



# Towards Sentence Level Imagined Speech Generation from EEG signals

Sparsh Rastogi<sup>1</sup>, Harsh Dadwal<sup>1</sup>, Khushboo Modi<sup>1</sup>, Jatin Bedi<sup>1</sup>, Jasmeet Singh<sup>1</sup>

<sup>1</sup>Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, India

srastogi.be22@thapar.edu, hdadwal.be22@thapar.edu, kmodi.be22@thapar.edu,  
jatin.bedi@thapar.edu, jasmeet.singh@thapar.edu

## Abstract

Brain-Computer Interfaces (BCIs) have emerged as alternative means of communication for individuals with speech & motor impairments. These systems enable patients to express themselves without any articulation, by decoding speech from neural activity. However, most of the existing studies rely on invasive surgical procedures, with limited studies using non-invasive signals for phoneme or word level classification, thus covering a short vocabulary. To the best of our knowledge, this study presents the first demonstration of a framework for sentence level imagined speech synthesis from non-invasive electroencephalography (EEG) signals. Our model uses an Efficient-Net based masked auto-encoder approach for learning feature embeddings from EEG signals which are then used for fine-tuning BERT for next token generation. For this study, Large Spanish Speech EEG Dataset has been used with a mixed subject approach for both training & evaluation purposes, resulting into a 48.92% accuracy.

**Index Terms:** neural speech synthesis, brain computer interface, electroencephalography(EEG) signals, imagined speech, masked auto-encoder

## 1. Introduction

Speech is a crucial mode of communication, enabling the transfer of complex ideas and emotions through coordinated sound patterns produced by the synchronized activity of the brain and vocal tract muscles, like the larynx, tongue, and lips [1]. However, conditions such as Amyotrophic Lateral Sclerosis (ALS), locked-in syndrome etc. impair muscular movement which is necessary for speech articulation. Restoring expressive capacity in population suffering from these conditions is a critical challenge. In recent years, Brain Computer Interfaces (BCIs) have emerged as an alternative means of communication for people suffering from these diseases. BCI systems analyze the neural activity of the person in order to decode their thoughts.

Several studies have been proposed to synthesize speech from intracranial electrocorticography (ECoG) signals [2, 3, 4, 5, 6]. The initial studies on this topic relied on conventional architectures like 3D Convolutional Neural Networks (CNNs) [2] and bidirectional Long Short Term Memory (Bi-LSTMs) [3] with some more recent works also employing advanced architectures like transformers for speech spectrogram reconstruction using an encoder-decoder framework [5, 7]. However, acquisition of ECoG signals is a highly invasive procedure requiring surgical intervention, in order to implant a chip in the cortex for signal acquisition, thus hindering the large scale adaptability of these methods. Some other studies proposed methodologies for imagined speech synthesis using minimally invasive signals like stereotatic EEG (sEEG) [6, 8, 9] but the acquisi-

tion of sEEG also requires the implantation of electrodes in the brain in order to acquire signals, thus reducing its feasibility for practical applications.

On the other hand, non-invasive modalities such as electroencephalography (EEG) or magnetoencephalography (MEG), offer safer alternatives for speech based BCI systems without requiring any kind of surgical intervention. However, these signals have a low signal to noise ratio due to the presence of inherent noise in the form of various artifacts like muscle movement, eye blink, cardiac activity *etc.*, thus making information extraction from them a difficult task. However, with the recent advancements in the field of signal processing and development of various sophisticated techniques for noise filtering and signal enhancement [10, 11], the usage of EEG and MEG signals for speech decoding and synthesis has become more feasible.

Recent studies have proposed the use of non invasive MEG signals for speech decoding as well as reconstruction using approaches like CNNs [12] and covariance filters [13] for mel spectrogram reconstruction. A 2024 study proposed, a Squeeze-former based architecture for spoken speech synthesis [14]. Despite being a non-invasive procedure, acquisition of MEG signals requires extensive setup and isolation from external magnetic interference, due to which it is performed in a magnetically shielded room [15], making it an expensive procedure. On the other hand, EEG offers a more portable and economical alternative for neural speech synthesis in real world settings. Although some studies have tried to utilize EEG signals for imagined speech synthesis, most of them focused mainly on phonetic [16], vowel [17, 18] or word level classification [19] covering a limited set of vocabulary, thus unable to capture the rich and diverse semantics of any language. Furthermore, most of these approaches employed a mere classification based approach, rendering them ineffective for prediction/generation of words outside the training corpora, limiting its usage in a real world setting. In this study, we propose a novel deep learning framework for sentence level imagined speech synthesis from non-invasive EEG signals. Our model takes raw EEG signal as input, processes it to extract meaningful features and utilizes them for imagined speech generation. Furthermore, our model is generalizable across different subjects and is robust to inter-subject variability. To the best of our knowledge and based on the extensive exploration of literature, this is the first demonstration of a framework for sentence level imagined speech synthesis from non-invasive EEG signals. Our model is made up of three major components - EEG Feature Encoder, Semantic Masked Auto-Encoder (SMA) and Neural Speech Synthesizer (NSS). EEG Feature Encoder extracts high-level feature embeddings from raw EEG signals which are then fed to SMA which uses a masked representation learning based reconstruction approach

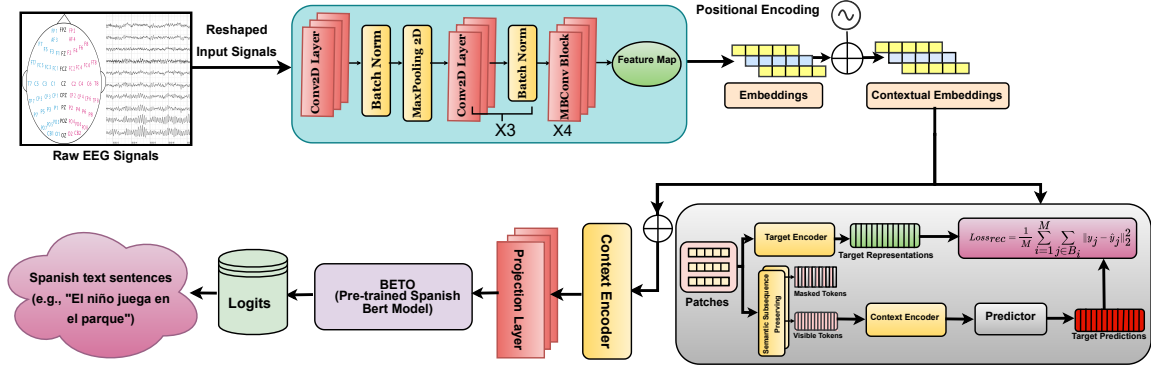


Figure 1: An overview of the architecture of the proposed framework

to understand the underlying representation of the data in order to capture the context and semantics of the language. The representations learned using this approach are further shared with the NSS which uses them to generate imagined speech, formulating the problem as a machine translation task. The major contributions of this work could be summarized as follows:

- We propose a model for sentence level imagined speech synthesis from EEG signals. To the best of our knowledge, this is the first demonstration of a framework for sentence level imagined speech synthesis from non-invasive EEG signals.
- We have employed a generation based approach in order to capture the rich semantics and large vocabulary of the language, as opposed to the conventional classification based approach used in the literature.
- The proposed model exhibits robustness to inter-subject variability and generalizes well to new subjects as well as unseen words and sentences.

## 2. Proposed Method

### 2.1. Model Architecture

The overall architecture of the proposed model has been illustrated in Fig. 1. The model consists of three major components - EEG Feature Encoder, SMA and NSS. In the training phase, the features are extracted from the input EEG data by the EEG Feature Encoder and are passed to the SMA to learn the context representations of the underlying data using masked representation learning. These learned representations are further fed to NSS to provide it with better context while generating speech and improve its overall performance. However, at the time of inference only EEG Feature Encoder and NSS are used, while the SMA is skipped using the residual connection as depicted in Fig. 1. Each of the individual components along with their detailed working have been discussed in the following subsections.

#### 2.1.1. EEG Feature Encoder

This module takes a raw EEG signal represented as multivariate time series data as input  $X \in \mathbb{R}^{L \times C}$ , where L denotes the sequence length of input time series data and C represents the number of EEG channels. This module utilizes an EfficientNet [20] based architecture comprising of simple convolution blocks followed by Mobile Inverted Bottleneck Convolution (MBCConv) blocks with squeeze and excite mechanism for

feature extraction. The simple convolution blocks made up of 2D Convolution, Batch Normalization and Max Pooling layers extract the spatial features across various channels. These features are further refined by the MBCConv block, which leverages depthwise convolution layers to capture temporal dependencies, expands the feature space via an inverted bottleneck for efficient filtering, and enhances discriminative power through squeeze-and-excite mechanism, thereby optimizing feature extraction. The extracted embeddings are concatenated with positional encodings to preserve the sequential structure of the data and forwarded to the next module.

#### 2.1.2. Semantic Masked Auto-Encoder

SMA is based on [21], [22] and utilizes an approach of Masked Signal Modeling similar to Masked Image Modeling. A semantic subsequence preserving based approach has been employed for generating the masked tokens in order to preserve the semantic information in the data [23]. After applying masking on the contextual embeddings, visible tokens are fed to the context encoder which follows a transformer based architecture and tries to reconstruct the masked tokens using multi-head self attention [24]. Simultaneously, all the tokens, both masked and visible are also passed to the target encoder which share the same architecture as the context encoder and projects the tokens into another embedding space. Context encoder is followed by predictor network which is another transformer block and projects the reconstructed masked tokens into the same embedding space as the target encoder. The overall objective is to minimize the loss between the projections of the original tokens, and the reconstructed tokens. Instead of computing reconstruction loss, directly on the reconstructed tokens, a cross attention based approach has been utilized in order to avoid overfitting and representation collapse.

#### 2.1.3. Neural Speech Synthesizer

After the training is complete, NSS receives the final learned context representations from the context encoder of the SMA, which are utilized at the inference time to understand the context of the input EEG signals, for improved performance. At inference time, it directly receives the EEG feature embeddings from the EEG Feature Encoder using the residual connection. The overall problem has been framed as a Machine Translation task using EEG embeddings, where BETO [25], a BERT model pre-trained on a large corpus of Spanish text, has been fine-tuned for speech generation from input EEG embeddings

using a supervised learning approach [26].

## 2.2. Training Objectives

A two-stage training process has been employed to ensure effective representation learning and speech generation. The first stage focuses on self-supervised learning through Semantic Masked Autoencoder, which enables the model to develop a rich latent space by reconstructing masked portions of the input data. The reconstruction loss has been defined as a L2 loss (Eq. 1) calculated on the ground truth embeddings generated by the target encoder ( $y_j$ ) and the embeddings predicted by the predictor network using the tokens reconstructed by the context encoder ( $\hat{y}_j$ ) over  $N$  sets of  $M$  masked tokens.

$$\mathcal{L}_{\text{rec}^*} = \frac{1}{M \times N} \sum_{q=1}^N \sum_{i=1}^M \sum_{j \in B_i} \|y_j - \hat{y}_j\|^2 \quad (1)$$

A Variance-Invariance-Covariance regularization (VICReg) [27] has been added to regularize the reconstruction loss (Eq. 2) to avoid representational collapse of the network [23].

$$\mathcal{L}_{\text{reg}^*} = \lambda \text{Loss}_{\text{rec}^*} + \mu v(R) + \gamma c(R) \quad (2)$$

where  $v(R)$  and  $c(R)$  represent the variance and covariance losses on the representations, respectively, while  $\lambda$ ,  $\mu$ , and  $\gamma$  are hyperparameters that regulate the trade-off among these loss components.

The second stage leverages these learned representations to fine-tune the NSS, optimizing it for next token generation using a cross entropy based loss function (Eq. 3) calculated upon the one-hot encoded vector of ground truth ( $y_{\text{true},i}$ ) and the predicted probability distribution ( $y_{\text{pred},i}$ ) at position  $i$  over a vocabulary of size  $C$  and sequence of length  $L$ .

$$\mathcal{L}_{\text{gen}^*} = \frac{\sum_{i=1}^L \left( - \sum_{c=1}^C y_{\text{true},i}^{(c)} \log y_{\text{pred},i}^{(c)} \right) \cdot 1(y_{\text{true},i} \neq \text{pad\_id})}{\sum_{i=1}^L 1(y_{\text{true},i} \neq \text{pad\_id})} \quad (3)$$

Thus, the overall loss function could be expressed as Eq. 4. This approach ensures that the generated speech is contextually informed and coherent while benefiting from the structured knowledge acquired during pre-training.

$$\mathcal{L} = \min(\mathcal{L}_{\text{reg}^*} + \mathcal{L}_{\text{gen}^*}) \quad (4)$$

## 3. Experimental Setup

### 3.1. Dataset

For the purpose of this study, Large Spanish Speech EEG Dataset comprising high-quality EEG embeddings from 56 healthy participants (31 female and 25 male) all right-handed, with ages between 20-25 years has been used [28]. The participants were made to listen to 30 Spanish sentences and asked to imagine speaking the perceived sentence in their mind without any articulation. Corresponding EEG signals were captured using a SynsAmps RT 64-channel Amplifier EEG system at a sampling frequency of 1 kHz following the standard 10–20 system configuration. Each participant was made to perform 6-7 trials for each sentence, with each trial lasting 11 seconds (5s for perception, 1s for preparation and 5s for imagined speech production) preceded by a 5 second rest, resulting into an extensive collection of 312 EEG recordings per sentence.

### 3.2. Data Preprocessing

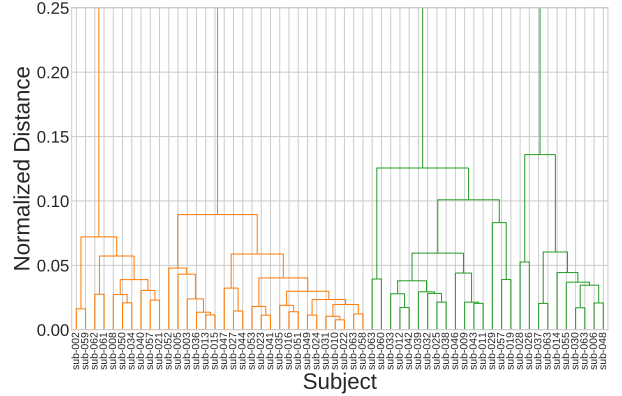


Figure 2: Dendrogram representation of clusters obtained using hierarchical cluster based on Dynamic Time Warping between EEG signals of different subjects

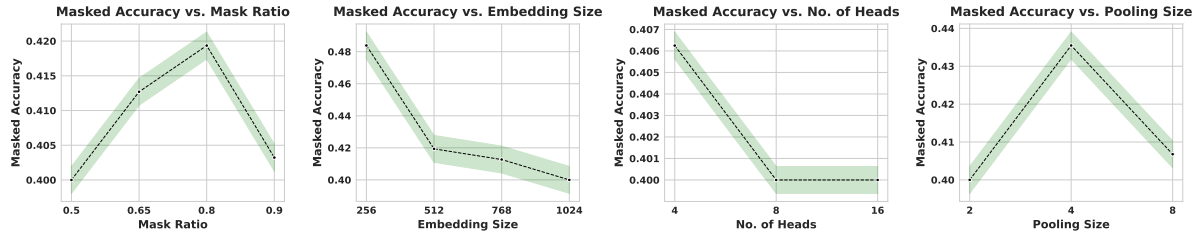
A data preprocessing approach similar to [29] was adopted for artifact removal and signal enhancement. Independent Component Analysis (ICA) was applied to remove artifacts such as eye blinks, channel noise, heartbeats, and muscle activity. To ensure a full-rank decomposition, 45 components were utilized. Prior to ICA computation, EEG signals were band-pass filtered between 1–100 Hz using a finite impulse response (FIR) filter with a Hamming window. The extended Infomax ICA algorithm was employed, and an automatic classification method identified artifact components with high confidence. The raw EEG signals were then band-pass filtered between 2–50 Hz, re-referenced using the average of 64 channels, and artifact components were removed via ICA-based projection. Finally, the cleaned signals were downsampled to 250 Hz to enhance computational efficiency for deep neural networks. Since, the acquired data had two varying trial lengths (10s for upto subject 18 and 11s for rest of the participants), a linear interpolation based approach was employed for data augmentation to extend all the trials to 11s (2750 data points) in order to ensure uniform trial lengths.

In order to construct a generalizable model, we tried to incorporate maximum possible diversity in the training set. To ensure this, a Dynamic Time Warping (DTW) approach was employed to calculate the similarity between the EEG signals of various participants, followed by a hierarchical clustering approach to group the subjects with similar EEG signals into the same cluster resulting into four clusters overall (Fig. 2). After that, the data from each cluster was split into training and validation set in the ratio 80:20. This approach ensures the robustness of the model to inter-subject variability, by capturing maximum possible variability in the data.

## 4. Results and Discussion

### 4.1. EEG Feature Encoding

Various architectures like CNNs, EEGNet [30], Temporal Multi-Channel Vision Transformer (TMC-ViT) [31] and EfficientNet were explored for the extraction of feature embeddings from the input EEG signals. The performance of these architectures were evaluated using cosine similarity computed between the original signal and the signal reconstructed from the feature embeddings extracted by these models. The results for the same

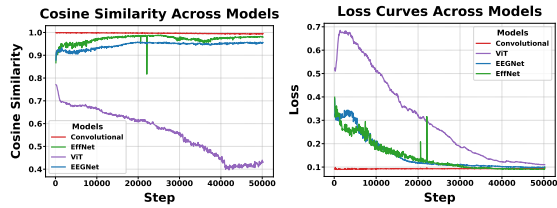


(a) Change in masked accuracy with masking ratio (b) Change in masked accuracy with embedding size (c) Change in masked accuracy with number of heads (d) Change in masked accuracy with pooling size

Figure 3: Assessment of model performance under different hyperparameter configurations

are depicted in Fig. 4. As evident from both cosine similarity (Fig. 4a) and loss (Fig. 4b), EfficientNet gives the best performance for encoding raw EEG signals into high level embeddings. EEGNet considered as the baseline model for most of the tasks using EEG, performs better than other models but lags behind EfficientNet. On the other hand, ViT converges earlier than the other models but its performance is slightly lower, whereas CNN falls into a representation collapse as evident from the results where its loss begins from close to zero and keeps rising and cosine similarity start from one and keeps reducing. On the basis of these results, EfficientNet has been adopted for EEG feature extraction in the final model.

#### 4.2. Learning Masked Representations



(a) Cosine Similarity (b) Loss

Figure 4: Ablation Results of Considered Feature Extractors

For the masked representation learning task, the performance evaluation has been done using cosine similarity as well as using masked accuracy which is defined as the ratio of the masked tokens correctly predicted by the fine-tuned BERT model to the total number of masked tokens. Since this module, comprises of multiple components and hyperparameters, it exhibits sensitivity to variations in them, with each hyperparameter influencing the model’s behavior in a distinct manner. Extensive ablation studies were conducted by varying the values of different hyperparameters, and the results obtained are illustrated in Fig. 3. During ablation studies, all hyperparameters except the one under investigation are set to their worst-performing values to isolate and assess the individual impact of the selected hyperparameter on overall model performance.

As observed from Fig. 3a, an increase in the masking ratio leads to improved performance up to a saturation point (0.8), beyond which a decline is noted, indicating that reconstructing more tokens facilitates learning richer contextual information. However, sufficient information must be provided for reconstruction, as the decline in accuracy beyond the saturation point indicates inadequate information for effective reconstruction. Similarly, for the dimensions of the embedding space (Fig. 3b) and the number of attention heads (Fig. 3c) in the transformer block, we observe that the performance reduces with the

increasing complexity which indicates the fact that as we go into higher dimensional spaces, the noise in the data also gets amplified leading to poor performance. Moreover, with the increase in the size of dimensional embeddings, the training time and computational cost also increases as inferred from Table 2. Similarly for the pooling size, we can observe from Fig. 3d and Table 1 that the performance initially rises with increase in pooling size and drops later which could be explained using the fact, that with a very small pooling size, the model captures the unnecessary and extra fine details of the input signals, leading to overfitting whereas with a too large pooling window, low-level details tend to get missed. After extensive experimentation, we determined that the model gives the most optimal performance for a mask ratio of 0.8 with a pooling size of 4, embedding size of 256 and 8 attention heads in the transformer block.

S. No.	Pooling Size	Training Masked Accuracy	Validation Masked Accuracy	Training Time (in hr)
1	2	40.30%	22.73%	5.27
2	4	<b>43.52%</b>	<b>25.27%</b>	3.92
3	8	40.81%	22.30%	<b>3.31</b>

Table 1: Performance of different pooling sizes on training and validation masked accuracy after 100 epochs, along with training time.

S. No.	Embedding Size	Masked Accuracy	Validation Masked Accuracy	Training Time (in hr)
1	256	<b>48.92%</b>	<b>27.27%</b>	<b>3.97</b>
2	512	42.2%	26.07%	4.37
3	768	41.34%	25.47%	4.83
4	1024	40.12%	22.73%	5.85

Table 2: Impact of different embedding sizes on masked accuracy and validation masked accuracy after 100 epochs, along with training time

## 5. Conclusion

This study proposes a generalizable framework for sentence level imagined speech synthesis from EEG signals. This is the first study on the topic, that cover a relatively larger corpora and is also able to construct words and sentences beyond those present in the training set. Using self-supervised representation learning and fine-tuning pretrained BERT model, we achieved significant results. Though our approach is currently limited to Spanish, however in our further studies we plan to extend it to other languages and multilingual translation. The methods proposed in this study will be insightful for future studies and accelerate the development of non-invasive speech BCI systems.

## 6. References

- [1] S. Saha, K. A. Mamun, K. Ahmed, R. Mostafa, G. R. Naik, S. Darvishi, A. H. Khandoker, and M. Baumert, "Progress in brain computer interface: Challenges and opportunities," *Frontiers in Systems Neuroscience*, vol. 15, 2021. [Online]. Available: <https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2021.578875>
- [2] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.
- [3] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.
- [4] J. Berezutskaya, Z. V. Freudenburg, M. J. Vansteensel, E. J. Aarnoutse, N. F. Ramsey, and M. A. J. van Gerven, "Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models," *J. Neural Eng.*, vol. 20, no. 5, Sep. 2023.
- [5] S. Komeiji, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, and T. Tanaka, "Feasibility of decoding covert speech in ecog with a transformer trained on overt speech," *Scientific Reports*, vol. 14, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:267548238>
- [6] J. Chen, X. Chen, R. Wang, C. Le, A. Khalilian-Gourtani, E. Jensen, P. Dugan, W. Doyle, O. Devinsky, D. Friedman *et al.*, "Transformer-based neural speech decoding from surface and depth electrode signals," *Journal of Neural Engineering*, 2025.
- [7] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker, "A neural speech decoding framework leveraging deep learning and speech synthesis," *Nat. Mach. Intell.*, vol. 6, no. 4, pp. 467–480, Apr. 2024.
- [8] M. Angrick, M. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff, "Towards closed-loop speech synthesis from stereotactic eeg: A unit selection approach," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1296–1300.
- [9] X. Wu, S. Wellington, Z. Fu, and D. Zhang, "Speech decoding from stereo-electroencephalography (seeg) signals using advanced deep learning methods," *Journal of Neural Engineering*, vol. 21, no. 3, p. 036055, jun 2024. [Online]. Available: <https://dx.doi.org/10.1088/1741-2552/ad593a>
- [10] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from eeg signals: a review," *Sensors*, vol. 19, no. 5, p. 987, 2019.
- [11] B. Kalita, N. Deb, and D. Das, "Aneeg: leveraging deep learning for effective artifact removal in eeg data," *Scientific Reports*, vol. 14, no. 1, p. 24234, 2024.
- [12] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (meg) signals," *Frontiers in neuroscience*, vol. 14, p. 290, 2020.
- [13] V. Verkhlyutov, V. Vvedensky, K. Gurtovoy, E. Burlakov, and O. Martynova, "Speech recognition from meg data using covariance filters," in *Biologically Inspired Cognitive Architectures Meeting*. Springer, 2023, pp. 904–911.
- [14] J. Kwon, D. Harwath, D. Dash, P. Ferrari, and J. Wang, "Direct speech synthesis from non-invasive, neuromagnetic signals," in *Proc. Interspeech 2024*, 2024, pp. 412–416.
- [15] A. L. Fred, S. N. Kumar, A. Kumar Haridhas, S. Ghosh, H. Purushothaman Bhuvana, W. K. J. Sim, V. Vimalan, F. A. S. Givo, V. Jousmäki, P. Padmanabhan *et al.*, "A brief introduction to magnetoencephalography (meg) and its clinical applications," *Brain sciences*, vol. 12, no. 6, p. 788, 2022.
- [16] Y. V. Varshney and A. Khan, "Imagined speech classification using six phonetically distributed words," *Frontiers in Signal Processing*, vol. 2, p. 760643, 2022.
- [17] M.-O. Tamm, Y. Muhammad, and N. Muhammad, "Classification of vowels from imagined speech with convolutional neural networks," *Computers*, vol. 9, no. 2, p. 46, 2020.
- [18] M. Li, S. H. Pun, and F. Chen, "A preliminary study of classifying spoken vowels with eeg signals," in *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2021, pp. 13–16.
- [19] M. Asghari Bejestani, G. R. Mohammad Khani, V. Nafisi, and F. Darakeh, "Eeg-based multiword imagined speech classification for persian words," *BioMed Research International*, vol. 2022, no. 1, p. 8333084, 2022.
- [20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [21] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmm: a simple framework for masked image modeling," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9643–9653.
- [22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988.
- [23] N. Mohammadi Foumani, G. Mackellar, S. Ghane, S. Irtza, N. Nguyen, and M. Salehi, "Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5544–5555.
- [24] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
- [25] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *PMLADC at ICLR 2020*, 2020.
- [26] S. Clinchant, K. W. Jung, and V. Nikoulina, "On the use of BERT for neural machine translation," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, and K. Sudoh, Eds. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 108–117. [Online]. Available: <https://aclanthology.org/D19-5611/>
- [27] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.
- [28] C. V. Araya, C. Mendez-Orellana, and M. Rodriguez-Fernandez, "'large spanish eeg,'" 2024.
- [29] C. Valle, C. Mendez-Orellana, C. Herff, and M. Rodriguez-Fernandez, "Identification of perceived sentences using deep neural networks in eeg," *Journal of neural engineering*, vol. 21, no. 5, p. 056044, 2024.
- [30] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
- [31] R. V. Godoy, G. J. Lahr, A. Dwivedi, T. J. Reis, P. H. Polegato, M. Becker, G. A. Caurin, and M. Liarokapis, "Electromyography-based, robust hand motion classification employing temporal multi-channel vision transformers," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 200–10 207, 2022.