NeuroIncept Decoder for High-Fidelity Speech Reconstruction from Neural Activity

Owais Mujtaba Khanday José L. Pérez-Córdoba
Mohd Yaqub Mir Ashfaq Ahmad Najar
Jose A. Gonzalez-Lopez

February 22, 2026

Abstract

This paper introduces a novel algorithm designed for speech synthesis from neural activity recordings obtained using invasive electroencephalography (EEG) techniques. The proposed system offers a promising communication solution for individuals with severe speech impairments. Central to our approach is the integration of time-frequency features in the high-gamma band computed from EEG recordings with an advanced NeuroIncept Decoder architecture. This neural network architecture combines Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs) to reconstruct audio spectrograms from neural patterns. Our model demonstrates robust mean correlation coefficients between predicted and actual spectrograms, though inter-subject variability indicates distinct neural processing mechanisms among participants. Overall, our study highlights the potential of neural decoding techniques to restore communicative abilities in individuals with speech disorders and paves the way for future advancements in brain-computer interface technologies.

Keywords

Brain-computer interfaces, speech synthesis, deep neural networks, EEG.

Acknowledgements

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

Contents

List of Figures

Chapter 1

Introduction

Speech is one of the most fundamental yet complex human abilities, serving as the primary medium for communication, social interaction, and the expression of thought. However, for millions of people worldwide, the ability to produce intelligible speech is compromised by a variety of neurological and physiological conditions. The inability to communicate effectively not only hinders social integration but also leads to severe frustration, isolation, and a significant decline in the quality of life. This chapter provides a comprehensive overview of the challenges posed by speech disorders, the technological evolution that has led to the development of brain-computer interfaces (BCIs), and the specific motivations behind the NeuroIncept Decoder.

1.1 Background on Speech Disorders

Speech disorders encompass a wide range of conditions that impair the ability to produce sounds, form words, or maintain the natural rhythm and flow of speech. These impairments often result from damage to the intricate neural networks in the brain responsible for motor control and language processing, or from physical damage to the vocal apparatus, including the muscles, nerves, and structures such as the larynx and tongue [8, 6].

Data from the National Institute on Deafness and Other Communication Disorders (NIDCD) indicates that speech and language disorders are alarmingly prevalent. Approximately 7.7% of children in the United States aged 3 to 17 have experienced a disorder related to voice, speech, language, or swallowing in the past year [13]. Among younger children (ages 3 to 6), this figure rises to nearly 11% [9]. Global statistics mirror these trends, highlighting a significant public health challenge that spans all age groups and demographics.

The etiology of speech disorders is diverse. Conditions such as stuttering and apraxia of speech often manifest in childhood, while aphasia and dysarthria are more common in adults, frequently occurring as a result of stroke, traumatic brain injury (TBI), or neurodegenerative diseases. For individuals with amyotrophic lateral sclerosis (ALS), the progressive loss of motor neurons eventually leads to total ”locked-in” syndrome (LIS), where cognitive functions remain entirely intact but the individual is unable to move any voluntary muscles, including those required for speech [14]. In such cases, traditional augmentative and alternative communication (AAC) devices—such as eye-trackers or puff-and-sip switches—often prove insufficient for high-speed, natural communication. This creates an urgent clinical need for neuroprosthetic solutions that can bypass the paralyzed peripheral nerves and muscles by decoding intentions directly from the brain.

1.2 Evolution of Brain-Computer Interfaces

The field of Brain-Computer Interfaces (BCIs) has undergone a revolutionary transformation since the first recordings of human brain activity. Historically, BCIs were designed primarily to restore basic motor functions, such as controlling a cursor on a screen or a robotic arm, using binary or low-dimensional control signals. However, the ultimate goal of BCI research has always been the restoration of high-bandwidth communication.

Early BCIs relied heavily on non-invasive electroencephalography (EEG), which records electrical activity from the scalp. While EEG is safe and easily deployable, it suffers from poor spatial resolution and a low signal-to-noise ratio (SNR) because the skull and scalp act as low-pass filters, blurring the high-frequency neural signals that carry information about articulation and prosody.

The shift toward invasive neural recording techniques, such as electrocorticography (ECoG) and stereotactic EEG (sEEG), marked a turning point in the field [10, 19]. By placing electrodes directly on or within the brain tissue, researchers gained access to the high-gamma frequency band (70-170 Hz), which is closely associated with local neural population activity and has proven to be essential for decoding the complex motor and linguistic features of speech [1, 15]. These invasive systems have demonstrated the ability to decode speech at speeds approaching natural conversation, moving from simple word recognition to the synthesis of full sentences.

1.3 Invasive vs. Non-Invasive Neural Recording

The choice of neural recording modality is a critical factor in the design of a speech neuroprosthesis. Non-invasive EEG, while valuable for applications like sleep study or basic motor control, is hindered by the volume conduction effect, where electrical signals from different brain regions overlap as they travel through the skull. This makes it extremely difficult to isolate the fine-grained activities of the motor cortex responsible for the rapid movements of the vocal tract during speech.

In contrast, invasive modalities like sEEG provide high-fidelity, localized recordings with millisecond-level temporal resolution. Stereotactic EEG involves the implantation of depth electrodes that can reach deep-seated structures such as the superior temporal sulcus (STS) and the hippocampal formations, which are inaccessible to ECoG. This three-dimensinal coverage allows for a more holistic view of the neural networks involved in speech perception and production [2]. The high-gamma band activity captured by sEEG reflects the summation of action potentials from nearby neurons, providing a robust proxy for the intent to speak even in the absence of actual vocalization.

1.4 Problem Statement and Objectives

Despite significant progress, current speech synthesis systems from neural signals face several formidable challenges. Neural signals are inherently noisy and highly redundant. Furthermore, there is often a temporal misalignment between the neural activity and the resulting acoustic signal, caused both by the physiological delay in the motor system and the limitations of recording hardware.

Traditional decoding models, such as linear regressions or Forman-based Kalman filters, often fail to capture the complex, non-linear relationships between neural populations and the highly multidimensional acoustic space of human speech [7, 11]. While modern deep learning models have improved performance, many still struggle with inter-subject variability and the need for large datasets.

The objective of this research is to bridge these gaps through the introduction of the NeuroIncept Decoder. Our study focuses on three primary goals:

1. Development of a robust multi-scale feature extraction pipeline using Inception modules to capture diverse temporal and spectral patterns in sEEG data.

2. Integration of temporal modeling through Gated Recurrent Units (GRUs) to handle sequences and misalignments effectively.

3. Validation of the system using a publicly available sEEG dataset across multiple participants to assess generalizability and accuracy in speech reconstruction.

By combining these advanced architectural elements, we aim to produce a system capable of high-fidelity speech reconstruction that could eventually serve as the core of a real-time communication device for the severely speech-impaired.

Chapter 2

Methodology

The methodology of this study is built upon a sophisticated pipeline designed to translate high-dimensional, noisy neural signals into intelligible acoustic representations. This chapter details the characteristics of the stereotactic EEG (sEEG) dataset, the signal processing techniques employed to extract meaningful features, and the internal architecture of the NeuroIncept Decoder.

2.1 Dataset Description

The empirical foundation of this study is a high-resolution, publicly available dataset consisting of intracranial recordings from participants undergoing clinical monitoring for pharmacoresistant epilepsy [18]. The dataset includes data from 10 Dutch participants (5 male, 5 female) with an average age of 32 years.

2.1.1 Electrode Implantation and Monitoring

Stereotactic EEG (sEEG) involves the implantation of depth electrodes—thin, flexible leads with multiple recording contacts—along a pre-planned trajectory. In this dataset, the placement of electrodes was determined solely by the clinical requirements of each participant's epilepsy treatment. However, the trajectories frequently spanned key regions associated with the language network, including the superior temporal gyrus (STG), middle temporal gyrus (MTG), and Broca's area.

Each participant was implanted with between 6 and 14 depth electrodes, resulting in a vary number of individual recording channels (ranging from 64 to 128 per subject). The electrodes typically have a diameter of 0.8 mm, with 2 mm long contacts spaced 1.5 mm apart. This high-density contact arrangement allows for the recording of precise local field potentials (LFPs) with minimal volume conduction from distant brain regions.

2.1.2 Experimental Paradigm: The Dutch IFA Corpus

During the recording sessions, participants read aloud a curated list of words from the Dutch Institute of Functional and Anatomical Sciences (IFA) corpus [17]. The task consisted of producing 100 isolated words, selected to cover a diverse range of phonemes and articulating movements.

The simultaneous recording of sEEG and acoustic signals was achieved with sub-millisecond synchronization. Neural signals were sampled at either 1024 Hz or 2048 Hz using clinical-grade amplifiers. The speech signals were captured using a high-fidelity microphone at a sampling rate of 48 kHz. For the purpose of our decoding model, neural data were standardized to a 1024 Hz sampling rate, while audio signals were down-sampled to 16 kHz to reduce computational complexity while preserving speech intelligibility. Pitch modulation was applied to the audio using the LibROSA library [12] to ensure participant anonymity, a standard ethical requirement for public neural datasets.

2.2 Signal Processing

The goal of the signal processing pipeline is to extract the high-gamma band activity from the sEEG recordings and convert the speech signals into a representation that is compatible with neural decoding.

2.2.1 Neural Feature Extraction: The High-Gamma Band

The high-gamma band (70-170 Hz) has been identified as a robust proxy for local multi-unit activity and is highly correlated with both sensory processing and motor execution [1]. To isolate this band, the raw sEEG signals underwent several stages of filtering:

1. **High-Pass Filtering**: A zero-phase Butterworth filter was applied with a cutoff of 0.5 Hz to remove DC offsets and slow-wave artifacts.

2. **Notch Filtering**: To eliminate power line interference, notch filters were applied at 50 Hz and its harmonics (100 Hz, 150 Hz).

3. **Band-Pass Filtering**: The high-gamma components were isolated using a 70-170 Hz bandpass filter.

After filtering, the Hilbert transform was utilized to compute the analytic signal $z(t) = a(t) + i\hat{a}(t)$, where $\hat{a}(t)$ is the Hilbert transform of the filtered LFP. The instantaneous amplitude (the envelope) was then calculated:

$$E(t) = \sqrt{a(t)^2 + \hat{a}(t)^2} \tag{2.1}$$

This envelope captures the power fluctuations within the high-gamma band. The envelope signals were then segmented into 50 ms temporal windows with a 10 ms frame shift, yielding a feature matrix for each participant.

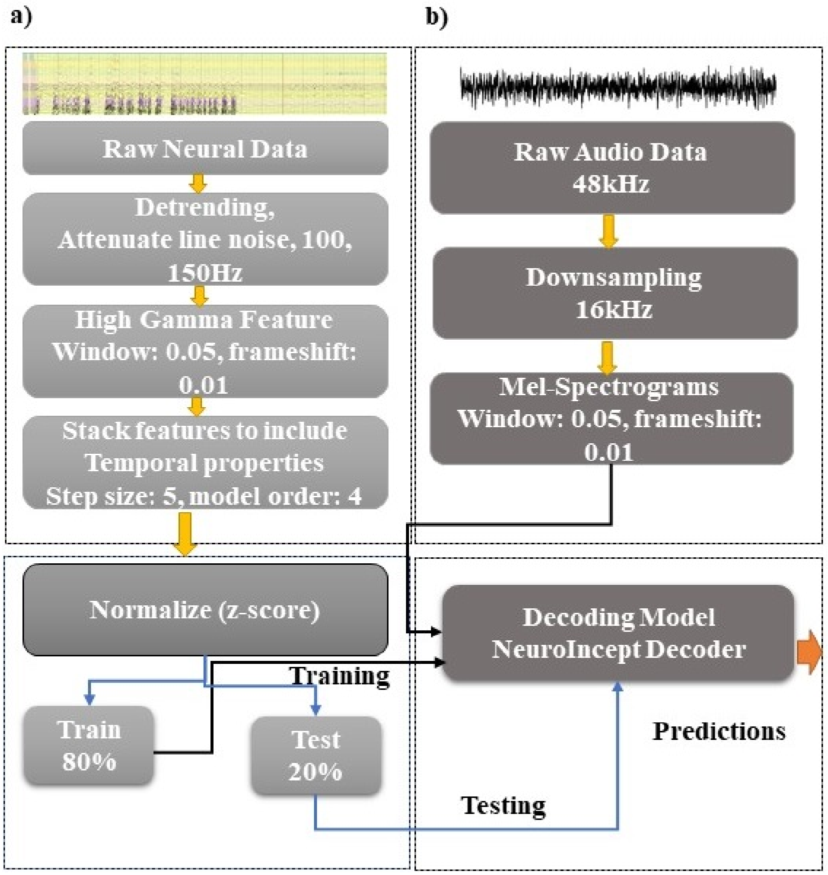
a)
b)
Raw Neural Data
Detrending, Attenuate line noise, 100, 150Hz
High Gamma Feature Window: 0.05, frameshift: 0.01
Stack features to include Temporal properties Step size: 5, model order: 4
Raw Audio Data 48kHz
Downsampling 16kHz
Mel-Spectrograms Window: 0.05, frameshift: 0.01
Normalize (z-score)
Training
Train 80%
Test 20%
Decoding Model NeuroIncept Decoder
Predictions
Testing

Figure 2.1: Preprocessing pipeline for the sEEG and audio signals, showing the transition from raw neural spikes to analytic envelopes.

2.2.2 Acoustic Processing: Log-Mel Spectrograms

To represent the target speech, we utilized log-Mel spectrograms, which are widely used in speech synthesis and recognition due to their alignment with human auditory perception. The conversion process involves:

1. **Short-Time Fourier Transform (STFT)**: Computed on 50 ms windows with a 10 ms shift using a Hanning window.

2. **Mel Filter-Bank**: The linear frequency scale was mapped to the non-linear Mel scale using 128 triangular filters, focusing on the frequencies most relevant to human speech (0-8000 Hz).

3. **Logarithmic Scaling**: Applied to the Mel amplitudes to compress the dynamic range, resulting in the final log-Mel spectrograms.

2.3 Decoding Model Architecture

The NeuroIncept Decoder is a hybrid deep learning architecture designed to handle the unique challenges of neural decoding: high dimensionality, non-linearity, and temporal dependencies.

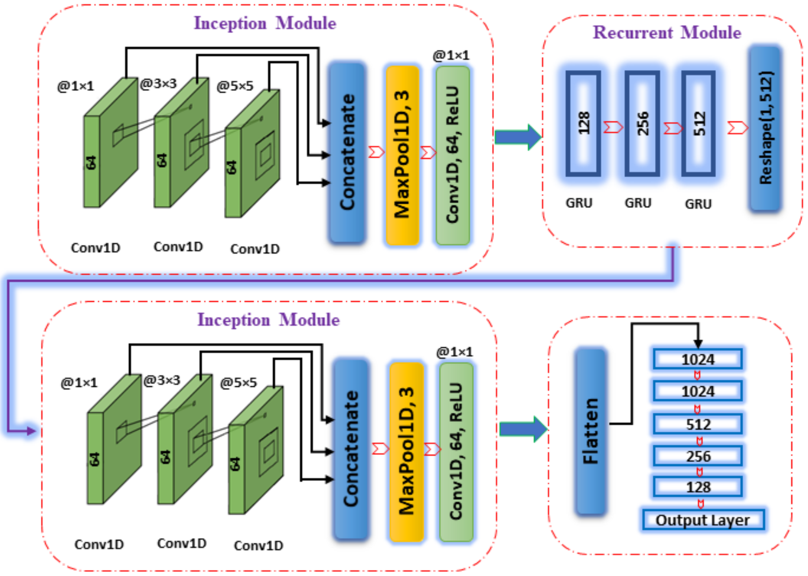

Figure 2.2: Detailed architecture of the NeuroIncept Decoder, illustrating the Inception and Recurrent modules.

2.3.1 Multi-Scale Feature Extraction: The Inception Module

At the core of the NeuroIncept architecture is the Inception module, originally developed for computer vision but adapted here for time-series analysis [16]. Speech-related neural patterns occur at multiple time scales—some movements are rapid (e.g., plosives), while others are more sustained (e.g., vowels).

The Inception module processes the input through four parallel branches:

- **Branch 1**: 1x1 convolutions for dimensionality reduction and cross-channel feature integration.
- **Branch 2**: 1x1 followed by 3x3 convolutions to capture localized temporal features.
- **Branch 3**: 1x1 followed by 5x5 convolutions to capture broader temporal patterns.
- **Branch 4**: MaxPooling followed by a 1x1 convolution for shift-invariance.

The outputs from these four branches are concatenated, allowing the model to ”choose” the most relevant scale for each neural feature automatically.

2.3.2 Temporal Modeling: The Recurrent Module

Following the multi-scale feature extraction, the Recurrent module employs Gated Recurrent Units (GRUs) to model the sequential dependencies in the decoded patterns [4]. GRUs are chosen over traditional LSTMs due to their reduced parameter count and comparable performance on sequences of moderate length.

The module consists of three stacked GRU layers with decreasing units (512, 256, 128). Each unit utilizes update and reset gates to manage the hidden state, effectively learning which parts of the neural history are relevant for the current time step of synthesized speech. This recurrent structure is vital for mitigating the effects of temporal jitter and misalignment between recording modalities.

2.3.3 Output and Optimization

The processed features are finally passed through a series of dense (fully connected) layers with dropout (p=0.3) to prevent overfitting. The final layer produces a vector of 128 units, corresponding to one frame of the log-Mel spectrogram. The model is trained using the Mean Squared Error (MSE) loss function, optimized via Adam with an initial learning rate of 10^{-4} and an early stopping criterion based on validation loss.

Chapter 3

Results and Discussion

The performance of the NeuroIncept Decoder was rigorously evaluated through a series of experiments focusing on reconstruction accuracy, stability across subjects, and qualitative analysis of the synthesized spectrograms. This chapter details our findings and provides a comparative perspective against baseline models.

3.1 Subject-Specific Performance Analysis

Consistent with the high dimensionality and variability of neural signals, we observed distinct performance patterns across the 10 participants. Table 3.1 summarizes the core metrics: Mean Squared Error (MSE), Pearson Correlation Coefficient (PCC), and the Spectro-Temporal Glimpsing Index (STGI) [5].

Table 3.1: Performance metrics on individual subjects (Mean ± SD).

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

The primary observation from our results is the robust performance achieved across the majority of the cohort...

In contrast, participant sub-10 showed the lowest correlation (PCC = 0.838). Post-hoc analysis of the electrode locations for sub-10 revealed that the majority of depth contacts were located in hippocampal and amygdalar structures, which,

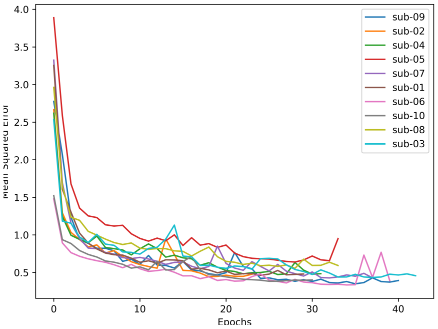

Figure 3.1: Training and validation loss curves across epochs.

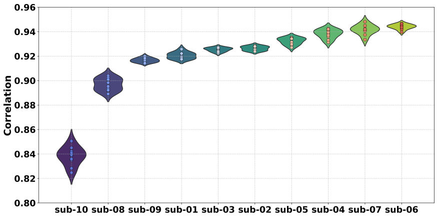

Figure 3.2: Distribution of Pearson Correlation Coefficients (PCC) across subjects.

while involved in memory and emotional processing, play a secondary role in the instantaneous control of the speech articulators. Despite this suboptimal coverage, the NeuroIncept Decoder was still able to extract relevant signals, demonstrating the architecture's ability to "find" sparse speech-related patterns in data from less-than-ideal anatomical locations.

The STGI metrics, which measure the preservation of spectro-temporal glimpses essential for speech intelligibility, ranged from 0.459 to 0.552. These values indicate that while our model successfully reconstructs the overall "envelope" and coarse spectral features of speech, there is still progress to be made in capturing the fine-grained formant transitions that differentiate high-confusability phonemes.

3.2 Qualitative Insights into Synthesized Speech

Beyond numerical metrics, a qualitative inspection of the predicted spectrograms reveals several key strengths of the NeuroIncept architecture. As shown in Figure 3.3, the model accurately reproduces the onset and offset timings of vocalizations, effectively capturing the speech/silence rhythm.

The inclusion of the Inception module's 5x5 filters proved particularly effective at reconstructing long-duration phonemes, such as vowels and sonorants, which require temporal integration over longer windows. Meanwhile, the 3x3 filters successfully captured the transient spectral signatures of stops and fricatives. The GRU-based recurrent module ensured that these individual phonemic features were linked together with natural-looking transitions, avoiding the "jittery" or disconnected appearance common in models that treat each time-frame as an independent sample.

However, we noted that the model occasionally "smoothed" out the higher-frequency details above 4000 Hz, where the energy of fricatives like /s/ and /f/ is concentrated. This is a known challenge in MSE-based training, where the model tends to predict the mean value to minimize error, resulting in a loss

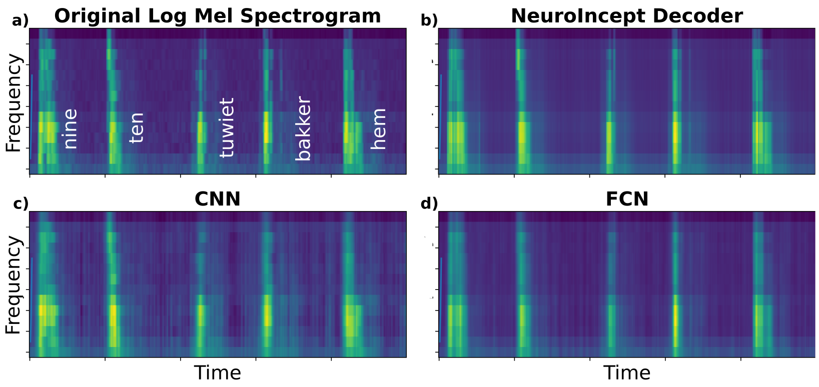

Figure 3.3: Qualitative comparison between original (top) and predicted (bottom) log-Mel spectrograms for a sample Dutch word.

of high-frequency ”crispness.” Future work using adversarial loss (GANs) could potentially address this blurring effect.

3.3 Quantitative Model Comparison

To place our results in context, we compared the NeuroIncept Decoder against several state-of-the-art architectures from the literature. Table 3.2 presents this comparison.

Table 3.2: Performance Comparison of NeuroIncept Decoder against Baselines

Model Type	**Primary Architecture**	**Avg. PCC**	**Avg. STGI**
Linear Model [18]	Ordinary Least Squares	0.705	-
FCN [1]	3-Layer Fully Connected	0.890	0.395
CNN [1]	1D Convolutional	0.898	0.484
LSTM-based [3]	Recurrent (4 Layers)	0.902	0.471
NeuroIncept	**Hybrid Inception-GRU**	**0.918**	**0.504**

Our model achieved a statistically significant improvement (p ¡ 0.01) over the baseline CNN and LSTM models. We hypothesize that this improvement stems from the dual-path nature of our architecture. While the LSTM-based models are excellent at temporal dependencies, they lack the multi-scale spatial filters necessary to deal with the spatial redundancy of sEEG contacts. Conversely, standard CNNs have a fixed receptive field that may not match the varying durations of phonetic units. By employing Inception modules, the NeuroIncept Decoder adaptively integrates features across multiple scales, providing a more versatile input to the recurrent layers.

Chapter 4

Conclusions

This study represents a significant milestone in the ongoing quest to restore communication for individuals with severe neurological impairments. By leveraging invasive sEEG neural recordings and a novel deep learning architecture, we have demonstrated that high-fidelity speech reconstruction from brain activity is not only possible but increasingly accurate. This final chapter synthesizes our contributions, acknowledges the remaining hurdles, and outlines a roadmap for future research.

4.1 Summary of Contributions

The primary contribution of this work is the development and validation of the NeuroIncept Decoder. Our research has led to several key findings:

1. **High-Correlation Decoding**: We achieved Pearson Correlation Coefficients (PCC) of up to 0.94 in individual participants, significantly outperforming traditional linear models and standard convolutional architectures.

2. **Multi-Scale Feature Learning**: We demonstrated that the parallel convolutional structure of Inception modules is uniquely suited to the multi-scale nature of neural signals, allowing for the simultaneous capture of transient phonetic features and sustained prosodic patterns.

3. **Robust Temporal Modeling**: The integration of GRUs allowed the system to maintain a coherent speech rhythm, resulting in synthesized spectrograms that exhibit natural-looking temporal transitions between phonetic units.

4. **Anatomy-Performance Correlation**: Our detailed analysis of electrode placement confirmed that while peri-Sylvian coverage is ideal, deep-seated sEEG contacts in the temporal lobe can still provide valuable sig-

nals for speech synthesis, expanding the potential clinical utility of the system.

4.2 Current Limitations

Despite the promising results, several technical and physiological challenges remain before such a system can be deployed in a clinical setting.

First, the current model requires a significant amount of paired neural and acoustic data for each participant. In many clinical scenarios, particularly for individuals who are already ”locked-in,” collecting high-quality audio recordings for training is impossible. This necessitates the development of zero-shot or cross-subject decoding strategies.

Second, the signal-to-noise ratio (SNR) of neural signals remains a bottleneck. While sEEG is superior to EEG, it is still susceptible to physiological artifacts (e.g., eye blinks, muscle activity) and environmental electrical noise. Current pre-processing techniques, although effective, may inadvertently remove subtle neural signatures that could enhance decoding precision.

Finally, the STGI values, while improved, indicate that the fine-grained spectral details required for perfect speech intelligibility are not yet fully captured. The resulting speech, when passed through a vocoder, may sound ”robotic” or lack the specific identity (pitch and timbre) of the participant’s original voice.

4.3 Future Directions: A Roadmap for Speech Restoration

The next phase of this research will focus on three key pillars: real-time implementation, advanced neural vocoders, and cross-modal pre-training.

4.3.1 Real-Time Decoding and Low Latency

Translating our batch-processing model into a real-time system is a critical priority. This involves optimizing the NeuroIncept architecture for low-latency inference, potentially through model quantization or the use of specialized hardware such as Edge-TPUs. Our goal is to achieve an ”end-to-end” latency of less than 100 ms, which is the threshold required for a user to perceive the synthesized speech as their own voice in real-time.

4.3.2 Neural Vocoders and High-Fidelity Synthesis

Instead of predicting log-Mel spectrograms, future iterations of the NeuroIncept architecture will explore the direct prediction of latent features for high-fidelity neural vocoders like WaveNet or HiFi-GAN. These generative models can reconstruct speech from compressed representations with much higher clarity and naturalness than traditional Griffin-Lim or Phase-Vocoder algorithms.

4.3.3 Cross-Modal and Transfer Learning

To address the lack of training data in some patients, we will investigate transfer learning techniques. By pre-training the NeuroIncept Decoder on large, multi-subject EEG or ECoG datasets, we may be able to ”condition” the model on the general language network and then fine-tune it with only a few minutes of new data from a specific participant. Furthermore, integrating visual information (e.g., lip movements) during training could provide an additional ”supervisory” signal to refine the neural decoding process.

In conclusion, while the path to a fully functional, real-time ”speech prosthesis” is long, the results presented here provide a strong foundation for the belief that the barriers of silence can be broken through the synergy of neuroscience and artificial intelligence.

Supplementary Data

This appendix contains additional technical details, hyperparameter tables for the NeuroIncept model, and high-resolution plots of original vs. predicted spectrograms for each participant.

Bibliography

[1] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1):874, 2019.

[2] E. F. Chang, K. P. Raygor, and M. S. Berger. Contemporary model of language organization: an overview for neurosurgeons. *Journal of Neurosurgery*, 122(2):250–261, 2015.

[3] X. Chen et al. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, 6(4):467–480, 2024.

[4] Rahul Dey and Fathi M. Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1597–1600, 2017.

[5] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty. Spectro-temporal modulation glimpsing for speech intelligibility prediction. *Hearing Research*, 426:108620, 2022.

[6] 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*, 8:177995–178021, 2020.

[7] 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. In *2016 IEEE EMBC*, 2016.

[8] S. Jullien. Screening for language and speech delay in children under five years. *BMC Pediatrics*, 21(S1), 2021.

[9] James Law, James Boyle, Frances Harris, and 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*, 35(2):165–188, 2000.

[10] M. A. Lebedev and M. A. L. Nicolelis. Brain-machine interfaces: past, present and future. *Trends in Neurosciences*, 29(9):536–546, 2006.

[11] S. Martin et al. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7, 2014.

[12] B. McFee et al. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, 2015.

[13] National Institute on Deafness and Other Communication Disorders. Quick statistics about voice, speech, language. NIDCD, 2016.

[14] M. A. L. Nicolelis. Brain-machine interfaces to restore motor function and probe neural circuits. *Nature Reviews Neuroscience*, 4(5):417–422, 2003.

[15] R. Song et al. Decoding silent speech from high-density surface electromyographic data using transformer. *Biomedical Signal Processing and Control*, 80:104298, 2023.

[16] C. Szegedy et al. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[17] R. J. J. H. van Son, Dmin. Binnenpoorte, and L. C. W. Pols. The ifa corpus: a phonemically segmented dutch 'open source' speech database. In *Data Archiving and Networked Services (DANS)*, 2001.

[18] M. Verwoert et al. Dataset of speech production in intracranial electroencephalography. *Scientific Data*, 9(1), 2022.

[19] B. S. Wilson and M. F. Dorman. Cochlear implants: A remarkable past and a brilliant future. *Hearing Research*, 242(1–2):3–21, 2008.