IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Decoding the Mind: Neural Differences and Semantic Representation in Perception and Imagination Across Modalities

Owais Mujtaba Khanday[1], *Marc Ouellet*[2], *José L. Pérez-Córdoba*[1], *Asma Hasan Sbaih*[1], *Laura Miccoli*[2], *Jose A. Gonzalez-Lopez*[1]

[1]Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
[2]Dept. of Experimental Psychology, University of Granada, Spain

{owaismujtaba, mouellet, jlpc, asmasbaih, lauramiccoli, joseangl}@ugr.es

Abstract

This study undertakes an analysis of neural signals related to perception and imagination concepts, aiming to enhance communication capabalities for individuals with speech impairments. The investigation utilizes publicly available Electroencephalography(EEG) data acquired through a 124-channel ANT Neuro eego Mylab EEG system (ANT Neuro B.V., Hengelo, Netherlands). The dataset includes 11,554 trials from 12 participants. The proposed convolutional neural network (CNN) model outperformed others in classifying the EEG data as being from the perception or the imagined speech task conditions, achieving a test accuracy of 77.89%. Traditional machine learning models, including Random Forest(RF), Support Vector Classifier (SVC), and XGBoost, showed tendencies to overfit, resulting in low accuracies. as for the semantic decoding, unfortunately, the different models performed at the chance level.

Index Terms: Speech decoding, EEG, BCI, Semantic decoding

1. Introduction

Deciphering the neural mechanisms behind speech processes and converting them into synthetic speech can greatly enhance the quality of life for individuals with speech impairments. In the US alone, it is estimated that 5% of children have speech disorders and nearly 2 million people suffer from brain injuries affecting language and comprehension [1, 2, 3]. Patients with neurodegenerative disorders that affect the muscular activity involved in articulation but do not affect cognitive functions, such as for patients with amyotrophic lateral sclerosis (ALS), could significantly benefit from speech neuroprosthetic technology. This technology enables the decoding of speech from neural activity [4]. Compared to other communication-enhancing technologies, neuroprostheses offer the potential for more natural interactions, even for patients with severe motor-control limitations [5, 6].

However, decoding speech from brain signals presents numerous challenges. These include the complexity of neural signals, with subtle differences in brain activity corresponding to various aspects of speech processing, and individual variability in neural responses [7, 8]. High-resolution and high signal-to-noise ratio data acquisition methods, such as ECoG (electrocorticography), sEEG (stereoelectroencephalography), and MEAs (microelectrode arrays), are invasive and primarily used in clinical settings. This invasiveness limits their scalability and universal applicability. Non-invasive EEG, on the other hand, faces its own challenges, including low signal-to-noise ratio, inter-trial variance, and model dependence on individual participants [9]. Various methodologies have been developed to determine neural tracking of speech, including decoding speech directly from brain signals and aligning brain signals with speech features in a common representation [10, 11, 12]. A significant challenge in neural signal decoding is the participant-specificity of results [13].Researchers in [9] introduced an Encoder-Decoder framework for speech synthesis from EEG signals using generative residual units (GRU, HiFi-GAN, and HuBERT) but, reported high character error rates (CER) of 83% for imagined speech and 78.82% for spoken speech a vocabulary pool of 12 words (22 characters). Similarly, the differential analysis approach in [14] to decode Spanish words and vowels, tested on 15 participants with a vocabulary pool of six words (five vowels), achieved modest accuracies of 29.8% for words and 33.6% for vowels, indicating participant specific outcomes. Anumanchipalli et al. [15] developed a neural decoder that translated ECoG recordings of 60 words from five participants to audible speech using a bidirectional long short-term memory (LSTM) model, reporting word error rates (WER) of 53%. Défossez et al. [16] combined a deep CNN with pre-trained Wav2Vec2.0, demonstrating high performance on MEG data of 19 participants with an accuracy of 70.7%, though lower effectiveness on EEG data with an accuracy of 25.7%, Furthermore, Chen et al. [17] compared 3D ResNet, 3D Swim Transformer, and LSTM frameworks for decoding spectrograms from brain signals, with 3D ResNet achieving the highest Pearson correlation coefficients (PCC) of 0.806 in a study involving 48 participants.Direct comparisons across studies are challenging due to variations in tasks, participant numbers, stimuli, and trial requirements. High-gamma band activity (70–170 Hz) is effective for decoding overt speech, while lower-frequency dynamics are more important for imagined speech [18].

The primary objective of this research is to evaluate state-of-the-art brain-computer interface (BCI) models for their efficacy in decoding neural activity related to perception versus imagination tasks. In order to do so, we used an EEG dataset [19] where participants were asked, in a first time, to perceive a semantic item (presented as a written word, a picture or an auditory word) and, in a second time, to imagine that same semantic stimulation. This study also assesses the participant sensitivity of these models, while also endeavoring to develop a subject-invariant model capable of generalizing across all participants. we aim to decode the semantic categories of the stimuli in both tasks (perception and imagination), thereby striving to achieve robust performance irrespective of individual variability.

2. Methods

2.1. Dataset Description

For this study, we employed a publicly available dataset [19] containing EEG recordings of 12 participants engaged in language perception and imagination tasks.The data were collected

10.21437/IberSPEECH.2024-26

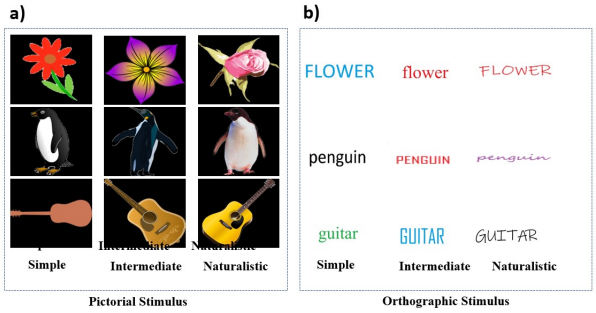

Figure 1: *(a) Pictorial and (b) orthographic stimulus from [19].*

Table 1: *Number of trials by participant.*

Participant ID	Perception	Imagination	Total
sub-14	891	891	1782
sub-15	774	773	1547
sub-11	378	377	755
sub-03	450	450	900
sub-17	450	450	900
sub-08	450	450	900
sub-10	450	450	900
sub-16	450	450	900
sub-12	450	450	900
sub-19	315	315	630
sub-18	270	270	540

using a 124-channel ANT Neuro eego Mylab EEG system (ANT Neuro B.V., Hengelo, Netherlands)As depicted in Figure 1, the stimuli comprised three different categories: flower, penguin, and guitar. These categories were selected based on their semantic distance, computed using Word2Vec [20], and their syllable length to ensure consistency.

In the perception phase, participants were presented with a stimulus belonging to one of three sensory modalities: visual (picture or text) or auditory (audio). For the auditory task, participants listened to recordings of the words spoken in different voices. The trial starts with a masking image for 500 ms (or a masking sound during 1000 ms in the auditory task). After that, the perception task started with the presentation of the stimulus for 3000 ms (2000 ms in the auditory task). Then, the mask was presented again for the same duration. Next, the imagination phase began, lasting 4000 ms. During this phase, participants were asked to mentally recreate the stimulus they had encountered in the perception phase. Because there was an issue with the data of participant sub-13, it was discarded from the analyses. The number of trials for each participant are listed in Table 1. More details about the dataset can be found in [19].

2.2. Signal Processing

Raw EEG signals recorded during both perception and imagination tasks were subjected to rigorous pre-processing to improve their quality and remove artifacts as shown in Figure 2. The preprocessing pipeline involved bandpass filtering within the range of 0.5 Hz to 150 Hz, referencing to a common average and notch filtering to remove power line noise harmonics (50 and 100Hz). Subsequently, independent component analysis (ICA) was employed, decomposing the signals into 20 components to optimize the separation of underlying sources. Post-ICA, signal normalization was achieved via z-score standardization, ensuring amplitude consistency across the dataset. Data for trials were segmented into 1-second intervals from stimulus onset, and models were trained on these segments. To capture essential information pertinent to model decoding, various feature extraction techniques were utilized. Power spectral density (PSD) analysis was conducted for each trial(1-second) to quantify the distribution of signal power across frequency bands, including delta (0.1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-150 Hz), as illustrated in Figure 2. Additionally, the Morlet wavelet transform was applied to extract 50 frequency-domain features for each trial(1-sec.), providing detailed insights into the temporal and frequency characteristics of the trial. Also for each trial complementary statistical measures (mean, std, skewness, kurtosis) were computed to provide further descriptive information about the EEG data segments, thereby facilitating a comprehensive analysis and interpretation in subsequent modeling tasks. These features were used with each segment and fed to the model for decoding.

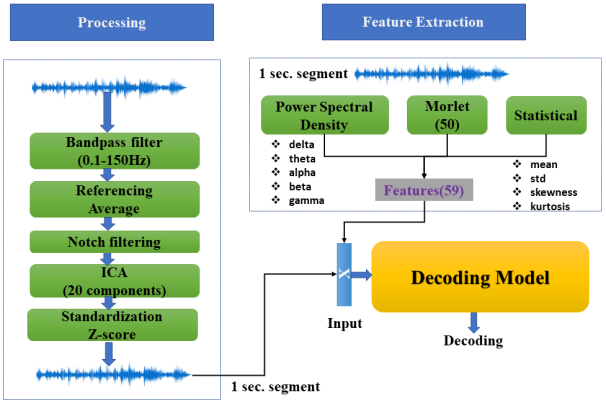

Figure 2: *Preprocessing and feature extraction.*

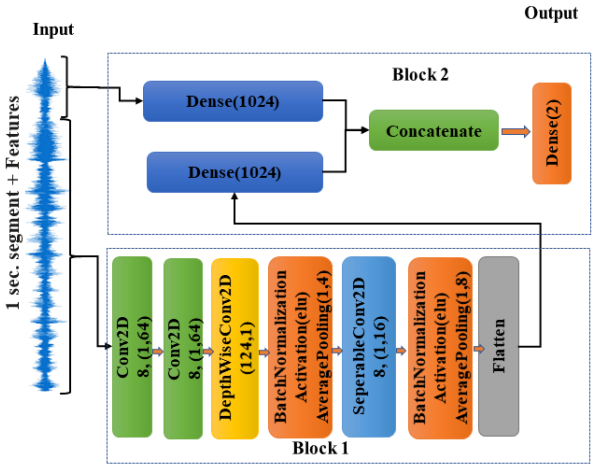

Figure 3: *CNN based model architecture.*

2.3. Model Architecture

We introduce a convolutional neural network (CNN)-based model Fig. 3 designed to process two distinct input streams: EEG data segments and features derived from Morlet wavelet transformations, along with statistical features extracted from the same segments. The architecture consists of two blocks. Block 1 processes the raw EEG data through a series of Conv2D layers, followed by DepthwiseConv2D and SeparableConv2D layers, all employing ReLU activation functions. Each layer is accompanied by Batch Normalization and Average Pooling to

enhance model regularization and reduce dimensionality. The output of Block 1 is then flattened and passed to Block 2. In Block 2, the processed EEG features from Block 1 and Morlet and statistical features are passed to two distinict Dense layers. The outputs from both streams are subsequently concatenated and passed to the final output layer for classification. The model is trained using Adam optimizeer with sparese categorical cross entropy loss function for 100 epochs. Early stopping with a patience of 5 is used to avoid overfitting.

2.4. Classification analysis

A diverse set of deep learning based models EEGNet[21], DeepConvNet[22] and EEGNetSSVEPN[23] were selected based on their reported efficacy in brain-computer interface (BCI) decodings and were implemented to decode the neural activity.

Traditional machine learning models, including RF [24] , SVC [25], and XGBoost [26], were also trained for neural activity decoding tasks. To optimize these models, a comprehensive grid search was conducted to identify the most effective hyperparameters. For XGBoost, the hyperparameters max_depth = [10, 20, 30] and n_estimators = [50, 100, 200] were explored. In the case of SVC, the hyperparameter ranges included C = [0.1, 1, 10, 100], gamma = [0.01, 0.1, 1], and kernel = [linear, rbf, poly, sigmoid].For the Random Forest model, the parameters investigated encompassed n_estimators = [20, 30, 50, 100, 500, 1000], max_depth = [10, 20, 30], max_features = [sqrt, log2], and min_samples_split = [10, 15, 20, 25, 30].

Implementation details can be found at `https://github.com/owaismujtaba/SemanticDecoding.git`

2.5. Evaluation and metrics

The dataset for decoding neural activity into perceptual, imaginative states, and semantic categories was combined across all participants. Each participant's data was split 80-20 for training and testing. Post-training, models were evaluated on unseen data, both collectively and individually, to assess overall and participant-specific accuracy. Performance was measured using accuracy, precision, recall, F1-score, and confusion matrix.

3. Results

3.1. Perception and Imagination

The performance metrics of each model architecture, highlighting their generalization efficacy for decoding neural activity related to perception and imagination tasks, are presented in Table 2. The CNN-based model demonstrated a superior test accuracy of 77.89% with a relatively low loss of 0.4817, indicating robust generalization capabilities. In contrast, traditional machine learning models, including Random Forest, SVC, and XGBoost, exhibited substantial overfitting, with test accuracies of 68.85%, 51.21%, and 50.77%, respectively, and inapplicable loss metrics. Among deep lraening based model architectures, EEGNetSSEVPN achieved the highest training accuracy at 80.14%, but its test accuracy of 76.08% and loss of 0.52 suggest marginal overfitting. DeepConvNet and EEGNet demonstrated commendable performance, with DeepConvNet achieving a test accuracy of 74.78% and a loss of 0.53, while EEGNet recorded a test accuracy of 75.56% with a loss of 0.49, indicating consistent yet slightly inferior performance relative to the CNN model. Table 3 delineates the participant-level performance metrics of various neural decoding models, emphasizing their capacity for participant-independent accuracy. The proposed CNN model emerges as the most effective, achieving a mean accuracy of 76.89% alongside a Standard Deviation(SD) of 5.61 across all participants, which suggests both high performance and substantial consistency across different participants. DeepConvNet has a mean accuracy of 74.64% with a standard deviation of 10.15, indicating considerable variability and reduced reliability in subject-independent applications. EEGNet, with a mean accuracy of 75.26% and SD 8.15, demonstrates improved consistency relative to DeepConvNet, though its accuracy remains marginally lower. Lastly, EEGNetSSEVPN records a mean accuracy of 74.04% and a SD of 8.42, reflecting moderate performance characterized by a moderate level of consistency. Figure 4 show the confusion matrix on the test data using the CNN model. The model was good at classifying the EEG data corresponding to the imagination tasks (precision: 75.4%) and perception tasks (recall: 76.2%). These metrics indicate that the model is fairly balanced in terms of identifying both perception and imagination classes correctly.

Table 2: *Performance on aggregate test data*

Model	**Train**		**Test**	
	Accuracy	**Loss**	**Accuracy**	**Loss**
DeepConvNet [22]	76.22	0.46	74.78	0.53
EEGNet [21]	75.95	0.49	75.56	0.49
EEGNetSSVEPN [23]	80.14	0.42	76.08	0.52
CNN (proposed)	77.48	0.47	77.89	0.48
Random Forest	1.00	NA	68.85	NA
SVC	1.00	NA	51.21	NA
XGB	1.00	NA	50.77	NA

Table 3: *Participant Level Performance Metrics*

Model Name	**Mean Accuracy**	**Standard deviation**
DeepConvNet [22]	74.64	10.15
EEGNet [21]	75.26	8.15
EEGNetSSEVPN [23]	74.04	8.42
CNN(Proposed)	76.82	5.61

3.2. Semantic Decoding

All the models in section 3.1 were trained for decoding the semantic categories from the neural activity. Table 4 shows the comparison of various models. Even if the Random Forest (RF) model standed out as the most balanced model, it exhibited precision and recall scores that indicated a performance at a chance level, leading to an overall accuracy of 32.44% for the classification into the three semantic categories. While DeepConv and EEGNet demonstrated higher recall for specific classes, their precision suffered, indicating potential challenges in accurately identifying all instances. Conversely, models like SVC and XGB displayed lower overall performance, particularly in balancing both metrics. EEGNet model showed high recall for one class but struggled with precision and recall for other categories, underscoring the need for a model that can effectively generalize across all classes. Overall, the Random Forest was the most balanced model, but failed to classify the semantic categories. Nevertheless, it seems that alternative models could be used in the selection of specific items, which could lead to the selection of a certain model based on its particular strengths.

Model	Accuracy	Precision			Recall		
		Flower	Guitar	Penguin	Flower	Guitar	Penguin
DeepConvNet [22]	35.42%	33.03%	36.01%	30.0%	18.23%	80.87%	0.43%
EEGNet [21]	34.08%	34.1%	35.29%	0.0%	98.99%	0.72%	0.0%
CNN	36.03%	60.0%	35.98%	0.0%	0.38%	99.88%	0.0%
EEGNetSSVEPN [23]	35.94%	100.0%	36.0%	30.56%	0.13%	98.56%	1.59%
RF	35.42%	33.03%	36.01%	30.0%	18.23%	80.87%	0.43%
SVC	33.78%	33.92%	32.74%	32.64%	88.23%	4.45%	6.8%
XGBoost	32.14%	34.2%	33.47%	29.14%	31.77%	29.96%	35.17%

Table 4: *Comparison of model performance on semantic decoding*

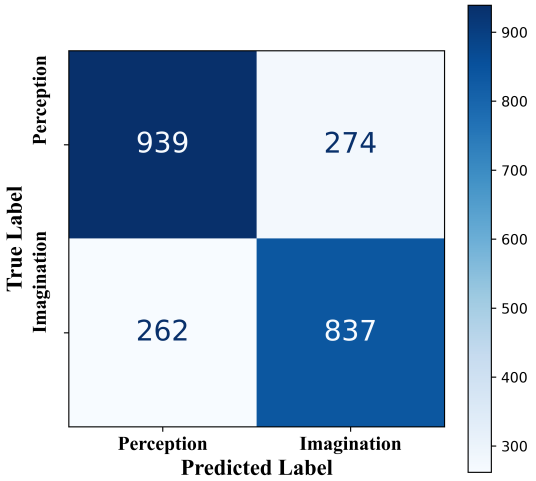

Figure 4: *Confusion matrix of CNN Model on test data*

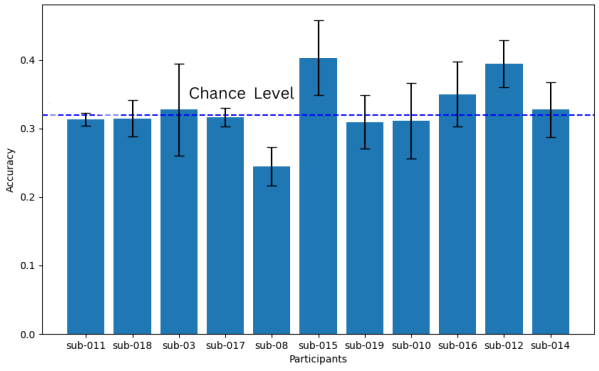

Figure 5: *Subject level accuracy.*

The low accuracy in the models is primarily due to high semantic similarity among the dataset's categories. The authors used Word2Vec latent space to assess semantic distance, representing words as vectors. Distances between vector pairs, all less than 0.2, indicate the categories are very closely related. This closeness makes it challenging for the models to differentiate between classes, leading to confusion during classification. The models may struggle to learn meaningful distinctions, resulting in overall low performance.Another reason why the models failed in the classification of the semantic categories might also be linked to the characteristics of the task used by [19]. Instead of asking the participants to focus on the meaning of the stimuli, participants were asked to recall the perceptual characteristics of previous stimulus (e.g., imagining the auditory stimulus with the same voice as in the presentation phase). This strategy might have fostered the participants to orient their attention on the perceptual characteristics of the stimuli and to not pay much attention to their meaning. Figure 5 illustrates the performance metrics of the Random Forest (RF) model across individual participants. The analysis reveals that the model achieves the lowest accuracy on participant 8, with a value of 25.4%, whereas the highest accuracy was observed for participant 15 at 39.6%. Taking into account the SD, it becomes clear that the model did not perform above the chance level.

4. Conclusion

This study undertook a comprehensive exploration of the complex decoding of neural signals associated with both perceptual, imaginative and semantic decoding tasks. A wide range of methodologies and computational models were meticulously assessed, with a particular focus on developing a participant-invariant model. An extensive protocol for signal processing and feature extraction was employed to enhance the quality of the EEG data. Subsequently, multiple machine learning and deep learning models were trained and evaluated for their effectiveness in decoding these neural signals. The proposed CNN-based model demonstrated superior accuracy and is participant invariant in tasks involving perception and imagination, achieving a test accuracy of 77.89%, thereby indicating robust generalization capabilities across all participants. In contrast, traditional machine learning models, including Random Forest, SVC, and XGBoost, exhibited significant overfitting, as reflected in their comparatively lower test accuracies. In the domain of semantic decoding, the Random Forest model displayed a balanced performance across various semantic categories, but none of the models achieved a classification that was above the chance level. We hypothesize that this failure might be linked to the design of the experiment, which included semantic items that are closely related and a task that emphasized an orientational focus on the perceptual characteristics of the items instead of their semantic content. More research will help determine if these models improve when these variables are controlled.

5. Acknowledgements

This work was supported by grant PID2022-141378OB-C22 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

6. References

[1] G. Alsebayel, M. Nasri, C. P. Myers, G. M. Troiano, E. Hatamimajoumerd, S. Ostadabbas, K. Allison, and C. Harteveld, "Articumotion: Towards assessing motor speech disorders via gamifica-

tion," in *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, 2024, pp. 232–247.

[2] National Institute on Deafness and Other Communication Disorders, "Quick statistics about voice, speech, language," 2024, accessed: 2024-06-14. [Online]. Available: https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

[3] J. Law, J. Boyle, F. Harris, A. Harkness, and C. Nye, "Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature," *International journal of language & communication disorders*, vol. 35, no. 2, pp. 165–188, 2000.

[4] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE access*, vol. 8, pp. 177 995–178 021, 2020.

[5] J. Pearson, "The human imagination: the cognitive neuroscience of visual mental imagery," *Nature reviews neuroscience*, vol. 20, no. 10, pp. 624–634, 2019.

[6] S. Koch Fager, M. Fried-Oken, T. Jakobs, and D. R. Beukelman, "New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science," *Augmentative and Alternative Communication*, vol. 35, no. 1, pp. 13–25, 2019.

[7] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, 2018.

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

[9] Y.-E. Lee, S.-H. Lee, S.-H. Kim, and S.-W. Lee, "Towards voice reconstruction from EEG during imagined speech," *arXiv e-prints*, January 2023. [Online]. Available: https://arxiv.org/abs/2301.07173

[10] S. Martin, I. Iturrate, J. d. R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis," *Frontiers in neuroscience*, vol. 12, p. 367292, 2018.

[11] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, 2018.

[12] M. Naddaf, "Mind-reading devices are revealing the brain's secrets," *Nature*, vol. 626, no. 707, pp. 706–708, 2024. [Online]. Available: https://www.nature.com/articles/d41586-024-00481-2

[13] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from EEG-A comprehensive review," *Frontiers in Neuroscience*, vol. 15, p. 642251, 2021.

[14] V. R. Carvalho, E. M. A. M. Mendes, A. Fallah, T. J. Sejnowski, L. Comstock, and C. Lainscsek, "Decoding imagined speech with delay differential analysis," *Frontiers in Human Neuroscience*, vol. 18, p. 1398065, 2024.

[15] 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.

[16] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1097–1107, 2023.

[17] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devin-sky, Y. Wang, and A. Flinker, "A neural speech decoding framework leveraging deep learning and speech synthesis," *Nature Machine Intelligence*, pp. 1–14, 2024.

[18] T. Proix, J. Delgado Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky *et al.*, "Imagined speech can be decoded from low-and cross-frequency intracranial EEG features," *Nature communications*, vol. 13, no. 1, p. 48, 2022.

[19] H. Wilson, M. Golbabaee, M. J. Proulx, S. Charles, and E. O'Neill, "EEG-based BCI dataset of semantic concepts for imagination and perception tasks," *Scientific Data*, vol. 10, no. 1, p. 386, 2023.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] 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.

[22] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[23] N. Waytowich, V. J. Lawhern, J. O. Garcia, J. Cummings, J. Faller, P. Sajda, and J. M. Vettel, "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *Journal of neural engineering*, vol. 15, no. 6, p. 066031, 2018.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[25] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[26] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785