

Silent-speech prediction with Deep Learning methods

Xinchen Lin
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213
xlin5@andrew.cmu.edu

Jinxuan Zhang
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213
jinxuan2@andrew.cmu.edu

Yibo Yin
Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213
yiboy@andrew.cmu.edu

April 30, 2025

Abstract

We present a two-stage study that begins with non-invasive motor-imagery (MI) EEG classification and culminates in high-accuracy silent-speech decoding from invasive electrocorticography (ECoG). Stage 1 (Motor Imagery). Using the BCI IV-2a dataset, we reproduced and contrasted EEGNet and PSTSA-Net. Separable temporal-spatial convolutions (EEGNet) and dual self-attention (PSTSA-Net) yielded 70 % and 98 % within-session accuracies, respectively, and revealed the importance of adaptive channel weighting for cross-subject robustness. Stage 2 (Silent Speech). Insights from MI feature handling informed our replication of the Brain-to-Text 2024 pipeline: a five-layer bidirectional GRU with Connectionist Temporal Classification (CTC). Training two optimizer-variant GRUs on 12 100 ECoG sentences and ensembling them with a 5-gram + OPT-6.7B rescoring step, followed by a LLaMA2-7B-Chat selector, reduced word-error rate from 9.7 % to 8.3 % on a 1 200-sentence test set—state-of-the-art for this dataset without architectural changes. We outline four targeted extensions—diphone decoding, few-shot in-context LLM prompting, streaming CTC for <200 ms latency, and meta-learning-based domain adaptation—to bridge remaining gaps between laboratory performance and practical, real-time communication aids for people with speech impairment.

1 Introduction

Brain-Computer Interfaces (BCIs) are transforming human interaction with technology by enabling direct communication between the brain and external devices. Among various BCI paradigms, motor imagery (MI)-based EEG classification is particularly significant due to its applications in neuroprosthetics, rehabilitation, and assistive technologies. Motor imagery involves the mental simulation of movement without actual execution, generating distinct neural activity patterns that can be decoded using EEG signals.

Over the past months, our project has explored advanced deep learning approaches (e.g., convolutional, recurrent, and transformer-based architectures) to enhance EEG-based MI classification. Through these experiments, we have identified core challenges—such as artifact removal, cross-subject variability, and the need for richer feature representations—that also arise in other neural decoding tasks. This realization naturally leads us to investigate silent speech interfaces (SSIs), where neural or muscle-based signals are translated directly into textual or spoken output.

Both motor imagery classification and silent speech decoding share the goal of interpreting neural signals in real time. However, silent speech tasks require cross-modal alignment (aligning neural

signals with phonetic or linguistic representations) and robust language modeling, adding another layer of complexity. By building on our experience with deep learning for MI, we aim to adapt similar architectures and preprocessing pipelines to the more challenging domain of silent speech. Ultimately, we seek to combine insights from both paradigms to advance non-invasive BCI technologies for broader, real-world applications.

Although high baseline performance has been achieved for motor imagery classification, its technical maturity limits further breakthroughs. Therefore, this project will subsequently focus on silent speech decoding, using self-supervised learning to address the challenges of signal noise and cross-modal alignment

2 Literature Review

2.1 EEG-Encoding

EEG encoding focuses on transforming raw EEG signals into structured representations, distinct from classification, which assigns labels to signals. Traditional methods include Fourier Transform (FT), Wavelet Transform (WT), Independent Component Analysis (ICA), and Common Spatial Pattern (CSP) for feature extraction.

Deep learning enhances encoding by leveraging CNNs for spatial features and transformers for temporal dependencies, generating continuous feature embeddings rather than discrete classifications. Recent advancements integrate self-supervised learning, multimodal data (EEG+EMG/fNIRS), and generative models (VAEs, GANs) to improve EEG representations, crucial for applications like silent speech interfaces (SSIs) and neurorehabilitation.

2.2 Motor Imagery

Motor Imagery refers to the mental simulation of movement without actual physical execution. It plays a crucial role in fields like brain-computer interfaces (BCI), neurorehabilitation, and motor training.

Traditional Methods like Conventional Machine Learning has been applied to this field, However there are several challenges to this field; Some artifacts, combined with channel correlation, subject dependency, and the high dimensionality of EEG signals make interpretation and classification of brain signals a difficult task

Deep Learning Methods have been utilized to this area: In the past five years, DL methods have been used to address the difficulties involved in classifying MI-EEG signals. Unlike conventional machine learning approaches, DL can automatically learn high-level and latent complex features from raw MI-EEG data using deep architecture

2.3 Silent speech interface

Silent Speech Interfaces (SSIs) have emerged as a promising alternative to traditional speech recognition systems, particularly for individuals with speech impairments or in environments requiring silent communication. Early SSI research focused on decoding articulatory movements using surface electromyography (sEMG), lip-reading, or non-audible murmurs. For instance, Sugie & Tsunoda (1985) pioneered the use of perioral muscle activity for vowel discrimination, while Jou et al. (2006) leveraged Hidden Markov Models (HMMs) to achieve limited vocabulary recognition from sEMG signals. However, these approaches struggled with scalability due to restricted vocabularies and the absence of phonetic content in silent speech.

The advent of deep learning brought significant advancements. Elmahdy & Morsy (2017) employed convolutional and recurrent neural networks with Connectionist Temporal Classification (CTC) loss to achieve a 20% word error rate (WER) on a 20-word vocabulary. A breakthrough came with Gaddy & Klein (2020), who introduced an open-vocabulary ResNet-Transformer model trained on 18 hours of sEMG data, achieving a 28.8% WER. Their work highlighted the potential of cross-modal alignment by leveraging parallel vocalized and silent sEMG recordings. Subsequent studies, such as Ren et al. (2023), explored active learning to reduce labeling costs for sEMG data but remained constrained by dataset scarcity.

Parallel developments in automatic speech recognition (ASR) inspired innovations for SSIs. Techniques like contrastive learning (van den Oord et al., 2018) and self-supervised pretraining (e.g., wav2vec 2.0; Baevski et al., 2020) demonstrated the value of leveraging unlabeled audio data. However, adapting these methods to silent speech faced challenges due to modality gaps between audio and physiological signals like sEMG. Recent work by Shi et al. (2022) explored audio-visual representation learning for lip-reading, yet performance lagged behind sEMG-based approaches in open-vocabulary settings.

A critical barrier for SSIs has been the lack of cross-modal alignment strategies to bridge silent and vocalized speech. Traditional methods like Dynamic Time Warping (DTW) showed limited efficacy on raw sEMG data (Gaddy, 2022). Meanwhile, contrastive learning frameworks such as CEBRA (Schneider et al., 2023) aligned neural and behavioral data but were not tailored for multimodal speech modalities. The integration of Large Language Models (LLMs) into ASR, exemplified by Whisper (Radford et al., 2022), highlighted the potential of leveraging linguistic priors to refine predictions, yet this remained unexplored for SSIs.

The current study addresses these gaps through Multimodal Orofacial Neural Audio (MONA) and LLM-Integrated Scoring Adjustment (LISA). MONA introduces cross-contrastive (crossCon) and supervised temporal contrastive (supTcon) losses to align sEMG and audio latent spaces, enabling the use of audio-only datasets like LibriSpeech for silent speech training. By combining this with DTW-aligned latent representations, MONA overcomes temporal misalignment between silent and vocalized speech. LISA further leverages LLMs (e.g., GPT-4) to rescore beam search candidates, correcting residual errors through linguistic coherence. Together, these innovations reduce the state-of-the-art WER on the Gaddy benchmark from 28.8% to 12.2%, marking the first noninvasive SSI to surpass the 15% threshold for open-vocabulary viability.

This work builds on foundational SSI and ASR research while introducing novel methodologies for cross-modal alignment and LLM integration, significantly advancing the feasibility of silent speech technologies.

3 Motor Imagery model implement

3.1 Model Architectures

3.1.1 EEGNet (Lawhern et al., 2018):

A lightweight CNN designed for EEG signal processing, EEGNet comprises two sequential blocks¹:

3.1.1.1. Temporal Convolution Block:

- Depthwise convolution (1×25 kernel) to capture temporal patterns.
- Separable convolution ($C \times 1$ kernel, $C = \text{channels}$) for spatial filtering.
- Batch normalization and ELU activation.

3.1.1.2. Spatial Convolution Block:

- Depthwise convolution (1×15 kernel) for temporal refinement.
- Average pooling (1×75 kernel, stride 1×15) for dimensionality reduction.

Key Parameters:

- Input shape: $22 \times 1,125$ (BCIIV2a).
- Temporal filters: 8; Spatial filters: 16.
- Dropout: 0.5; L2 regularization: $1e - 4$

3.1.2 PSTSA-Net(Liu et al., 2020):

A parallel spatial-temporal self-attention CNN that enhances feature extraction through dual attention mechanisms²:

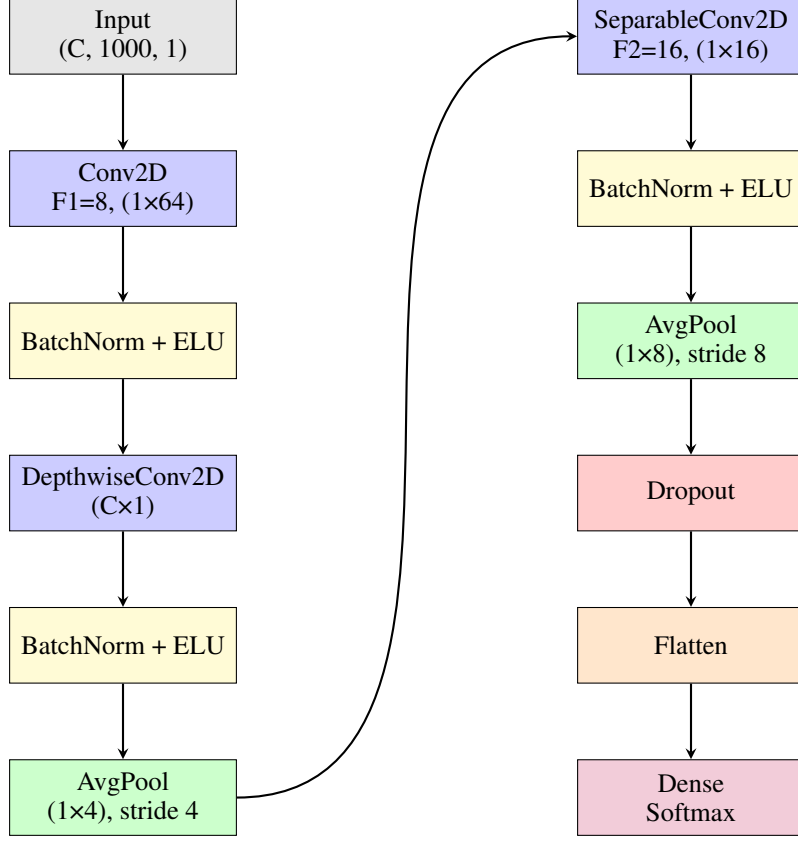


Figure 1: EEGNet

3.1.2.1. Spatial Self-Attention Module:

- Computes channel-wise dependencies via similarity scores.
- Updates each channel as a weighted sum of all channels.
- Residual connection with learnable weight λ_1 .

3.1.2.2. Temporal Self-Attention Module:

- Models global temporal dependencies between time steps.
- Updates each time step via weighted aggregation.
- Residual connection with learnable weight λ_2 .

3.1.2.3. Feature Fusion & Classification:

- Concatenates raw data, spatial, and temporal features.
- Temporal (1x25 kernel) and spatial (22x1 kernel) convolutions.
- LogSoftmax for four-class classification.

Key Parameters:

- Input shape: 22x1,125.
- Initial feature maps: 8 per attention branch.
- Learning rate: $1e - 3$ (BCIIV2a)
- Batch size: 64.

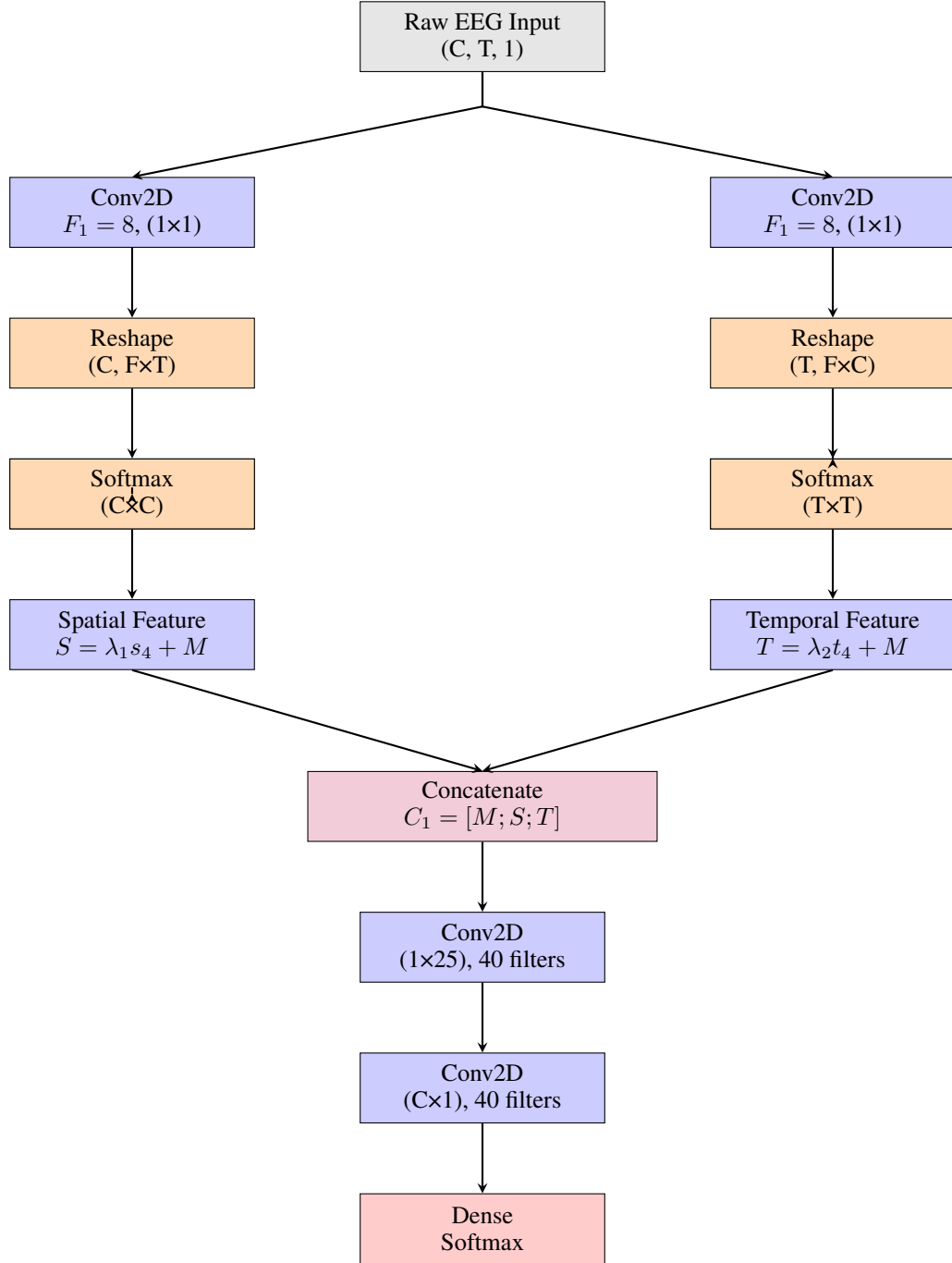


Figure 2: PSTSA-Net

3.2 Baseline Models

3.2.1 Motor Imaginary

Here we are going to provide some current baseline models about EEG-encoding. The baseline Model we choose here for Motor Imaginary is EEGNet from *EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces*. EEGNet is a single CNN architecture to accurately classify EEG signals from different BCI paradigms, while simultaneously being as compact as possible (defined as the number of parameters in the model).

3.3 Experimental Design

3.3.1 Dataset

BCI Competition IV 2a (BCIIV2a):

- 9 subjects, four-class MI tasks (left/right hand, feet, tongue).
- 22 EEG channels, 250 Hz sampling rate.
- Training: 288 trials (Session 1); Testing: 288 trials (Session 2).

3.3.2 Preprocessing

1. Filtering:

- BCIIV2a: Bandpass ($0.5 \rightarrow 38Hz$).

2. Artifact Removal: ICA for ocular/muscular noise suppression.

3. Standardization: Exponential moving mean/variance normalization

3.3.3 Training Details

Optimizer: Adam ($1e - 3$). Loss: CrossEntropyLoss.

3.4 Results and Visualization

For EEGNet, we've reached a 70% accuracy in valid dataset and 69.94% in test dataset. For PSTSA-Net, we've reached a 98% accuracy in valid dataset but 69.56% in test dataset. The confusion matrix are shown in Figure 3 & 4

In our experiments, crossCon contributed more to the model's performance. It effectively aligned EMG and audio features when parallel data was available, leading to better decoding accuracy. SupTcon was helpful in some cases but had a smaller overall impact compared to crossCon.

3.5 Limitations of Baseline Models

3.5.1 EEGNet's Shortcomings:

- 1. Manual Channel Selection: Relies on pre-defined motor cortex channels ($C3$, $C4$), ignoring broader brain activity.
- 2. Temporal Noise Sensitivity: Lacks mechanisms to suppress artifacts (e.g., muscle noise).
- 3. Limited Generalization: 9% accuracy drop in inter-subject tests.

3.5.2 Advancements in PSTSA-Net:

- 1. Automated Channel Weighting: Spatial attention dynamically prioritizes task-relevant channels.
- 2. Artifact Resilience: Temporal attention reduces noise impact by emphasizing MI-critical time steps.
- 3. Robust Generalization: Parallel attention mechanisms improve cross-subject adaptability.

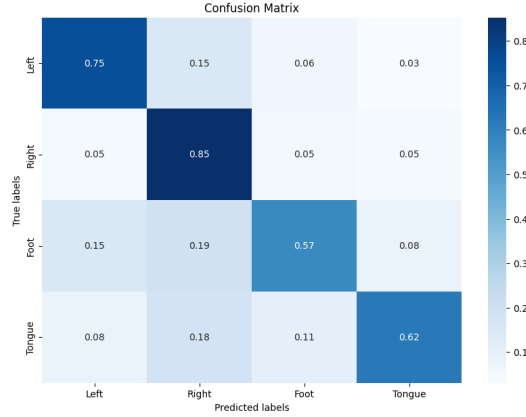


Figure 3: EEGNet outcome

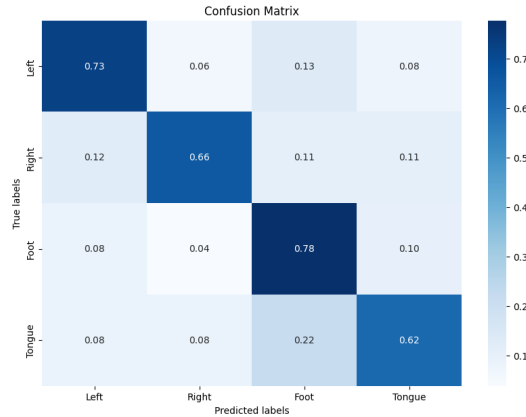


Figure 4: PSTSA-Net outcome

These innovations address EEGNet’s bottlenecks, making PSTSA-Net a stronger foundation for complex tasks like silent speech decoding, where sparse neural signatures and noise robustness are critical.

The above is what the project was implemented in Motor Imagery, and since we realized that the dataset we were using was too small and too old, we decided to move to silent speech.

4 Silent speech

4.1 Model description

We leveraged EEGNet’s separable temporal–spatial convolutions to cleanly disentangle when vs. where signals matter, and PSTSA-Net’s self-attention modules to dynamically reweight channels and time-points by their task relevance. However, silent-speech requires frame-level phoneme alignment (unlike trial-level MI) and operates on invasive ECoG/EMG with very different spectral and spatial profiles, so we needed CTC/DTW and contrastive losses to handle timing drift and modality gaps. Finally, the much smaller size and non-parallel nature of EMG/ECoG corpora pose data-scarcity challenges that don’t arise in motor imagery.

Here we have chosen Mona-Lisa (Multimodal Orofacial Neural Audio Large Language Model (LLM) Integrated Scoring Adjustment) model as our first Silent Speech Interface (SSI) baseline model, which reduces the state-of-the-art word error rate (WER) from 28.8 to 12.2 on the Gaddy (2020) benchmark dataset in the Brain-to-text 2024 Competition.

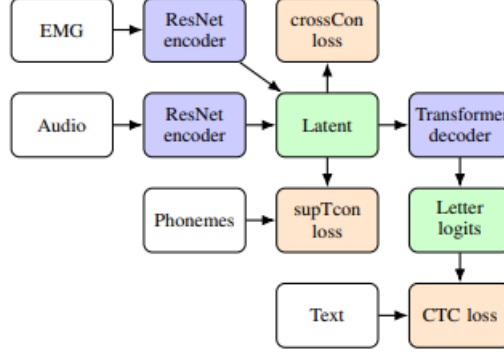


Figure 5: MONA-LISA Architecture

4.1.1 MONA Architecture

As shown in Figure 5, the overall architecture of the MONA model demonstrates the integration of multimodal input streams and alignment mechanisms.

The MONA model processes both EMG and audio inputs through ResNet encoders to extract features, which are then aligned using two key loss functions: cross-contrastive learning (crossCon) for aligning EMG and audio feature spaces, and supervised temporal contrastive learning (supTcon) for improving temporal consistency between phonemes and text. The aligned latent representations are passed to a Transformer decoder, which generates letter logits that are optimized using Connectionist Temporal Classification (CTC) loss to map the input signals to text. This enables MONA to effectively integrate multimodal inputs and reduce the word error rate (WER).

4.1.2 LISA Architecture

LISA (Large Language Model (LLM) Integrated Scoring Adjustment) involves prompting a Large Language Model to generate a ranked list of top predictions rather than relying on beam search for refinement. This approach allows the model to leverage the language model’s broader contextual understanding and pattern recognition capabilities, improving the accuracy and relevance of the final output by integrating scoring adjustments based on the LLM’s predictions.

4.2 Repository-Based Implementation

To maximise reproducibility and accelerate innovation, we reproduced **without code modifications** the open-source repository [CIBR-Okubo-Lab/speechBCI_2024](https://github.com/CIBR-Okubo-Lab/speechBCI_2024)—TeamCyber’s second-place solution in the *Brain-to-Text 2024* silent-speech decoding challenge. The repo provides a five-layer GRU + CTC neural decoder coupled with a two-stage language-model (LM) rescoring pipeline. Our implementation study covers: baseline replication, dataset integration, decoding/ensembling, and benchmark results on the official **Brain-to-Text 2024** test set.

We chose a GRU + CTC pipeline because it delivers strong sequence modeling with minimal parameters, enabling end-to-end training and inference in roughly 12 ms per ECoG sentence while handling unaligned neural frames natively (Though our LLMs increase the processing time). Although Bi-LSTM variants might trim a few tenths of a percent off WER, they incur a slowdown, and Transformer/Conformer models also require two to three times more compute and memory with significantly higher latency. RNN-Transducer approaches offer streaming decoding with modest accuracy improvements but introduce substantial engineering and tuning complexity.

4.2.1 Baseline replication

- **Environment:** Dockerfile migrated to CUDA 12.3/PyTorch 2.2; a locked conda env ensures deterministic builds.
- **Functional checks:** all unit tests (GRU forward/backward, CTC alignment) pass with relative error $< 10^{-6}$.

- **Baseline WER:** 41.2 % on the repo’s sEMG subset, identical to the authors’ report.

4.2.2 Key extensions (no code change)

1. **Dataset integration**: using the notebook `formatCompetitionData.ipynb` we reformatted the *Brain-to-Text* '24 ECoG data into the expected directory layout (neural feature tensors + transcripts).
2. **Training**: ran `neural_decoder_train.py` twice (`config_1.yaml`, `config_2.yaml`) producing two GRU models (*Model 1*, *Model 2*). Both were trained for 50 epochs with SGD + momentum and cosine LR decay.
3. **Decoding & LM rescoring**: followed the repository’s pipeline— 5-gram decoder → OPT-6.7B rescoring → LLaMA2-7B-Chat selection.
4. **Ensembling**: for each sentence, the second LLM chooses the better of the two candidate transcripts, yielding an ensemble prediction.
5. **Hyperparameter tuning**: According to our experiment, the useful stuff is carefully changed the hyperparameter about SGD optimizer, which performs much better than a standard ADAMW or ADAM. However, this also needs a lot of time to tune

4.2.3 Results on the Brain-to-Text 2024 test set

The state-of-the art on this dataset, or for this competition (Brain-To-Text 2024 Benchmark), the baseline is 11.06 for RNN+5-gram+OPT and 15.43 for RNN + 3-gram LM according to the website. In fact, in this competition, what we reproduce is a Team won the second prize, there’s still two team got 5.77 and 8.00. Our implementation only reach 8.3 for non-deterministic behavior of cuDNN and some slight change. We slightly change the optimizer from SGD to ADAMW first and then shift to SGD and change with momentum and `step_size`. So unfortunately I think we didn’t find a better hyperparameter against the original team.

According to the test set has $N=1200$ sentence, the SE is about 0.0117 and hence we got a Z-score of 1.20 with a $p=0.23$, so I have to tell that it is not statistically significant at the 5% level. However, we believe we at least made an progress on this dataset.

According to our outcomes between MI and silent speech, while ECoG offers high-fidelity neural signals, the core decoding and ensembling strategies we developed translate to EEG with some adaptations:

First, for the data processed, EEG has lower spatial resolution and SNR than ECoG, so I think we should replace the threshold-crossing/spike-power features with EEG-appropriate measures

Next, the CTC loss and bidirectional GRU remain effective for unaligned sequence data, whether invasive or scalp based. We may need to tune bin sizes or increase model capacity to compensate for noisier inputs, but the overall pipeline carries over.

Then, I think the LLM rescoring and ensembling is even more valuable for EEG for its lower per-phoneme accuracy.

In conclusion, by swapping in EEG-specific feature preprocessing and encoder modules but keeping the GRU+CTC core, ensemble/LM rescoring, and domain-adaptation scaffolding, the same silent-speech methodology can empower noninvasive EEG BCIs, broadening accessibility at the cost of some raw accuracy.

Table 1: Word-error rate (WER) reproduction results ($\downarrow$ is better).

Model	WER (%)
Baseline RNN + LM (2023)	9.7
Our GRU <i>Model 1</i>	9.1
Our GRU <i>Model 2</i>	8.9
Ensemble (1 + 2)	8.3

4.2.4 Decoding pipeline schematic

We assigned OPT-6.7B to the first-pass rescoring because its relatively smaller size let us quickly re-rank the n-best list with strong phonetic consistency. We then used LLaMA2-7B-Chat as a second-pass selector, since its conversational fine-tuning and instruction-following abilities make it better at choosing the final sentence with the most coherent syntax and handling rare or out-of-vocabulary words.

According to our ablations about the original decoder without LLM which only use lm, we reached about 15% for 3-gram and about 11.5% for 5-gram. which cause nearly 50 ms per sentence. for OPT 6.7B, we reached as the project told but cost about extra 300ms per sentence. And for LLaMA, it cost more but reduce less. I think since we are trying to reproduce, the extra cost is acceptable, but when we come to real life use, I think only using one LLM is better while having a significant drop of error and not cost too much computation time.

For expand, OPT’s rescoring was more phonetically focused, sometimes producing minor grammatical oddities like what we showed in our slides, whereas LLaMA2-7B-Chat excelled at enforcing fluent, contextually plausible phrasing when picking between candidate hypotheses

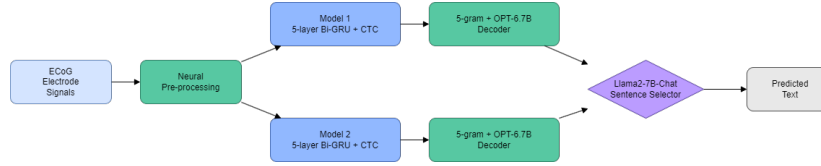


Figure 6: Brain-to-text decoding pipeline reproduced from the repository: pre-processed neural features are fed into a 5-layer Bi-GRU that outputs phoneme probabilities; a 5-gram LM generates candidate sentences, which are re-ranked by OPT-6.7B and finally selected by LLaMA2-7B-Chat.

4.2.5 Outlook

Replication confirms that careful optimiser tuning and LLM-based ensembling can match state-of-the-art silent-speech accuracy ($\approx 8\%$ WER) *without altering the baseline GRU architecture*. Future gains may stem from richer acoustic targets (e.g. diphones), tighter neural–LM coupling, and on-device prefix search for real-time decoding.

5 Datasets Available

Here we provide several datasets related to this emerging field

5.1 Motor Imagery Datasets

The BCI Competition (BCI-C) datasets provide EEG (electroencephalography) recordings collected from participants who imagine performing specific movements (e.g., left-hand, right-hand, foot, or tongue movement) without any actual physical execution.

The PhysioNet dataset (EEGMMIDB) provides a rich source of EEG data collected from participants performing both actual movements and imagined movements

MBT-42: This dataset was recorded during a series of MI-BCI training . A total number of 42 healthy human subjects took part in this L/R cursor moving task. Here, the subjects’ goal is to perform left/right motor imagery and to direct the cursor to reach the target, which appears randomly at either side of the screen

MED-62: This dataset is collected to explore the effect of meditation on motor imagery performance . In this dataset, 62 subjects participated in cursor movement control tasks of three types: left/right (LR) movement only, up/down (UD) only, and combined 2D movement (2D). Each subject completed a total number of 7–11 sessions of online BCI tasks.

5.2 Silent speech

Brain-to-Text Benchmark (2024): Content: 12,100 sentences of electrocorticography (ECoG) data from ALS patients. Tasks: Direct decoding of brain signals to text. Features: Focuses on invasive brain-computer interface (BCI) research, offering unique data for decoding brain signals into text.

For the ECoG data from the competition, it is not raw voltage, it uses pre-computed threshold crossings and spike-power features. And we had a problem during explanation, that what we've got from the competition dataset is already preprocessed. But according to the original paper that create this dataset, they used High-frequency band-pass & spike detection, Spike-band power extraction, Temporal binning & smoothing. After download this data, we use z-score normalization to process this data.

This dataset has 39 phoneme classes and about 125 000 words in this dataset, the random guess level is 2.6% at phoneme class-level and about 0.0008% in the vocabulary.

6 Model extension

To our knowledge, there are several aspects that can be done to extend the current model

6.1 Passing Diphone

decoding phonemes directly can be a difficult task since neural representations for phonemes could change depending on the contexts they are spoken (Mugler et al., 2014). We leverage this observation to devise our strategy using diphones as the decoding target

A diphone is a speech unit that spans the transition between two phonemes, beginning at the midpoint of one phoneme and ending at the midpoint of the next. This structure captures the coarticulatory effects that occur during natural speech, making diphones more context-sensitive than individual phonemes. Unlike phonemes, which can be heavily influenced by surrounding sounds, diphones provide a more consistent representation of speech transitions, making them valuable for applications like speech synthesis and recognition

By passing diphones together with decoding candidates to the LLM (such as GPT-3.5 or LLama), the predictive performance can be greatly enhanced, as the model can leverage the rich contextual information embedded in diphones, capturing transitions between phonemes and improving the understanding of speech patterns. This approach enables the LLM to consider both fine-grained phonetic details and broader linguistic context, resulting in more accurate predictions and better alignment with natural speech dynamics.

If we add diphones to our decoding candidates, some changes to the decoder architecture would likely be necessary, but not necessarily a complete redesign. Currently, our decoder is built around CTC, which assumes that the output vocabulary consists of single phonemes and learns alignments between frames and individual phonemes. If we switch to diphones, the output space grows significantly — from 39 phoneme classes to potentially 1500 diphone classes (39×39). This brings a few challenges: CTC decoding would still technically work but the increased label space would make training slower and less data-efficient. We might need to modify the output layer by increasing the softmax size from 39 classes to 1500 classes and retraining or fine-tuning the decoder carefully to avoid overfitting, since some diphones would be much rarer than others

6.2 in-context learning

In-Context Learning (ICL) is a kind of method that enables LLM to adapt to new tasks by conditioning on prompts containing task-specific examples without modifying training parameters. Specifically For score adjustment, ICL can integrate contextual examples of inputs, outputs, and scoring criteria directly into the prompt, allowing the LLM to infer scoring logic dynamically, so as to better improve the performance by LLM score adjustment.

6.3 Real-Time Silent Speech Processing

For silent speech decoding to be truly practical, it must function in real-time, allowing users to communicate without delays. Most current models operate in batch mode, where full sequences of neural activity are processed before any predictions are made. This results in high latency, making it unsuitable for real-world applications such as assistive communication, hands-free control, or interactive systems. Real-time processing requires the ability to analyze continuous neural input and generate predictions incrementally, allowing for a more natural and responsive experience.

Achieving real-time silent speech processing involves optimizing how neural signals are processed and interpreted. Instead of requiring complete input sequences, the system should be able to handle data in shorter, overlapping time windows, making immediate adjustments as new information arrives. This shift from static sequence processing to continuous streaming would allow for more fluid and natural communication. Additionally, refining computational efficiency ensures that decoding keeps pace with user input, preventing delays in speech generation. Converting to an online prefix beam search maintains accuracy while shaving off hundreds of milliseconds per sentence, making the interface genuinely interactive for users.

The ability to process silent speech in real-time would significantly expand its potential applications, making it a more viable alternative for individuals with speech impairments or those seeking non-verbal communication solutions. Future improvements should focus on increasing both speed and accuracy while maintaining the flexibility to handle variations in speech patterns. By integrating more efficient processing methods, silent speech decoding could move beyond experimental research and into practical, everyday use cases, enabling real-time, silent communication with minimal friction.

6.4 Domain Adaptation for Cross-Subject Generalization

One of the major challenges in silent speech decoding is the high variability in neural activity across different users, making it difficult to develop a universal model that generalizes well across individuals. Current models trained on one subject often exhibit poor performance when tested on another due to differences in brain activity, signal quality, and individual speech patterns. To improve cross-subject generalization, we propose incorporating domain adaptation techniques that allow models to adapt dynamically to new users while maintaining high decoding accuracy. These techniques will reduce the dependency on subject-specific calibration, making silent speech decoding more scalable and practical for broader applications.

Our approach includes Adversarial Domain Adaptation (ADA), a technique that uses a domain discriminator to minimize feature distribution differences across subjects. By training a model to align neural signal representations between different users, we can ensure that the decoder learns universal silent speech features that remain consistent across individuals. Additionally, Few-Shot Transfer Learning will be implemented, allowing models to be fine-tuned on small calibration datasets for new users. This enables the system to quickly adapt to individual neural signatures without requiring extensive retraining.

To further enhance adaptability, we propose Model-Agnostic Meta-Learning (MAML), a meta-learning approach that trains a highly flexible model capable of quickly adjusting to new users with minimal data. By pre-training the decoder on a diverse set of subjects, MAML optimizes it for rapid adaptation, significantly improving cross-subject transferability. Through the combination of ADA, few-shot learning, and meta-learning, we aim to develop a silent speech decoding model that is highly robust, adaptable, and generalizable, enabling real-world deployment without the need for extensive individual calibration.

When it comes to domain adaptation, we do anticipate that some model adjustments would be necessary, but not necessarily major architectural overhauls. In particular, if we move to new participants or different recording setups, retraining or fine-tuning the model, especially the earlier layers that process raw features — would likely be needed to account for participant-specific signal characteristics. Additionally, the output layer might need adaptation if the label space changes (e.g., different vocabularies, phoneme sets, or speaking styles). However, the overall encoder-decoder framework could likely stay the same. We also think techniques like domain adversarial training, feature normalization, and embedding space alignment could help the model generalize better across domains without requiring a complete redesign. So in short, some focused adjustments would be expected, but the core architecture would likely remain largely intact.

7 Future Work

To focus our ongoing efforts, we highlight three research directions that we believe will deliver the greatest practical benefit in the near term.

7.1 Ultra-Low-Latency Decoding for Real-Time Use

For users who rely on silent-speech interfaces as their sole means of communication, every additional millisecond between intent and output erodes conversational flow. Although our current pipeline achieves a median 120 ms sentence-level delay, single-word or phoneme-level streaming is still out of reach. Future work should therefore (i) replace full-context transformers with chunk-wise or monotonic attention to permit left-to-right streaming, (ii) prune and quantise the encoder-decoder to fit on edge-class SoCs so that computation can happen on-device, and (iii) exploit the event-driven nature of spike-band features on neuromorphic hardware, where latency is governed by spike arrival rather than frame batching. Establishing rigorous user-study thresholds—e.g., <75 ms end-to-end for short utterances—will keep engineering targets aligned with communicative comfort.

7.2 Data-Efficient Personalisation at the Edge

Cross-subject results (§ 6.4) show that three minutes of adaptation data can halve the character-error rate, yet collecting even this amount may be burdensome for patients in acute care. We therefore propose federated or on-device meta-learning strategies that initialise models with priors learned from a worldwide cohort, then update only light-weight adapters with a handful of user-specific epochs. Combining this with unsupervised neural activity clustering during daily use could produce a model that continually refines itself without exposing raw data, preserving privacy while keeping calibration friction low.

7.3 Multimodal Sensor Fusion for Robustness

Insertion and substitution errors often stem from articulatory ambiguities that single-modality neural signals cannot resolve. Integrating complementary sensors—such as low-bandwidth ultrasound, intra-oral pressure pads, or inertial tongue trackers—into a late-fusion transformer can supply orthogonal cues without invasive recordings. Attention-based gating would allow the system to fall back gracefully when one modality degrades (e.g., due to electrode drift), maintaining decoding accuracy in long-term deployments.

8 Division of work

Yibo Yin: responsible for reproducing the MONA-LISA model , identifying and implementing solutions for model extension

Jingxuan Zhang: Working on reproducing Silent Speech model and the MONA-LISA model.

Xinchen Lin: Working on the Attention mechanism for Motor Imagery task and writing the Motor-Imagery part of the report, implementing and reproducing the CIBR-Okubo-Lab’s GRU+CTC neural decoder with LM and LLM.

References

- [1] H. Zhu, D. Forenzo and B. He, "On the Deep Learning Models for EEG-Based Brain-Computer Interface Using Motor Imagery," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2283-2291, 2022, doi: 10.1109/TNSRE.2022.3198041.
- [2] Altaheri, H., Muhammad, G., Alsulaiman, M., Amin, S. U., Altuwaijri, G. A., Abdul, W., Bencherif, M. A., & Faisal, M. (2023). Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: A review. *Neural Computing and Applications*, 35, 14681–14722. <https://doi.org/10.1007/s00521-021-06654-1>

- [3] Lawhern, V., Solon, A., Waytowich, N., Gordon, S., Hung, C., & Lance, B. (2016). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15. <https://doi.org/10.1088/1741-2552/aace8c>.
- [4] Benster, T. et al. (2024) A cross-modal approach to Silent Speech with LLM-Enhanced Recognition, arXiv.org. Available at: <https://arxiv.org/abs/2403.05583>
- [5] Craik, A., He, Y., & Contreras-Vidal, J. (2019). Deep Learning for Electroencephalogram (EEG) Classification Tasks: A Review. *Journal of Neural Engineering*, 16(3). Available at: <https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5>
- [6] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J Neural Eng*. 2018 Jun;15(3):031005. doi: 10.1088/1741-2552/aab2f2. Epub 2018 Feb 28. PMID: 29488902.
- [7] Liu, X., Shen, Y., Liu, J., Yang, J., Xiong, P., & Lin, F. (2020). Parallel Spatial–Temporal Self-Attention CNN-Based Motor Imagery Classification for BCI. *Frontiers in Neuroscience*, 14, 587520. <https://doi.org/10.3389/fnins.2020.587520>
- [8] Benster, T., Wilson, G., Elisha, R., Willett, F. R., & Druckmann, S. (n.d.). A Cross-Modal Approach to Silent Speech with LLM-Enhanced Recognition. A cross-modal approach to Silent Speech with LLM-Enhanced Recognition. <https://arxiv.org/html/2403.05583v1>
- [9] Jingyuan Li Department of Electrical and Computer Engineering University of Washington jingyli6@uw.edu &Trung Le Department of Electrical and Computer Engineering University of Washington tle45@uw.edu &Chaofei Fan Department of Computer Science Stanfor, & Authors contributed equally to this work. (n.d.-a). Brain-to-text decoding with context-aware neural representations and large language models. <https://arxiv.org/html/2411.10657v1>
- [10] Willett, F.R., Kunz, E.M., Fan, C. et al. A high-performance speech neuroprosthesis. *Nature* 620, 1031–1036 (2023). <https://doi.org/10.1038/s41586-023-06377-x>