IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

#neural2speech: Decoding Speech and Language from the Human Brain

Xabier de Zuazo[1], *Vincenzo Verbeni*[2], *Li-Chuan Ku*[2], *Ekain Arrieta*[1], *Ander Barrena*[1], *Anastasia Klimovich-Gray*[3], *Ibon Saratxaga*[1], *Eva Navas*[1], *Eneko Agirre*[1], *Nicola Molinaro*[2,4]

[1]HiTZ Center, University of the Basque Country – UPV/EHU, Spain
[2]Basque Centre on Cognition, Brain and Language – BCBL, Spain; [3]University of Aberdeen, UK
[4]Ikerbasque, Basque Foundation for Science, Bilbao, Spain

xabier.dezuazo@ehu.eus, v.verbeni@bcbl.eu, l.ku@bcbl.eu, ekain.arrieta@ehu.eus

Abstract

Can speech be decoded from brain activity? The #neural2speech project will leverage breakthroughs in cognitive neuroscience and natural language processing to address this compelling question by means of robust neural decoders. Specifically, brain-to-speech decoders will be designed to reconstruct both perceived and produced speech from non-invasive brain recordings, namely functional magnetic resonance imaging and magnetoencephalography data. By integrating deep learning techniques and large language models, not only does #neural2speech seek to deepen our understanding of language processing in the human brain – with a particular focus on multilingual processing –, but it also aims to pave the way for the development of innovative communication aids that can help individuals affected by speech impairments. The potential applications are vast, with the promise to revolutionize clinical neuroscience and human-computer interactions.

Index Terms: Brain-computer interface, neural speech decoding, magnetoencephalography, functional magnetic resonance imaging

1. Introduction

The ongoing rapid development of artificial intelligence (AI) makes it more and more feasible to interpret neural signals [1], with important implications for both clinical neuroscience and our overall understanding of brain functions. The idea of using AI to explore human cognition builds upon the hypothetical similarity between the two systems in terms of design and information processing. Indeed, the architecture of artificial deep neural networks (DNNs) [2] such as recurrent neural networks (RNNs) [3] and convolutional neural networks (CNNs) [4] is inspired by the organization of neurons in living organisms, despite not being exact replicas of their biological counterparts. These DNNs contain many single processing units that are ordered in hidden yet interconnected layers and are powerful in learning the latent features of – for instance – images [5], mirroring the processes through which biological visual pathways handle sensory inputs. Can AI be equally effective in decoding human speech and language based on its underlying neural activity? Combining neural response data with DNNs has been shown to allow for various applications, including the development of models that can reconstruct speech and picture stimuli from brain activity [6, 7]. To date, the most promising results have been generated by encoder-decoder architectures [8]. While one must be careful in assuming that cognitive processes can be directly analyzed based on the internal mechanisms of the computational models used for decoding [9], such models do appear to develop cognitively plausible internal representations [10, 11, 12]. Crucially, encoders use large language models (LLMs, e.g., Transformer architectures or GPT [13, 14]) to perform complex language comprehension tasks, such as next-word prediction. To this aim, LLMs leverage linguistic information at different hierarchical levels – from syntax to semantics. Provided that both LLMs and human brains can predict words from context [10, 15], making use of LLMs as predictors of speech-related neural activity represents a compelling avenue for researchers who aim to address and develop new hypotheses about language processing in the brain.

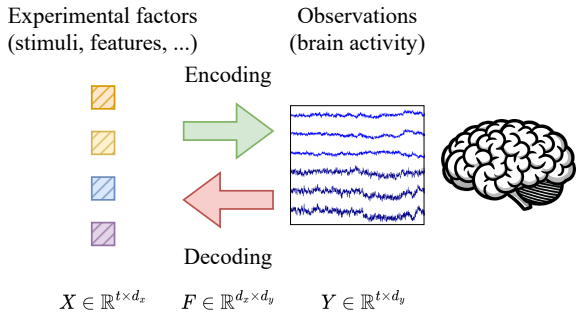

Figure 1: *Encoding-decoding approaches model the relationship between experimental factors (X), such as stimuli or behavioral tasks, and observed brain activity (Y). F symbolizes the mathematical functions used for encoding or decoding.*

Figure 1 provides a visual summary of the rationale behind encoding-decoding approaches [8]. Encoding models predict brain activity based on known stimuli or task conditions, offering insights into how information is processed and represented in the brain. Conversely, decoding models reconstruct these stimuli from brain activity. Together, these complementary approaches can enhance our understanding of cognitive functions and improve our ability to interpret intricate brain data in a meaningful way.

Recent studies have shown that, by embracing an encoding-decoding approach, it is possible to model neural activity collected via electrocorticography (ECoG) [16], an invasive recording technique that requires intracranial electrode implantation. Specifically, RNNs have been used to decode the intended speech of patients affected by articulatory impairments, showing a median word error rate of 27.6% (text decoding) and 54.4% (synthesized speech) on a large vocabulary (around 40,000 words). Even though the overall performance in synthesized speech reconstruction was not optimal, these results prompted the need for further research on speech decoding from non-invasive brain recordings, since non-surgical techniques are safer, more practical, and applicable to a broader popula-

10.21437/IberSPEECH.2024-52

tion. Indeed, speech decoding has also been attempted on neural data collected via functional magnetic resonance imaging (fMRI) [17], electroencephalography (EEG) [18], and magnetoencephalography (MEG) [19, 20].

The distinctive value of #neural2speech lies in its focus on decoding language from data collected during both (1) speech perception and (2) speech production tasks. This approach will allow us to fully harness the potential of decoding techniques, namely to a) advance scientific understanding of language processing and b) drive future clinical applications.

1.1. Speech perception

Understanding the neural encoding of perceived speech is a key objective of cognitive neuroscience, and the use of non-invasive brain imaging techniques during speech perception tasks provides valuable details about the neural activity that underlies the processing of different speech properties. While it is well established that the primary auditory cortex is active during auditory language processing, several non-auditory cortical regions contribute to the overall encoding of listened speech [21, 11]. Nonetheless, brain responses to speech are mostly skewed towards the auditory cortex, an instance that complicates the possibility of investigating how higher-order properties of language – such as semantics and syntax – are neurally encoded. It must be noted that speech decoding per se appears to be feasible by leveraging purely auditory information. For instance, Défossez and colleagues demonstrated that, in their decoding task, relatively high accuracy scores can be achieved by using speech self-supervised models like wav2vec 2.0 [18]: these models are trained similarly to LLMs, but on speech data, as they primarily rely on acoustic information to create representations of the stimuli. Conversely, our replication experiments showed that using sBERT [22] sentence embeddings to encode stimuli resulted in very low accuracy scores, thus suggesting that speech decoding relies almost exclusively on acoustic information. These outcomes highlight the need to develop new methods that capture the relationship between the non-acoustic properties of speech stimuli – most importantly, their semantic representations – and the neural activity that underpins their processing.

Within a brain-to-speech decoding context, the use of speech stimuli's semantic properties holds significant promise for addressing key research questions in cognitive neuroscience. Given the unique linguistic landscape of the Basque Country, home to a large population of balanced (Basque-Spanish) bilinguals, one of the objectives of #neural2speech is to apply decoding algorithms to bilingual speech perception datasets, with the ultimate goal of investigating the distribution of semantic networks within the bilingual brain. Indeed, the way bilingual brains handle semantic representations across different languages is still a matter of debate across the scientific community. Some theories argue that bilingual brains are characterized by separate (language-specific) semantic networks, a notion backed by findings that show how brain damage can affect the understanding of concepts in one language only, leaving the others unaffected [23, 24]. Conversely, another set of theories suggests that the same brain regions are involved in processing meaning, regardless of the language that conveys it. This idea is supported by evidence of cross-language interference, such as longer processing times for false cognates when learning a second language [25]. Within this context, brain-to-speech decoding will provide important contributions to the field of bilingual neurolinguistics.

1.2. Speech production

Decoding speech production from brain activity presents unique challenges and opportunities for the future of clinical neuroscience.

Unlike speech perception tasks, speech production involves the active engagement of complex motor planning, auditory feedback, and execution processes within the brain [26]. Moreover, myographic artifacts caused by the contraction of face muscles produce significant interference, which complicates the extraction of neural signals purely related to language [27, 28, 29]. From this perspective, effective artifact removal techniques [30, 31] are crucial for isolating the relevant components of brain activity, yet current methods – such as Independent Component Analysis (ICA) – must be implemented with great care in order to avoid the accidental removal of valuable neural signals [32, 33]. Finally, an additional challenge is represented by the design of appropriate experimental tasks, since the accurate decoding of speech production requires that a) subjects are presented with a diverse set of stimuli, as well as b) precise data labeling, and c) fine-grained synchronization between speech and neural data.

MEG data and deep learning classification techniques have been recently employed to decode isolated spoken words or sentences [19, 34, 35, 36], progressively advancing our understanding of the brain dynamics involved in producing speech. In our MEG pilot studies, we analyzed the classification of phonetic units (phones) during active speech production. Interestingly, our findings suggest that produced speech can be decoded from brain signals with greater accuracy than perceived speech, a result that highlights the wealth of information provided by the integrative neural processes that link cognition with verbal outputs. These preliminary observations will be critical for the development of brain-computer interfaces specifically designed to restore communication.

2. Main goals

The primary purpose of #neural2speech is to leverage advancements in neuroscience, natural language processing (NLP), and speech technology to improve our understanding of the brain mechanisms involved in speech perception and production. Specifically, we intend to analyze brain activation patterns and decode speech across multiple levels of the linguistic hierarchy, from phonetics [11] to semantics [17], extracting meaningful acoustic and linguistic information [18] directly from neural data.

Our goals are ambitious and multifaceted: first, we aim to provide significant contributions to the scientific community by deepening our understanding of the brain circuitry involved in language processing; second, we aim to set the foundation for the development of cutting-edge methodologies to generate synthetic speech from brain signals, an advancement that will make daily interactions more accessible and enriching to people affected by speech impairments.

More specifically, the objectives of #neural2speech include: a) replicating the speech decoding accuracy scores obtained in previous neuroimaging studies; b) testing new decoding solutions on publicly available neuroimaging datasets; c) collecting multimodal neuroimaging data (MEG and fMRI) during natural speech perception tasks to evaluate language decoding in Spanish-Basque bilinguals; 4) developing the first (public) speech production MEG dataset, with data derived from experimental tasks that satisfy the optimal requirements to test a

neural decoding algorithm; 5) testing a newly-developed multimodal contrastive model optimized for the decoding of both perceived and produced speech.

In addition, our technology will be applied to an ongoing study on dyslexia that focuses on the development of a classification model (DyslexNet) based on neural speech processing data. Research is ongoing, but promising results have already been published [37].

3. Methodology

In the present section, we will describe the collaborative framework that brought #neural2speech into existence, as well as the resources utilized for the implementation of the project.

3.1. Team

The project requires joint collaboration between specialists in different disciplines. Experts in multimodal neuroimaging optimization and computational neuroscience from the Basque Centre on Language, Brain and Cognition (BCBL), the University of Aberdeen, and the University of Connecticut will lead the collection of speech perception and production data. The process will be coordinated with the other parties involved in #neural2speech, with the ultimate objective of collecting data that is optimized for the modeling goals of the project. Researchers specialized in language technology and speech signal processing from the Basque Centre for Language Technology (HiTZ Zentroa), the Ixa Research Group, and the Signal Processing Laboratory (Aholab) of the University of the Basque Country (UPV/EHU) will lead the development of the multimodal contrastive model that will be used to decode speech from brain activity. HiTZ Zentroa will also take an active part in the development of a protocol to decode speech production data collected under different experimental conditions, with a particular focus on the prospect of decoding intended speech when articulation is compromised.

BCBL and the University of Connecticut are already working on an NSF/AEI-granted project to develop a neural network model of human speech recognition across multiple languages. On the other hand, BCBL, UPV/EHU, and the University of Aberdeen have been working on separate projects focused on (i) neural models applied to speech processing and (ii) the classification of developmental dyslexia based on neural data.

The partners involved in the project (BCBL, HiTZ Zentroa UPV/EHU, University of Connecticut, University of Aberdeen) have a longstanding record of key scientific contributions and are exceptionally well-connected with other renowned experts throughout the world. Their combined expertise will contribute to enhancing the visibility, research positioning, and international competitiveness of the Basque Country. The work conducted within #neural2speech will open new opportunities for research and innovation in the area of computational neuroscience; given its extensive network of collaborations, the team will be in a competitive position to seek additional EU funding – in the form of, for instance, ERC Synergy grants, Doctoral Training Networks and the Pathfinder funding scheme, among others. The team members will participate in international activities (Marie Curie forums, Researchers' Night, Pint of Science, International Day of Women and Girls in Sciences, and Week of Science of UPV/EHU). Publication of results in international peer-reviewed journals (i.e., Nature Neuroscience, Science, eLife) will follow the open data principle underlying Horizon Europe; manuscripts will be deposited in institutional or thematic repositories. If this is not possible, the team will adopt a green (vs gold) model. All members of the team will disseminate the results of the project at international conferences (e.g., Society for Neuroscience, Interspeech, ICASSP).

3.2. Neuroimaging

Speech decoding is usually performed on data derived from a single non-invasive technique, be it M/EEG or fMRI [19, 17, 18]. However, each neuroimaging method comes with its own limitations: fMRI generates recordings of brain activity with a highly detailed (millimeter-order) spatial resolution, but its temporal resolution (lower than $1\,\mathrm{s}$) is too low to track neural responses to the acoustic and phonological changes that characterize speech; on the other hand, MEG is optimal for tracking fast neuronal activity – as it is marked by a millisecond-order temporal resolution –, at the cost of a sparser spatial resolution (lower than $1\,\mathrm{cm}$). To tackle this issue, in #neural2speech, we will feed our decoding algorithm with a richer set of speech perception data derived from both fMRI and MEG, with the ultimate objective of capturing the complex spatiotemporal dynamics of speech-related brain activity. Conversely, due to practical limitations, the neural signals related to speech production will solely consist of MEG data.

The MEG facility of BCBL hosts a 306-sensor Megin Triux Neo® scanner (204 planar gradiometers and 102 magnetometers, arranged in a helmet configuration) equipped with ARMOR technology for instant recovery. The MEG device also includes an integrated 128-channel EEG system for simultaneous MEG and EEG recording, which can be acquired at a sampling rate of up to $5\,\mathrm{kHz}$ in either AC or DC. The facility includes active shielding to reduce external noise and MaxFilter™ software, which filters artifacts and both internal and external sources of noise.

BCBL also houses a Siemens 3T MAGNETOM PRISMAfit, an MR scanner that allows BCBL researchers to conduct functional MRI (BOLD and perfusion ASL) studies. This system uses the Siemens Total Imaging Matrix (TIM) 4G technology. In addition, the BCBL MRI lab is equipped with multiple stimulation and response MR-compatible peripherals that allow for all types of fMRI studies. An MR-compatible BrainAmp Plus system with up to 64 electrodes (Brain Products) is also available for simultaneous EEG-fMRI experiments.

3.3. Language models

Deep learning-based self-supervised large models – such as speech encoders and transformer-based language models (LMs) – have created new opportunities to study the relationship between language and brain signals, extending the influence of NLP to the field of neurolinguistics.

The project involves the use of both audio encoders and LLMs. For instance, Wav2vec 2.0 [38, 39] – a powerful speech encoder – can be used to generate rich, high-level embeddings directly from raw audio files. Speech embeddings can then be combined with brain activity data to train models that predict words from brain activity. Similarly, transformer-based language models – such as BERT [40] and GPT [14, 41] – excel at capturing linguistic structure, a feature that makes them ideal for generating candidate word sequences during brain-to-speech decoding.

Training such models and analyzing large datasets of brain activity demand substantial computational resources. Access to high-performance computing infrastructures allows us to train new models and process large volumes of data more efficiently.

Specifically, this project relies on the computational resources made available by HiTZ Zentroa – including 20 A100 GPUs – and the IKUR-IKA project, which grants access to the Hyperion cluster – with 72 A100 GPUs hosted by the DIPC Supercomputing Center.

4. Expected results

The #neural2speech project aims to decode speech from brain data related to both speech production and perception.

Production-wise, we will gather more than 100 hours of MEG data across different experimental conditions – ranging from overt speech production to imagined speech. This large dataset is expected to be sufficient for training advanced neural models to perform decoding tasks – especially for imagined speech, which poses distinct challenges due to the absence of overt vocal and articulatory signals.

Regarding speech perception, we will gather more than 120 hours of brain data distributed across two different neuroimaging modalities: MEG and fMRI. Our primary goal is to replicate the results of Tang et al. [17], with the added objective of leveraging the decoding capabilities of multimodal language models and the unique opportunities provided by bilingual communities in the Basque Country. Specifically, we expect to identify – by means of cross-language decoding – the extent by which a common semantic space exists between Spanish and Basque. Importantly, the development of advanced fusion techniques will be crucial for the effective integration of MEG and fMRI signals.

5. Acknowledgements

This project is funded by the Basque Government under the IKUR program (IKUR-IKA-23/18). The authors would like to thank the DIPC Supercomputing Center for their technical support when accessing the Hyperion cluster.

6. References

[1] W. Huang, H. Yan, K. Cheng, C. Wang, J. Li, Y. Wang, C. Li, C. Li, Y. Li, Z. Zuo, and H. Chen, "A neural decoding algorithm that generates language from visual activity evoked by natural images," *Neural Networks*, vol. 144, pp. 90–100, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021003117

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[3] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth annual conference of the international speech communication association*, 01 2014, pp. 1964–1968.

[4] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *CoRR*, vol. abs/1511.08458, Nov 2015, 10 pages, 5 figures. [Online]. Available: https://api.semanticscholar.org/CorpusID:9398408

[5] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," *Advances in neural information processing systems*, vol. 26, p. 2553–2561, 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/f7cade80b7cc92b991cf4d2806d6bd78-Paper.pdf

[6] Y. Benchetrit, H. Banville, and J.-R. King, "Brain decoding: toward real-time reconstruction of visual perception," 2024. [Online]. Available: https://arxiv.org/abs/2310.19812

[7] Y. Yang, H. Jo, Y. Duan, Q. Zhang, J. Zhou, W. H. Lee, R. Xu, and H. Xiong, "MAD: Multi-alignment MEG-to-Text decoding," 06 2024. [Online]. Available: https://arxiv.org/abs/2406.01512

[8] J.-R. King, L. Gwilliams, C. Holdgraf, J. Sassenhagen, A. Barachant, D. Engemann, E. Larson, and A. Gramfort, "Encoding and Decoding Framework to Uncover the Algorithms of Cognition," in *The Cognitive Neurosciences*. The MIT Press, 05 2020, vol. 6, pp. 691–702. [Online]. Available: https://doi.org/10.7551/mitpress/11442.003.0076

[9] O. Guest and A. E. Martin, "On logical inference over brains, behaviour, and artificial neural networks," *Computational Brain & Behavior*, vol. 6, no. 2, pp. 213–227, Jun 2023. [Online]. Available: https://doi.org/10.1007/s42113-022-00166-x

[10] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "The neural architecture of language: Integrative modeling converges on predictive processing," *Proceedings of the National Academy of Sciences*, vol. 118, no. 45, p. e2105646118, 2021.

[11] L. Gwilliams, J.-R. King, A. Marantz, and D. Poeppel, "Neural dynamics of phoneme sequences reveal position-invariant code for content and order," *Nature Communications*, vol. 13, no. 1, p. 6606, Nov 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-34326-1

[12] R. Antonello, A. Vaidya, and A. Huth, "Scaling laws for language encoding models in fMRI," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 21 895–21 907. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/4533e4a352440a32558c1c227602c323-Paper-Conference.pdf

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, p. 6000–6010, 2017.

[14] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," *arXiv*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[15] G. Tuckute, N. Kanwisher, and E. Fedorenko, "Language in brains, minds, and machines," *Annual Review of Neuroscience*, vol. 47, no. Volume 47, 2024, pp. 277–301, 2024. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-120623-101142

[16] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, I. Zhuravleva, A. Tu-Chan, K. Ganguly, G. K. Anumanchipalli, and E. F. Chang, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, Aug 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06443-4

[17] J. Tang, A. LeBel, S. Jain, and A. G. Huth, "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nature Neuroscience*, vol. 26, no. 5, pp. 858–866, May 2023. [Online]. Available: https://doi.org/10.1038/s41593-023-01304-9

[18] 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, Oct 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00714-5

[19] D. Dash, P. Ferrari, D. Heitzman, and J. Wang, "Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 5531–5535.

[20] M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, and F. P. de Lange, "A hierarchy of linguistic predictions during natural language comprehension," *Proceedings of the National Academy of Sciences*, vol. 119, no. 32, p. e2201968119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2201968119

[21] J. M. Correia, C. Caballero-Gaudes, S. Guediche, and M. Carreiras, "Phonatory and articulatory representations of speech production in cortical and subcortical fMRI responses," *Scientific Reports*, vol. 10, no. 1, p. 4529, 2020.

[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[23] M. Paradis, "On the representation of two languages in one brain," *Language Sciences*, vol. 7, no. 1, pp. 1–39, 1985.

[24] A. Ku, E. A. Lachmann, and W. Nagler, "Selective language aphasia from herpes simplex encephalitis," *Pediatric neurology*, vol. 15, no. 2, pp. 169–171, 1996.

[25] T. Dijkstra, J. Grainger, and W. J. van Heuven, "Recognition of cognates and interlingual homographs: The neglected role of phonology," *Journal of Memory and Language*, vol. 41, no. 4, pp. 496–518, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749596X99926542

[26] F. H. Guenther, *Neural Control of Speech*. The MIT Press, 07 2016. [Online]. Available: https://doi.org/10.7551/mitpress/10471.001.0001

[27] S. Aristei, A. Melinger, and R. A. Rahman, "Electrophysiological chronometry of semantic context effects in language production," *Journal of cognitive neuroscience*, vol. 23, no. 7, pp. 1567–1586, 2011.

[28] R. Fargier, A. Buerki, S. Pinet, F.-X. Alario, and M. Laganaro, "Word onset phonetic properties and motor artifacts in speech production EEG recordings," *Psychophysiology*, vol. 55, no. 2, p. e12982, 2018.

[29] S. Vanhatalo, J. Voipio, A. Dewaraja, M. D. Holmes, and J. W. Miller, "Topography and elimination of slow EEG responses related to tongue movements," *NeuroImage*, vol. 20, no. 2, pp. 1419–1423, 2003.

[30] B. S. Philip, G. Prasad, and D. J. Hemanth, "Non-stationarity removal techniques in MEG data: A review," *Procedia Computer Science*, vol. 215, pp. 824–833, 2022, 4th International Conference on Innovative Data Communication Technology and Application. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922021561

[31] G. P. Beril Susan Philip and D. J. Hemanth, "A systematic review on artifact removal and classification techniques for enhanced MEG-based BCI systems," *Brain-Computer Interfaces*, vol. 10, no. 2-4, pp. 99–113, 2023. [Online]. Available: https://doi.org/10.1080/2326263X.2023.2233368

[32] M. Bourguignon, N. Molinaro, M. Lizarazu, S. Taulu, V. Jousmäki, M. Lallier, M. Carreiras, and X. De Tiège, "Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud," *NeuroImage*, vol. 216, p. 116788, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920302755

[33] S. Tuomaala, S. Autti, S. F. Cotroneo, P. Lioumis, H. Renvall, and M. Liljestrom, "Automated speech artefact removal from MEG data utilizing facial gestures and mutual information," *bioRxiv*, pp. 2024–09, 2024.

[34] D. Dash, P. Ferrari, S. Malik, A. Montillo, J. A. Maldjian, and J. Wang, "Determining the optimal number of MEG trials: A machine learning and speech decoding perspective," in *Brain Informatics*, S. Wang, V. Yamamoto, J. Su, Y. Yang, E. Jones, L. Iasemidis, and T. Mitchell, Eds. Cham: Springer International Publishing, 2018, pp. 163–172.

[35] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 489–493.

[36] J. Kwon, D. Harwath, D. Dash, P. Ferrari, and J. Wang, "Direct speech synthesis from non-invasive, neuromagnetic signals," in *Interspeech 2024*, 2024, pp. 412–416.

[37] A. Klimovich-Gray, G. Di Liberto, L. Amoruso, A. Barrena, E. Agirre, and N. Molinaro, "Increased top-down semantic processing in natural speech linked to better reading in dyslexia," *NeuroImage*, vol. 273, p. 120072, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811923002185

[38] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[39] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[40] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.