SHINE: Sequential Hierarchical Integration Network for EEG and MEG

Anonymous Author(s)
Affiliation
Address
email

Abstract

How natural speech is represented in the brain constitutes a major challenge for cognitive neuroscience, with cortical envelope-following responses playing a central role in speech decoding. This paper presents our approach to the Speech Detection task in the LibriBrain Competition 2025, utilizing over 50 hours of magnetoencephalography (MEG) signals from a single participant listening to LibriVox audiobooks. We introduce the proposed Sequential Hierarchical Integration Network for EEG and MEG (SHINE) to reconstruct the binary speech-silence sequences from MEG signals. In the Extended Track, we further incorporated auxiliary reconstructions of speech envelopes and Mel spectrograms to enhance training. Ensemble methods combining SHINE with baselines (BrainMagic, AWavNet, ConvConcatNet) achieved F1-macro scores of 0.9155 (Standard Track) and 0.9184 (Extended Track) on the leaderboard test set.

1 Introduction

How natural speech is represented in the brain remains a major challenge in cognitive neuroscience. Over the past two decades, neuroscientists have made seminal contributions, progressively elucidating the pivotal role of cortical envelope-following responses—wherein the auditory cortex tracks the amplitude modulations of acoustic signals (1; 2). These responses constitute a foundational neural mechanism for speech decoding, namely, the extraction of perceived speech from neural signals. In recent years, the rapid advancement of machine learning techniques has spurred some researchers aimed at decoding speech features (such as temporal envelopes and Mel spectrograms) from non-invasive neural recordings, such as magnetoencephalography (MEG) (3) and electroencephalography (EEG)(4).

This year, the Parker Jones Neural Processing Lab (PNPL) at the University of Oxford hosted the LibriBrain Competition 2025 (5). The organizers sourced stimuli from LibriVox, presented them to a single participant, and acquired over 50 hours of MEG data (6). Within the competition's **Speech Detection** task, participants were tasked with training a model to **distinguish speech vs. silence** based on brain activity measured by MEG during a listening session. This challenge task comprises two tracks: the Standard Track permits the use of only the LibriBrain training dataset, whereas the Extended Track allows incorporation of any external data.

In this task, our team developed several innovative strategies as follows:

1. **Task Reformulation:** We restructured the conventional binary classification framework into a sequence reconstruction task, focusing on reconstructing a binary sequence directly from MEG signals. This reformulation more effectively captures the temporal dynamics of the MEG signals.

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

2. **SHINE model:** We proposed SHINE (Sequential Hierarchical Integration Network for EEG and MEG), an advanced neural architecture originally developed for reconstructing speech envelopes and Mel spectrograms from EEG signals.

3. **Multi-Feature Reconstruction Strategy (only in Extended Track):** To augment model training, we incorporated the reconstruction of speech envelopes and Mel spectrograms as auxiliary tasks alongside the primary binary sequence reconstruction during the training stage.

2 Methods

2.1 Datasets

The LibriBrain dataset (6) used for the competition contained non-invasive MEG recordings acquired from one healthy participant listening to over 50 hours of audiobooks, all sourced from LibriVox. The MEG recordings were acquired from 306 sensors covering the whole brain. Neural data were minimally filtered (e.g., to remove line noise and drift) and downsampled to 250 Hz. The data were standardly split into train, validation, and test sets. Though this standard split, all data were permissible for model training during the competition. For the competition, the organizer reserved an additional competition **holdout** split, which included disjoint subsets of data to be used to update the leaderboard during the competition and to decide the final ranking of submissions.

2.2 Task Reformulation

In the speech detection task, we needed to train a model to distinguish speech vs. silence based on brain activity measured by MEG during a listening session. The baseline routine provided by the organizers employed 0.8-second epochs of MEG data to predict the label at the central sampling point—designating it as speech (label 1) or silence (label 0). However, given the critical role of contextual information in speech decoding (7), we posited that decoding should leverage longer neural signal segments to better encapsulate temporal dependencies. This approach aligned with established sequence-to-sequence (seq2seq) paradigms in EEG-based speech decoding, such as the reconstruction of speech envelopes (8) or Mel spectrograms (9; 10; 11; 4).

Specifically, we adopted a comparable strategy, utilizing 30-second epochs of MEG signals to reconstruct a 30-second binary sequence, wherein 0 denoted silence and 1 denoted speech. To optimize this process, we employed the negative Pearson correlation coefficient as the loss function, enhancing the model's ability to capture temporal dynamics effectively.

Speech perception involves hierarchical processing in the brain, with MEG signals reflecting neural responses to various levels of speech features (12). We hypothesized that these features could help to finish the binary sequence reconstruction task. To this end, we incorporated auxiliary reconstructions of the speech envelope and Mel spectrograms in the **Extended Track** to facilitate learning of the primary binary sequence. During the training stage, the envelope, a 10-sub-band Mel spectrograms, and the ground-truth binary sequence were concatenated along the sub-band dimension to form a composite 12-sub-band representation. In the validation stage, only the binary sequence was reserved to calculate the Pearson correlation coefficient with the target binary sequence.

2.3 Model

A **S**equential **H**ierarchical **I**ntegration **N**etwork for **E**EG and MEG (SHINE), drawing inspiration from previous works (8; 4), was developed for the binary sequence reconstruction task. In this framework, the input comprised 30-second MEG signals, while the output is a binary sequence (speech: 1 and silence: 0). The architecture of SHINE consisted of six stacked blocks (Figure 1a, with each block encompassing four distinct components (Figure 1b,. Preceding these six blocks, two linear layers were employed to extract initial features, thereby compressing the MEG dimensionality from 306 channels to 64. Subsequently, following the blocks, a convolutional neural network (CNN) and a long short-term memory (LSTM) module were integrated to derive higher-order global sequence features.

The first part of each block was the CNN stack consisting of 5 convolutional layers. The second part was a simple fully connected linear layer. The third part was the output context layer consisting of a zero-padding, a temporal convolutional layer, LeakyReLU activation function, and layer normalization. The last part was a self-attention layer. As shown in Figure 1c, each of the first four layers

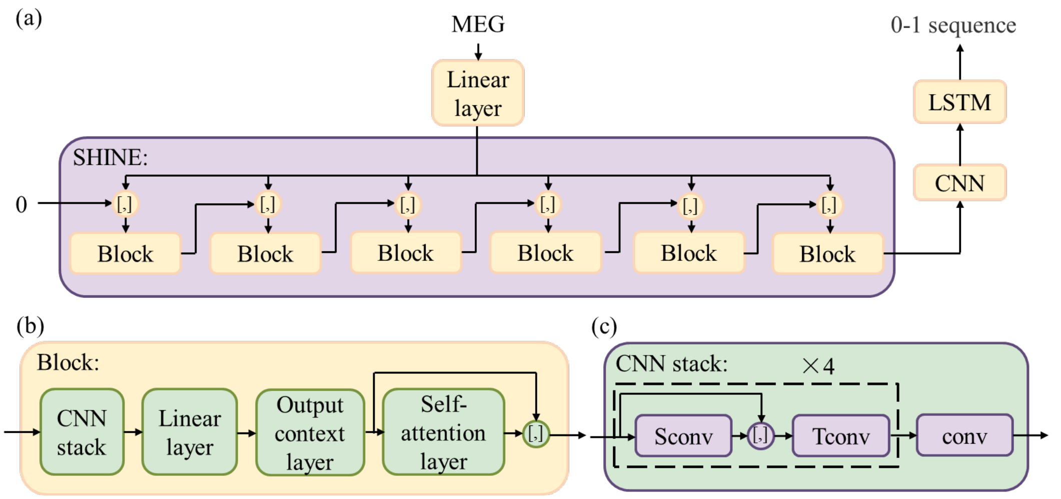

Figure 1: Structure of the proposed SHINE. (a) the overall network architecture of SHINE. (b) four different parts in each block. (c) the network architecture of the CNN stack.

in the CNN stack contains Sconv and Tconv. Sconv was a pointwise convolution followed by LLP (Layer normalization, LeakyReLU activation function, and zero-padding). Tconv was a temporal convolution with groups equal to the input channel number followed by LLP. The last layer in the CNN stack, conv, contained a simple temporal convolutional layer followed by LLP.

The architecture extensively leveraged concatenation operations to facilitate the hierarchical integration of speech-relevant information embedded within MEG signals. As shown in Figures 1a and 1b, the MEG after the initial linear layer, the output of the output context layer, and the output of the self-attention layer were concatenated along the channel dimension and used as the input for the next block. Additionally, in the CNN stack, the output of Sconv was concatenated with the initial input along the channel dimension.

2.4 Training and Validation

We combined the officially split training and validation sets to enable subsequent cross-validation analyses. This integrated corpus comprised over 50 hours of MEG recordings spanning 92 distinct sessions. Considering temporal autocorrelations (TAs) (13) prevalent in MEG signals within individual sessions, which might undermine the model's ability to discern speech-relevant features, we adopted a "leave-session-out" splitting to mitigate overfitting to session-specific TAs features (14; 15). In each training iteration, 8 sessions were randomly selected from the 92 to constitute the validation set, with the remainder allocated to training.

The neural networks were implemented with PyTorch and trained on an NVIDIA A800 GPU. Optimization was performed via the AdamW algorithm to maximize predictive correlation, with an initial learning rate of 10^{-3} and a weight decay coefficient of 0.01. Training was capped at a maximum of 20 epochs. The loss function comprised the negative Pearson correlation coefficient between predicted and ground-truth sequences.

2.5 Testing

We designated the officially split test dataset as the "**Local test**", while the subset reserved for leaderboard updates upon submission was termed the "**leaderboard test**". Performance metrics for our used models were reported across both datasets. The pilot experiment revealed declined reconstruction performance at the start and the end of the reconstructed sequence. Accordingly, we discarded the initial and terminal 5-second segments from seq2seq outputs prior to evaluation. The metric used in the task was the F1-macro score, details in Section **A.1**.

2.6 Baseline Models and Ensemble

To enhance model performance, we implemented an ensemble learning framework. Within this ensemble, in addition to the proposed SHINE model, we incorporated several established models for EEG signal reconstruction, including ConvConcatNet (4), AWavNet (16), and BrainMagic (3). We hypothesized that their inclusion would augment the ensemble's diversity, thereby improving overall robustness.

Throughout the course of the competition, we iteratively refined the hyperparameters and random seeds of the SHINE, as well as those baseline models. Ultimately, for each track, we archived over 200 trained models to enable the final ensemble.

3 Results

We compared the performance of SHINE with baseline models. Each model underwent hyperparameter fine-tuning to optimize the F1-macro score. Table 1 delineated the highest scores attained by each model on the leaderboard test set.

Table 1: F1-macro score on the Local test and Leaderboard test for the Standard and Extended Track

Model	Standard Track		Extended Track	
	Local test	Leaderboard test	Local test	Leaderboard test
BrainMagic	0.8996	0.8912	0.8999	0.8951
AWavNet	0.8974	0.8881	0.8993	0.8905
ConvConcatNet	0.8988	0.8917	0.9007	0.8956
SHINE	**0.9067**	**0.9015**	**0.9097**	**0.9045**

For the Extended Track, we augmented each model by concurrently reconstructing Mel spectrograms and speech envelopes during the training stage. These results demonstrated that such auxiliary training strategy enhanced overall model efficacy.

Finally, we ensembled over 200 models by refining the hyperparameters and random seeds of four mentioned models for each track. This integrative approach yielded macro-averaged F1 scores (F1-macro) of **0.9155** in the Standard Track and **0.9184** in the Extended Track.

4 Discussion

The proposed SHINE network achieves efficient extraction of MEG features through iterative concatenation, with its core architecture predicated on CNN. Prior research posits that CNN functions analogously to matched filters, thereby enabling the capture of signal "templates" from low signal-to-noise ratio EEG and MEG signals (17). This property may underpin one of the key factors contributing to the SHINE network's commendable performance in the present competition, further underscoring the important role of CNNs in neural signal decoding (18; 19; 20; 21; 22; 23).

In recent EEG decoding works, architectures such as Transformers (24; 25) and Mamba (26) have been used to improve performance. However, constrained by the time limit in this competition, our similar explorations failed to yield performance gains. Such avenues nonetheless hold promise as prospective directions for future enhancements.

5 Conclusion

In this work, we proposed a novel SHINE model for reconstructing binary speech-silence sequences from MEG signals, demonstrating superior performance in the LibriBrain Competition 2025. Through task reformulation into a seq2seq framework, integration of auxiliary speech features in the Extended Track, and ensemble learning with diverse baselines, our approach achieved high F1-macro scores, underscoring the importance of capturing temporal dynamics and hierarchical neural processing in speech detection. Future work should explore architectures like Transformers or Mamba to achieve better results.

References

[1] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, pp. 13 367–13 372, 2001.

[2] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[3] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1097–1107, 2023.

[4] 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)*, 2024, pp. 113–114.

[5] 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. P. Jones, "The 2025 PNPL competition: Speech detection and phoneme classification in the libribrain dataset," no. arXiv:2506.10165, 2025, arXiv:2506.10165 [cs].

[6] 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," no. arXiv:2506.02098, 2025, arXiv:2506.02098 [cs].

[7] 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.

[8] B. Accou, J. Vanthornhout, H. Van Hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Scientific Reports*, vol. 13, no. 11, p. 812, 2023.

[9] L. Bollens, C. Puffay, B. Accou, J. Vanthornhout, H. Van Hamme, and T. Francart, "Auditory EEG decoding challenge for ICASSP 2024," *IEEE Open Journal of Signal Processing*, pp. 1–12, 2025.

[10] C. Fan, S. Zhang, J. Zhang, E. Liu, X. Li, G. Zhao, and Z. Lv, "DMF2Mel: A dynamic multiscale fusion network for EEG-driven mel spectrogram reconstruction," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25, New York, NY, USA, 2025, pp. 6977–6985.

[11] C. Fan, S. Zhang, J. Zhang, Z. Pan, and Z. Lv, "SSM2Mel: State space model to reconstruct mel spectrogram from the EEG," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[12] B. Wang, X. Xu, Z. Zhang, H. Zhu, Y. Yan, X. Wu, and J. Chen, "Self-supervised speech representation and contextual text embedding for match-mismatch classification with EEG recording," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 111–112.

[13] X. Xu, B. Wang, B. Xiao, Y. Niu, Y. Wang, X. Wu, H. Cheng, and J. Chen, "The impacts of temporal autocorrelations on EEG decoding," *Biomedical Signal Processing and Control*, vol. 113, p. 108783, 2026.

[14] C. Puffay, B. Accou, L. Bollens, M. Jalilpour Monesi, J. Vanthornhout, H. Van Hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.

[15] I. Rotaru, S. Geirnaert, N. Heintz, I. Van de Ryck, A. Bertrand, and T. Francart, “What are we really decoding? unveiling biases in EEG-based decoding of the spatial focus of auditory attention,” *Journal of Neural Engineering*, vol. 21, no. 1, p. 016017, 2024.

[16] Y. Fang, H. Li, X. Zhang, F. Chen, and G. Gao, “Cross-attention-guided wavenet for mel spectrogram reconstruction in the ICASSP 2024 auditory EEG challenge,” in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 7–8.

[17] L. Stanković and D. Mandić, “Convolutional neural networks demystified: A matched filtering perspective-based tutorial,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 6, pp. 3614–3628, 2023.

[18] B. Accou, M. Jalilpour Monesi, J. Montoya, H. Van hamme, and T. Francart, “Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network,” in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1175–1179.

[19] Z. Qiu, J. Gu, D. Yao, and J. Li, “Exploring auditory attention decoding using speaker features,” in *INTERSPEECH 2023*, 2023, pp. 5172–5176.

[20] Z. Qiu, J. Gu, D. Yao, J. Li, and Y. Yan, “BMMSNet: Bidirectional mapping and multilevel similarity comparison for EEG-speech match-mismatch problem,” in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 117–118.

[21] Z. Qiu, D. Yao, and J. Li, “StreamAAD: Decoding spatial auditory attention with a streaming architecture,” in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 1–5.

[22] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, “EEG-based detection of the locus of auditory attention with convolutional neural networks,” *eLife*, vol. 10, p. e56481, 2021.

[23] X. Xu, B. Wang, Y. Yan, X. Wu, and J. Chen, “A DenseNet-based method for decoding auditory spatial attention with EEG,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway, 2024, pp. 1946–1950.

[24] L. Bollens, B. Accou, H. Van hamme, and T. Francart, “Contrastive representation learning with transformers for robust auditory EEG decoding,” *Scientific Reports*, vol. 15, no. 1, p. 28744, 2025.

[25] Q. Ni, H. Zhang, C. Fan, S. Pei, C. Zhou, and Z. Lv, “DBPNet: Dual-branch parallel network with temporal-frequency fusion for auditory attention detection,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 1–9.

[26] C. Fan, H. Zhang, Q. Ni, J. Zhang, J. Tao, J. Zhou, J. Yi, Z. Lv, and X. Wu, “Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding,” *Information Fusion*, p. 102946, 2025.

A Appendix

A.1 Evaluation Criteria

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K} 2 \cdot \frac{\mathrm{Precision}_k \cdot \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k} \tag{1}$$

This is the unweighted average of per-class F1 scores, where the F1 score is the harmonic mean of Precision and Recall. Here, each class k is speech or silence for Speech Detection. For reference, Precision and Recall can be defined in terms of True Positives (TP), False Positives (FP), and False Negatives (FN), each of which is an integer count:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$