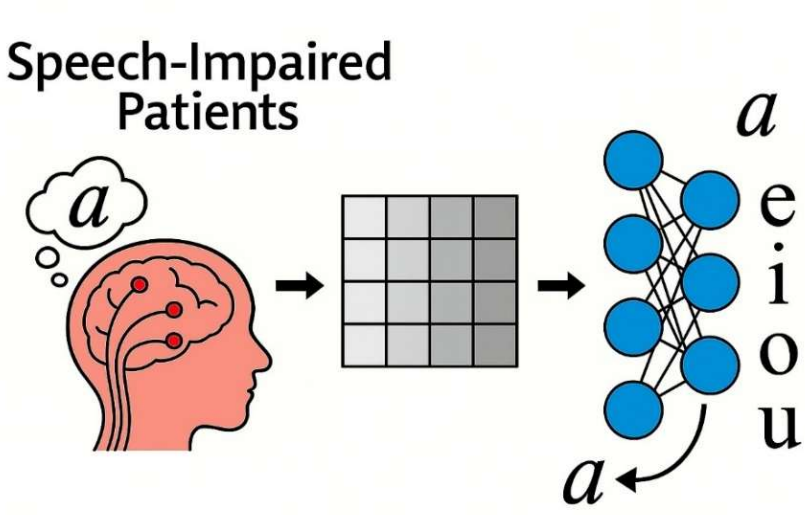
Executive Summary - Decoding of Speech Features from Single Neuron Recordings from the Human Brain

By: Eadan Schechter,
Idan Kanat
Supervisors: Dr. Ariel Tankus
Functional Neurosurgery Unit
Tel Aviv Sourasky Medical Center ("Ichilov"),
Department of Neurology and Neurosurgery, School of Medicine, Tel Aviv University, Tel Aviv, Israel,
Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
Prof. Neta Rabin
School of Industrial & Intelligent Systems Engineering, Tel Aviv University, Israel
October 2024 - July 2025



Executive Summary

Individuals with neurological disorders, e.g. ALS, brain stem stroke or brain injury, may experience significant speech impairments, leaving them unable to communicate even their most basic needs. Brain–computer interfaces (BCIs) offer hope for restoring communication by decoding speech-related electrical signals directly from the human brain. However, a significant technical challenge remains - accurately translating single-neuron recordings into meaningful speech features in real time, across various individuals with diverse neurological conditions and levels of impairment. This project aimed to address this challenge by developing and comparing deep learning (DL) models for offline decoding of vowel articulations directly from the electrical activity of single neurons in the human brain. Our study population consisted of epilepsy patients undergoing invasive neurosurgical monitoring. The ultimate goal is to implement decoders for speech restoration in completely-paralyzed individuals and allow them to speak again.

This project was a collaborative effort between the School of Industrial & Intelligent Systems Engineering at Tel Aviv University and Dr. Tankus from Tel Aviv Sourasky Medical Center (Ichilov). It built on Dr. Tankus' earlier work[1], which demonstrated the feasibility of classifying two vowels from single-neuron recordings in one patient. Our project expanded on this foundation by decoding a broader range of phonemes across multiple patients.

The first steps of the project included exploratory data analysis to examine class balance and firing rate patterns across vowels and patients. Although clear separability was not evident at this stage, the absence of distinct discriminative patterns highlighted the need for more advanced modeling using artificial neural network architectures.

In our project, the input data consisted of articulations of the five English vowels by seven patients. Each such articulation was referred to as a trial, which consisted of a binned firing rate matrix, structured as 20 time bins by C channels, with the number of channels varying across patients according to the clinical implantation of electrodes. This format introduced non-tabular and variable-dimensional input, demanding architectural flexibility and careful modeling design.

Our modeling framework incorporated three distinct deep learning architectures, each adapted for decoding single-neuron recordings. The first was a lightweight convolutional neural network (CNN) based on Tankus et al. (2024) [1], designed to capture local and deeper temporal patterns across neural channels using a series of temporal convolutional layers followed by ReLU activations. The second model was an EEGNet-inspired CNN, adapted from Lawhern et al. (2018) [2], originally developed for EEG classification. We tailored its architecture to our setting using domain-specific convolution operations that extract fine-grained temporal patterns while preserving spatial structure. Batch normalization, Exponential Linear Unit (ELU) activations, average pooling, and dropout were applied for regularization and improving generalization. The third model was a gated recurrent unit (GRU) network adapted from Willett et al. (2023) [3], aimed at capturing sequential

dynamics in multichannel time-series data. It consisted of stacked unidirectional GRU layers, with the final hidden state of the final GRU layer passed to a fully connected layer for classification.

All models were implemented in PyTorch and trained using dropout and the Adam optimizer with Cross-Entropy loss. Furthermore, all models ended with a fully connected layer outputting class logits. These logits were processed by a SoftMax layer via the cross-entropy loss function. Extensive hyperparameter optimization was conducted for each architecture using an inner-loop Optuna-based search, exploring parameters such as dropout rates, hidden dimensions, learning rates, and more. This step was essential to reduce overfitting, and also due to the limited sample sizes and patient-specific variability in the dataset. The search followed best practices and early experimentation to span a diverse, well-reasoned configuration space. Training and validation metrics were tracked using the Weights & Biases API.

To evaluate these models in a consistent and statistically reliable way, we adopted a stepwise training and validation procedure. All models were trained on a patient-specific setting, and initially, on the full five-class classification task using a standard train/validation/test split for each patient. This phase provided a baseline performance profile and enabled us to identify potentially discriminable vowel pairs. Given the limited success on the 5-class task, we refined our approach to focus on simpler, pairwise binary classification problems the CNN models could partially distinguish and trained dedicated binary classifiers on each pair. This targeted strategy allowed us to isolate patient-specific discriminative signals and improve interpretability.

To ensure robust generalization and support statistical inference, we used a Double Cross-Validation framework for the binary classification tasks, which cleanly separated hyperparameter tuning from final evaluation. This approach minimized overfitting and provided unbiased performance estimates.

Our comparative evaluation of three deep learning architectures - CNN, EEGNet-Inspired CNN, and GRU, revealed several consistent patterns in model performance across patients and vowel pairs. First, Convolutional models (i.e: CNN & EEG-Inspired CNN) demonstrated the most reliable and statistically significant performance. In particular, the CNN frequently achieved the highest or near-highest mean accuracy, often accompanied by low p-values ($p < 0.05$), indicating robust vowel pair discrimination. The EEGNet-Inspired CNN also delivered competitive performance, excelling on most reported vowel pairs (9 of 19), notably outperforming the CNN for certain patients.

Second, the GRU model performance varied greatly across patients. While it matched or exceeded CNN accuracy for selected pairs, it failed to converge or to deliver meaningful results for others.

Third, model performance varied substantially across vowel pairs and patients, reflecting both the heterogeneity in neural signal quality and intrinsic differences in vowel pair discriminability. In some cases, classification accuracies exceeded 70% and even 85%, demonstrating the presence of discriminative neural information; in others, test accuracies remained close to chance (50%).

Lastly, none of the models tested emerged as universally optimal; rather, performance varied considerably depending on the specific patient and vowel pair, reinforcing the need for patient-specific modeling and adaptation. Similarly, no single model consistently outperformed the others across all vowel pairs for more than one specific patient.

Our findings underscore the importance of considering individual variability in electrode placement, brain anatomy and physiology. As a result, further analysis and larger datasets will be required to fully account for these sources of heterogeneity and to optimize performance across diverse patient populations.

Overall, this project demonstrates the feasibility of decoding speech features from single-neuron recordings using deep learning - specifically CNN and GRU-based architectures - laying the foundation for future communication-restoring BCIs. Our work extends existing literature by evaluating DL models not previously applied to single-neuron activity recorded in the human brain for speech decoding. Results showed that deep learning methods, particularly convolutional architectures, can effectively support speech neuroprosthesis development in clinical populations. Nevertheless, further validation, refinement, and patient-specific adaptation are required before real-world BCI deployment. These results directly address our project goal of benchmarking deep learning methods for decoding vowel articulations from single-neuron recordings of epilepsy patients.

Based on these findings, we recommend several key directions for future development. First, expanding the collection of single-neuron recordings from a broader and more diverse patient cohort will be essential for improving model robustness and generalizability. Personalized approaches should be prioritized, with future BCI systems adapting model architectures to each patient's unique brain anatomy, physiology, and electrode placements. Finally, comprehensive ethical and data governance frameworks, including protocols for patient consent, privacy, and data ownership, are vital to maintaining public trust and patient welfare. Collectively, these actions will help translate the technical advances demonstrated in our project into practical, patient-centered communication solutions for individuals living with complete paralysis.

References:

1. Tankus, A., Stern, E., Klein, G., et al. (2024). A Speech Neuroprosthesis in the Frontal Lobe and Hippocampus: Decoding High-Frequency Activity into Phonemes. Neurosurgery, 96(2): 356–364
2. Lawhern, V. J., Solon, A. J., Waytowich, N. R., et al. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering, 15(5): 056013.
3. Willett, F. R., Kunz, E. M., Fan, C., et al. (2023). A high-performance speech neuroprosthesis. Nature, 620(7973): 117–124.