MEGFormer: enhancing speech decoding from brain activity through extended semantic representations

Maria Boyko[1,2][0000−0001−5839−1014], Polina Druzhinina[1,2,3][0000−0002−8458−8539], Georgii Kormakov[1][0000−0002−7728−0392], Aleksandra Beliaeva[4][0009−0005−4627−3666], and Maxim Sharaev[1,2][0000−0002−5670−2891]

[1] Center for Applied AI, Skolkovo Institute of Science and Technology, Moscow, Russian Federation Maria.Boyko@skoltech.ru
[2] BIMAI-Lab, Biomedically Informed Artificial Intelligence Laboratory, University of Sharjah, Sharjah, United Arab Emirates
[3] Artificial Intelligence Research Institute (AIRI)
[4] Lomonosov Moscow State University, Moscow, Russian Federation

Abstract. Even though multiple studies have examined the decoding of speech from brain activity through non-invasive technologies in recent years, the task still presents a challenge as decoding quality is still insufficient for practical applications. An effective solution could help in the advancement of brain-computer interfaces (BCIs), potentially enabling communication restoration for individuals experiencing speech impairments. At the same time, these studies can provide fundamental insights into how the brain processes speech and sound. One of the approaches for decoding perceived speech involves using a self-supervised model that has been trained using contrastive learning. This model matches segments of the same length from magnetoencephalography (MEG) to audio in a zero-shot way. We improve the method for decoding perceived speech by incorporating a new architecture based on CNN transformer. As a result of proposed modifications, the accuracy of perceived speech decoding increases significantly from the current 69% to 83% and from 67% to 70% on publicly available datasets. Notably, the greatest improvement in accuracy is observed in longer speech fragments that carry semantic meaning, rather than in shorter fragments with sounds and phonemes. Our code is available at https://github.com/maryjis/MEGformer/

Keywords: Decoding speech · Contrastive Learning · Brain-computer interface · CNNtransformer · MEG

1 Introduction

Brain-Computer Interfaces (BCIs) are increasingly being looked upon as promising avenue to identify and potentially restore lost abilities in individuals affected by neurological conditions. Among multitude of applications, a significant portion is dedicated to exploring and enhancing the potential for improving and

restoring communication abilities.[3] [4] [1] Recent advancements in invasive neural interfaces have showcased the capacity to decode speech at remarkable speeds of 62 and 78 words per minute approaching the speed of natural conversation. [14] [8] Invasive neural interfaces enhance BMI performance but come with technical challenges like electronic limitations, signal quality variations over time, the current impracticality for home use, and the need for surgical intervention in the human body, with unexplored health consequences. [6]

Here we focus on the application of non-invasive MEG-BCI for investigating speech perception in healthy individuals aiming to analyze this process utilizing high-quality signals across varied sized groups. This exploration may shed light on the nature and structure of language representations while offering insight into the neural mechanisms underlying speech production and comprehension for further research.

To the best of our knowledge, the only method [4] for speech decoding from MEG data involves training a unified architecture on a diverse participant cohort, rather than individual patients, and utilizing deep speech representations rather than a limited set of interpretable features like sounds, phonemes, words, and etc. [1] In [4], the model receives a segment of MEG recording as input and aligns it with the respective audio segment using contrasting learning with CLIP and a CNN encoder for brain activity decoding. However, the CNN encoder's limitation lies in its failure to consider extended interactions among brain activity patterns, which could be beneficial for creating complex semantic representations. Thus, replacing the CNN encoder block with a CNN transformer block has the potential to enhance speech decoding performance.

In this paper, we introduce the first transformer-based approach for speech decoding from non-invasive magnetoencephalography data, **MEGFormer**. The model consists of two primary components: a brain module uses CNN transformer encoder block for processing MEG data, and an audio encoder based on a pre-trained wav2vec 2.0 model for generating audio representations. Our analysis incorporates two open MEG datasets of different sizes, comprising recordings of brain activity during passive listening to short stories, along with the corresponding audio files. With contrastive learning techniques, the model is able to predict brain activity patterns based on the audio segments. The proposed method refines existing techniques, offering a superior and more adaptable solution. Our contribution is threefold:

- We introduce a new architecture MEGFormer that shows superior performance compared with recent state-of-the-art methods on open datasets [4]
- To the best of our knowledge, we are the first to propose a transformer-based architecture to encode brain signals for speech decoding task
- We demonstrate how the model's performance improves with longer speech segments, highlighting its ability to accurately decode complex speech representations linked to high-level perception or even speech comprehension as compared to simple sounds or phonemes.

2 Methods

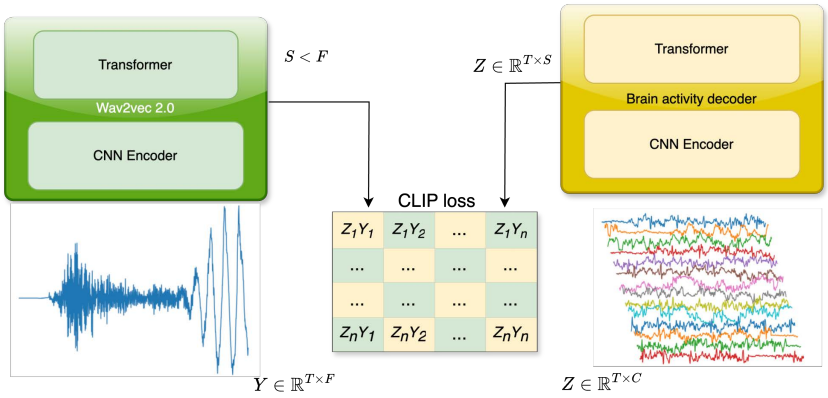

Fig. 1. Illustration of the proposed MEGFormer architecture for decoding speech based on MEG brain activity. The proposed model is trained to map corresponding audio representations extracted from wav2vec model **(left side)** to brain representations received from CNN Transformer **(right side)** with CLIP loss.

2.1 Problem Formulation

Our goal is to decode perceived speech from brain signals - multivariate time series MEG data recorded from many scalp channels. We divide all brain recordings and audio signals into segments of equal length. Each segment is characterized by a multivariate time-series with dimensions $T \times C$, where T is the number of time points and C is the number of channels. The latent representation of speech is a vector with $T \times F$ shape, where F is the number of features. Hence, our objective is to find the best decoding function D, that matches brain activities with audio representations:

$$D : \mathbb{R}^{T \times C} \Rightarrow \mathbb{R}^{T \times F} \tag{1}$$

2.2 Model Architecture

We introduce MEGFormer architecture (Fig. 1), which is inspired by preceding state-of-the-art work by Defossez [4] and allows the modulation of both local and long interactions, which may be crucial for improving the quality of brain activity representations. Furthermore, enhancing alignment between the brain activity encoder and the wav2vec architecture will ensure that the generated brain representations closely match audio representations. Therefore, we suggest

incorporating a new transformer block into the Brain encoder module to improve the model's capabilities in extracting more precise features and encouraging cross-modal alignment.

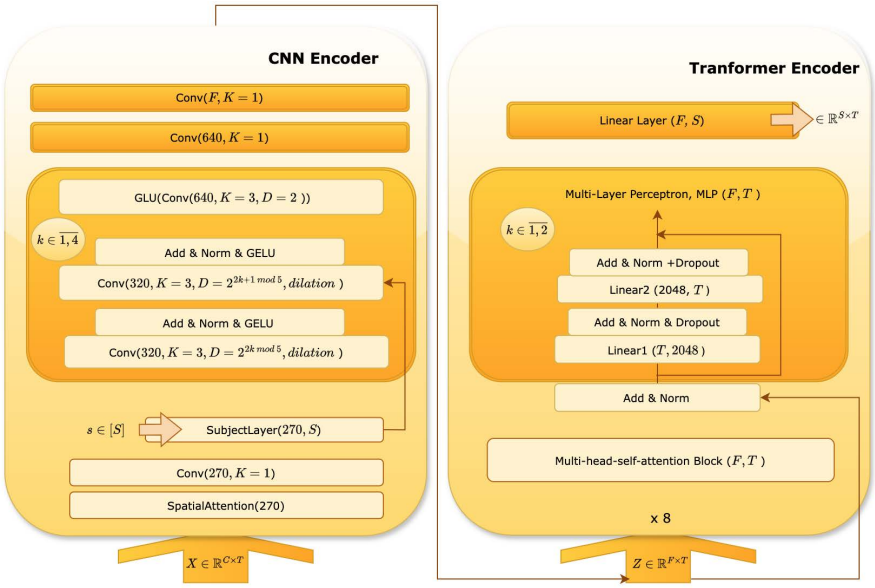

Fig. 2. Illustration of the proposed Brain activity encoder, composed of a CNN encoder followed by a transformer encoder. The CNN encoder consists of 4 convolutional blocks with residual connections, batch normalization, and GLU activation. The transformer block comprises MHSA and MLP blocks. S is the subject index which passes to Subject Layer. The CNN encoder block takes MEG data as input X with the size CxT. Here T represents the number of time points, while C refers to the number of channels. It then returns a latent representation with a size of SxT, where S is the output feature size that is equivalent to the feature size obtained from wav2vec.

Audio encoder As an audio encoder, we take wav2vec 2.0 model (specifically, wav2vec2-large-xlsr-53 version). Wav2vec 2.0 model is trained with convolution and transformer blocks to learn latent audio representations [2]. In [9] the authors have shown the efficacy of mapping such speech representations with brain activities.

Brain activity encoder Brain activity encoder obtains MEG signals from all sensors and utilizes a spatial attention layer to map them onto 270 channels. Additionally, following the methodology presented in [4], a Subject Layer is introduced to facilitate the alignment of representations to the particular subject. The received representations are passed through both CNN encoder layers and transformer encoder layers (see Fig. 2). The choice of these layers is intentional, as CNN encoder layers excel at capturing local contextual information in brain

activities, while transformer layers focus on long-range interactions. The CNN layers are implemented with four convolutional blocks, each containing three convolutional layers, following the described approach [4]. The first two convolutions within each block include residual skip connections and progressively increasing dilation rates. Each of them is followed by applying a BatchNorm layer and a GELU activation function. The third convolution within the block excluded both a residual connection and normalization, instead utilizing a GLU activation, reducing channel count by half. Finally, two 1x1 convolutions are performed, separated by a GELU activation function.

The transformer block comprises a Multi-head-self-attention block (MHSA) and Multi-Layer Perceptron (MLP) [13]. The embeddings obtained from the CNN encoder are treated as tokens for each timestamp. Each token, firstly, passed through the multi-head-self-attention block formulated as:

$$SA_i = softmax(\frac{Q \times K^T}{\sqrt{d_k}})V \tag{2}$$

$$MSA_i = Concat(SA_1, SA_2, .., SA_n)W_0 \tag{3}$$

Q, K, V is a query, key, and value matrices obtained by multiplication embedding on corresponding projection matrix: W_q, W_k, W_v.W_0 is projection all concatenated heads to original embedding shape d_{model}. We employ 8 attention heads with a model dimension d_{model} equals 264. Position embedding preceding the transformer block is omitted, as it is integrated within the spatial attention layer. The MLP layer with batch normalization and non-linearity followed by the MHSA with an inner dimension of 2048.

Contrastive Loss. We employ the 'CLIP' loss [11], well-regarded for its capacity to align latent representations across different modalities, in the process of mapping brain activities to sound representations. Using CLIP loss, brain recordings X are mapped to their sound representations Y, with the model f_{clip} predicting match probabilities. This is computed by the dot product of X's latent representation Z with Y, followed by softmax normalization: $\hat{p}_j = \frac{e^{\langle Z, \hat{Y}_j \rangle}}{\sum_{j'=1}^{N} e^{\langle Z, \hat{Y}_{j'} \rangle}}$. The 'CLIP' loss is refining the model's precision in predicting the correct sound representation from brain activity, thus increasing both accuracy and robustness formulated as:

$$L_{\text{CLIP}}(p, \hat{p}) = -\sum_{j=1}^{N} p_j \log(\hat{p}_j), \tag{4}$$

3 Experiments

Datasets and preprocessing We evaluate our methods on two publicly available MEG datasets: Gwilliams[5] and Schoffelen[12]. Gwilliams dataset comprises raw magnetoencephalography (MEG) recordings from 27 English speakers who underwent a two-hour session of listening to speech narratives. Each healthy

volunteer was recorded over two sessions using a 208 axial-gradiometer MEG scanner at a sampling rate of 1000 Hz while listening to four stories from the Manually Annotated Sub-Corpus (MASC). The dataset further provides information about the temporal alignment of audio signals and brain activity collected via MEG, linked to corresponding phonemes and words at each time step. From the Schoffelen dataset, we used data obtained from 96 Dutch-speaking individuals engaged in listening to audio speech. MEG data were captured with a 275-channel axial gradiometer system (CTF) operating at a sampling frequency of 1200 Hz. MEG data from each dataset was resampled to 120 Hz and filtered within a range from 0.1 to 40 Hz. The data was standardized, and values exceeding 20 standard deviations were restricted to minimize the influence of significant outlier samples. A segment is defined as a distinct N-second brain recording paired with its corresponding audio representation. All segments acquired from the participants were divided into training, validation, and testing sets, with 70%, 20%, and 10% of the data allocated to each split, respectively. Given that a segment can be associated with multiple participants, we designated each segment exclusively to one split across multiple iterations. Thus, we ensure the absence of identical segments among the splits.

We also propose to segment audio signals into distinct non-overlapping sound blocks with the same window and stride values instead of fragmenting them into sentence blocks with a 3-second window and a 0.3-second stride, as commonly seen in related research [4]. This segmentation strategy enables the creation of segments that could start with a word from one sentence and end with words from another sentence. Consequently, a segment may not correspond to a single sentence within our context. This strategy increases the number of segments by 50%. Furthermore, this modification facilitates dataset adjustment, resulting in a more balanced distribution among the training, validation, and testing sets with different segment lengths.

Experimental Setup The experiments were carried out using PyTorch and were trained for 20 epochs on RTX 6000 GPUs. The Adam optimizer was applied with a learning rate of 3e-4, implementing early stopping based on the best validation loss. The assessment of the model's ability to identify corresponding audio segments was based on the evaluation of top-10 accuracy and top-1 accuracy metrics, which means that the target segment was present among 10 and 1 predicted ones.

4 Results

The proposed architecture MEGFormer outperforms the current state state-of-the-art method ([4], CNN approach) achieved top-10 accuracy 83.99 % compared to 69.66 % in Gwilliams dataset and 70.49 % compared to 67.89 % in Schoffelen dataset. (Table 1)

The preprocessing steps result in a minor improvement in quality and also help to decrease the model overfitting. The introduced approach of segmenting sound blocks based solely on sound characteristics, disregarding sentence bound-

Table 1. Performance comparison of the proposed method with other state-of-the-art methods on Gwilliams and Schoffelen dataset at 3s speech segments.

	Gwilliams		Schoffelen	
	top-10 acc	top-1 acc	top-10 acc	top-1 acc
[4] (base)	69.66	40.12	67.89	37.40
+prepossessing	70.38	41.13	68.66	37.74
+splitting segments via sound	77.05	47.81	-	-
[4] (base) (depth =6, dropout =0.2)	77.78	49.02	68.85	38.04
MEGFormer (ours)	**83.99**	**55.08**	**70.49**	**39.43**
Vanilla transformer	64.47	33.27	53.75	31.64
TimesNet	72.81	43.37	46.32	29.94

aries, led to an expansion in the sample size, thereby enhancing the model's quality. In the case of the Gwilliams dataset, the sample size increased from 203152 to 314922 segments for training, yielding a top-10 accuracy of 77.05. However, implementing this approach for the Schoffelen dataset posed challenges due to the dataset's configurations and the initial annotating encompassing audio-text mapping based on sentence boundaries. Due to the smaller size of neuroimaging datasets compared to other fields, improving data through augmentation methods (such as segment splitting adjustments) and simplifying the base architecture have been successful in improving the overall performance of the base model [4]. This enhancement has resulted in an improvement in top-10 accuracy from 69.66% to 77.78 % and from 67.89 % to 68.85 % on Gwilliams and Schoffelen datasets respectively.

Additionally, we experimented with the basic Vanilla transformer variation and the current leading approach in time series analysis, TimesNet. However, both models failed to outperform CNN and transformer CNN approaches in the speech decoding task. This observation indicates that to effectively decode brain activity from MEG data, it is advantageous to utilize a strategy that combines convolutional and transformer blocks. This approach allows for addressing both local context and long-range interactions.

Previously [4], the authors opted to use only segments that were 3 seconds in duration. We conducted a study where we explored the relationship between the length of brain and audio segments and their impact on speech decoding quality. Segments of 3s, 4s, 7s were selected, and our results demonstrate that a longer segment length leads to improved decoding quality (Fig. 3). We suppose that the enhancements are associated with the properties of MEG data. A single electrode in MEG can receive signals from a vast number of neurons spanning a wide area of the brain, whereas discrete small regions in the brain may be responsible for interpreting particular words or phrases. As a result, the creation of longer segments representing advanced speech concepts requires the synchronization of numerous brain regions, allowing for a more accurate capture of these distinctions in the MEG signal.

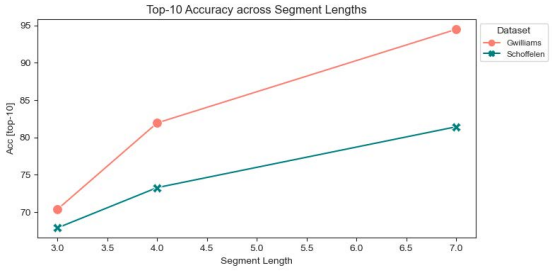

Fig. 3. The performance of the model significantly increases with longer segment lengths. On the X-axis are segment lengths, and on the Y-axis is the model quality depending on segment length.

Closeness of phonemes audio to MEG signal To explain the improvement of accuracy with the segment's length increasing, we compared the shape of the original audio of phonemes and received a MEG response. The Dynamic Time Warping (DTW) algorithm was used to compare the raw series in the time domain. [10]. Because the value of DTW depends on the length of audio and MEG signal, a normalized version (nDTW) was used (See Eq. 1 in supplementary materials). The intuition behind this is "higher nDTW - closer the MEG sensor data to audio". In Fig. 1 (see supplementary materials), we measured nDTW for a random subset of subjects in the Gwilliams dataset on sensors positioned over the auditory cortex, using MEG segments of three different lengths: 0.8s, 3s, and 4s. As a result, the value of nDTW becomes greater with the increasing length of the segment that can cause a boost in performance for the proposed architecture. Also, we used Short Time Fourier Transform (STFT) to compare the spectral characteristics of various sensors. Spectral characteristics play an important role in understanding the frequency band of received signal[15]. We estimated the STFT for the same sensors as for nDTW. Then, The distance (mean square error) between each pair of sensors was measured and averaged on all phonemes (see Fig. 3 in supplementary materials). From this, we can make two conclusions: the total difference between spectral characteristics of sensors is decreasing by order, but the pattern (ratio) between each pair is preserved. Thus, the following interpretation emerges: with an increase in the length of the MEG segment, the different parts of the brain could be interpreted for the model as the cluster of audio receptive field.

5 Discussion & Conclusion

Decoding speech directly from brain signals is a relatively new and emerging task. In this study, we propose a new architecture MEGFormer, transformer-based model for decoding perceived speech which improves previous state-of-the-art model performance on two open datasets. Furthermore, we demonstrate that the model's performance dramatically improves with longer speech segments.

With the development of larger datasets for speech production, it could be adapted for this task without significant modifications. Thus, our work represents a critical advancement toward building a foundational model for brain recordings. Recent studies have utilized limited vocabularies, whereas our model demonstrates a zero-shot performance with an unrestricted one. This capability is evidenced by the model's ability to generalize to new words that do not overlap with the training set, thus eliminating the need for additional training.

In practical applications, such as aiding individuals with speech impairments, our model can be adapted to use a limited vocabulary set, significantly enhancing performance and making it suitable for tasks like issuing predefined commands. Our approach also holds potential for understanding differences between healthy controls and patients with Auditory processing disorder which can lead to the development of brain interfaces for them [7]. We believe that MEGFormer could be used for developing more accurate and reliable BCI systems which is the goal of our future research.

Acknowledgments. This study was funded by Internal Skoltech grant for BIMAI-Lab, Biomedically Informed Artificial Intelligence Laboratory (Skoltech - University of Sharjah joint laboratory).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anumanchipalli, G.K., Chartier, J., Chang, E.F.: Speech synthesis from neural decoding of spoken sentences. Nature **568**(7753), 493–498 (2019)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
3. Chaudhary, U., Mrachacz-Kersting, N., Birbaumer, N.: Neuropsychological and neurophysiological aspects of brain-computer-interface (bci) control in paralysis. The Journal of physiology **599**(9), 2351–2359 (2021)
4. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.R.: Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence **5**(10), 1097–1107 (2023)
5. Gwilliams, L., Flick, G., Marantz, A., Pylkkänen, L., Poeppel, D., King, J.R.: Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. Scientific Data **10**(1), 862 (2023)
6. Haci, D., Liu, Y., Ghoreishizadeh, S.S., Constandinou, T.G.: Key considerations for power management in active implantable medical devices. In: 2020 IEEE 11th Latin American Symposium on Circuits & Systems (LASCAS). pp. 1–4. IEEE (2020)
7. Lartseva, A., Dijkstra, T., Buitelaar, J.K.: Emotional language processing in autism spectrum disorders: a systematic review. Frontiers in human neuroscience **8**, 991 (2015)
8. Metzger, S.L., Littlejohn, K.T., Silva, A.B., Moses, D.A., Seaton, M.P., Wang, R., Dougherty, M.E., Liu, J.R., Wu, P., Berger, M.A., et al.: A high-performance

neuroprosthesis for speech decoding and avatar control. Nature **620**(7976), 1037–1046 (2023)
9. Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., King, J.R., et al.: Toward a realistic model of speech processing in the brain with self-supervised learning. Advances in Neural Information Processing Systems **35**, 33428–33443 (2022)
10. Molina, M., Tardón, L.J., Barbancho, A.M., De-Torres, I., Barbancho, I.: Enhanced average for event-related potential analysis using dynamic time warping. Biomedical Signal Processing and Control **87**, 105531 (2024)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
12. Schoffelen, J.M., Oostenveld, R., Lam, N.H., Uddén, J., Hultén, A., Hagoort, P.: A 204-subject multimodal neuroimaging dataset to study language processing. Scientific data **6**(1), 17 (2019)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
14. Willett, F.R., Kunz, E.M., Fan, C., Avansino, D.T., Wilson, G.H., Choi, E.Y., Kamdar, F., Glasser, M.F., Hochberg, L.R., Druckmann, S., et al.: A high-performance speech neuroprosthesis. Nature **620**(7976), 1031–1036 (2023)
15. Yang, Y., Tarr, M.J., Kass, R.E.: Estimating learning effects: A short-time fourier transform regression model for meg source localization. In: International Workshop on Machine Learning and Interpretation in Neuroimaging. pp. 69–82. Springer (2013)