Bypassing Direct Reconstruction: Speech Detection from MEG via Large-Scale Audio Retrieval

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decoding speech from non-invasive brain signals is challenging. For the LibriBrain 2025 Speech Detection task, we propose a novel two-step framework that bypasses direct reconstruction. First, a contrastive learning model retrieves the matching speech segment for the given test MEG from a large-scale audio library (LibriVox). Second, a speech detection model generates the binary silence/speech sequence directly from this retrieved audio. With this approach, our team **Sherlock Holmes** achieved first place in the extended track (F1-score: 0.962), demonstrating that leveraging external audio databases is a highly effective strategy.

1 Introduction

Speech perception involves transforming auditory inputs into increasingly abstract language representations (1; 2; 3; 4). Accordingly, non-invasive magnetoencephalography (MEG) or electroencephalography (EEG) recordings during speech perception have been shown to capture hierarchical features of the speech (5; 6; 7; 8; 9). Numerous studies have successfully related M/EEG with speech. These efforts can be broadly categorized into two paradigms: regression tasks and match-mismatch tasks (10). In regression tasks, neural networks are employed to reconstruct speech features directly from M/EEG segments, such as envelopes, mel-spectrograms, and et al. (5; 11; 12). In match-mismatch tasks, neural networks learn to identify the target speech segment a subject is listening to from a predefined pool of candidates by maximizing the similarity between M/EEG segments and speech segments in a latent space (7; 13; 14). Collectively, these studies have made significant contributions toward developing non-invasive Brain-Computer Interfaces (BCIs) based on speech decoding.

In the LibriBrain Competition 2025 Speech Detection task, participants are required to train a model to distinguish between speech and silence based on brain activity measured by MEG (15). In this setup, the label '0' corresponds to silence and '1' to speech. This task can be viewed as a regression problem, aiming to reconstruct a binary 0/1 sequence from MEG signals. However, reconstructing dynamic speech features from M/EEG data is challenging due to the relatively low signal-to-noise ratio (SNR). For instance, the accuracy (measured by the Pearson correlation coefficient between decoded and target speech features) of decoding mel-spectrograms from EEG is typically below 0.2 (12; 16; 17). Our own experiments indicate that even with the superior signal quality of MEG, the accuracy for decoding mel-spectrograms remains around 0.4. This level of accuracy is currently insufficient to support the synthesis of intelligible speech. In contrast, match-mismatch tasks benefit from the constraint provided by the candidate set. Previous research has shown that models can identify the target speech segment from a pool of over 1000 candidates based on 3-second MEG recordings, achieving an average accuracy of 41%, which highlights a promising avenue for decoding speech from non-invasive brain activity (7).

Inspired by these findings, we proposed a two-step decoding approach for the extended track of the Speech Detection task, as illustrated in Figure 1. In the first step, we employed a match-mismatch task to identify the speech segment corresponding to the test MEG, from a large-scale dataset LibriVox. In the second step, we performed the speech detection task on the matched speech segment. Our approach ultimately achieved first place on the track.

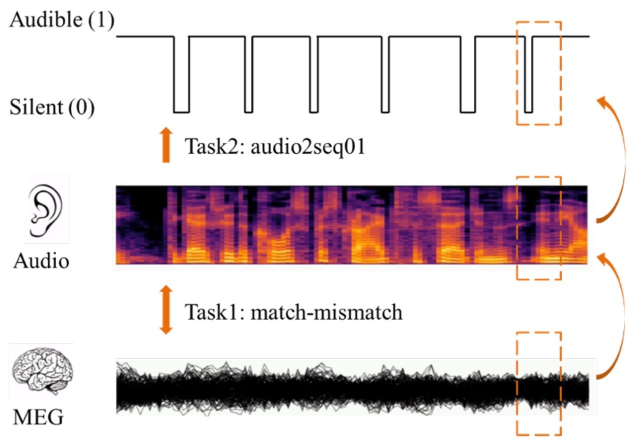

Figure 1: Overall framework of our approach.

2 Methods

2.1 Step1: MEG-Speech Match-mismatch

The objective of this step is to align MEG recordings with speech segments. We adopted a contrastive learning framework, as depicted in Figure 2. For a segment of MEG data $X \in \mathbb{R}^{C \times T}$, where C represents the number of MEG channels and T represents the time samples. A CNN-based MEG encoder was utilized for extracting neural features $Z \in \mathbb{R}^{H \times T}$. Meanwhile, a pretrained Wav2vec 2.0 model [1] was used to obtain the speech representation (extracted from the outputs of its ninth hidden layer). This representation is subsequently projected via a linear layer to obtain features $F \in \mathbb{R}^{H \times T}$. Given a batch of N samples, let $\mathcal{Z} = \{Z^1, Z^2, \ldots, Z^N\}$ denote the MEG features and $\mathcal{F} = \{F^1, F^2, \ldots, F^N\}$represent the speech representations. The InfoNCE (Information Noise-Contrastive Estimation) loss is employed (18), aiming to maximize the similarity between matched pairs (Z^i, F^i) while minimizing the similarity between mismatched pairs (Z^i, F^j) for $j \neq i$. The loss is formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(Z^i, F^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(Z^i, F^j)/\tau)} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes the similarity measure, and τ is a temperature parameter that modulates the sharpness of the distribution. The $\text{sim}(Z^i, F^i)$ is calculated as:

$$\text{sim}(Z^i, F^i) = \frac{1}{H}\sum_{k=1}^{H} \text{corr}(z_k^i, f_k^i) \tag{2}$$

where $\text{corr}(\cdot, \cdot)$ is the Pearson correlation between two vectors.

2.2 Step2: Speech detection

In this step, we train a speech detection model. This model takes the mel-spectrogram of a speech segment as input and outputs a binary sequence (0 for silence, 1 for speech). The model is implemented using a deep CNN. The network parameters are optimized using the negative Pearson correlation coefficient as the loss function.

[1]We used wav2vec2-base-960h from `https://huggingface.co/facebook/wav2vec2-base-960h`

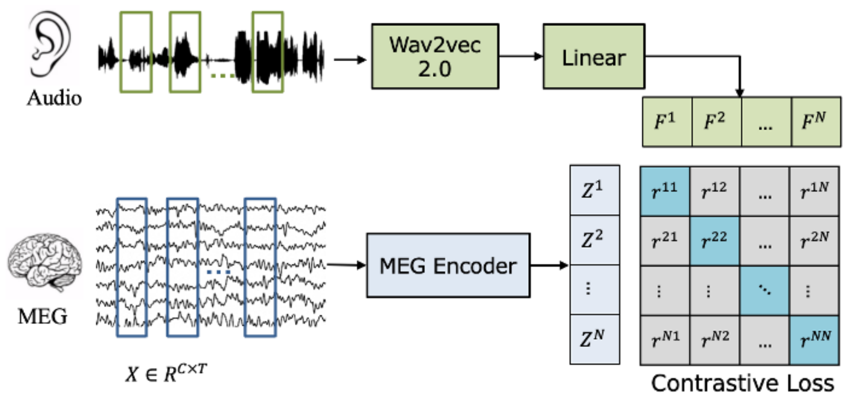

Figure 2: The contrastive learning framework for the match-mismatch task.

3 Experiment

3.1 Data Preparation

Our framework requires MEG data and its temporally aligned speech signals for training. Following the speech source URLs provided in the organizers' paper (19), we downloaded all corresponding audiobooks from LibriVox, which we refer to as Libriaudio. The actual audio stimuli presented to subjects during the MEG recording sessions are denoted as MEGaudio. Analysis of the dataset's event.tsv file, specifically by comparing the 'timemeg' and 'timechapter' columns, revealed that MEGaudio was generated by inserting silent segments into the original Libriaudio. For instance, in the session 'sub-0_ses-1_task-Sherlock1_run-1_proc-bads+headpos+sss+notch+bp+ds_meg.h5', the corresponding MEGaudio contained 171 additional silent segments compared to the original Libriaudio. The duration distribution of these extra silent segments is shown in Figure A.1a (median duration ≈ 0.03 s). Cumulatively, these segments added approximately 5 seconds of silence to the MEGaudio. As illustrated in Figure A.1b, we used the timing information from the event.tsv file to synthesize the MEGaudio by inserting silent segments of corresponding durations into the Libriaudio at the specified timestamps. This synthesized MEGaudio was subsequently used for training both the MEG-Speech match-mismatch model and the Speech Detection model.

3.2 Model Training

Data from session 9 and 10 of the Sherlock 1 were used as the local validation set, while data from session 11 and 12 served as the local test set. All remaining data constituted the training set.

For the MEG-speech match-mismatch model, the ConvConcatNet architecture (12) was employed as the MEG encoder. The dimensionality of the latent space was set to 8 to reduce computational cost during the subsequent testing phase. Following previous work, the MEG data and corresponding speech for each session were segmented into non-overlapping 3-second windows. The model was trained using the Adam optimizer (20) with a learning rate of 1×10^{-3} and a batch size of 256. The temperature parameter τ in the InfoNCE loss was set to 0.015. Training was stopped if the Top-10 accuracy on the validation set failed to improve for 5 consecutive epochs.

For the speech detection model, the ConvConcatNet was also used. The model input was the mel-spectrogram of the MEGaudio, and the target labels (binary 0/1 sequences) were derived from the event.tsv file. Segment length was set to 30 seconds to ensure that each segment contained both speech and silence periods. The model was trained using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64. The negative Pearson correlation coefficient was used as the loss function. Training was stopped if the validation loss did not decrease for 5 consecutive epochs. The optimal binarization threshold for distinguishing speech from silence was determined via a grid search on the validation set.

All models were trained on an HPC node equipped with 8 A800 GPUs.

3.3 Model Testing

3.3.1 Retrieving Matching Speech from LibriVox

During testing, our goal was to retrieve the speech segment from the LibriVox corpus that matched the given test MEG. We downloaded a large-scale subset of LibriVox, comprising approximately 60% of its total data ($\sim$10,000 audiobooks). Each audiobook was split into non-overlapping 5-second segments. As an example, the chapter "A Continuation of the Reminiscences of John Watson MD" from "A Study In Scarlet (Version 6)" (hereafter *studyinscarlet13*), with a duration of 27 minutes and 31 seconds, was split into 330 segments.

The holdout test MEG data had a total duration of 2243 seconds. It was segmented using a 5-second sliding window with a 0.1-second stride, resulting in 22,380 MEG segments. The matching process is illustrated in Figure A.2a. For each of the 330 speech segments from *studyinscarlet13*, we identified the most similar MEG segment from the pool of 22,380 test segments, recording its index. This produced a sequence of 330 indices, which we term the Matched MEG ID Sequence (MMIS). For the vast majority of LibriVox audiobooks, which do not correspond to the holdout MEG data, the MMIS shows no discernible pattern. However, for the matching audio, a large subset of the MMIS should form a monotonically increasing sequence, reflecting the temporal order of the MEG data (Figure A.2b).

To identify this subset, we computed the Longest Ascending Subsequence (LAS) of the MMIS. Experimental results indicated that among the $\sim$10,000 downloaded audiobooks, only *studyinscarlet13* yielded an LAS length exceeding a manually set threshold of 20. This audio was identified as matching the final portion of the holdout test MEG, starting from the 1398-s mark. No matching audio was found for the MEG data preceding 1398 s.

3.3.2 Generating the Binary Sequence from Speech

Based on the analysis in Section 3.1, which indicated that most extra silent segments were inserted between sentences, we segmented the *studyinscarlet13* audio into 241 sentences according to its text transcript. Using the trained MEG-Speech match-mismatch model, the first 126 sentences were confirmed to match the test MEG data. Silent segments of corresponding durations were inserted between these sentences so that the total duration matched that of the MEG data after 1398 s. Finally, the trained speech detection model was applied to generate the binary 0/1 sequence, which served as the decoded output for the MEG signal after 1398 s.

For the initial 1398 s of the test MEG, no audio file from our LibriVox subset produced an MMIS with an LAS length greater than 20. We hypothesize that the corresponding audio might reside in the remaining 40% of the LibriVox corpus we did not download, or originate from another source outside LibriVox. An interesting observation was that segments from audiobook "The Darkest Hour" appeared frequently in matches for the preceding 1398 s, but not in sequential order. For this initial portion, we employed a simple regression approach, similar to the method used by Team SHINE (as discussed in the provided Discord thread), utilizing a basic CNN+LSTM network to reconstruct the binary sequence directly from the MEG signals.

Submitting the prediction result comprising of the above two parts to the extended track, we obtained an F1-score of 0.962, securing first place on the leaderboard.

4 Conclusion

We presented a two-step framework for speech detection from MEG signals. By reframing the problem as an audio retrieval task followed by speech analysis, we circumvented the limitations of direct feature regression from noisy neural data. Our method successfully identified the target audio from a vast pool of candidates and generated accurate binary sequences, winning the LibriBrain competition. This work validates the potential of using match-mismatch tasks to advance non-invasive BCIs.

References

[1] I. DeWitt and J. P. Rauschecker, "Phoneme and word recognition in the auditory ventral stream," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. E505–E514, 2012.

[2] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen, "The hierarchical cortical organization of human speech processing," *Journal of Neuroscience*, vol. 37, no. 27, pp. 6539–6557, 2017.

[3] D. Poeppel, "The neuroanatomic and neurophysiological infrastructure for speech and language," *Current Opinion in Neurobiology*, vol. 28, pp. 142–149, 2014.

[4] S. K. Scott and I. S. Johnsrude, "The neuroanatomical and functional organization of speech perception," *Trends in Neurosciences*, vol. 26, no. 2, pp. 100–107, 2003.

[5] 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.

[6] A. M. Chan, E. Halgren, K. Marinkovic, and S. S. Cash, "Decoding word and category-specific spatiotemporal representations from MEG and EEG," *NeuroImage*, vol. 54, no. 4, pp. 3028–3039, 2011.

[7] 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.

[8] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[9] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel, "Cortical tracking of hierarchical linguistic structures in connected speech," *Nature Neuroscience*, vol. 19, no. 1, pp. 158–164, 2016.

[10] 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.

[11] B. Wang, X. Xu, L. Zhang, B. Xiao, X. Wu, and J. Chen, "Semantic reconstruction of continuous language from MEG signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 2190–2194.

[12] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops*, 2024, pp. 113–114.

[13] M. Jalilpour Monesi, B. Accou, J. Montoya-Martinez, T. Francart, and H. Van Hamme, "An LSTM based architecture to relate speech stimulus to EEG," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 941–945.

[14] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing Workshops*, 2024, pp. 111–112.

[15] 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].

[16] H. Li, Y. Fang, X. Zhang, F. Chen, and G. Gao, “Cross-attention-guided WaveNet for EEG-to-mel spectrogram reconstruction,” in *Proceedings of Interspeech*, 2024, pp. 2620–2624.

[17] M. Sakthi, A. Tewfik, and B. Chandrasekaran, “Native language and stimuli signal prediction from EEG,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 3902–3906.

[18] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” no. arXiv:1807.03748, 2018, arXiv:1807.03748.

[19] 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].

[20] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2015, arXiv:1412.6980.

A Supplementary Material

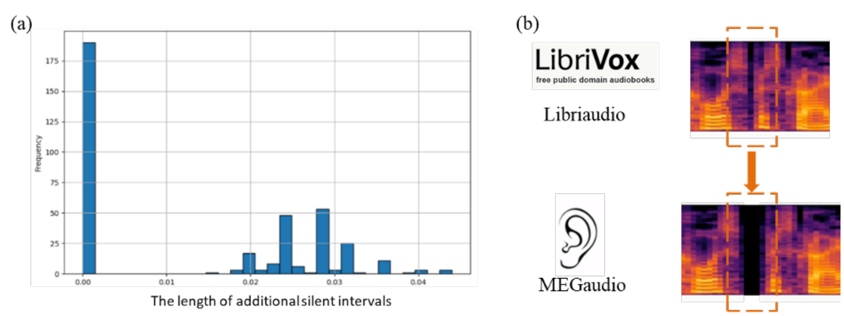

Figure A.1: (a) Duration distribution of the extra silent segments for an example session. (b) Synthesizing MEGaudio by inserting silent segments into the original Libriaudio.

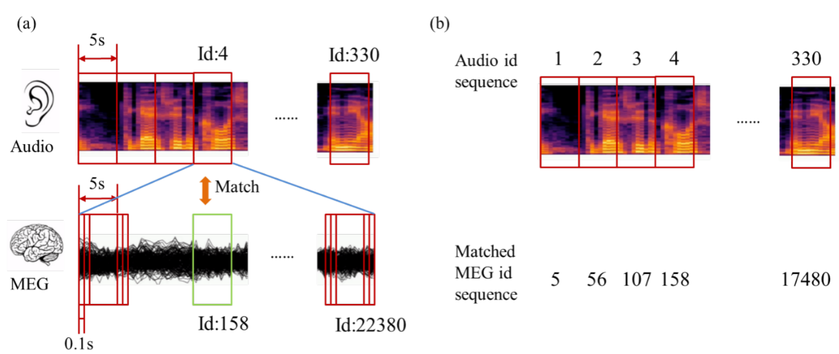

Figure A.2: (a) During testing, speech segments from the LibriVox corpus are matched against segments from the holdout MEG data. (b) The ideal MMIS for the matching audio is a monotonically increasing sequence.