All-for-One: Combining Small Convolutional Models for MEG-Based Speech Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decoding speech from non-invasive brain recordings remains a major challenge for brain–computer interfaces. The LibriBrain Competition (NeurIPS 2025) (Landau et al., 2025) addresses this question, with the aim of classifying moments of speech and silence from magnetoencephalography (MEG) recordings. To accomplish this task, we adapted a convolutional neural network (CNN) previously developed to model a similar task (Défossez et al., 2023). Instead of training a single large model, we took an ensemble modeling approach and trained multiple small convolutional networks and combined their outputs using a lightweight transformer. Our results show that in this task, small convolutional architectures, when aggregated, can achieve strong performance.

1 Introduction

The 2025 LibriBrain competition aims to catalyze progress in non-invasive speech decoding by providing a large-scale MEG dataset and standardized benchmarks (Landau et al., 2025). The competition consisted of a classification task that classified speech versus silence from the MEG signals of participants listening to continuous speech. We based our model on BrainMagick (Défossez et al., 2023), which demonstrated the potential of CNNs for decoding speech from MEG signals. One specificity of our team was that we worked with limited computational resources (two 4GB Quadro T1000 GPUs). This prevented us from training large end-to-end architectures. Despite these computational constraints, we show that small convolutional models trained independently can achieve strong individual performance. Second, we demonstrate that combining several such models through a transformer yields synergistic improvements, outperforming the best individual network.

2 Material and Methods

2.1 Dataset and Task

We used the LibriBrain dataset (Özdogan et al., 2025), comprising over 50 hours of preprocessed MEG recordings (306 channels, 250 Hz) from a single participant listening to LibriVox audiobooks. The input data x is a MEG epoch of 306 sensor × T samples tensor. For each input data x we have a corresponding label y, where $y = 0$ if silence, $y = 1$ if speech, $y = 2$ if speech onset and $y = 3$ if speech offset. "Speech onsets" corresponded to the first 25 samples, or 100 ms, of each speech utterance (Hamilton et al., 2018). Similarly, "speech offsets" corresponded to the first 25 samples, or 100 ms, of silence after each speech utterance. Input data was sampled from the full MEG signals with a stride of 1 sample.

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

Training data consisted in the standard LibriBrain training dataset (51.57 hours) minus the 8th chapter of the 4th book that contained annotations errors ("*0,8,Sherlock4,1*"). Validation and test data were the LibriBrain validation (0.36 hours) and test (0.38 hours) dataset respectively.

2.2 Model Architecture

We used a two-stage architecture in which (1) multiple small CNNs generated class logits that were (2) subsequently combined by a transformer for final prediction.

The CNN architecture was based on BrainMagick(Défossez et al., 2023) and adapted for small-scale GPU setups. We used the same spatial attention layer and the same temporal convolutional structure. We decreased the width of the convolutional blocks from 320 to 64 channels. We increased the number of convolutional blocks with residual connections from 5 to 7. Finally, we added a classifier head at the end. The final architecture consisted in a spatial attention pooling layer with 128 channels, a layer of temporal convolution with 64 channels and kernel size 1, a layer of temporal convolution with 32 channels and kernel size 1, 7 blocks of temporal convolution with kernel size 3 and dilation from 1 to 64 with residual skip connections, a classification head of 3 fully connected layers with 256, 32 and 2 channels respectively.

We trained four first-stage CNNs varying in window length and label set: 3 s, 5 s, and 7 s windows for binary speech/silence classification, and a fourth 5 s model distinguishing speech, silence, speech onset, and speech offset (Hamilton et al., 2018) (see Table 1).

The transformer received as input the concatenated logits from all CNNs. It used positional encoding and operated on sequences of length 200 with a batch size of 32, a model dimension of 64, four attention heads, two layers, and a dropout rate of 0.2.

2.3 Regularization

The final prediction of the transformer model was further regularized to account for the global statistics of the training dataset: silence segments of less than 100 ms were replaced by speech and speech segments of less than 300 ms were replaced by silence.

2.4 Training details

Due to limited computational resources, each model was trained only once without hyperparameter optimization. For the CNNs, we used a cross-entropy loss and the AdamW optimizer (learning rate = 3×10^{-4}, batch size = 128). Validation loss was computed every 1k steps. Training was stopped after 10k steps without improvement in the validation loss. For the transformer, we used a binary cross-entropy with logits loss and the AdamW optimizer (learning rate = 3×10^{-4}, batch size = 32), training until the full training dataset had been traversed once.

3 Results

All four convolutional models achieved strong individual performance, with F1-macro scores consistently above 85%, largely beating the competition benchmark (F1-macro = 68%) on the test set. Combining their logits through the transformer led to a consistent improvement, reaching an

Table 1: Speech detection performance (F1-macro, %) of individual convolutional models, their transformer ensemble, and the final regularized system on the LibriBrain dataset.

Model	# labels	Window size [s]	F1-macro (test)	F1-macro (held-out)
Individual CNN 1	2	2	86.9	–
Individual CNN 2	2	5	88.8	–
Individual CNN 3	2	7	87.2	–
Individual CNN 4	4	5	86.6	–
+ transformer ensemble	–	–	89.3	–
+ regularization	–	–	89.5	90.3

F1-macro of 89.3%. Applying the corpus-level regularization further increased performance to 89.5% on the test data and 90.3% on the held-out evaluation set.

4 Discussion

Small convolutional networks remain efficient baselines for MEG-based speech decoding, even under computational constraints. When combined through a lightweight transformer, these models form an ensemble that combines the complementary representations learned by each CNN, resulting in improved performance. Our suggests that ensemble-based strategies allow to enhance performance without requiring large models, providing a simple approach to non-invasive speech decoding.

References

Landau, G.; Özdogan, M.; Elvers, G.; Mantegna, F.; Somaiya, P.; Jayalath, D.; Kurth, L.; Kwon, T.; Shillingford, B.; Farquhar, G.; others The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. *arXiv preprint arXiv:2506.10165* **2025**,

Défossez, A.; Caucheteux, C.; Rapin, J.; Kabeli, O.; King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* **2023**, *5*, 1097–1107.

Özdogan, M.; Landau, G.; Elvers, G.; Jayalath, D.; Somaiya, P.; Mantegna, F.; Woolrich, M.; Jones, O. P. LibriBrain: Over 50 Hours of Within-Subject MEG to Improve Speech Decoding Methods at Scale. *arXiv preprint arXiv:2506.02098* **2025**,

Hamilton, L. S.; Edwards, E.; Chang, E. F. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology* **2018**, *28*, 1860–1871.