Hessisches Kompetenzzentrum
für Hochleistungsrechnen

# *Learning Temporal and Spectral Patterns from EEG Spectrograms using Deep Learning*



Project Manager
Maik Pfefferkorn

Researchers
Suli Wang, Konstantin Preußer and Keivan Ahmadi

Principal Investigator
Prof. Dr. Ing. Rolf Findeisen

Project Term
2025 - 2026

Clusters
Lichtenberg II Cluster Darmstadt

Software
Python, PyTorch

Additional Software
CUDA

Institute
Fachbereich Elektrotechnik und Informatik

University
Technische Universität Darmstadt

## Introduction

Understanding the neural mechanisms underlying language production remains a central challenge in neuroscience. In particular, translating neural activity directly into linguistic output, such as words or sentences, requires identifying reliable mappings between dynamic brain signals and structured language representations. Recent advances in electroencephalography (EEG) offer promising opportunities for this endeavor. EEG provides high temporal resolution, enabling the capture of rapid neural dynamics associated with language planning and production. At the same time, progress in signal processing has improved the extraction of informative features from noisy, high- dimensional recordings. This project approaches EEG-to-text and EEG-to-language decoding through the development of machine learning and deep learning models. These models are designed to learn complex temporal and spatial patterns in EEG data and map them onto linguistic representations. By leveraging data-driven architectures capable of modeling sequential and high-dimensional signals, the project aims to establish robust computational frameworks for decoding language directly from non-invasive neural recordings.

## Methods

In the first subproject (Master's thesis by Suli Wang), a hierarchical multi-scale self-distillation network (MSDNet) with test-time adaptation was applied to decode isolated words from intracranial EEG recordings of twelve subjects from the Du-IN

dataset. The objective was to model subject-specific and cross-subject neural dynamics underlying word production. The MSDNet architecture combines three core components: (1) hierarchical multi-scale decomposable mixing to capture neural activity across multiple temporal scales; (2) layer- wise self-distillation to emulate cortical top-down modulation and improve representational stability and robustness; and (3) test-time adaptation, a parameter-free strategy that enables online adaptation to distributional shifts during inference without requiring additional training. The second subproject (Master's thesis by Konstantin Preußer) focused on EEG-to-text decoding at both the word and sentence levels using the ZuCo 1.0 and ZuCo 2.0 datasets. At the word level, several encoder–decoder architectures were designed, trained, and systematically evaluated. The encoder components were implemented using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer- based models to capture spatial, temporal, and contextual dependencies in the EEG signals. Across architectures, multi-layer perceptrons (MLPs) served as decoders for mapping latent neural representations to lexical outputs. At the sentence level, existing approaches to EEG-to-text decoding were reviewed and reimplemented. These models typically combined multi-layer transformer encoders for neural signal representation with pre-trained large language models as text decoders. Building on these frameworks, dedicated training and validation protocols were developed and empirically assessed, including strategies for encoder pre-training to improve generalization. In addition, a refined model architecture was proposed that explicitly incorporates distinct frequency bands of neural activity within the transformer encoder, enabling a more structured representation of oscillatory dynamics relevant for sentence-level decoding.

## Results

The MSDNet with test-time adaptation outperformed established baseline methods for nine of the twelve test subjects. Ablation analyses further demonstrated that the full model configuration, integrating hierarchical multi-scale decomposable mixing, layer-wise self-distillation, and test-time adaptation, yielded superior performance compared to reduced variants. Specifically, performance decreased when employing only multi-scale decomposable mixing, only self-distillation, or their combination without test-time adaptation, underscoring the complementary contribution of all three components. At the word level, decoding accuracy decreases, as expected, with increasing vocabulary size. Across all evaluated architectures, only modest improvements over chance level were observed, with the transformer-based encoder achieving the best overall performance. Moreover, decoding performance varied systematically across lexical categories: content words such as nouns and verbs were more reliably classified than function words (e.g., articles, prepositions, and conjunctions), which carry comparatively less semantic information. At the sentence level, previously reported results from the literature could only be replicated when evaluation was conducted in teacher-forcing mode. In contrast, autoregressive evaluation resulted in a



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN TECHNISCHE UNIVERSITÄT DARMSTADT UNIKASSEL VERSITÄT Philipps Universität Marburg

substantial degradation of performance, leading to outputs that were effectively meaningless. Notably, model performance on genuine EEG input was comparable to that obtained with random input signals. This finding suggests shortcomings in the validation protocols adopted in earlier studies. Specifically, the evaluated models appear to rely predominantly on frequent token transitions and distributional regularities in the training corpus rather than on information contained in the EEG signals. The proposed refined architecture, despite incorporating frequency-specific representations, did not resolve this fundamental limitation.

## Discussion

This work highlights differences between invasive and non-invasive neural decoding. For intracranial EEG, the MSDNet with test-time adaptation achieved measurable word-level decoding across subjects. Ablation studies suggest that multi-scale modeling, self- distillation, and online adaptation contribute to capturing neural dynamics. While these results are encouraging, performance varied across participants and the small dataset limits conclusions about generalization, indicating that further improvements and larger studies are needed to assess broader applicability. In contrast, scalp EEG proved insufficient for reliable word- or sentence-level decoding. Accuracy remained near chance, and sentence generation failed under autoregressive evaluation, suggesting that prior reports likely reflected teacher-forcing or dataset-level patterns rather than true EEG–language mapping. These findings indicate that progress will rely less on model complexity and more on signal fidelity, dataset size, and rigorous evaluation. Future work should focus on high-resolution recordings, and larger and more diverse datasets, while leveraging adaptive deep learning models.

## *Last Update:* 2026-03-30 10:08