MEBM-Speech: Multi-scale Enhanced BrainMagic for Robust MEG Speech Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose **MEBM-Speech**, a multi-scale enhanced neural decoder for speech activity detection from non-invasive magnetoencephalography (MEG) signals. Built upon the BrainMagic backbone, MEBM-Speech integrates three complementary temporal modeling mechanisms: a multi-scale convolutional module for short-term pattern extraction, a bidirectional LSTM (BiLSTM) for long-range context modeling, and a depthwise separable convolutional layer for efficient cross-scale feature fusion. A lightweight temporal jittering strategy and average pooling further improve onset robustness and boundary stability. The model performs continuous probabilistic decoding of MEG signals, enabling fine-grained detection of speech versus silence states—an ability crucial for both cognitive neuroscience and clinical applications such as monitoring residual speech perception and detecting speech intention in patients. Comprehensive evaluations on the **LibriBrain Competition 2025 Track 1** benchmark demonstrate strong performance, achieving an average $\mathrm{F1}_{\mathrm{macro}}$ of 89.3% on the validation set and comparable results on the official test leaderboard. These findings highlight the effectiveness of multi-scale temporal representation learning for robust MEG-based speech decoding.

1 Introduction

Understanding how the human brain encodes and processes acoustic and linguistic information underlying speech perception is a long-standing goal in cognitive neuroscience and brain–computer interface (BCI) research. Recent advances in magnetoencephalography (MEG) decoding open new possibilities for non-invasively mapping brain dynamics to acoustic and linguistic features (1; 2). Accurately decoding speech-related neural activity not only deepens our understanding of language processing but also holds important clinical implications. Real-time detection of speech and silence states from MEG can enable monitoring of residual speech perception in patients with neurological impairments and facilitate detection of speech intention in locked-in or aphasic individuals, contributing to communication-restoring BCIs (3).

The LibriBrain Competition 2025 (4; 5) provides a large-scale benchmark for decoding acoustic and linguistic features from non-invasive brain recordings. In particular, Track 1 focuses on detecting speech versus silence segments from MEG data while participants listen to natural audiobook stimuli. This task requires the model to learn the complex temporal dynamics underlying speech and silence auditory processing and generalize effectively across speakers and contexts.

To address this task, we propose a novel model termed MEBM-Speech, specifically designed for Track 1 of the LibriBrain Competition.

Our approach introduces three key innovations:

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

1. **Decoding Strategy:** Unlike the official baseline method that formulates the task as frame-wise binary classification, our model performs end-to-end probabilistic decoding, predicting a continuous probability for each frame within a temporal window, followed by adaptive thresholding to determine speech versus silence regions. A similar continuous decoding perspective has proven effective in recent EEG-to-audio reconstruction work (6), which motivates our design choice for temporally coherent probabilistic modeling.
2. **Model Architecture:** We augment the BrainMagic (2) backbone with a short-term multi-scale convolutional module for capturing fine-grained temporal features, a BiLSTM layer for modeling long-term dependencies, and a depthwise separable convolution layer for efficient feature fusion.
3. **Training Protocol:** We adopt a 100 Hz downsampled MEG signal representation and use only the gradient (`grad`) channels for prediction. This design significantly reduces computational and memory costs while preserving critical spatial-temporal information, leading to faster convergence without degrading performance.

These innovations enable MEBM-Speech to achieve competitive performance on both local and online evaluation sets, demonstrating its effectiveness in non-invasive MEG-based speech detection.

2 Methods

2.1 Decoding Strategy

In the speech detection task, the auditory stimuli alternate continuously between speech and silence states, and the corresponding MEG signals dynamically adjust to these changing acoustic states. Instead of treating each frame as an independent binary classification target, we formulate the task as a continuous probabilistic decoding problem, where the model predicts a time-varying probability sequence $\mathbf{P} \in [0,1]^{1\times T}$ representing the likelihood of speech activity at each time point.

This design better aligns with the intrinsic continuity of brain activity: neural responses to speech do not abruptly switch between binary states but evolve smoothly in time with overlapping temporal integration windows. By learning continuous probabilities, the model can capture gradual onset and offset transitions, improving temporal precision and robustness to boundary ambiguity. During inference, an adaptive thresholding scheme is applied to $\mathbf{P}$ to obtain final binary predictions for speech versus silence segments.

To further enhance temporal robustness, we apply a mild temporal jittering when generating training labels. For each phoneme onset, the starting point is randomly shifted within the range $[\text{onset} - 2, \text{onset} + 2]$ frames (corresponding to ± 20 ms at 100 Hz), ensuring the model remains invariant to small onset misalignments inherent in MEG signals.

During training, the model is optimized using a mean squared error (MSE) loss between the predicted probability sequence and the ground-truth binary speech labels. For model selection, the five checkpoints with the lowest validation losses are retained. Each candidate model is then evaluated on the validation set using 99 classification thresholds ranging from 0.01 to 0.99 (with a step size of 0.01). The model-threshold combination achieving the highest F1_{macro} score is selected as the final configuration for subsequent testing.

2.2 Model Architecture

As illustrated in Figure 1, the proposed MEBM-Speech architecture extends the BrainMagic framework by incorporating enhanced short-term temporal modeling and multi-scale feature integration. Given the MEG input $\mathbf{X} \in \mathbb{R}^{C_{\text{in}}\times T}$, where C_{in} denotes the number of MEG sensor channels, and T represents the total number of temporal samples, the model first passes through a spatial attention module composed of multiple linear and activation layers, which dynamically recalibrates channel responses to produce a spatially refined representation $\mathbf{H}_s \in \mathbb{R}^{D\times T}$, with D being the dimensionality of the projected feature space. The refined representation is then processed in parallel by three temporal branches: (1) 5 BrainMagic (BM) encoders that capture mid-term contextual dependencies, (2) 12 multi-scale convolutional blocks that extract fine-grained local features across multiple receptive fields, and (3) a BiLSTM that models long-range temporal dependencies. The outputs of these three branches are concatenated along the feature dimension to form a unified temporal representation

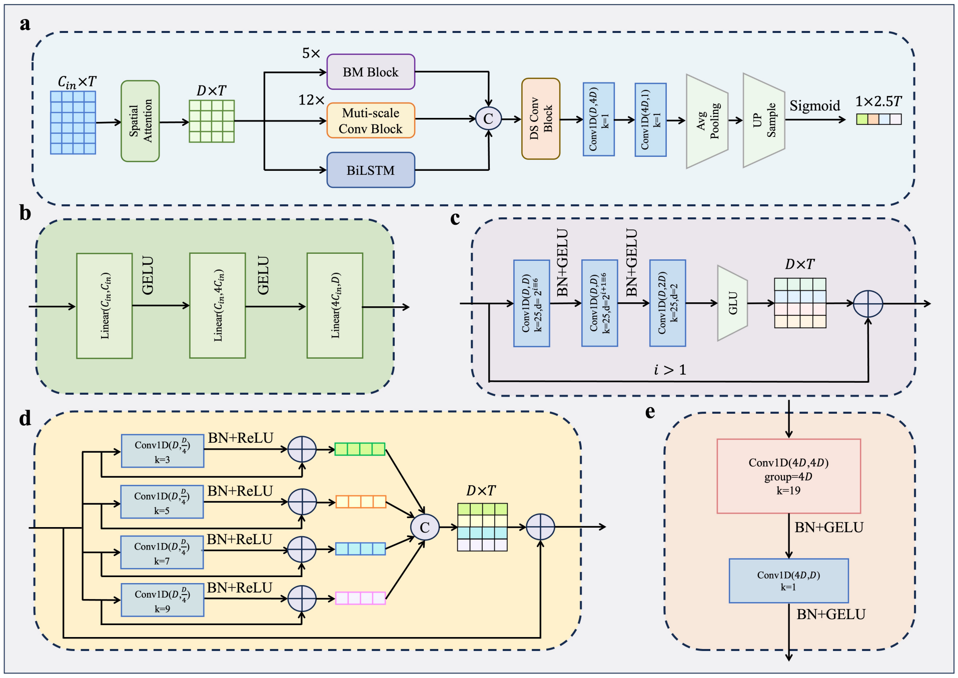

Figure 1: Overall architecture of the proposed MEBM-Speech model. **(a)** The complete processing pipeline. **(b)** The *spatial attention module* enhances sensor-level representations by learning spatial relevance weights across MEG channels. **(c)** The *BM encoder* extracts mid-term contextual features from spatially weighted signals. **(d)** The *short-term multi-scale convolutional module* captures fine-grained temporal dependencies using multiple receptive fields. **(e)** The *depthwise separable convolutional layer* further refines temporal representations with lightweight channel-wise and pointwise filtering.

encompassing multi-scale contextual information. A depthwise separable convolutional block is then applied to efficiently fuse and refine the concatenated features while reducing inter-channel redundancy. Subsequent average pooling provides temporal smoothing and stabilizes boundary estimation. Since the MEG data are downsampled to 100 Hz, the final probabilistic predictions are temporally upsampled via linear interpolation, yielding a continuous per-sample probability of speech activity. This hierarchical fusion framework enables MEBM-Speech to capture transient neural responses and sustained temporal dependencies simultaneously, thereby enhancing the model's ability to detect speech-related neural activity in continuous MEG recordings with high temporal fidelity and robustness.

3 Experiments

3.1 Experimental Setup

The offline validation set was constructed using the official validation and test sessions (*Sherlock1, sessions 11–12*) to approximate the holdout distribution defined by the LibriBrain challenge. For consistency, both the training and validation data followed the same preprocessing pipeline. The continuous MEG signals of each session were first downsampled to 100 Hz and normalized independently along the temporal dimension. Only the *grad* channels were retained for subsequent processing, resulting in the input dimensionality of $C_{\text{in}} = 204$. For each sample, MEG segments were extracted using a 12-second window ($T = 1200$) with a 6-second step size, yielding overlapping windows to augment the effective training data volume. Each 12-second segment was normalized separately to maintain session-level stability.

The proposed MEBM-Speech model was implemented in PyTorch and trained on a single NVIDIA A800 GPU (80 GB). The network comprised approximately 10.3 M trainable parameters and con-

Table 1: Results and ablation analysis on the local validation set under six random seeds (0–5). Metrics include $F1_{macro}$ and Acc_{macro} (mean ± std).

Model Variant	$\mathbf{F1_{macro}}$ **(%)**	$\mathbf{Acc_{macro}}$ **(%)**
Full Model	**89.34±0.24**	89.25±0.37
w/o BM Encoder	88.36±0.13	88.29±0.24
w/o Multi-scale Conv	89.17±0.19	88.98±0.28
w/o BiLSTM	89.21±0.17	**89.27±0.22**
w/o BM Encoder + Multi-scale Conv	87.91±0.13	87.76±0.38
w/o Multi-scale Conv + BiLSTM	89.17±0.21	89.20±0.24
w/o BiLSTM + BM Encoder	85.59±0.20	85.47±0.18

verged within 10 epochs, requiring roughly 20 minutes of training. The intermediate feature dimension was set to $D = 128$ with a dropout rate of 0.01. All convolutional operations employed `padding='same'` to preserve temporal resolution. The average pooling layer employed a window size of 31 and a stride of 15. Training was performed using the AdamW optimizer with a learning rate of 1×10^{-3}. Model selection and hyperparameter tuning were conducted using the offline validation set derived from the *Sherlock1 Session 11–12* data.

3.2 Results and Ablation

We report the performance of the proposed MEBM-Speech model and its ablated variants on the offline validation set. All results are averaged over six random seeds $0, 1, 2, 3, 4, 5$ for reproducibility. Evaluation metrics include $F1_{macro}$ (%) and Acc_{macro} (%). Table 1 summarizes the averaged results across all seeds. The full MEBM-Speech model achieves an average $F1_{macro}$ of 89.34% and Acc_{macro} of 89.25% on the validation set. When submitted to the official online test server, it attains a comparable performance of approximately 89% $F1_{macro}$, demonstrating strong generalization to unseen sessions.

Removing the BM Encoder leads to the largest performance drop, confirming its central role in capturing mid-term spatiotemporal representations and encoding discriminative brain dynamics. Excluding the Multi-scale Convolution branch or the BiLSTM results in moderate yet consistent degradation, suggesting that both modules contribute complementary temporal features—short-term fine-grained cues from convolutional filters and long-range contextual dependencies from recurrent modeling. The fusion of these components within the full architecture yields the most balanced and robust decoding performance across all evaluation metrics.

Overall, these results validate the effectiveness of integrating multi-scale temporal extraction, recurrent modeling, and hierarchical fusion for MEG-based speech activity detection.

4 Conclusion

In this work, we introduce MEBM-Speech, a multi-scale enhanced decoding framework tailored for MEG-based speech activity detection. By combining the BM encoder, multi-scale convolutional branch, and BiLSTM with a depthwise separable fusion layer, the model effectively integrates complementary temporal dynamics across multiple scales. Our probabilistic decoding strategy with temporal jittering and pooling enhances robustness to onset misalignment and label noise. Ablation analyses confirm that each module contributes uniquely to performance, with the BM encoder providing strong mid-term representations and the BiLSTM capturing long-range dependencies. The resulting system achieves approximately 89% $F1_{macro}$ on both local and online benchmarks, demonstrating strong generalization across sessions.

Future work will focus on extending the framework toward real-time speech decoding, exploring cross-subject adaptation for broader generalization, and exploring MEG-based speech detection during speech production, moving closer to practical speech–BCI applications for communication restoration. These directions represent important steps toward practical, non-invasive brain–speech decoding systems for communication restoration.

References

[1] Moses, D.A., Mesgarani, N., Leonard, M.K. & Chang, E.F. (2016) Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *Journal of Neural Engineering*, **13**(5):056004.

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

[3] Dash, D., Ferrari, P., Dutta, S. & Wang, J. (2020) NeuroVAD: Real-time voice activity detection from non-invasive neuromagnetic signals. *Sensors (Basel, Switzerland)*, **20**(8):2248.

[4] Landau, G., Özdogan, M., Elvers, G., Mantegna, F., Somaiya, P., Jayalath, D., Kurth, L., Kwon, T., Shillingford, B., Farquhar, G., Jiang, M., Jerbi, K., Abdelhedi, H., Ramos, Y.M., Gulcehre, C., Woolrich, M., Voets, N. & Jones, O.P. (2025) The 2025 PNPL Competition: Speech Detection and Phoneme Classification in the LibriBrain Dataset. *arXiv*, 2506.10165. doi:10.48550/arXiv.2506.10165.

[5] Özdogan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., Woolrich, M. & Jones, O.P. (2025) LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *arXiv*, 2506.02098. doi:10.48550/arXiv.2506.02098.

[6] Xu, X., Wang, B., Yan, Y., Zhu, H., Zhang, Z., Wu, X. & Chen, J. (2024) ConvConcatNet: A deep convolutional neural network to reconstruct mel spectrogram from the EEG. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 113–114. doi:10.1109/ICASSPW62465.2024.10626859.