0.1 Introduction

Speech disorders, resulting from damage to muscles, nerves, or vocal structures, impair sound production and can severely impact communication and quality of life **jullien2021screening**; **Gonzalez2020**. These conditions, including stuttering and apraxia, are prevalent worldwide, especially among children, with 7.7% of US children aged 3-17 and 11% of those aged 3-6 experiencing speech-related disorders annually **bib1**; **bib2**; **bib3**. Neuroprostheses (also known as brain-computer interfaces), which interface with the nervous system to restore lost functions, offer innovative solutions for communication impairments, particularly in conditions like amyotrophic lateral sclerosis (ALS), where speech muscles are affected but cognition remains intact **Gonzalez2020**; **neuro1**; **neuro2**; **neuro3**. These devices, whether implanted or external, enable more natural and effective interactions than traditional communication aids, providing critical support for individuals with severe motor limitations **bib4**; **bib5**.

Automatic speech recognition (ASR) and speech synthesis are the two main approaches for decoding speech from neural signals. The ASR-based approach **asr1**; **asr2**; **asr3** utilizes specialized software and models to convert neural signals into textual representations, while the speech synthesis approach aims to generate audible audio directly from the neural signals. In this work, we focus on this second approach. Traditional approaches for decoding and synthesizing speech spectrograms from neural signals, such as linear models and formant synthesis with Kalman filters, achieved limited performance with low-quality synthesized audio and Pearson correlation coefficients (PCC) below 0.7 **bib6**; **bib7**; **bib8**. Recent advances leverage deep neural networks (DNNs), including convolutional and recurrent architectures, which improve intermediate representations and quality of speech synthesis **bib9**; **bib10**; **bib11**. However, challenges remain due to the noisy and redundant nature of neural signals, low signal-to-noise ratios (SNRs), and temporal misalignments between neural and speech signals, making it difficult to accurately capture speech-related patterns like prosody and articulation **asr2**; **a3**. Thus, speech synthesis from neural signals remains an open problem.

In this study, we introduce a novel technique for decoding audible speech directly from neural signals. The main novelty of our work is the development of a novel DNN architecture, NeuroIncept, which leverages a multi-scale feature extraction pipeline through the incorporation of Inception modules. As outlined in Section 0.2.3, this architecture enables the model to capture a wide range of temporal and spectral patterns, facilitating the analysis of both fine-grained details and broader trends in neural data. This design addresses a critical limitation in prior approaches, such as **bib9**, which rely on uniform filter sizes and may overlook the multiscale patterns crucial for precise decoding. Furthermore, our NeuroIncept decoder integrates temporal modeling via gated recurrent units (GRUs)**gru**, ensuring robust handling of neural signal misalignments while effectively capturing temporal dependencies. By combining Inception modules with recurrent mechanisms, NeuroIncept delivers a comprehensive and adaptive feature representation, offering significant advancements over traditional

methods.

0.2 Methodology

0.2.1 Dataset Description

The dataset used in this study is publicly available **verwoert2022dataset** and consists of stereotactic EEG (sEEG) recordings from 10 Dutch participants (5 Male and 5 Female; average age: 32 years) with pharmacoresistant epilepsy. Depth sEEG electrodes were implanted in the participants as part of their clinical treatment. The placement of the electrode was determined solely based on clinical requirements, primarily targeting the superior temporal sulcus, the hippocampus, and the inferior parietal gyrus. As a result, the number and locations of the electrodes varied between the participants. sEEG signals were recorded at either 2048 Hz or 1024 Hz, synchronized with participants' speech (F_s = 48 kHz), while they read aloud a list of 100 words from the Dutch IFA corpus **ifaCorpus**. The sEEG recordings were subsequently down-sampled to 1024 Hz, while speech signals were down-sampled to 16 kHz for further analysis. To ensure participant anonymity, pitch modulation of the audio recordings was applied using the LibROSA library **librosa**.

0.2.2 Signal Processing

The sEEG signals for each participant were parameterized as time-frequency features extracted from the high-gamma band (70-170 Hz), as shown in Figure 1a. This band was chosen because previous studies have shown that it contains information related to speech and language production and perception **a3**; **band1**; **band2**. The raw sEEG data was first detrended to remove linear trends. A bandpass filter (70-170 Hz) was then applied to isolate the high-gamma-frequency components, while a notch filter targets line noise (50 Hz) and its two first harmonics (100 Hz, 150 Hz), further refining the signal. After filtering, the Hilbert transform was utilized to compute the analytic signal, which enables the extraction of the signal envelope capturing amplitude fluctuations within the high-gamma band. The processed sEEG signals underwent segmentation into overlapping temporal windows of 0.05s, with a frame-shift of 0.01s. Within each window, the mean amplitude was computed, producing a feature matrix for subsequent analysis. To incorporate temporal dynamics, each 0.05s window was further expanded by integrating the features from both the current and neighboring time windows. This process was achieved using a sliding window approach with a model order of 4 and step size of 5, enhancing temporal resolution and capturing dependencies across multiple time intervals.

The audio signals, on the other hand, were converted into logarithmic Mel-scaled spectrograms (logMel) with 128 spectral bins, as illustrated in Figure 1b. The logMel spectrograms were extracted from 0.05 s overlapping windows with a frame shift of 0.01 s using a Hanning window. The neural data for each

participant was then normalized by z-score standardization, which enhances the comparability between data points and optimizes the subsequent model training procedures.

0.2.3 Decoding Model Architecture

The NeuroIncept Decoder architecture shown in Figure 2, is designed to efficiently process and analyze sequential data by combining the complementary strengths of CNNs and GRUs. Central to this architecture are two distinct, yet synergistic, modules: an Inception module, which serves as the primary feature extractor, and a Recurrent module, which is responsible for temporal pattern recognition.

Inception Module: The Inception Module **inception** of the NeuroIncept Decoder architecture acts as the primary feature extractor, adeptly processing input sequence data through multiple convolutional filters with varying kernel sizes: 1x1, 3x3 and 5x5. Each filter serves a unique role: the 1x1 convolution reduces the dimensionality of the data, preserving essential spatial information while streamlining computation; the 3x3 and 5x5 convolutions capture medium- and large-scale patterns, respectively, from the input sequences. The outputs from these operations are concatenated followed by the MaxPooling operation and the 1x1 convolutional layer to further reduce the spatial dimensions while integrating the pooled features. This approach allows the Inception Module to capture diverse temporal/spectral patterns from the sequence, thus enabling the model to process fine-grained and broader features in the neural data. In contrast, approaches such as **bib9**; **bib10** use uniform filter sizes, which may fail to detect multiscale patterns critical for accurate decoding.

Recurrent Module This module leverages GRU-based recurrent neural networks to capture temporal dependencies **gru**. The first GRU layer, consisting of 128 units, processes the extracted features and returns a complete sequence of hidden states. This ensures that the succeeding GRU layers can operate with a full temporal context. As the data flows through successive GRU layers, the model progressively learns more intricate temporal patterns. The final GRU layer, with 512 units, produces a single output that encapsulates the entire sequence into a condensed summary. The Reshape layer then reformat this output into a tensor with a one time-step and 512 features, preparing the data for subsequent processing.

The Recurrent Module output is passed through the second Inception module, then flattened into a one-dimensional vector. This vector is sent through several dense layers, gradually reducing sizes from 1024 to 128 units. These layers are designed to refine the extracted features and yield the model's ultimate output. The implementation of our NeuroIncept Decoder architecture can be found at https://github.com/owaismujtaba/NeuroInceptDecoder.

0.2.4 Evaluation

The dataset for each participant, consisting of 5 minutes of simultaneous sEEG and audio signals while the participant read aloud a list of 100 Dutch words, was divided into training and validation sets in an 80-20% ratio. In the training stage, the parameters of NeuroIncept architecture were optimized using the Mean Squared Error (MSE) loss function. To address the potential for overfitting, early stopping was implemented with a patience of 5 epochs. For evaluation, 1,000 samples were selected from the 20% validation set for each participant using 10-fold cross-validation. The performance of the proposed method was evaluated using the Pearson Correlation Coefficient (PCC) to quantify the similarity between the spectrograms generated by the NeuroIncept Decoder and the original audio spectrograms. Additionally, the Spectral Temporal Glimpsing Index (STGI) **?? edraki2022spectro** was employed to assess how effectively the predicted spectrograms retained the temporal and spectral characteristics of the original audio signals. As this study is primarily focused on accurately decoding neural signals into log-Mel spectrogram representations of audio, subjective listening tests were not conducted. Future research will address this limitation by incorporating vocoder systems to reconstruct audio waveforms from the decoded spectrograms, enabling more comprehensive assessments of neural-to-audio decoding performance.

$$\text{STGI} = \frac{1}{T \cdot F} \sum_{t=1}^{T} \sum_{f=1}^{F} \mathcal{M}(t, f) \tag{1}$$

Where $\mathcal{M}(t, f)$ is defined by the local SNR threshold θ:

$$\mathcal{M}(t, f) = \begin{cases} 1 & \text{if SNR}(t, f) > \theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

0.3 Results

In this section, we present the results achieved by the proposed NeuroIncept Decoder model in the speech synthesis task outlined above. First, in Section 0.3.1, we present the speech quality metrics obtained by our system for each subject. Then, in Section 0.3.2, we perform a quantitative and qualitative comparative analysis with other models from the literature.

0.3.1 Detailed Results

Table 1 shows the average and standard deviation (std) for each metric obtained by our model among the ten participants in the database described in Section 0.2.1. The MSE values range from 0.400 to 0.788, reflecting differences in prediction accuracy. Higher correlation values suggest strong relationships and a high likelihood of accurate audio reconstruction from neural data. The small

Table 1: Performance metrics on individual subjects.

Participant	MSE	PCC		STGI	
		Value	STD	Value	STD
sub-01	0.445	0.921	0.003	0.511	0.004
sub-02	0.511	0.926	0.002	0.477	0.005
sub-03	0.506	0.925	0.002	0.502	0.005
sub-04	0.522	0.938	0.004	0.479	0.005
sub-05	0.594	0.932	0.003	0.502	0.003
sub-06	0.409	0.944	0.002	0.552	0.004
sub-07	0.788	0.942	0.004	0.511	0.006
sub-08	0.652	0.897	0.005	0.526	0.005
sub-09	0.400	0.917	0.002	0.459	0.004
sub-10	0.498	0.838	0.007	0.522	0.004

standard deviations of the correlations indicate consistent performance between participants. The STGI values range from 0.502 for sub-09 to 0.552 for sub-06, with minimal standard deviations, demonstrating stable individual scores.

Detailed results of the distribution of each metric for the participants are shown in Figs.3 and **??**. In particular, Fig. 3 depicts the distribution of the Pearson correlation coefficients computed between the original audio spectrograms and the reconstructed spectrograms in the test set. Our model demonstrates robust performance across all participants, as indicated by high correlation coefficients ranging from 0.83 in sub-10 to 0.93 in sub-06. The variation in individual correlation coefficients reflects differences in neural activity between participants, which can be attributed to the varying number of electrodes and their implantation sites. In particular, Sub-06 and Sub-07 have a higher number of electrodes implanted in the Broca and Wernicke regions of the brain compared to other participants, regions traditionally associated with cognitive processes of speech and language **chang2015contemporary**. In contrast, the number of implanted electrodes in Sub-10 is relatively low, particularly in the Broca and Wernicke regions. This limited electrode coverage may affect the effectiveness of monitoring or stimulating these critical areas involved in language processing.

Fig. **??** shows a box plot diagram of the distribution of the STGI metric for the reconstructed spectrograms compared to the original audio for each participant. This metric indicates how well portions of speech signals can be understood in noisy environments. Sub-06 again is the participant with the best results, with an average STGI of 0.55, indicating better speech intelligibility, while Sub-09 has the lowest average of 0.45. The wide range of STGI values highlights differences in results between participants, which, as discussed before, could be due to variations in electrode placement, experimental conditions, or other factors.

0.3.2 Comparison with other models

Table 2 presents a performance comparison of the proposed NeuroIncept Decoder model against other well-known models in the literature. The comparison

Table 2: Performance Comparison of NeuroIncept Decoder

Model	PCC	STGI
LR **verwoert2022dataset**	0.7050	-
FCN **band1**	0.8907	0.3947
CNN **band1**	0.8988	0.4839
NeuroIncept Decoder	**0.9179**	**0.5040**

is based on two metrics: PCC and STGI. The Linear Regression (LR) model, as reported by Verwoert et al. **verwoert2022dataset**, achieved a PCC of 0.70, but no STGI was provided. The fully connected network (FCN) model, proposed by Band et al. **band1**, obtained a PCC of 0.89 and an STGI of 0.3947. Similarly, the CNN-based model, also from Band et al. **band1**, demonstrated a PCC of 0.89 and a higher STGI of 0.48. The NeuroIncept Decoder outperformed all other models, achieving the highest PCC of 0.91 and the best STGI of 0.50, indicating its superior performance in both metrics.

Finally, Fig. 4 shows example spectrograms produced by various models of neural activity along with the original audio recorded from participants. Our model more accurately predicts key speech characteristics, such as formants, while spectrograms generated by other techniques are smoother, suggesting lower synthesis accuracy.

0.4 Conclusions

This study represents a significant step forward in neural decoding, demonstrating the potential to reconstruct high-quality speech directly from neural activity using advanced deep learning techniques. Our approach employs high-gamma features extracted from invasive EEG data and maps them to logMel audio features using a novel neural network architecture- the NeuroIncept Decoder. Our technique achieved high correlations between predicted and actual audio spectrograms, with values ranging from 0.83 to 0.94, highlighting its robustness in capturing key audio features. However, the modest STGI values achieved by our method indicate room for improvement in modeling the intricate temporal dynamics of speech processing in the brain.

Future research will explore pretraining strategies for deep learning models trained on limited data, particularly approaches that utilize EEG signals linked to words or phrases without paired audio. This direction aims to enhance the system's ability to generate real-time speech from neural data, further advancing the practical applications of neural decoding in speech restoration and brain-computer interfaces.

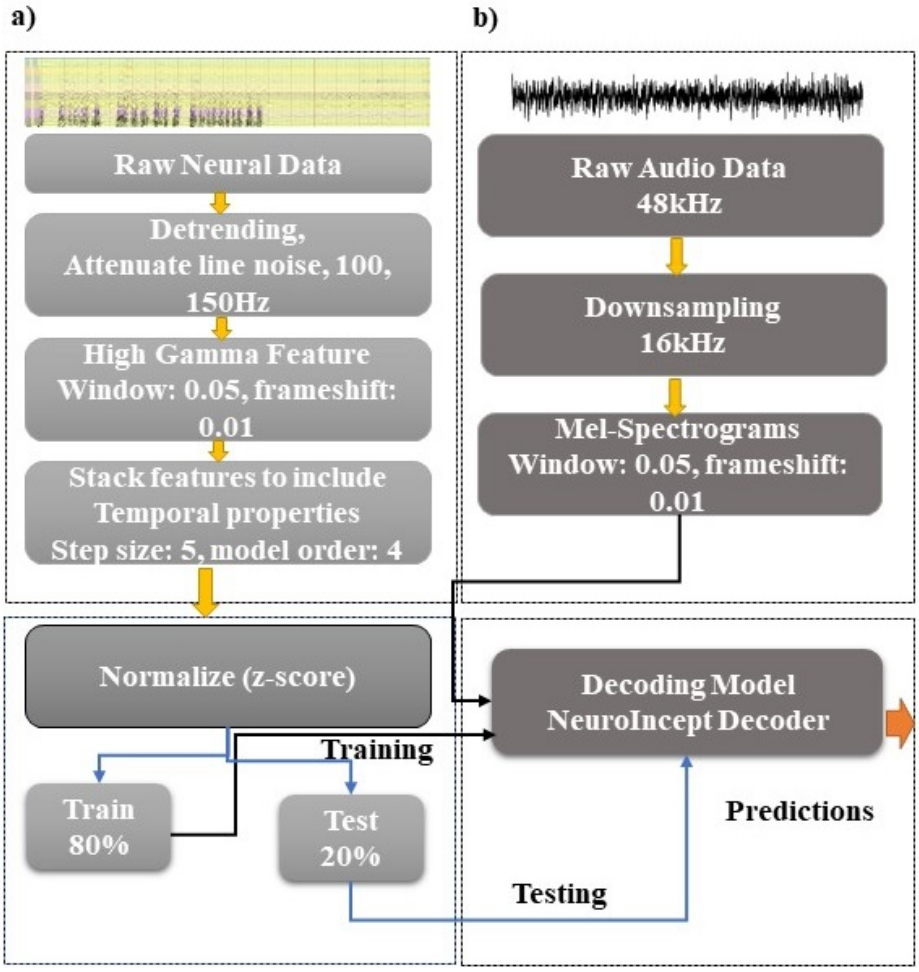

Figure 1: Preprocessing pipeline for the sEEG and audio signals.

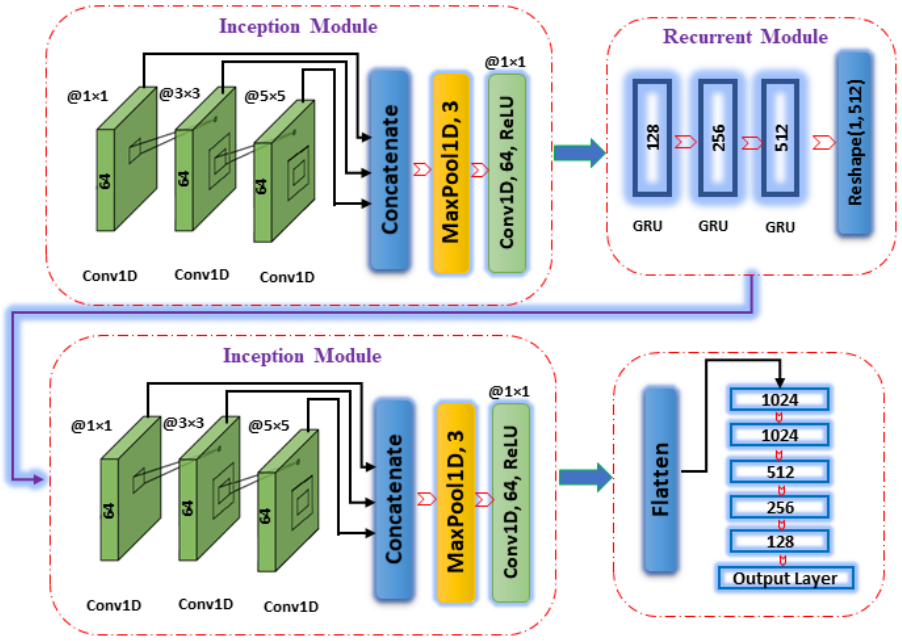

Figure 2: NeuroIncept Decoder model architecture.

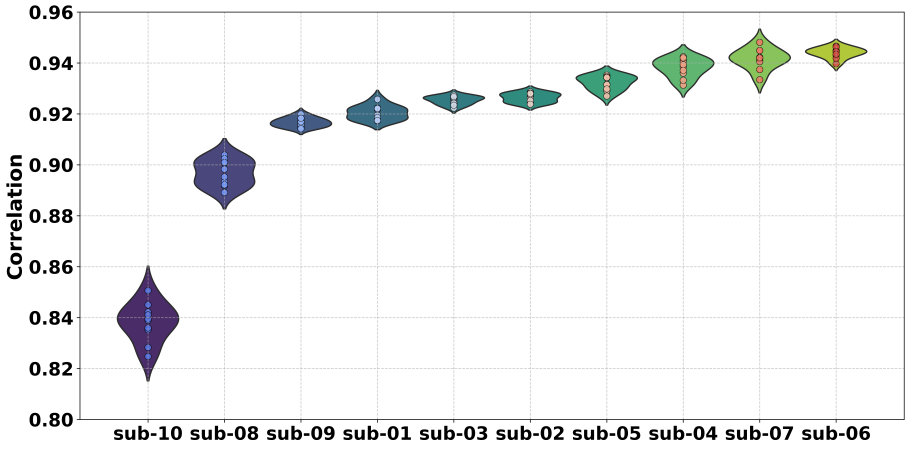

Figure 3: Pearson correlation between predicted and original spectrograms.

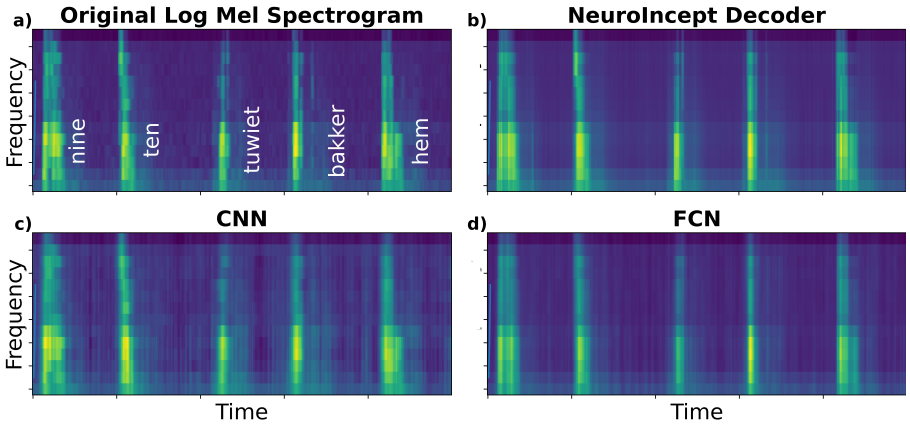

Figure 4: Examples of logMel spectrograms for: (a) natural speech recorded by the participants, (b) speech generated by the proposed NeuroIncept model, (c) the CNN model, and (d) the FCN model. The words are same for all the plots

Bibliography

[1] S. Jullien, "Screening for language and speech delay in children under five years," BMC Pediatrics, vol. 21, no. S1, Sep. 2021, doi: https://doi.org/10.1186/s12887-021-02817-7.

[2] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martin Donas, J. L. Perez-Córdoba, and A. M. Gomez, "Silent Speech Interfaces for Speech Restoration: A Review," IEEE Access, vol. 8, pp. 177995–178021, 2020, doi: https://doi.org/10.1109/access.2020.3026579.

[3] National Institute on Deafness and Other Communication Disorders, "Quick Statistics About Voice, Speech, Language," NIDCD, May 19, 2016. [Online]. Available: https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language. [Accessed: 2024-08-01].

[4] J. Law, James Boyle, Frances Harris, A, "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, Apr. 2000, doi: https://doi.org/10.1080/136828200247133.

[5] Y.-E. Lee, S.-H. Lee, S.-H. Kim, and S.-W. Lee, "Towards Voice Reconstruction from EEG during Imagined Speech," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 5, pp. 6030–6038, Jun. 2023, doi: https://doi.org/10.1609/aaai.v37i5.25745.

[6] M. A. L. Nicolelis, "Brain–machine interfaces to restore motor function and probe neural circuits," Nature Reviews Neuroscience, vol. 4, no. 5, pp. 417–422, May 2003, doi: https://doi.org/10.1038/nrn1105..

[7] M. A. Lebedev and M. A. L. Nicolelis, "Brain–machine interfaces: past, present and future," Trends in Neurosciences, vol. 29, no. 9, pp. 536–546, Sep. 2006, doi: https://doi.org/10.1016/j.tins.2006.07.004.

[8] B. S. Wilson and M. F. Dorman, "Cochlear implants: A remarkable past and a brilliant future," Hearing Research, vol. 242, no. 1–2, pp. 3–21, Aug. 2008, doi: https://doi.org/10.1016/j.heares.2008.06.005.

[9] J. Pearson, "The human imagination: The cognitive neuroscience of visual mental imagery," Nature Reviews Neuroscience, vol. 20, no. 10, Aug. 2019, doi: https://doi.org/10.1038/s41583-019-0202-9.

[10] 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, Jan. 2019, doi: https://doi.org/10.1080/07434618.2018.1556730.

[11] C. Herff et al., "Brain-to-text: decoding spoken phrases from phone representations in the brain," Frontiers in Neuroscience, vol. 9, Jun. 2015, doi: https://doi.org/10.3389/fnins.2015.00217.

[12] D. A. Moses et al., "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," New England Journal of Medicine, vol. 385, no. 3, pp. 217–227, Jul. 2021, doi: https://doi.org/10.1056/nejmoa2027540.

[13] F. R. Willett et al., "A high-performance speech neuroprosthesis," Nature, vol. 620, no. 7976, pp. 1031–1036, Aug. 2023, doi: https://doi.org/10.1038/s41586-023-06377-x.

[14] C. Herff, G. D. Johnson, L. Diener, J. J. Shih, D. J. Krusienski, and T. Schultz, "Towards direct speech synthesis from ECoG: A pilot study," Aug. 2016, doi: https://doi.org/10.1109/embc.2016.7591004.

[15] S. Martin et al., "Decoding spectrotemporal features of overt and covert speech from the human cortex," Frontiers in Neuroengineering, vol. 7, May 2014, doi: https://doi.org/10.3389/fneng.2014.00014.

[16] M. Angrick et al., "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," Communications Biology, vol. 4, no. 1, Sep. 2021, doi: https://doi.org/10.1038/s42003-021-02578-0.

[17] F. H. Guenther et al., "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis," PLoS ONE, vol. 4, no. 12, p. e8218, Dec. 2009, doi: https://doi.org/10.1371/journal.pone.0008218.

[18] X. Chen et al., "A neural speech decoding framework leveraging deep learning and speech synthesis," Nature Machine Intelligence, vol. 6, no. 4, pp. 467–480, Apr. 2024, doi: https://doi.org/10.1038/s42256-024-00824-8.

[19] Y. Hong, S. Ryun, and Chun Kee Chung, "Evoking artificial speech perception through invasive brain stimulation for brain-computer interfaces: current challenges and future perspectives," Frontiers in Neuroscience, vol. 18, Jun. 2024, doi: https://doi.org/10.3389/fnins.2024.1428256.

[20] B. Accou, J. Vanthornhout, H. V. Hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," Scientific Reports, vol. 13, no. 1, p. 812, Jan. 2023, doi: https://doi.org/10.1038/s41598-022-27332-2.

[21] 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, Apr. 2019, doi: https://doi.org/10.1038/s41586-019-1119-1.

[22] M. Verwoert et al., "Dataset of Speech Production in intracranial Electroencephalography," Scientific Data, vol. 9, no. 1, Jul. 2022, doi: https://doi.org/10.1038/s41597-022-01542-9.

[23] R.J.J.H. van Son, D.M. Binnenpoorte, van, and L. C. W. Pols, "The IFA corpus: a phonemically segmented dutch 'open source' speech database," Data Archiving and Networked Services (DANS), Sep. 2001, doi: https://doi.org/10.21437/eurospeech.2001-484.

[24] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proceedings of the 14th Python in Science Conference, 2015, doi: https://doi.org/10.25080/majora-7b98e3ed-003.

[25] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," Scientific Reports, vol. 9, no. 1, p. 874, Jan. 2019, doi: https://doi.org/10.1038/s41598-018-37359-z.

[26] R. Song et al., "Decoding silent speech from high-density surface electromyographic data using transformer," Biomedical Signal Processing and Control, vol. 80, p. 104298, Feb. 2023, doi: https://doi.org/10.1016/j.bspc.2022.104298.

[27] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015, doi: https://doi.org/10.1109/cvpr.2015.7298594.

[28] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 2017, pp. 1597-1600, doi: 10.1109/MWSCAS.2017.8053243.

[29] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Spectro-temporal modulation glimpsing for speech intelligibility prediction," Hearing Research, vol. 426, p. 108620, Dec. 2022, doi: https://doi.org/10.1016/j.heares.2022.108620.

[30] E. F. Chang, K. P. Raygor, and M. S. Berger, "Contemporary model of language organization: an overview for neurosurgeons," *Journal of Neurosurgery*, vol. 122, no. 2, pp. 250–261, Feb. 2015, doi: https://doi.org/10.3171/2014.10.JNS132647.