A Penny for Your Thoughts: Decoding Speech from Inexpensive Brain Signals

Quentin Auster Chuang Ma Kateryna Shapovalenko Demaio Sun
Carnegie Mellon University
{qja, chuangm, kshapova, cleons}@andrew.cmu.edu
(Ordered alphabetically by last name)

Abstract

We investigate whether neural networks can approximate a decoding function converting brain signals in the form of EEG recordings into speech (brain-to-speech decoding). Given EEG data recorded while a subject listened to audio, we train our model using a contrastive CLIP loss that takes in the embeddings generated by our models from passing through the EEG data and embeddings from the audio passed through a pre-trained transformer-based English speech model. We contribute three proposed alterations to the current state of the art architecture, two of which improved performance in our experiments: (i) adding an attention mechanism to the subject layer (0.29% improvement relative to baseline), (ii) personalizing the spatial attention score for each subject (1.28% improvement relative to baseline), and (iii) using a dual path RNN in combination with attention layers (6.13% reduction in performance relative to baseline). Our results are promising for applications in brain-computer interaction, such as speech-impaired accessibility.

1 Introduction

Human eardrums react to small and rapid changes in air pressure. This simple reaction is the basis for the richness of sound that humans experience. Despite the simplicity, it is able not only to capture sounds, but to filter and distinguish between many sounds occurring simultaneously based on frequency tuning in the cochlea [7]. That is, from a single sound wave reaching the eardrum, humans are able to separate information from multiple sources and focus on those of interest. Speech cues are represented at the subcortical level in the brain, and these representations are shaped by short- and long-term auditory experiences [5].

However, information from the auditory brainstem is distributed across multiple cortex regions. These regions are tuned to extract specific features of speech—lower level sound representations are converted to higher level speech representations, which can be used for speech perception [5]. Rich acoustic feature representations stored in core auditory regions are sent to non-core auditory regions in the lateral parts of the superior temporal gyrus for mapping onto phoneme representations [5].

Given the relationship between auditory experiences and neural representations of speech, we hypothesize that the process of decoding auditory language from brain signals can be approximated using deep neural networks. Specifically, we aim to predict sequences of English words a person has just heard from sequences of brain activity recorded by Electroencephalography (EEG). The current state of the art for speech decoding uses Magnetoencephalography (MEG) data, but EEG is much cheaper to record than MEG. Therefore, the ability to accurately decode speech from EEG would be significant. In particular, drawing connections between brain activity and speech could help to improve accessibility for speech-impaired individuals, as well as improve brain-computer interfaces.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

In this paper, we work with EEG data that was recorded while subjects listened to a chapter from Alice In Wonderland [2]. We propose model architecture alterations to the work from Défossez et al. (Meta) [4]. In particular, we expand their approach of creating subject-specific layers that can account for variation between subjects. Our model uses subject-specific spatial attention, subject-specific attention, convolutional layers, and recurrent layers. In experimentation on a subset of the subjects used in the original paper, our alterations show promising results relative to Meta's baseline in terms of word error rate (WER).

2 Literature Review

The field of speech decoding from brain activity is a rapidly evolving area of research, characterized by its challenges and the variety of approaches employed. We can gain insights into this field by focusing on two key aspects: (i) the type of brain signals used, and (ii) the types of models for speech decoding employed.

Types of brain signal recordings Research in this field predominantly uses either non-invasive methods like Electroencephalography (EEG), Magnetoencephalography (MEG), and Functional Magnetic Resonance Imaging (fMRI), or invasive methods such as Electrocorticography (ECoG). Non-invasive methods, while more accessible and less risky, often provide less rich signal data compared to invasive methods. Invasive methods, despite their higher risk and complexity, tend to yield more accurate results in speech decoding due to richer data quality. In our research, we concentrate on non-invasive methods (specifically EEG), each with unique characteristics. EEG offers excellent temporal resolution but falls short in spatial resolution, limiting its effectiveness in detailed spatial analysis of brain activity. In contrast, MEG provides high precision in both temporal and spatial aspects, generally leading to better decoding results. For instance, Défossez et al. [4] observed significantly improved performance using MEG data over EEG. fMRI, with its superior spatial resolution but limited temporal resolution, struggles in capturing the precise timing of events. The scanner noise in fMRI can also interfere with auditory responses. However, its high spatial resolution has been instrumental in achieving notable accuracy in speech decoding. This underscores the importance of spatial resolution in decoding speech perception, a factor that should be carefully considered when working with EEG signals.

Models for brain signal decoding The decoding of speech from brain activity has seen various approaches using combinations of convolutional, recurrent, and sequence-to-sequence architectures. Défossez et al. (2023) [4] utilized a transformative approach, employing a pre-trained transformer-based speech model (`wav2vec2` [1]) to process audio recordings. They combined this with a convolutional neural network featuring a subject-specific layer for extracting representations from MEG and EEG recordings. The use of a contrastive loss, CLIP Loss [8], allowed them to train a zero-shot decoding classifier effectively. Tang et al. (2023) [9] introduced an innovative brain signal decoder designed to reconstruct continuous language from fMRI across various tasks. Their model included a beam search decoder, generating candidate sequences of words, alongside a GPT-based language model, demonstrating the potential of integrating advanced language processing models in speech decoding. Zhang et al. (2018) [10] explored a combination of convolutional and recurrent neural networks to decode brain activities from motion imagery EEG recordings. They also incorporated an autoencoder layer to filter out background activity, highlighting the importance of noise reduction in enhancing decoding accuracy.

Table 1: Brain Signal Types and Models of Speech Decoding

Brain Signals	Characteristics of Brain Signals	Models for Speech Decoding
EEG	• Non-invasive, measures electrical activity generated by neurons in the brain. • High temporal, low spatial resolution. • Noise: (1) Environmental electrical noise such as power lines and electronic devices; (2) Physiological noise such as blinking and heartbeat; (3) Electrode contact issues.	• Défossez et al. (2023) [4]: Transformer-based speech model (`wav2vec2`) combined with a CNN for data processing. CLIP Loss [8] for training a zero-shot decoding classifier. Achieved 25.75% vocabulary-specific accuracy. • Zhang et al. (2018) [10]: Hybrid CNN-RNN (LSTM) network for feature learning from raw EEG signal and autoencoder layer for feature adaptation. XGBoost classifier for intent recognition with a classification accuracy of 95.53%.
MEG	• Non-invasive, measures magnetic fields produced by neuronal electrical activity. • High temporal and high spatial resolution. • Noise: (1) Environmental magnetic noise (needs highly shielded room); (2) Patient movement can affect data quality.	• Défossez et al. (2023) [4]: Employed the same approach as with EEG, achieving a higher vocabulary-specific accuracy of 66.69%.
fMRI	• Non-invasive, measures changes in blood flow and oxygenation associated with neural activity. • Low temporal, high spatial resolution. • Noise: (1) Physiological noise; (2) Scanner noise (loud during operation); (3) Head motion.	• Tang et al. (2023) [9]: Beam search decoder with a GPT-based language model for continuous language reconstruction from fMRI data. • Caucheteux et al. (2022) [3]: Utilized a GPT-2 based language model.

Drawing inspiration from successful models in MEG and fMRI research, particularly those utilizing advanced neural networks and language models, we can enhance EEG signal decoding. Adapting techniques such as the CNN-RNN hybrid model or integrating language models could be particularly beneficial.

3 Materials and Baseline Model

3.1 Data

We chose to work with EEG data as it is non-invasive and inexpensive to record compared to other brain signal recording methods such as MEG or fMRI. We rely on data from Brennan and Hale (2019) [2], which is also used by Meta - our baseline reference. This dataset contains EEG data collected using 62 sensors from 33 subjects, totaling approximately 6.7 hours of recordings.

EEG data Brennan and Hale recorded EEG data while participants listened to spoken prose from Chapter One of Alice in Wonderland. EEG was recorded for 49 participants. However, the recordings of 16 participants were not used due to noise in the recording of the data, poor results on a test of listening comprehension after recording, or both, leaving recordings from 33 subjects [2]. EEG was recorded from 62 sensors, or channels, including "VEOG" (Vertical Electrooculogram) and "AUD" (Auditory) channels. VEOG is used to measure vertical eye movements, which can create artifacts in EEG data [6]. The EEG data was recorded with sampling rate of 500 Hz (in the data, time stamps increment by 0.002 seconds) in BrainVision Core Data format (.vhdr, .vmark, and .eeg files provide the metadata, events collected, and EEG data with additional signals, respectively).

Audio data Audio data recording the reading out loud of Chapter One of Alice in Wonderland is provided in 12 wav files (segments). The paper's authors also provide a table with each word spoken in the chapter, its starting and ending time in its corresponding segment.

Pre-processing We follow the Meta's approach for EEG and audio pre-processing. For EEG data, we begin by loading the raw data, then apply baseline correction for signal stability and robust scaling for variance consistency. We further refine the data by clipping outliers below the 5th percentile and above the 95th percentile, and clamping values exceeding 20 standard deviations. Both EEG and audio data undergo standard normalization. We segment these data into three-second windows to focus on word-level. While initially experimenting with segment-level data to capture broader neural patterns, we faced challenges in model convergence, indicating the importance of segment length in EEG-audio studies. We leave further experimentation with segment length as an area for future work.

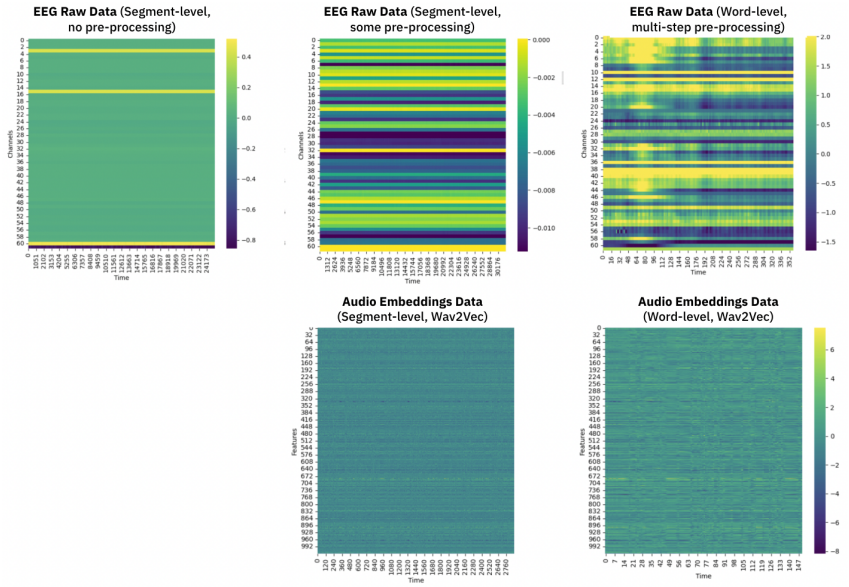

Figure 1: Pre-processing of EEG and Audio Data

3.2 Evaluation Metrics

Evaluating model performance presents certain challenges, primarily due to the scarcity of open datasets and limited availability of reproducible code in existing studies. To address these challenges, we have established specific criteria for our evaluation process.

As our model targets brain-computer interfaces where users think of words to control digital platforms, precise word-for-word accuracy isn't always necessary (i.e., capturing the intended meaning or command is often sufficient). Therefore, we introduce a dual metric approach: traditional metrics such as Word Error Rate General (WER General) and Levenshtein Distance for exact accuracy, and Word Error Rate Vocab (WER Vocab) for flexibility, ensuring predicted words fall within the target vocabulary:

- **Levenshtein Distance**: This metric indicates the average number of single-character edits (insertions, deletions, substitutions) required to change predicted sentences into target sentences, normalized by the number of words in the target sentences. For example, a value of 2.25 suggests that, on average, approximately 2 to 3 edits are needed per word to correct the predictions.

- **WER General**: This is calculated as the simple proportion of correctly identified words, where both position and order are crucial. For example, a WER General of 50% indicates that 50% of the words in the predictions were incorrectly identified compared to the target words, taking into account the exact sequence in which they appear. Additionally, we calculate *Accuracy General* as $100 - \textit{WER General}$, reflecting the percentage of correctly predicted words in their correct order.

- **WER Vocab**: This calculates the proportion of words in the predictions that are present in the target vocabulary, regardless of their position or order. For example, a WER Vocab of 40% indicates that 40% of the words in the predictions were not found in the target vocabulary. This metric is more lenient, focusing on the presence of predicted words within the overall pool of words used in the targets, without considering their specific sequence or placement. *Accuracy Vocab* is calculated as $100 - \textit{WER Vocab}$, representing the percentage of predicted words found in the target vocabulary, irrespective of sequence or placement.

Also, we are using train and validation loss to monitor model performance.

3.3 Baseline Model Performance

For our baseline, we refer to the study by Meta [4]. This paper serves as a foundational reference for several reasons: (i) it is the most recent paper in the field and builds upon all previous research, providing a comprehensive analysis of current advancements; (ii) the study demonstrates promising results, particularly in identifying speech segments from 3-second magneto-encephalography signals with up to 41% accuracy across participants, and up to 80% accuracy in the best cases (this level of performance is notable as it enables the decoding of words and phrases not present in the training set); (iii) the paper supports four different public datasets, offering a unique opportunity to compare model performances across various brain signals (EEG vs MEG), as both the source code and the dataset are publicly available, allowing for comparative studies.

As an initial step in our research, we have attempted to replicate the Meta's results with a few modifications to adapt the approach to our specific dataset and research objectives.

Table 2: Baseline Model Performance

Evaluation Criteria	Meta's Model [4]	Replicated Model	
Sample size (brennan2019)[2]	33 subjects	10 subjects	20 subjects
Train loss	N/A	5.46	5.22
Validation loss	N/A	5.42	5.32
WER General	N/A	98.10%	95.97%
WER Vocab	74.25%	71.81%	69.49%
Accuracy Vocab	25.75%	28.19%	30.51%

As demonstrated in the table, our replication of the Meta's model yielded comparable outcomes. Notably, WER Vocab in the replicated model is marginally lower than that of the original Meta's study. Slight improvement could be attributed to variations in individual EEG patterns among subjects. By focusing on a smaller group of subjects, or potentially those with clearer, less noisy EEG signals, we observed enhanced performance.

With these findings, we have effectively established a baseline for our future experiments, marked by a WER Vocab of 69.49%, or conversely, an Accuracy Vocab of 30.51%.

4 Model and Contributions

4.1 Problem Formulation and Training

Problem formulation Letting $X \in \mathbb{R}^{T \times C}$ be a segment of EEG recording with C channels and T time steps and $Y \in \mathbb{R}^{T \times F}$ be the representation of an audio recording (for instance, in the form of a Mel spectrogram or a hidden representation generated from a pre-trained speech model) with F channels and T time steps which are aligned with the EEG signal, we can represent the decoding of brain signal to speech as a function, $\mathbf{f} : \mathbb{R}^{C \times T} \rightarrow \mathbb{R}^{F \times T}$. That is, once the model $\mathbf{f}$ maps the EEG recording to the same space and time-alignment as the audio representation, the two states can be compared.

Contrastive loss While a regression loss such as mean-squared error could be used to train $\mathbf{f}$, Defossez et al. point out that this may not be appropriate [4]. Instead, they suggest the use of a contrastive loss, specifically CLIP Loss [8], which looks to find discriminative combinations of features. For a given segment of brain recording X, the authors select a sample $\bar{Y}_j, j \in \{1, ..., N\}$ of N examples where $\bar{Y}_1$ through $\bar{Y}_{N-1}$ are "negative" samples of audio recordings from the dataset and $\bar{Y}_N$ is the "positive" sample, i.e., the audio recording Y corresponding to X. The model $\mathbf{f}_{\text{clip}}$ then predicts probabilities $p_j = \mathbb{P}[\bar{Y}_j = Y], \forall j \in \{1, ..., N\}$ by mapping the brain activity X to the same space as the Mel spectrogram, where these representations can be compared by taking the softmax of the dot product of the $Z = \mathbf{f}_{\text{clip}}(X) \in \mathbb{R}^{F \times T}$:

$$\hat{p}_j = \frac{\exp(\langle Z, \bar{Y}_j \rangle)}{\sum_{k=1}^{N} \exp(\langle Z, \bar{Y}_k \rangle)}, \tag{1}$$

where $\langle ., . \rangle$ is the inner product over the dimensions of Z and $\bar{Y}$. From here, the function $\mathbf{f}_{\text{clip}}$ can be trained using the cross-entropy loss between p_j and $\hat{p}_j$.

Inference For inference and validation, we follow Meta in passing the hidden states generated from our model and from the pre-trained model through the final layer(s) of the pre-trained model.

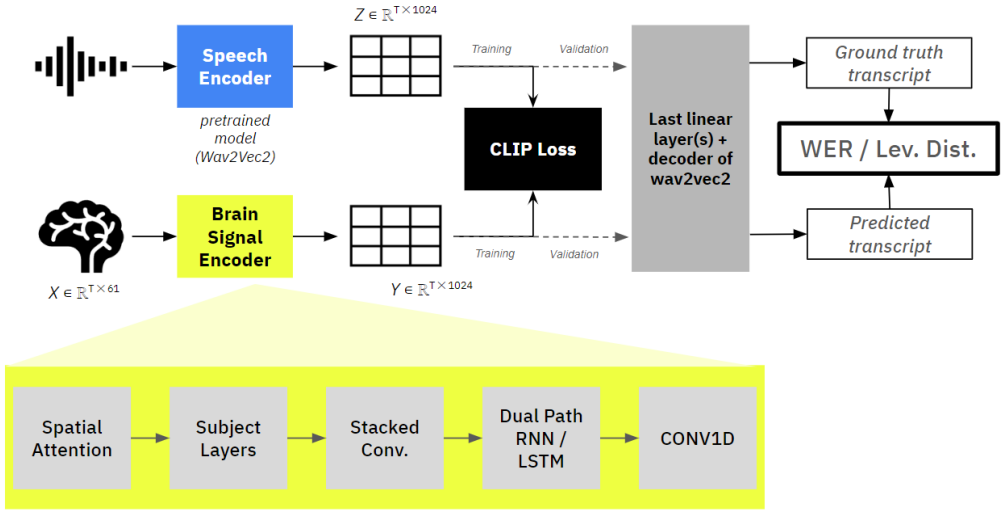

Figure 2: Problem Formulation and Model Architecture

4.2 Ablations

In order to build intuition about how model design choices affect performance, we ran three primary categories of ablations using Meta's existing codebase and varying the use of (1) Pre-processing (signal clamping), (2) Spatial Attention, (3) Subject Layers, and (4) Dual Path RNN. The detailed results are discussed in the Results section.

4.3 Model Architecture

After studying the specifics of EEG data and speech perception, and conducting a series of experiments, we gained insights into what might work effectively for this task. Consequently, we made modifications and enhancements to Meta's model architecture and utilized their existing codebase for training and evaluation. Specifically, we concentrated on the following components: (1) Adaptive Spatial Attention, (2) Subject-Specific Attention, (3) Stacked Convolutions, (4) Dual Path RNN with Attention, and (5) Final Convolutions.

Spatial attention Our model enhances Meta's approach by introducing a subject-specific attention mechanism for processing brain data. Like Meta, we utilize the MNE package function `mne.channels.find_layout` to project 3-dimensional sensor locations into 2 dimensions and normalize between $[0, 1]$. For each output channel, a learned function, parameterized in Fourier space, assigns values in the range $[0, 1]^2$, and spatial dropout is applied to mitigate overfitting. Unlike the original model, which applied uniform spatial attention across all subjects, our model incorporates a unique attention mechanism for each individual. This is achieved through the implementation of our `SubjectAttentionLayers` module. This module assigns specific attention weights to each subject, leveraging the `SubjectAttention` class to calculate these weights. This modification allows our model to better accommodate inter-individual variability in brain signal patterns, providing a more personalized analysis of neural data. By doing so, we aim to address the limitations observed in Meta's model, where attention weights did not consistently align with regions typically activated during auditory stimulation, as noted in their Extended Data Figure 5 [4]. Our approach, therefore, enhances the spatial attention mechanism's sensitivity to individual differences in brain activity.

Subject layers with attention Our model advances Meta's design by incorporating a specialized subject layer after the spatial attention layer for each participant. Crucially, we integrate an attention mechanism tailored to individual subjects using our `SubjectAttentionLayers` module. This module, leveraging both `SubjectLayers` and `SubjectAttention` classes, enables the model to focus on and learn unique brain signal patterns for each subject. This enhancement significantly boosts the model's ability to account for and adapt to the unique characteristics of individual neural data, offering a more personalized and accurate analysis.

Stacked convolutions In our model, we employ a series of stacked convolutional blocks, as conceptualized in the `ConvSequence` class. This approach closely follows Meta's architecture by using convolutional blocks, each comprised of two convolutional layers with residual connections, as implemented in our `ResidualBlock` class. These layers extend the output to $D_2 = 320$ channels, incorporating batch normalization to ensure model stability. The dilation of convolutions, achieved through the `ConvSequence` class follows the pattern $2^{2k mod 5}$ and $2^{2k+1 mod 5}$ for each convolutional layer, enhancing the receptive field of the network. The convolution sequence follows by a Gated Linear Unit (GLU) activation, effectively reducing the channel dimensions by half. This design, mirroring Meta's established framework, ensures the effective handling of complex neural data without any modification to the core convolutional architecture.

Dual path RNN with attention In our model, we enhance the Meta's Dual Path RNN. We utilize the `DualPathRNN_attention` class to implement a dual-direction LSTM structure, which allows for more nuanced neural signal processing in both forward and backward directions. This is crucial for capturing the intricacies of long sequence patterns in brain signals. Additionally, we incorporate an attention mechanism within each LSTM layer, as instantiated in the `SelfAttention` class. This attention mechanism is designed to focus on specific signal features, thereby strengthening the model's capability to identify relevant patterns across extended sequences. This combined approach of bidirectional LSTM layers with integrated attention mechanisms marks a substantial improvement over the traditional Dual Path RNN, enabling our model to process complex neural data with heightened accuracy and specificity.

Final convolutions In alignment with Meta's framework, our model employs a final convolutional stage consisting of two 1×1 convolutions. These convolutions transform the output to $2D_2$ channels and then to F channels to match the speech representation's dimensionality.

5 Results

In this section, we present the results obtained from our extensive ablation studies and experiments involving modifications to the model architecture, and we compare these results with the baseline (Meta's model). While we conducted a broad range of ablations, we report here only the most interesting ones. These specific ablations were instrumental in guiding our choice of the model architecture. To implement these selected approaches, we developed new classes within our modeling framework. The detailed outcomes of these experiments are summarized in the figure and the table below. Further analysis and interpretation of these results will be provided in the discussion section.

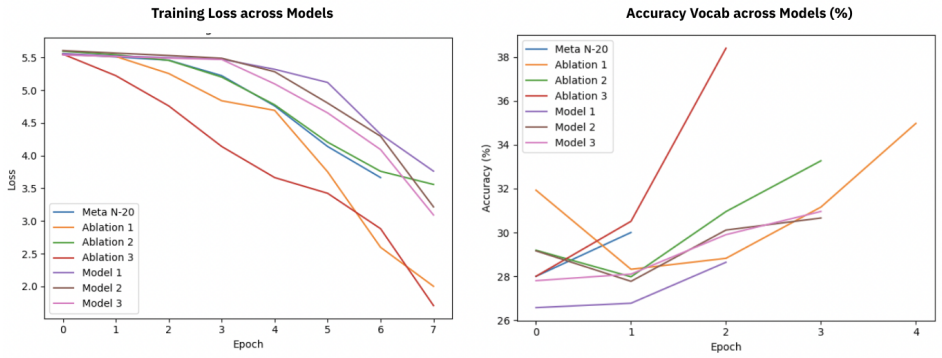

Figure 3: Training Loss and Validation Accuracy Across Models

Table 3: Results of Ablations and Model Architecture Changes

Models	Pre-processing (clamp value)	Spatial Attention	Subject Layers	Dual Path RNN	Accuracy Vocab
Meta N=33	20	General + dropout	Subject layer	LSTM	25.75%
Meta N=20	20	General + dropout	Subject layer	LSTM	30.51%
Ablation 1	100	default	default	default	34.98%
Ablation 2	100	w/o dropout	Subject layer + embedding	default	33.27%
Ablation 3	default	default	Subject layer + embedding	default	38.41%
New model 1	default	default	default	BLSTM + attention	28.64%
New model 2	default	default	Subject layer + attention	default	30.66%
New model 3	default	Subject-specific attention	default	default	30.96%

6 Discussion

Through extensive ablation studies and architectural modifications, we identified several critical elements that significantly influenced the effectiveness of our model. Notably, the three top improvements were the use of a larger clamp value, the integration of subject-specific layers, and the extensive application of attention mechanisms, along with optimized initial and final convolution processes. On the other hand, the absence of channel merger dropout emerged as a primary factor that negatively impacted performance. Alongside these findings, we also explore the limitations of our current study

and suggest potential avenues for future research to further enhance the model's capabilities and applicability.

Pre-processing (signal clamping) Our findings indicate a significant improvement in model performance with a clamp value of 100 compared to 20. A higher clamp value appears to enable the model to capture more complex patterns in the EEG data, possibly including finer details and nuances that a lower clamp value might miss. This capability is particularly beneficial in handling EEG data's inherent variability, including artifacts or unusual patterns.

Spatial attention The implementation of spatial attention in the model facilitates a focused analysis of pertinent channels (sensors), thereby enabling the selection of the most significant ones for subsequent processing and prediction tasks. In the initial design of the model, a uniform score matrix for channel evaluation is applied across all subjects. This aspect presents an opportunity for refinement and enhancement, suggesting that individualized adaptations of the score matrix for each subject could potentially yield more accurate and tailored results.

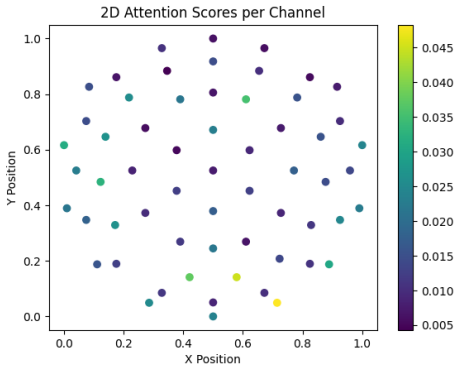

Figure 4: Spatial Attention Scores

Subject layers Integrating subject-specific layers contributed to addressing the individual differences in brain anatomy and function. These layers afforded the model a degree of personalization, enabling it to adapt to and capitalize on each subject's unique brain dynamics. Consequently, this reduced inter-subject variability, which is a common challenge in EEG data analysis. Subject layers also functioned as learnable embeddings, refining the representation of each individual's EEG data.

Dual path RNN We enhanced the original DualPathRNN with dual-direction LSTM layers and an added attention mechanism after each LSTM layer. This modification strengthened the model's ability to discern patterns in long sequences. The bidirectional LSTM layers capture time-related dependencies more effectively, while the attention mechanism focuses on relevant sequence parts, thus improving the model's accuracy in speech decoding from EEG data.

7 Future Work

While our results are good relative to the current baseline and state-of-the-art from Meta, we believe there is room for future work.

Expanding the subject pool for enhanced generalizability A primary limitation of our study was the constrained subject pool, utilizing only 20 of the 33 subjects used in Meta's research. This smaller dataset may not be entirely representative and could potentially increase overfitting risks. To improve the robustness and generalizability of our model, future work should focus on including a larger and more diverse subject pool. This expansion would allow for more comprehensive validation of our findings, ensuring that our model performs consistently across a wider range of individual brain patterns.

Exploring more diverse pre-processing techniques Our results indicate a potential for optimizing EEG signal preprocessing. Future research could experiment with varying the EEG segment duration, such as extending it to 5-, 10-, or 15-second intervals. This change might offer more contextual information per training example, possibly leading to improved model performance. Additionally, while we have not employed Independent Component Analysis (ICA) for noise reduction, its integration might be a valuable preprocessing step. Considering the trade-offs, incorporating ICA could be beneficial despite its potential to complicate spatial attention, especially given EEG's lower spatial resolution compared to MEG.

Incorporating language model for contextual interpretation Recognizing the complexity of speech perception, where not every word is fully processed and each word might have multiple meanings, integrating a language model could significantly enhance our system's predictive capabilities. Employing context-sensitive models, possibly based on transformer architectures, could provide a deeper interpretation of intended meanings. Such models could use surrounding neural patterns to fill in gaps in brain activity data or infer the context and meaning of partially detected words.

Maintaining a strong commitment to ethical guidelines Future endeavors in this domain should be dedicated to ensuring the privacy and security of brain data while maintaining cognitive liberty. It is crucial to focus on obtaining informed consent, responsibly using data, and engaging in continuous dialogue with ethicists and legal experts.

8 Conclusion

The ability to decode speech from brain signals, in particular from EEG recordings, is a difficult task—to this point, the state of the art has only reached approximately 26% accuracy. However, this task holds promise for improving brain-computer interfaces, especially if advances are made in the decoding of EEG, which is both non-invasive and cheap to record relative to other brain signal data types.

In this study, we performed ablations and experiments using the state of the art model from Meta as a backbone in order to understand the types of hyper-parameters and architectures might lead to improved performance. First, we found that results were sensitive to pre-processing steps such as signal clamping. Second we found that the inclusion of attention within the learnable subject-specific layers of the model helped capture and adapt to individualized patters across subjects and improved model performance over the baseline. Finally, we found that the use of spatial attention helped to mitigate the effects of poor EEG spatial resolution and improved performance relative to the baseline.

References

[1] Alexei Baevski et al. "wav2vec2.0:AFrameworkforSelf-Supervised LearningofSpeechRepresentations". In: *NeurIPS* (2020).

[2] J.R. Brennan and J.T. Hale. "Hierarchical structure guides rapid linguistic predictions during naturalistic listening". In: *PLoS ONE* 14.1 (2019). DOI: https://doi.org/10.7302/746w-g237.

[3] Alexandre Gramfort Jean-Rémi King Charlotte Caucheteux. "Deep language algorithms predict semantic comprehension from brain activity". In: *Nature* (2022).

[4] Alexandre Défossez et al. "Decoding speech perception from non-invasive brain recordings". In: *Nature Machine Intelligence* 5 (2023), pp. 1097–1107. DOI: https://doi.org/10.1038/s42256-023-00714-5.

[5] Lori L. Holt · Jonathan E. Peelle Allison B. Coffin · Arthur N. Popper Richard R. Fay. *Speech Perception*. Springer, 2022.

[6] Xiao Jiang, Gui-Bin Bian, and Zean Tian. "Removal of Artifacts from EEG Signals: A Review". In: *Sensors (Basel)* (2019).

[7] Andrew J. Oxenham. "How We Hear: The Perception and Neural Coding of Sound". In: *Annual Review of Psychology* 69 (2018), pp. 27–50. DOI: https://doi.org/10.1146/annurev-psych-122216011635.

[8] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *PMLR* 139 (2021).

[9] Jerry Tang et al. “Semantic reconstruction of continuous language from non-invasive brain recordings”. In: *Nature Neuroscience* 26 (2023), pp. 858–866.

[10] Xiang Zhang et al. “Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals”. In: *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 2018, pp. 1–10. DOI: 10.1109/PERCOM.2018.8444575.

Division of Work

In alphabetical order by last name:

- Quentin Auster: Co-led the project, conducted a literature review, created a custom data loader for EEG and audio, deployed Wav2Vec for audio embeddings, set up elements of the training pipeline, and drafted sections of the mid-term, presentation, and final report.
- Chuang Ma: Ran ablations using the Meta codebase, proposed and implemented new extensions to the original Meta models, created GitHub Repository, visualized model performance during training, and drafted sections of the final report.
- Kateryna Shapovalenko: Conceptualized and co-led the project, conducted a literature review, implemented a baseline using the Meta codebase, performed EDA, set up elements of the training pipeline, ran ablations, and drafted sections of the mid-term, presentation, and final report.
- Demaio Sun: Ran ablations using the Meta codebase, proposed and implemented new extensions to the original Meta models, and drafted sections of the presentation and final report.

Code

Our code is available here.