NeuroIncept Decoder for High-Fidelity Speech Reconstruction from Neural Activity

Owais Mujtaba

February 20, 2026

Contents

Chapter 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 . 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. 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. 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.

1.1 Methodology

1.1.1 Dataset Description

The dataset used in this study is publicly available 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 . 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.

1.1.2 Signal Processing

he sEEG signals for each participant were parameterized as time-frequency features extracted from the high-gamma band (70-170 Hz), as shown in Figure . This band was chosen because previous studies have shown that it contains information related to speech and language production and perception 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 subsection 1.1.1. 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.

1.2 Decoding Model Architecture

The NeuroIncept Decoder architecture shown in Figure , 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.

1.2.1 Inception Module

: The Inception Module 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 use uniform filter sizes, which may fail to detect multiscale patterns critical for accurate decoding.

1.2.2 Recurrent Module

This module leverages GRU-based recurrent neural networks to capture temporal dependencies [**?**]. 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.