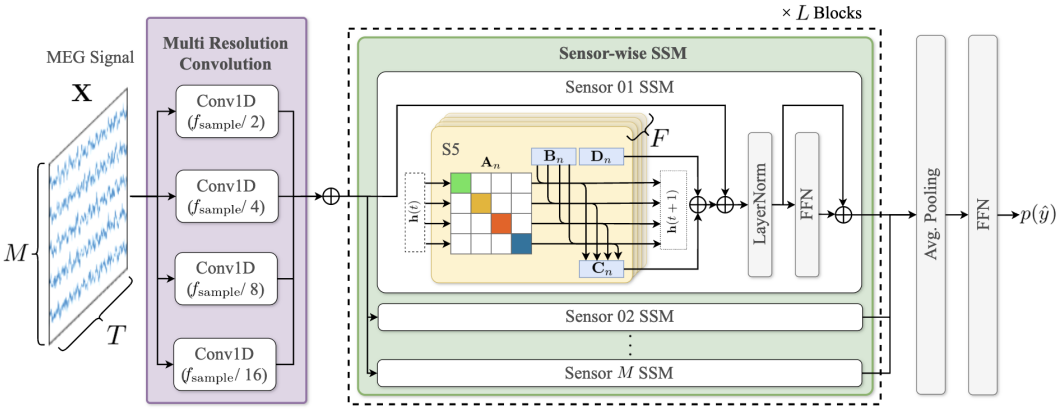

Figure 1: Model architecture of the proposed MEGState. Given a MEG signal, Mutli Resolution Convolution module extracts local temporal dependencies while Sensor-wise SSM models global spatial and temporal dependencies, respectively.

mapping from inputs $\mathbf{x}(t) \in \mathbb{R}^P$ to outputs $\mathbf{y}(t) \in \mathbb{R}^P$ through latent states $\mathbf{h}(t) \in \mathbb{R}^Q$ governed by

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \qquad \mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{Q\times Q}$ is the state matrix and $\mathbf{B} \in \mathbb{R}^{Q\times P}$, $\mathbf{C} \in \mathbb{R}^{P\times Q}$, $\mathbf{D} \in \mathbb{R}^{P\times P}$ are input/output projections. Among SSM variants, S5 [Smith et al., 2023] has proved especially effective for modeling continuous signals. S5 sets $\mathbf{A}$ to the HiPPO-N matrix Gu et al. [2022] to capture long-range temporal dependencies. Since HiPPO-N is real symmetric, it admits the diagonalization $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, which yields the decoupled form

$$\frac{d\tilde{\mathbf{h}}(t)}{dt} = \mathbf{\Lambda}\tilde{\mathbf{h}}(t) + \tilde{\mathbf{B}}\mathbf{x}(t), \qquad \mathbf{y}(t) = \tilde{\mathbf{C}}\tilde{\mathbf{h}}(t) + \mathbf{D}\mathbf{x}(t), \tag{2}$$

where $\tilde{\mathbf{h}}(t) = \mathbf{V}^{-1}\mathbf{h}(t)$, $\tilde{\mathbf{B}} = \mathbf{V}^{-1}\mathbf{B}$, and $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{V}$. Moreover, introducing a timescale vector $\mathbf{\Delta} \in \mathbb{R}+^Q$ and applying zero-order hold discretization Zhang and Chong [2007] gives the recurrence

$$\tilde{\mathbf{h}}_t = \bar{\mathbf{\Lambda}}\tilde{\mathbf{h}}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t, \qquad \mathbf{y}_t = \bar{\mathbf{C}}\tilde{\mathbf{h}}_t + \bar{\mathbf{D}}\mathbf{x}_t, \tag{3}$$

where $\bar{\mathbf{\Lambda}} = \exp(\mathbf{\Lambda}\mathbf{\Delta})$, $\bar{\mathbf{B}} = \mathbf{\Lambda}^{-1}\left(\bar{\mathbf{\Lambda}} - \mathbf{I}\right)\tilde{\mathbf{B}}$, $\bar{\mathbf{C}} = \tilde{\mathbf{C}}$, $\bar{\mathbf{D}} = \mathbf{D}$. In practice, we take $\mathbf{D}$ to be diagonal and learn $\mathrm{diag}(\mathbf{\Lambda})$, $\tilde{\mathbf{B}}$, $\tilde{\mathbf{C}}$, $\mathrm{diag}(\mathbf{D})$, and $\mathbf{\Delta}$.

2.2 Model Architecture

The overall architecture of the proposed model is depicted in Figure 1. It comprises two primary modules: Multi-Resolution Convolution and Sensor-wise SSM. Given a MEG sample $\mathbf{X} \in \mathbb{R}^{M\times T}$, where M and T represent the number of sensors and the sequence length, respectively, the network proceeds as follows.

First, the Multi-Resolution Convolution module extract fine-grained local temporal structures reflecting distinct cortical responses evoked by different phonemes. It consists of four parallel one-dimensional convolutional layers with kernel sizes of $f_{\text{sample}}/2$, $f_{\text{sample}}/4$, $f_{\text{sample}}/8$, and $f_{\text{sample}}/16$, where f_{sample} denotes the native sampling rate. The outputs from these convolutional layers are concatenated to yield $\mathbf{H} \in \mathbb{R}^{F\times M\times T}$, where F represents the frequency feature dimension.

Next, the Sensor-wise SSM module models global temporal dependencies in a sensor-specific manner. To handle the sparsity and high temporal resolution of MEG signals, we extend S5 [Smith et al., 2023], a variant of SSM well suited for modeling high-dimensional multivariate time series. The module comprises L hierarchically organized blocks, and the output $\tilde{\mathbf{H}}^{(l)} \in \mathbb{R}^{F\times M\times T}$ from the l-th block ($l = 1, \ldots, L$) is obtained as follows, where $\tilde{\mathbf{H}}^{(0)} = \mathbf{H}$:

$$\mathbf{H}'^{(l-1)} = \text{LayerNorm}\left(\text{SSM}\left(\tilde{\mathbf{H}}^{(l-1)}\right) + \tilde{\mathbf{H}}^{(l-1)}\right), \tag{4}$$

$$\tilde{\mathbf{H}}^{(l)} = \text{LayerNorm}\left(\text{FFN}\left(\mathbf{H}'^{(l-1)}\right) + \mathbf{H}'^{(l-1)}\right). \tag{5}$$

Algorithm 1 Sampling training data

Input: Dataset $(\mathcal{X}, \mathcal{Y}) = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^{N}$
Output: Training sample $(\tilde{\boldsymbol{X}}, \tilde{y})$
Randomly sample two labels $y_1 \sim \mathrm{Uniform}(\mathcal{Y})$, $y_2 \sim \mathrm{Uniform}(\mathcal{Y})$
$\mathcal{I}_1 \leftarrow \{\, i \in \{1, \ldots, N\} \mid y_i = y_1 \,\}$ ▷ All indices with label y_1
$\mathcal{I}_2 \leftarrow \{\, i \in \{1, \ldots, N\} \mid y_i = y_2 \,\}$ ▷ All indices with label y_2
Randomly sample subsets $\mathcal{I}_1' \subset \mathcal{I}_1$, $\mathcal{I}_2' \subset \mathcal{I}_2$ s.t. $|\mathcal{I}_1'| = |\mathcal{I}_2'| = N'$
$\bar{\boldsymbol{X}}_1 \leftarrow \mathrm{AVERAGE}_{i \in \mathcal{I}_1'}(\boldsymbol{X}_i)$
$\bar{\boldsymbol{X}}_2 \leftarrow \mathrm{AVERAGE}_{i \in \mathcal{I}_2'}(\boldsymbol{X}_i)$
$\tilde{\boldsymbol{X}} \leftarrow \alpha \bar{\boldsymbol{X}}_1 + (1-\alpha)\bar{\boldsymbol{X}}_2$ ▷ Mixup augmentation
$\tilde{y} \leftarrow \alpha y_1 + (1-\alpha) y_2$

Here, SSM$(\cdot)$, LayerNorm$(\cdot)$, and FFN$(\cdot)$ denote the S5 layer, layer normalization, and feed-forward network, respectively. Subsequently, the output $\tilde{\mathbf{H}}^{(L)}$ from the Sensor-wise SSM is aggregated to produce the predicted phoneme probability $p(\hat{y})$ corresponding to $\mathbf{X}$ as follows:

$$p(\hat{y}) = \mathrm{FFN}\left(\mathrm{AvgPool}\left(\tilde{\mathbf{H}}^{(L)}\right)\right), \tag{6}$$

where AvgPool$(\cdot)$ denotes average pooling over the temporal dimension, following the standard SSM architectures [Gu et al., 2022]. The model is trained using the cross-entropy loss function.

3 Experiments

3.1 Dataset and Pre-procesing

In this experiment, we used the publicly available LibriBrain dataset Özdogan et al. [2025], which contains MEG recordings of a single participant listening to audiobook narrations of Sherlock Holmes. The recordings were acquired with a MEGIN Triux™ Neo system using 306 sensors at 1 kHz, yielding 52.32 hours of data. Preprocessing involved Maxwell filtering [Taulu and Simola, 2006] and Signal Source Separation to remove sensor noise and external magnetic interference. A notch filter (50 Hz), Butterworth band-pass filter (0.1–125 Hz), and downsampling to 250 Hz were further applied. The dataset provides 39 ARPAbet-based phoneme labels temporally aligned with the auditory stimuli, where each label corresponds to a 0.5 s MEG segment. It was divided into training, validation, and test sets comprising 1,488,392, 11,289, and 12,051 samples, respectively. Models were trained, tuned, and evaluated on these splits, and further tested on the LibriBrain leaderboard set provided in the 2025 PNPL Competition [Landau et al., 2025].

3.2 Implementation Details

During training, we leverage smoothing and data augmentation in the sampling of the training samples to improve the signal-to-noise ratio (SNR) of MEG signals and mitigate phoneme-label imbalance. Concretely, at each sampling step within a training epoch, we independently draw two phoneme labels y_1 and y_2 uniformly from the MEG dataset $(\mathcal{X}, \mathcal{Y}) = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^{N}$. For each label y_k $(k \in \{1, 2\})$, we construct the index set of all samples with that label, $\mathcal{I}_k = \{i \in \{1, \ldots, N\} \mid y_i = y_k\}$, and uniformly sample a subset $\mathcal{I}_k' \subset \mathcal{I}_k$ of size N'. We then form a class-conditional prototype by averaging the corresponding MEG inputs, $\bar{\boldsymbol{X}}_k = \frac{1}{N'} \Sigma_{i \in \mathcal{I}_k} \boldsymbol{X}_i$, which acts as a denoised representative and thus enhances the SNR. Furthermore, by introducing a mixing coefficient $\alpha \in [0, 1]$, we address label imbalance (see Sec. 4) via a mixup-style convex combination [Zhang et al., 2018] of both the prototypes and their labels:

$$\tilde{\boldsymbol{X}} = \alpha \bar{\boldsymbol{X}}_1 + (1-\alpha)\bar{\boldsymbol{X}}_2, \quad \tilde{y} = \alpha \bar{y}_1 + (1-\alpha)\bar{y}_2. \tag{7}$$

The overall sampling procedure is summarized in Algorithm 1.

We trained the model with the AdamW optimizer ($\beta_1 = 0.9,\ \beta_2 = 0.999$) and a learning rate of 1.0×10^{-4}. The batch size was 32, and training proceeded for 50 epochs. For the Sensor-wise SSM module, we set the block size to $L = 2$. For sampling training data, we set the prototype size of $N' = 100$ and a mixing coefficient of $\alpha = 0.5$.

Table 1: Performance comparison on the LibriBrain [Özdogan et al., 2025] test set and leaderboard set. **Bold** values denote the best performances, while † indicates statistical significance compared to the baseline method ($p < 0.05$). Multi-Resol Conv. indicates Mutli-Rresolution Convolution module.

Model	**Test Set**			**Leaderboard Set**
	Acc. [%] ↑	**Kappa [%] ↑**	**Macro-F1 [%] ↑**	**Macro-F1 [%] ↑**
Baseline [Özdogan et al., 2025]	38.80±2.40	45.71±0.74	34.82±1.92	—
Ours (w/o Mulit-Resol. Conv.)	40.25±3.15	37.90±2.83†	34.77±1.83	—
Ours (w/o Sensor-wise SSM)	40.37±3.20	49.60±6.25	37.18±4.44	—
Ours (MEGState)	**45.53±1.88**†	**54.19±2.42**†	**41.11±2.20**†	**55.74** (68.41)

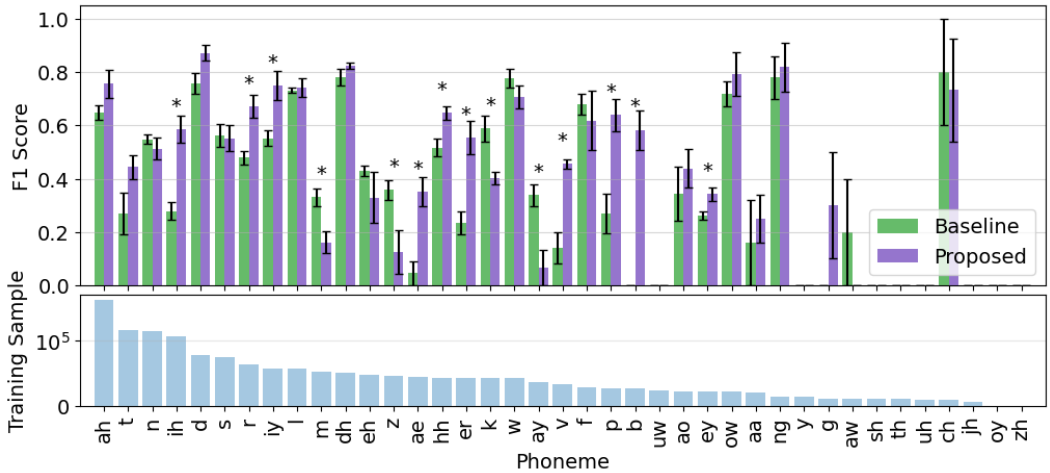

Figure 2: Quantitative comparison across phonemes. The upper panel shows the macro-F1 for each phoneme, while the lower panel indicates the number of training samples per phoneme. Error bars represent the standard error of the mean, and $*$ denotes statistical significance ($p < 0.05$).

4 Results

Table 1 presents the quantitative comparison between the proposed and baseline methods. Values in the table represent the mean and standard deviation obtained from over five distinct random seeds.

On the test set, the proposed method achieved balanced accuracy, Cohen's Kappa, and macro-F1 of 45.53%, 54.19%, and 41.11%, respectively, surpassing the baseline method by 6.73, 8.48, and 6.29 points. Here, all improvements were statistically significant ($p < 0.05$). Moreover, ablation studies revealed that removing either the Multi-Resolution Convolution module or the Sensor-wise SSM consistently degrades performance across all metrics on the test set, indicating that both modules contribute substantially to phoneme classification from MEG signals.

On the leaderboard set, the proposed method achieved a macro-F1 of 55.74%. Notably, the final leaderboard submission employed an ensemble strategy that selected the most probable label from predictions of five independently trained models, resulting in a higher macro-F1 of 68.41%.

Figure 2 further presents the phoneme-wise quantitative comparison. Each bar shows the macro-F1 score for individual phonemes, ordered by the number of training samples. As shown, the proposed method outperformed the baseline in terms of macro-F1 on 19 phonemes and achieved statistically significant improvements on 10 phonemes ($p < 0.05$).

5 Conclusion

In this work, we introduced MEGState, a novel architecture for phoneme decoding from MEG signals. By integrating a Multi-Resolution Convolution module to capture fine-grained local temporal dynamics and a Sensor-wise SSM to model long-range dependencies across individual sensors, our approach effectively mitigates the challenges of MEG signal sparsity and high temporal resolution. Comprehensive experiments on the LibriBrain dataset demonstrated that MEGState consistently outperformed the baseline across multiple evaluation metrics.

References

Nicholas S Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R Willett, Erin M Kunz, Chaofei Fan, Maryam Vahdati Nia, et al. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *ICML*, volume 235, pages 10041–10071, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. In *ICLR*, 2022.

R. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Gilad Landau, Miran Özdogan, Gereon Elvers, Francesco Mantegna, Pratik Somaiya, Dulhan Jayalath, Luisa Kurth, Teyun Kwon, et al. The 2025 PNPL Competition: Speech Detection and Phoneme Classification in the LibriBrain Dataset. In *NeurIPS Competition Track*, 2025.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Shunya Nagashima, Kanta Kaneda, Yuiga Wada, Tsumugi Iida, Misa Taguchi, Masayuki Hirata, and Komei Sugiura. ECoG Dual Context Network for Brain Machine Interfaces. *IEEE Access*, 13: 99252–99265, 2025.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 Hours of Within-Subject MEG to Improve Speech Decoding Methods at Scale. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Jimmy Smith, Andrew Warrington, and Scott Linderman. Simplified State Space Layers for Sequence Modeling. In *ICLR*, 2023.

Shuntaro Suzuki, Shunya Nagashima, Masayuki Hirata, and Komei Sugiura. Cortical-SSM: A Deep State Space Model for EEG and ECoG Motor Imagery Decoding. *arXiv preprint arXiv:2510.15371*, 2025.

Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine & Biology*, 51(7):1759, 2006.

Yiqian Yang, Yiqun Duan, Qiang Zhang, Hyejeong Jo, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Neuspeech: Decode neural signal as speech. *arXiv preprint arXiv:2403.01748*, 2024a.

Yiqian Yang, Hyejeong Jo, Yiqun Duan, Qiang Zhang, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Mad: Multi-alignment meg-to-text decoding. *arXiv preprint arXiv:2406.01512*, 2024b.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *ICLR*, 2018.

Zheng Zhang and Kil To Chong. Comparison between first-order hold with zero-order hold in discretization of input-delay nonlinear systems. In *ICCAS*, pages 2892–2896, 2007.