UNIVERSITY OF CALIFORNIA

Los Angeles

A Wearable Platform for Decoding
Single-Neuron and Local Field Potential Activity
in Freely-Moving Humans

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Electrical and Computer Engineering

by

Uros Topalovic

2022

ABSTRACT OF THE DISSERTATION

A Wearable Platform for Decoding Single-Neuron and Local Field Potential Activity in Freely-Moving Humans

by

Uros Topalovic
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Dejan Markovic, Chair

Advances in technologies that can record and stimulate deep-brain activity in humans have led to impactful discoveries within the field of neuroscience and contributed to the development of novel closed-loop stimulation therapies for neurological and psychiatric disorders. Human neuroscience research based on intracranial electroencephalography (iEEG) is conducted on voluntary basis during various stages of participant's disease treatment using both external (in-clinic) and implantable systems. In clinical practice, external systems serve as monitoring and testing ground for biomarker extraction and closed-loop neuromodulation, which are, once approved, translated into a compact and low compute resource implantable version for disorder treatment.

External systems allow recordings with fine spatiotemporal resolution at the expense of participant's mobility due to their large size, while implantable devices have reduced recording capabilities and they are not restricted to clinical environment. Due to high transmission and processing latencies across multiple devices, external systems have limited support for

testing computationally expensive online biomarker detection and machine-learning based closed-loop electrical stimulation paradigms including online stimulation programmability.

The motivation for this work comes from the need to extend capabilities of externalized systems, allowing more naturalistic (freely-moving) human neuroscience experiments with fine spatiotemporal resolution. Additionally, externalized systems should provide flexible and local hardware resources that can support real-time and moderately complex embedded neural decoders (biomarker extraction), which in turn could be used to trigger adaptive closed-loop stimulation with low latency. In order to demonstrate initial proof-of-concept technology, this work incorporates: 1. A small versatile neuromodulation platform that can be wearable and lightweight, supporting up to 16 depth electrode arrays; 2. A high-rate ($\sim$4 MB/s on all channels) interfacing of the analog sensing and stimulation front-ends with wearable hardware suitable for embedded machine learning algorithms including artificial neural networks (usually >100M multi-accumulate operations or MACs); 3. A state of the art, performance-driven, neural decoder, small enough to run on an embedded hardware and large enough to generalize across participants; 4. Real-time training and inference with millisecond latency; 5. Closing the loop from the decoder output to the stimulation engines.

Therefore, we developed a wearable, miniaturized, embedded, and external neuromodulation platform built from previously reported integrated circuits for sensing and stimulation, and interfaced with Edge Tensor Processing Unit (TPU) for real-time neural analysis. The Neuro-stack can record and decode single-neuron (32 channels), local field potential (LFP; 256 channels) activity, and deliver highly programmable current-controlled stimulation (256 channels) during stationary and ambulatory behaviors in humans. The TPU Dev Board was chosen because of the ability to perform 2 trillion MACs per second (64 × 64 MAC matrix at 480 MHz) using 2 W of power, with data bandwidth of 40 MB/s. Additionally, the system contains a field-programmable gate array (FPGA) for data pre-processing (filtering, down-sampling) and ARM-based microprocessor (TPU Dev Board) for data management, device control, and secure wireless access point. The Neuro-stack interfaces with the brain through commonly used macro- and micro-electrodes. The Neuro-stack validation includes in-vitro testing of recorded signal quality and measurement of system induced delays (e.g., closed-loop delay from sensing to stimulation site - 1.57 ± 0.19 ms). We provide in-vivo single-unit, LFP, iEEG, and stimulation delivery recorded (2 – 40 channels) from twelve human participants who had depth electrodes implanted for epilepsy evaluation. Among this data are also the first recordings of single-neuron activity during human walking.

To utilize hardware capabilities of the Neuro-stack, we developed a software decoder based on prerecorded human LFP data, which uses TensorFlow artificial neural network (sequential convolutional 1D and recurrent layers) to predict the outcome of a memory task from raw data with higher performance (F1-score 88.6 ± 5.5%) than current state of the art that use shallow machine learning methods (∼70%) under a latency constraint. To shorten the signal processing latency of our decoder, while keeping the accuracy high, the trained and tested model was then ported (coefficient quantization from 16-bit floating-point to 8-bit fixed-point) to the TPU co-processor to make the prediction in real-time on the Neuro-stack. Additionally, we utilized transfer learning approach to update the TPU model with coefficients that were fine-tuned to each participant in real-time. The Neuro-stack decoder was in-vitro validated as part of the time adaptive closed-loop stimulation delivery with pre-configured stimulation current parameters based on the LFP decoder outputs that were predictive of unsuccessful memory encoding. We also used the Neuro-stack to perform human in-vivo real-time binary prediction (69% F1-score) of memory task performance from medial temporal lobe (MTL) regions. Each inference step was executing 193M MACs in 2.8 ms on average, for total round-trip delay of 4.4 ms.

The Neuro-stack is a wearable and versatile neuromodulation platform, able to record and stimulate large number of iEEG and single-unit channels, and process the raw data using artificial neural networks in real-time. These functionalities were not available so far on a single low-latency device. Thus, the Neuro-stack can improve existing or allow completely new research studies. By using the Neuro-stack, researchers could, for example, determine

the neural mechanisms underlying human freely-moving behaviors (e.g., spatial navigation) to identify spatially selective neurons and their modulation by cognition that have been previously discovered only in animals. Also, the Neuro-stack decoder could be used to identify more complex multimodal biomarkers as well as to record and characterize their exact changes under stimulation with known or previously not possible parameters. This could lead to developing novel neuromodulation therapies for patients with brain disorders, while they participate in hospital trials resembling real world environments.

The dissertation of Uros Topalovic is approved.

Jonathan C. Kao

Danijela Cabric

Nanthia Suthana

Dejan Markovic, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my advisor Professor Dejan Markovic and my mentor Professor Nanthia Suthana for their guidance, leadership, patience that made my PhD program a great experience. Their passion and knowledge in the fields of engineering and neuroscience motivated me to pursue a research career in the area of neurotechnology.

Further, I want to thank the committee members: Professor Danijela Cabric and Professor Jonathan C. Kao for their valuable feedback on my research proposal and this dissertation.

I am grateful to my incredible colleagues in the Parallel Data Architecture group and Suthana Lab at UCLA for their willingness to collaborate and share their great knowledge. I thank Dr. Dejan Rozgic, Dr. Wenlong Jiang, Dr. Hariprasad Chandrakumar, Dr. Vahagn Hokhikyan, Dr. Sing Baris-Kazeruni, and Dr. Ahmed Alzuhair, who previously developed ASIC chips used in this work, and Dr. Ahmed Alzuhair, Sam Barclay, and Chenkai Ling for participanting in the development and testing of this work. I would like to thank Dr. Zahra M. Aghajan, Dr. Cory S. Inman, Dr. Matthias Stangl, Sabrina Maoz, and Jay Gill for their help during my translation to neuroengineering and neuroscience fields, as well as for the help with the human in-vivo validation. I am thankful to neurosurgeons Dr. Itzhak Fried, Dr. Aria Fallah, Dr. Ausaf Bari, to a neurologist Dr. Dawn S. Eliashiv, and to clinical research staff Guldamla Kalender, and Natalie Cherry for giving me an oppoutunity to test my work with human participants at the Ronald Reagan Hospital.

A big thank you goes to administrative staff at the deparment of Electrical and Computer Engineering and at the Semel Institute for Neuroscience and Human Behavior: Deeona Columbia, Ryo Arreola, Kyle Jung, Rochelle L. Landicho, and Sonja Hiller for making my journey through graduate program a smooth experience.

Finnaly, I want to thank my family and friends for their continuous support throughout this process, which otherwise would have not been possible.

A version of this work is presented in the preprint publication posted online [1]. At the time of this writing, the work is also being prepared for future publication in the Nature Neuroscience journal.

VITA

2010–2014	B.S., Electrical Engineering University of Belgrade, Serbia
2014–2016	M.S., Electrical Engineering University of Belgrade, Serbia
2014–2015	Embedded Security Engineer Intern Escrypt Inc., MI
2016	Digital ASIC Design Engineer NovelIC, Belgrade, Serbia
2016–2022	Graduate Student Researcher, Electrical and Computer Engineering University of California, Los Angeles

PUBLICATIONS

U. Topalovic, S. Barclay, C. Ling, A. Alzuhair, W. Yu, V. Hokhikyan, H. Chandrakumar, D. Rozgic, W. Jiang, S. Basir-Kazeruni, S. L. Maoz, C. S. Inman, J. Gill, A. Bari, A. Fallah, D. Eliashiv, N. Pouratian, I. Fried, N. Suthana, and D. Markovic, ”A wearable platform for closed-loop stimulation and recording of single-neuron and local field potential activity in freely-moving humans,” *BioRxiv*, 2022. *Nature Neuroscience* – In Review, 2022.

M. Stangl, **U. Topalovic**, C. S. Inman, S. Hiller, D. Villaroman, Z. M. Aghajan, L. Christov-Moore, N. R. Hasulak, V. R. Rao, C. H. Halpern, D. Eliashiv, I. Fried, and N. Suthana.

”Boundary-anchored neural mechanisms of location-encoding for self and others,” *Nature*, vol. 589, 2021.

U. Topalovic, Z. M. Aghajan, D. Villaroman, S. Hiller, L. Christov-Moore, T. J. Wishard, M. Stangl, N. R. Hasulak, C. S. Inman, T. A. Fields, V. R. Rao, D. Eliashiv, I. Fried, and N. Suthana, ”Wireless Programmable Recording and Stimulation of Deep Brain Activity in Freely Moving Humans,” *Neuron*, vol. 108, 2020.

Z. M. Aghajan, D. Villaroman, S. Hiller, T. J. Wishard, **U. Topalovic**, L. Christov-Moore, N. Shaterian, N. R. Hasulak, B. Knowlton, D. Eliashiv, V. R. Rao, I. Fried, and N. Suthana, ”Modulation of human intracranial theta oscillations during freely moving spatial navigation and memory,” *BioRxiv*, 2019.

1 Introduction

Recent advancements in neurotechnology have allowed not only improved diagnosis and monitoring of the brain disorders, but also their successful treatment. Emergence of implantable systems that can record and stimulate cortical or deep brain regions has proven to be effective in treating and evaluating abnormal brain activity in patients with neurologic or psychiatric disorders (e.g., epilepsy, Parkinson's disease, obsessive-compulsive disorder, tremor, etc.; Figure 1.1).

Understanding brain function and its relation to cognition and behavior (Figure 1.1) requires the integration of multiple levels of inquiry, ranging from the examination of single cells all the way up to the probing of human experience under naturalistic conditions. One major barrier that separates these approaches is the inability to record from single neurons during naturalistic behaviors in humans, which frequently involve full-body locomotion as well as twitches, gestures, and actions of the face and hands. This is problematic because behaviors that are studied in animal neurobiology are done almost exclusively in freely-moving animals (e.g., rodents) [2], [3]. Thus, major gaps remain between understanding findings from neuroscience studies in animals to those in humans.

In parallel with progress in neuroscience, the medical field has seen a significant increase in the use and development of therapies delivered through implanted neural devices to treat and evaluate abnormal brain activity in patients with neurologic and psychiatric disorders [4]–[7]. However, current implantable devices do not allow for the recording of single-neuron activity, nor do they allow for extensive customization of stimulation parameters (e.g., pulse

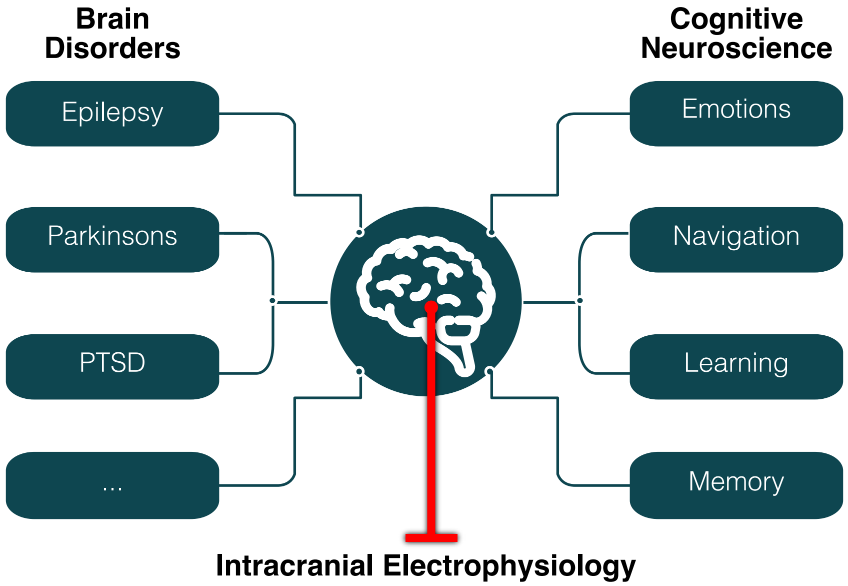


Figure 1.1: Intracranial electrophysiology drives both neuroscience and clinical research/treatments for brain disorders

shape, precise timing with respect to ongoing neural activity), capabilities which would significantly expand the types of research questions that can be investigated. Furthermore, there is a critical need for robust data analytic capabilities on these devices (e.g., using deep learning and artificial intelligence) to deal with the large and complex neural data in real-time. Finally, an additional impediment in developing new responsive neurostimulation treatments is the lack of a customizable bi-directional interface that can record simultaneously with stimulation (full-duplex) and thus "talk" with the brain at the speed of behavior and cognition.

Since neural mechanisms underlying specific behaviors or brain disorders can span across a large population of cells, often from widespread brain regions [8], [9], there is a need for neural devices to record from an increased number of channels across the brain. Further, there is a need for a sufficient temporal scale (< 1 ms) to capture both single-neuron and local field potential (LFP) activity. Importantly, such technology should have a minimal impact on a person's ability to move freely. Current neuroimaging techniques used in humans (e.g., functional magnetic resonance imaging [fMRI], scalp electroencephalography [EEG], magnetoencephalography [MEG]) have insufficient combined spatial and temporal resolution to record single-neuron activity. Intracranial electrophysiological studies, using micro-wire electrodes in epilepsy patients, can record LFPs and single-unit activity, however research participants must be tethered to large equipment and remain immobile. The high spatiotemporal resolution of LFPs (1 – 10 mm, $\geq$ 1 ms) and single-unit (10 – 50 µm, < 1 ms) recordings comes at the cost of brain coverage, which is mitigated, whenever possible, with a larger number of recording channels through clinically-guided implantation of 10 – 15 depth electrodes (i.e., in stereo-EEG [SEEG]).

In this realm, there are two possibilities for neuroscience studies to leverage clinical opportunities where individuals have electrodes implanted in their brains. The first is to use in-clinic research equipment or external systems (Figure 1.2) with immobile participants undergoing clinically indicated SEEG who participate in voluntary research studies while hospitalized. Stimulation research studies are similarly done bedside, primarily using open-loop stimulation [10], [11], although recent studies have begun to explore the use of closed-loop stimulation [12], [13]. Critically, the equipment used in these research studies is expensive (up to $200K), bulky, and does not allow for extensive on-device customization of stimulation or complex real-time analyses for closed-loop stimulation. Using external resources for online processing, however, can increase systems latency to several hundred milliseconds. The second option is to use FDA-approved commercially available neural devices already implanted (Figure 1.2) in several thousand individuals to treat epilepsy and movement disorders. These chronically implanted devices offer research participants mobility at the expense of using large macro-recording electrodes that cannot record single-unit activity, fewer channels (usually 4 bipolar), and lower sampling rates (250 Hz). Implanted systems are primarily developed as medical devices and thus do not offer full control over recordings, stimulation, and real-time processing of the data.

Here, we present a miniaturized bi-directional neuromodulation external device (Neuro-

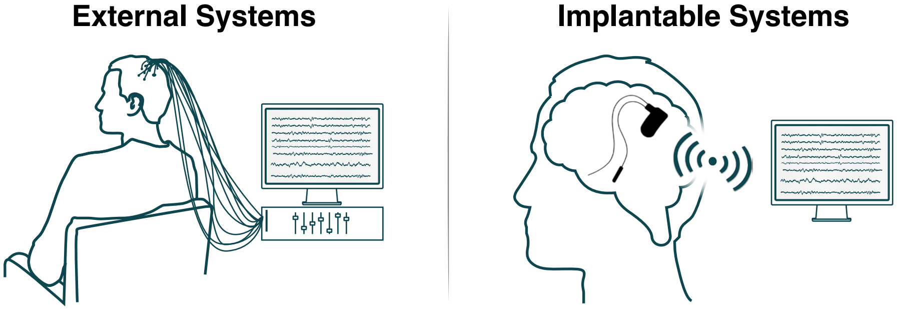


Figure 1.2: External and implantable systems for intracranial recording and stimulation.

stack) that can record up to 256-channel (128 monopolar/bipolar macro-recordings) iEEG and 32-channel single-unit/LFP activity from micro-wires during ambulatory behaviors in humans who have macro- and micro-wire depth electrodes implanted for clinical reasons. It offers a full wireless access and resources for embedded online processing with millisecond latency. Additionally, it can record and stimulate concurrently with highly programmable stimulation current parameters. These capabilities can be useful for future studies investigating the neural mechanisms underlying naturalistic behaviors in humans and developing novel neuromodulation therapies for patients with brain disorders that will be effective in real-world settings.

1.1 Dissertation Outline

- **Chapter 1: Introduction**. Describes the motivation behind neuromodulation systems and its impact on neuroscience and clinical care of patients with brain disorders.
  .

- **Chapter 2: Background** provides information about the human intracranial electrophysiology, existing recording and stimulation systems, and methods for biomarker extraction. The focus is on technologies with the ability to record human single-neuron and local field potential (LFP) activity. Then, we emphasize the importance of real-time processing and decoding of the neural activity in closed-loop systems that use deep brain stimulation. Finally, through prior work review, we identified the inability of the current devices to record wide-band neural activity during human freely-moving behaviors or to use computationally expensive neural processing algorithms in time-sensitive closed-loop experiments. Key challenges and requirements for an end-to-end solution that addresses the problem are presented.

- **Chapter 3: Proposed Sensing and Stimulation Device** chapter describes miniaturized wearable platform for intracranial sensing and stimulation complemented with wireless Wi-Fi capabilities and the TPU accelerator for online decoding of the neural activity. We describe the device and an embedded implementation of the real-time pipeline. This chapter outlines software and hardware design choices driven from the need for a practical device that could be used in clinical setting with human participants. Security of the wireless communication and interfaces to all commonly used electrodes are covered in this chapter. We also discuss use cases of the proposed system in an in-vivo experimental environment. Finally, we compare the Neuro-stack to other existing devices and discuss the choice of the TPU compared to other processing units with emphasis on performance and latency. In-vitro results are presented in this chapter.

- **Chapter 4: Proposed Neural Activity Decoder** chapter provides a software implementation of the artificial neural network model used to predict human memory performance from medial temporal lobe channels, including details about the experiment, neural dataset, training, and testing. This chapter further describes decoder translation into an online embedded version, which could run on the hardware described in Chapter 3. We describe a real-time transfer learning operation, which was used to fine-tune the model for each individual from which data were recorded. Finally, we compare this implementation to existing solutions. In-vitro results are presented in this chapter, which also include closed-loop testing based on the prerecorded LFP data and decoder outputs.

- **Chapter 5: Human In-Vivo Validation** chapter presents a human in-vivo ability of the complete system to record wide-band (single-unit and LFP) neural signals and to extract relevant correlations between the data and behavior using the decoder from Chapter 4. This chapter includes data acquired from twelve human participants implanted with depth electrodes for clinical epilepsy monitoring during resting state, stationary, and ambulatory behavioral tasks, which further validated proposed system in an actual clinical environment.

- **Chapter 6: Discussion** chapter covers several open topics not described in development and testing sections. Topics include further justification for the Neuro-stack development in the context of current research protocols and available systems. It also briefly describes the process and challenges of developing a complex hardware and software system in an academic environment, and its reproducibility.

- **Chapter 7: Conclusion** summarizes key results and contributions of the dissertation. Here, we discuss how proposed device and algorithm can help advance neuroscience research as well as to how it can be used for testing novel therapies for brain disorders. Finally, future work is discussed.

2 Background

2.1 Electrophysiology

Researchers and clinicians have been using various neuroimaging techniques in humans such as functional magnetic resonance imaging (fMRI), scalp electroencephalography (EEG), magnetoencephalography (MEG) (Figure 2.1). Each technique provides different variable for observation. The focus of this work is intracranial electroencephalography (iEEG), which allows higher spatiotemporal resolution and thus deeper regions of the brain may be examined. Although iEEG allows for recording activity within specific deep brain structures, previously listed techniques remain prominent methods to probe the human brain for both research and clinical care, as they are more readily available due to their non-invasive nature.

The type of the electrophysiological signal acquired depends on the design of the electrode and characteristics of the recording system. Electrophysiological signals represent measurements of extracellular field potential at sub-millisecond temporal resolution, but with varying spatial scales. For example, scalp EEG contact usually records superposition of the field across 1 – 10 cm, iEEG across 1 mm – 1 cm, and Single-unit recordings capture one to several neurons at the scale of 10 µm – 100 µm (Figure 2.1).

There are numerous different electrodes available (e.g., subdural strip, grid, depth electrodes) that are designed for a particular type of signals or the brain regions of interest. For example, strip electrodes that can record cortical regions are quite different in both geometry and electrical properties from the depth ones (Figure 2.2). All of these electrodes, however,

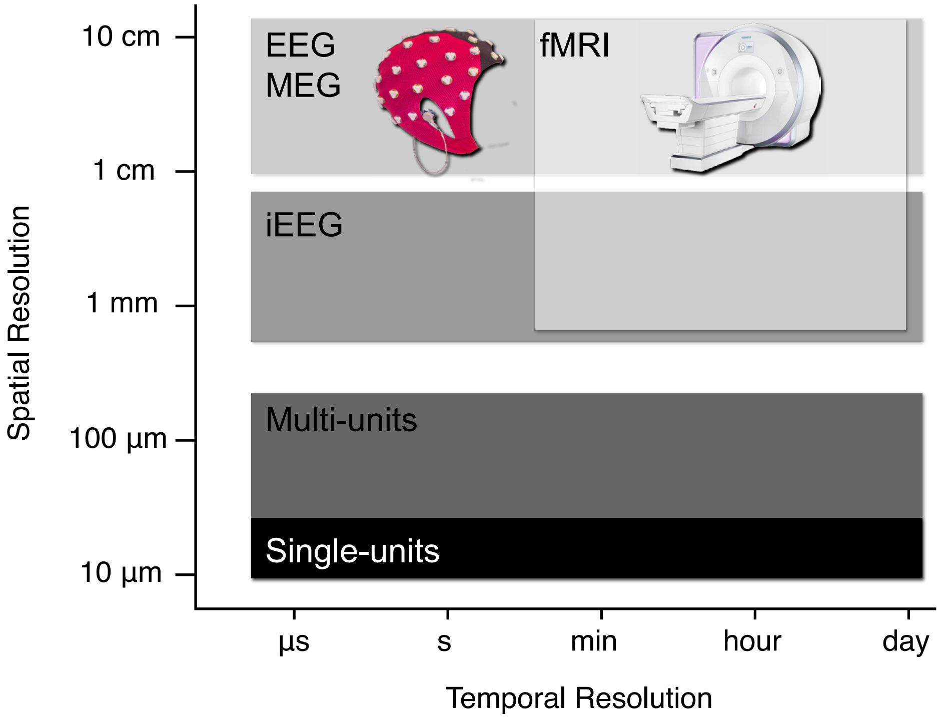


Figure 2.1: Spatiotemporal resolution of neural signals. Spatiotemporal comparison of different electrophysiological measurements and fMRI (Adapted from [14]).

contain macro-contacts with size on the order of millimeter with different spacing (∼mm – ∼cm). Here and for the rest of the text, we are only interested in depth electrodes, or more specifically Behnke-Fried macro-micro electrodes [15], [16] (Figure 2.2-bottom). This is because we were focused on cognitive functions within MTL regions, that could only be reached by macro- and macro-micro depth electrodes.

Micro-contacts of the Behnke-Fried electrode are located at the tips of isolated platinum-iridium micro-wires, which are coming out from the tip of the macro-electrode. Isolation is removed from each micro-wire tip, leaving contact area of 40 µm diameter. Such small contact area allows recording at submillimeter scale from very small neuronal population, from which single-units or action potentials from a single neuron can be isolated. Micro-wire bundle contains 8 wires for neural recordings and the 9th wire for reference use. The 9th wire is completely uninsulated. Macro portion of the Behnke-Fried depth electrode contains 4 – 12 contacts (usually 8).

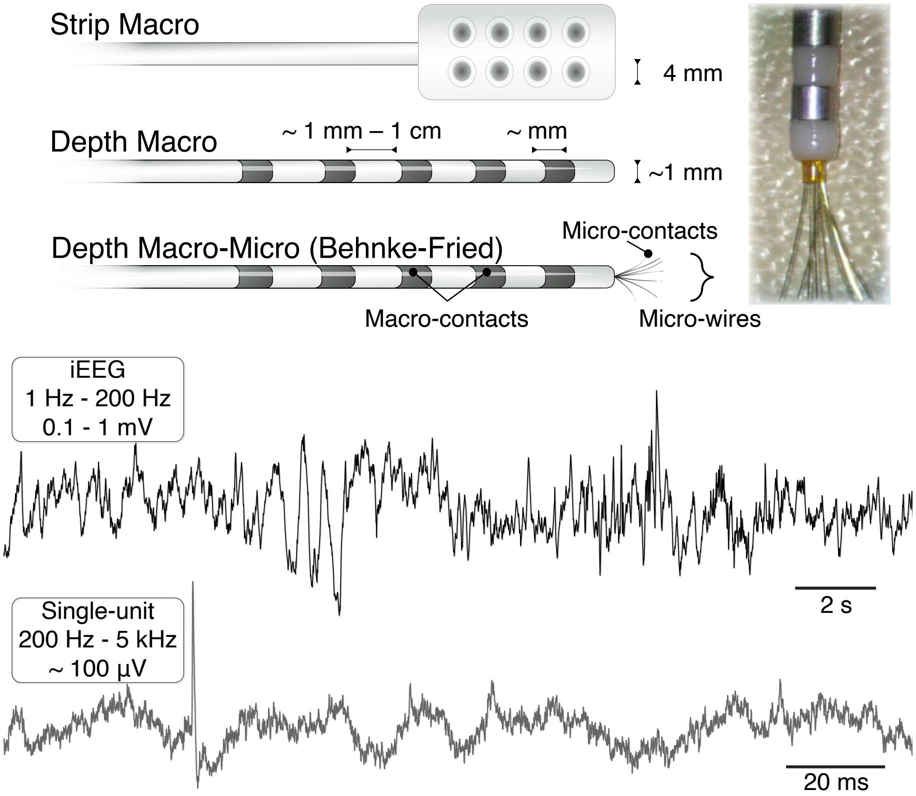


Figure 2.2: Depth electrodes needed to reach MTL. Illustration of subdural strip electrode (top), depth macro-electrode (middle), and Benhke-Fried electrode that includes macro- and micro-contacts (bottom). iEEG activity is obtained through macro-contacts (frequency range: 1 Hz – 200 Hz; voltage range: 0.1 mV – 1 mV). Single-unit activity is obtained through micro-contacts (frequency range: 200 Hz – 5 kHz; voltage range: ∼ 100 µV).

iEEG signals are acquired from macro-contacts, usually at sampling frequency of 1 kHz – 2 kHz, and are then downsampled to the band of interest, 1 Hz – 200 Hz (Figure 2.2). Single-unit signals are acquired from micro-contacts, usually at sampling frequency of 30 kHz, and are then downsampled to the band of interest, 200 Hz – 5 kHz (Figure 2.2). Recordings can be monopolar between selected micro/macro-contact and the reference or bipolar between any two micro/micro-contacts. Due to electrode and contact geometry, which affect electrode-tissue resistance and capacitance [17], macro-contacts cannot capture single-unit recordings even if sampled at 30 kHz, but are intended for iEEG recordings of large population of neurons. Neural signals obtained from micro-contacts at less than 200 Hz are often called local field potentials (LFP).

In literature, signal acquisition from depth electrodes is also called stereo electroencephalography (SEEG), which combines macro iEEG and micro single-unit/LFP recordings. To avoid confusion, strip or grid electrodes record electrocorticography (ECoG), which is cortical equivalent to depth iEEG [18].

2.2 Intracranial Sensing and Stimulation of Human Behavior

Once depth electrodes are implanted (Figure 2.3-left), neuroscientist have the option to record iEEG data with clinical equipment, already present in the hospital (e.g., Nihon Kohden; Figure 2.3-middle) for disease monitoring. Single-neuron research requires different devices able to record at high sampling rates from a high number of channels (e.g., Blackrock Microsystems or Neuralynx; Figure 2.3-right). These research-only devices are not part of the clinical treatment protocol and have to be acquired at the expense of researcher. In the rest of the text we will call these devices external systems.

Another option that is becoming more and more popular is the use of medical implantable systems for research purposes. These devices are designed for clinical treatment via closed-loop stimulation of patients during their regular day to day lives out of the hospital. Implants, however, can only offer iEEG or ECoG, but not single-unit nor LFP activity. In the rest of the text we will call these devices implantable systems.

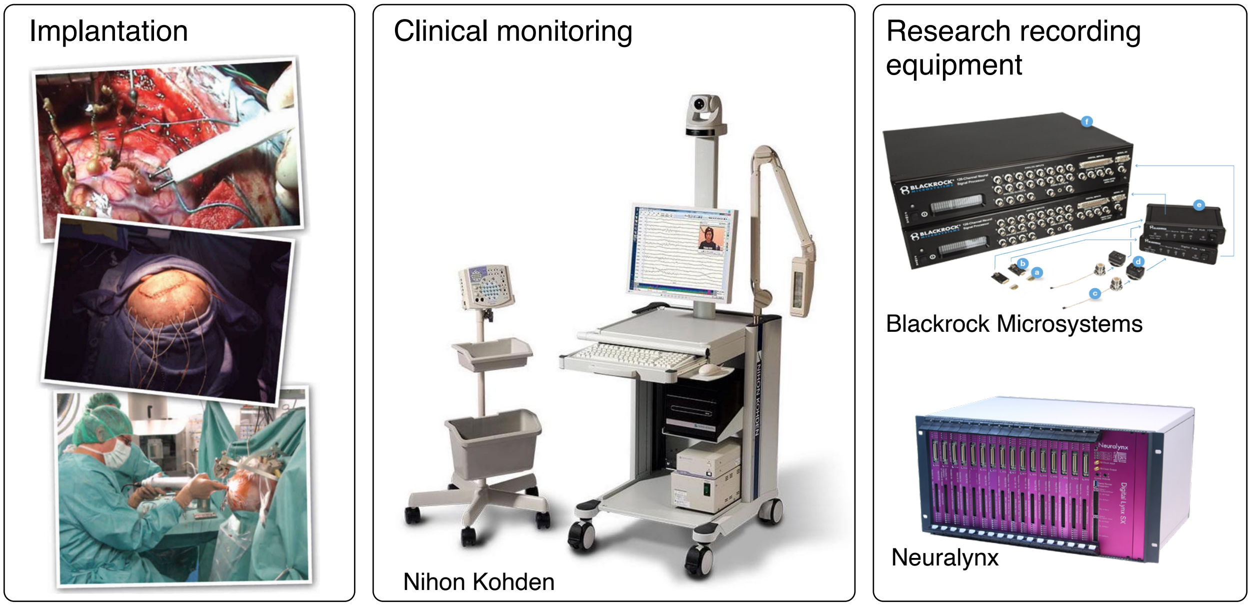


Figure 2.3: Examples of clinical and research external systems.

2.2.1 External Systems

External systems or in-clinic research equipment (e.g., Blackrock Microsystems [19], Neuralynx [20], Nihon Kohden [21], Ripple Neuro [22]) is used with immobile participants undergoing clinically indicated SEEG who participate in voluntary research studies while hospitalized (Figure 2.3). Example of an external system setup connected to depth electrodes is presented in Figure 2.4. Macro-electrodes must be connected to the clinical monitoring system (e.g., Nihon Kohden) at all times during patient's stay for clinical reasons. The splitter boxes are then used to record simultaneously with the research external systems. Micro-electrodes are used for research purposes only and each manufacturer provide a headstage to which other end of the electrode is being connected.

Blackrock and Neuralynx systems are widely used by research groups. For example, Blackrock NeuroPort recording system can record from up to 256 channels using a variety of electrodes at up to 30 kHz sampling rate. Multiple devices can be connected to achieve higher channel count. Numerous high impact results have been published using either of the

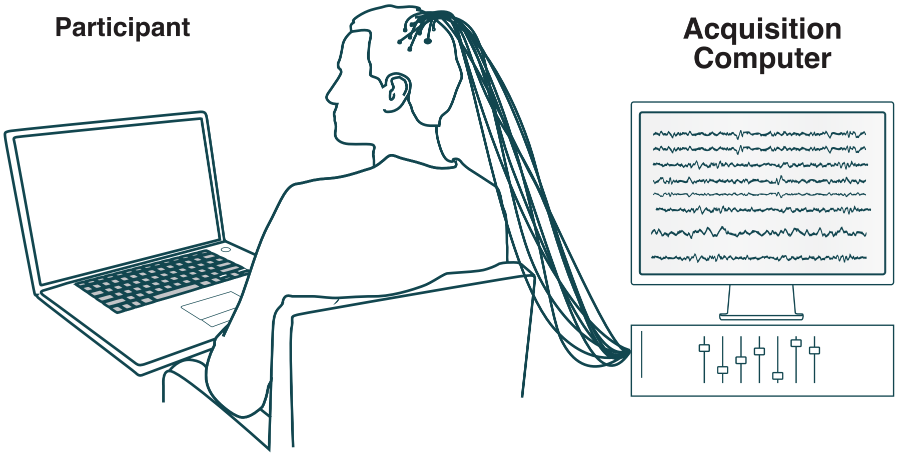


Figure 2.4: External system connected to the implanted electrodes in the hospital.

two systems (e.g., [23], [24]). Majority of the experiments are still designed in an open-loop fashion, meaning that there is no feedback informing the stimulus presentation nor the direct electrical stimulation based on ongoing neural activity. Stimulation research studies are similarly done bedside, primarily using open-loop stimulation [10], [11], [25]–[29], although recent studies have begun to explore the use of closed-loop stimulation [12], [13], [30], [31].

Both Blackrock NeuroPort and Neuralynx Cheetah systems have support for over the network Application Programming Interfaces (API) access from Windows machines, called CBMex and NetCom, respectively. CBMex is MATLAB based, while NetCom is .NET based. These allow automatized control of the open-loop and closed-loop experiments. We will revisit these functionallities in later sections.

2.2.2 Implantable Systems

The second option is to use FDA-approved commercially available neural devices already implanted in several thousand individuals to treat epilepsy and movement disorders (e.g., NeuroPace RNS® System [32] and Medtronic Percept™ [33]; Figure 2.5). The RNS System, for example, detects abnormal electrical activity in the brain and responds by delivering imperceptible levels of electrical stimulation to normalize brain activity before an individual experiences seizures. Since most of individuals implanted with RNS System suffer from pharmacoresistant epilepsy, the leads are usually placed within MTL regions (Figure 2.5-right).

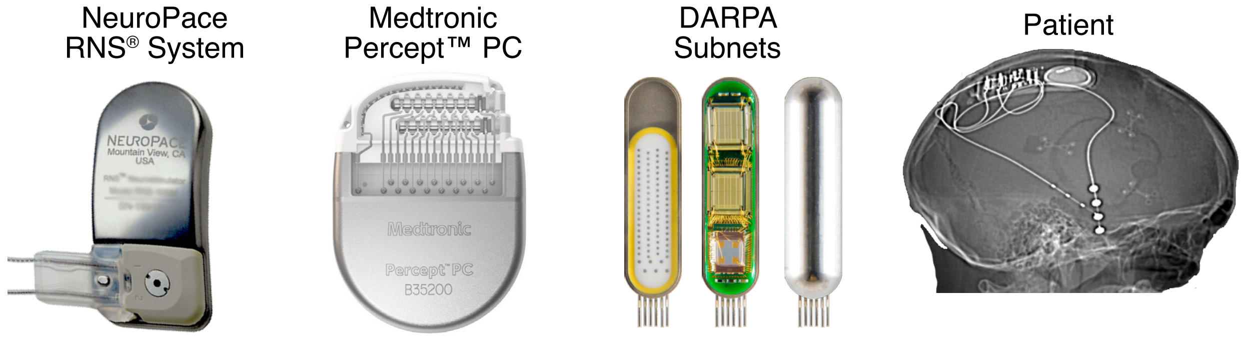


Figure 2.5: Implantable neuromodulation systems. From left to right: NeuroPace RNS® System; Medtronic Percept™; DARPA SUBNETS investigational implant; a brain scan of a implanted person.

These chronically implanted devices offer research participants mobility at the expense of using large macro-recording electrodes that cannot record single-unit activity, fewer channels (usually 4 bipolar), and lower sampling rates (250 Hz).

Other investigational devices such as the Medtronic Summit RC+S [34]–[36], allow for recording 16-channel iEEG activity at up to 1 kHz sampling rates (no single-units). However, they are not FDA-approved for clinical treatment and thus exist in only a handful of patients with an FDA investigational device exemption (IDE) approval, limiting their widespread use by the scientific community. There are also multiple other promising devices being developed, such as those coming from DARPA funded SUBNETS program (Figure 2.5) [37]. Research studies are increasingly adapting these clinical devices for research use [38] and have given rise to several impactful neuroscientific discoveries [39].

Some of the FDA-approved devices also offer custom closed-loop funcionalities such as delivery of electrical stimulation based on power or phase of the ongoing iEEG signals. These features have been exploited by researchers as part of the clinical trials aiming to find suitable treatments, similar to RNS System's epilepsy protocol, for disorders such as PTDS, binge eating, etc.

2.3 Extracting Neural Signatures of Human Behavior

Finding biomarker or neural signatures that are correlated with or are causally predictive of certain human disorder or behavior is one of the crucial steps in every neural pipeline. By now it should be evident that a neuroscientific investigation of a particular human behavior and the brain regions driving it, is restricted and dependent on individuals with a brain disorder that happens to be in the same area of the brain. On top of that, research is voluntary, and excluding implanted participants, it is carried out on a very restrictive timetable in a very research unfriendly and noisy environment, which the hospital room is. Having all this in mind, gathering human data is a challenging task, especially compared to animal studies, which are far ahead in behavioral findings. Because of that, researchers usually aim to acquire the iEEG data from a handful of participants and slightly more from single-unit participants due to lower yield.

Clever behavioral task design, appropriate target group of participants, and correct hypothesis are the desired outcome of every experiment. This means that first order analysis will likely be sufficient for extraction of neural signatures with significant correlation to behavior. However, sometimes the analysis requires more advanced techniques to uncover relationship between behavior and the data. Given the challenges of acquiring new data, similar approach can also be taken to redo the old data analysis and find novel insights using new tools and methods.

2.3.1 Conventional Methods

Conventional methods rely on transformation of the raw data into well-established measures that have proven to be most often correlated with behavior. For example, iEEG often exhibits power or phase changes in certain frequency sub-bands, while for single-unit data that measure is often spiking rate. So, the analysis almost always involves Fourier, Hilbert, or Wavelet Transformation depending on researchers' preference. Desired variables, time-locked with behavior, are extracted in time, frequency, or time-frequency domain and compared with baseline data. If there is no clear effect in one of the variables, then combination of multiple variables are checked against the behavior using some of the linear multivariate regression models. This concludes the conventional approach to analyzing the neural data.

2.3.2 Machine Learning Methods

Most of the experiments are designed to test one, very specific, kind of behavior against the baseline or opposite behavior. That means that neural analysis can often be rephrased into a binary classification problem that tries to separate neural data in two classes, behavior and non-behavior.

For example, in a verbal memory task, words are presented to the participants, sequentially one by one, and the question is which words will be remembered during recall after some time and how does that reflect in the neural data. Logistic Regression classifier proved to be the method that could easily extract the signatures of the behavior by trying to separate the data into remembered and forgotten classes [10]. Another group used similar approach, but with artificial neural network, to perform multiclass classification in order to decipher what participants wanted to say just by analyzing the data from the brain motor areas [40].

2.3.3 Real-Time Methods

Usual approach in real-time neural analysis, which is needed for closed-loop paradigms, is to first perform open-loop behavioral experiments, and then to extract signatures of the behavior using some of the described methods. Once a strong neural signature is known, implementation of the online search for it is straightforward. However, this approach does not always work as there are discrepancies in neural responses in each individual participant, which often require adjusting the parameters (e.g. thresholds) at the very least. Ideal scenario would be a closed-loop that can continuously respond to neural short- and long-term changes. However, online algorithms are in general more challenging to implement because they require causality, which can introduce frequency dependent distortions (e.g., nonlinear phase lag) into the data, while trying to replicate their non-causal offline counterparts.

External systems offer application programming interface (API) over the network that can be used in experimental paradigm to dynamically update presentation stimulus or electrical stimulation parameters. For example, Blackrock users can use third-party CBMex package to design experiments and closed-loop protocols within MATLAB on a Windows machine. Similarly, Neuralynx offers, as part of its software package Neuralynx Cheetah, NetCom library that can be in the background of every .NET application containing experimental protocols. The APIs offer full control over recording and stimulation functions. The experimental computers executing the tasks can connect to external devices over their wireless access point (Figure 2.6). Using API functions on the experimental computer can result in many milliseconds of delay depending on the processing method and the type of connection as part of an external loop. Manufacturers of external systems also provide intergrated hardware resources for digital signal processing (Auxiliary Hardware, Figure 2.6), which translets to millisecond delays as the neural samples do not leave the recording systems for custom processing.

Implantable system by default utilized closed-loop stimulation to treat brain disorders.

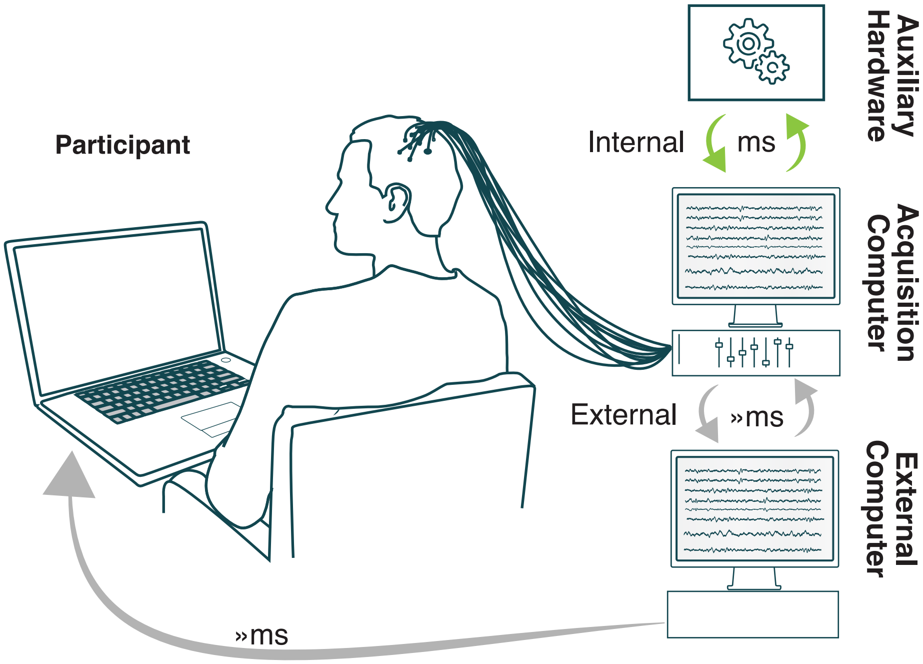


Figure 2.6: Examples of clinical monitoring and research recording and stimulation systems connected to the implanted electrodes in the hospital.

For example, RNS System detects bandpower increases over predefined thresholds to trigger stimulation in order to reduce epileptic seizures. Frequency bands, thresholds, and stimulation parameters are set beforehand, during neurologic assessment. With built-in causal analog filtering and online power spectrum calculations, the RNS System can be used to test custom closed-loop paradigms relying on power thresholding. Custom in this case means choosing custom frequency band, threshold, and stimulation current parameters such as amplitude, phase width, burst frequency, duration, etc. The device also offers phase-locked stimulation at specified frequency. Available implantable devices do not offer real-time access to neural samples and thus cannot be used for custom closed-loop paradigms other than built-in power thresholding.

2.4 Human Learning, Memory, and Navigation

Formation of human cognitive functions, such as learning, memory, and spatial navigation have been associated with medial temporal lobe regions, more specifically hippocampus and entorhinal cortex. Neural processing techniques for the MTL data does not differ from those used on the neural data from any other brain region. The difference is that MTL regions are located deep within the brain and requires depth electrodes to be reached (Figure 2.1C-bottom). Thus, cognitive functions are one of the least explored due to limited number of depth electrodes that can be implanted. There are few institutions in the world with the expertise to perform such implantation with decent yield. Discovery of units governing navigation through physical space (Figure 2.7), such as grid and place cells in animals have won the only Nobel prize for the field of neuroscience and has since caused increasing popularity of human spatial navigation research in order to bridge the gap between human and animal findings [41]. This is the reason behind the need for and importance of the technology that allows freely-moving human experiments.

Some researchers have utilized implants to show hexdirectional modulation of the iEEG bandpower across the space ([42]), and increased iEEG bandpower close to the space boundary ([39]). Others have recorded single-unit activity using external systems from neurosurgical participants navigating through virtual reality (VR) space to show evidence of grid cells in humans ([43]). Recordings of single-units and thus evidence of place and grid cells during real human navigation has not been possible so far.

Stationary experiments probing various forms of memory and learning formation are just as important as ambulating ones. Described systems and synchronization between measurements and behavior are all essential in experimental design for both types of experiments.

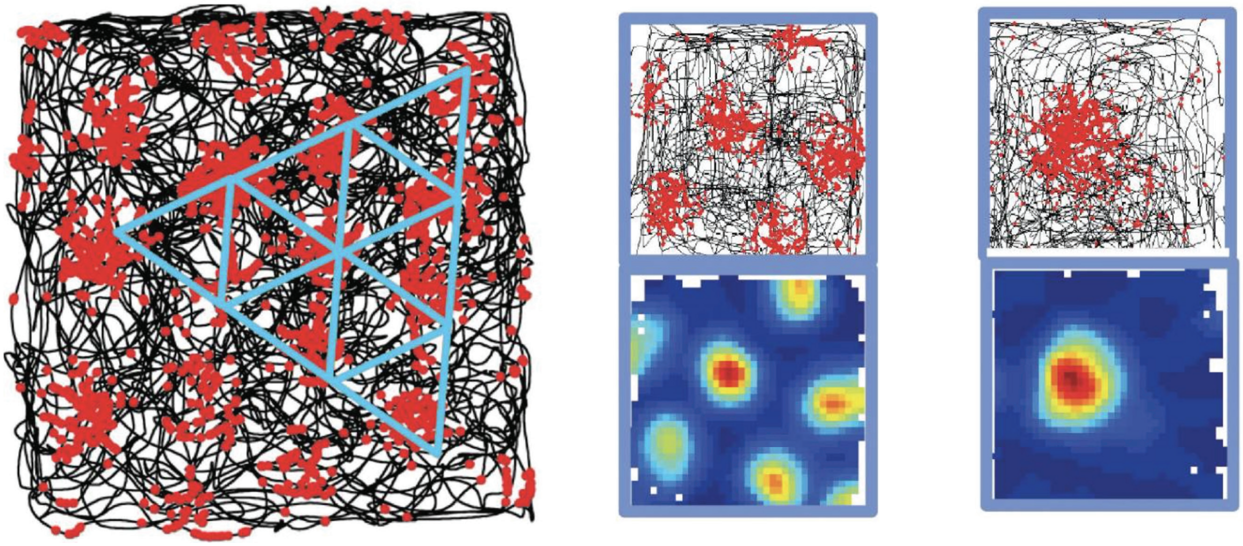


Figure 2.7: Neurons representing animal spatial navigation. Left: A physical space and navigation trace of an animal. Middle: Animal entorhinal grid cells. Right: Animal hippocampal place cells. Adapted from [41].

2.5 Review of Prior Work

The aim of previous sections was to introduce various engineering and neuroscience terms and concepts leading into discussion of the current progress of interdisciplinary efforts to explore functions of the human brain.

Within defined neuroscientific framework, we will discuss present state of the art in four key areas: 1. Recording and stimulation capabilities; 2. Participant's mobility; 3. Resources for custom closed-loop algorithms targeting memory enhancement; 4. Online algorithms for extracting signatures of memory formation.

External systems, Blackrock and Neuralynx, can support a large number (≤ 256) of both iEEG and single-unit/LFP electrodes at high sampling frequencies and high input dynamic ranges (Table 2.1). As discussed previously, implantable systems usually offer 4 bipolar recordings of iEEG at 250 Hz. While they offer completely mobile experiments they lack

Table 2.1: Comparison of external and implantable systems available for human use

	Externalized		**Implantable**	
	Blackrock® NeuroPort	Neuralynx Digital Lynx SX	NeuroPace RNS® System	Medtronic Percept™ PC
Channels	256	256	4	6
Samp. Freq.	30 kHz	40 kHz	250 Hz	250 Hz
Input AC Range	±8.192 mV	±132 mV	up to ±1.9 mV	X (~1 mV)
API	CBMex Win	NetCom Win	—	—
CL Resources	—	Zync 7000	STFT	STFT
CL Latency	Internal: 1 ms External: » ms	Internal: 4 ms External: » ms	Unknown	Unknown

number of channels and capability of capturing single-units. We also list ava

Here we also discuss available resources that researchers have used to develop closed-loop experiments based on both iEEG and single-unit data. Already described external system setup is the same, regardless on the target signal: iEEG, LFP, or single-unit data. As mentioned, implantable systems with on-chip resources for online filtering and power/phase extraction have extremely low latency, but have limited online processing and stimulation programmability. Implantable closed-loop trials are however very valuable for disorder treatments and are either already approved for medical use ([32], [44]), or are currently being explored as part of clinical trials (`https://www.clinicaltrials.gov`; NCT0401149 [2019], NCT03582891 [2018], NCT04152993 [2019], NCT05120625 [2021], NCT04558164 [2020], NCT04874220 [2021]). All other, external systems, require data transmission and closing the loop outside of the acquisition device.

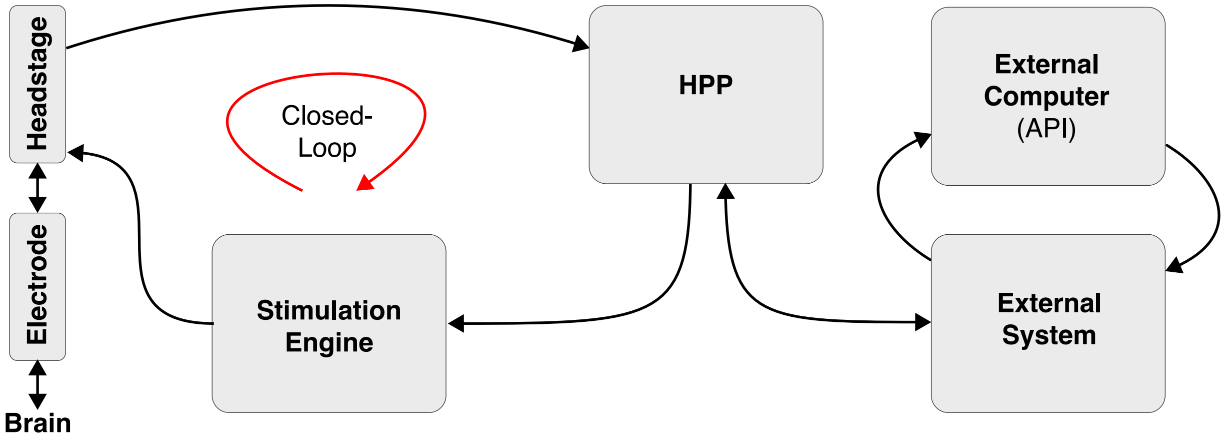


Figure 2.8: Simplified block diagram of external system's components, indicating internal and external closed-loop pathways.

Recently, external systems started offering local hardware resources for real-time analysis and closing the loop 'closer' to the analog front-ends. For example, Neuralynx Digital Lynx SX system now comes equipped with Hardware Processing Platform (HPP) for real-time, data processing and closed-loop neuromodulation. The HPP is a board from Xilinx family Zync®-7000 SoC, which combines dual-core 1 GHz ARM Cortex-A9 with an FPGA. Additionally, it contains 1 GB of DDR3 RAM and 16 MB of flash memory. When combined with NetCom API, it provides powerful framework for closed-loop paradigms (2.8). With these improvements, Neuralynx has the option of sub-millisecond internal closed-loop with digital signal processing algorithms (e.g., online filtering, power/phase spectrum, etc.) programmed in C/C++ or Verilog/VHDL and executed on the HPP. Processing algorithms that cannot be ported to the Zync-7000 board, have to be executed on an External Computer that uses API to communicate with the external system and close the loop (2.8). This can translate to larger latencies (∼100 ms) depending on the algorithm and the type of the connection between devices.

There are two general directions of modeling representations generated withing the brain about the external world. First are encoding models, which try to predict brain's representation based on the stimulus presented, and the second are decoding models, which try to

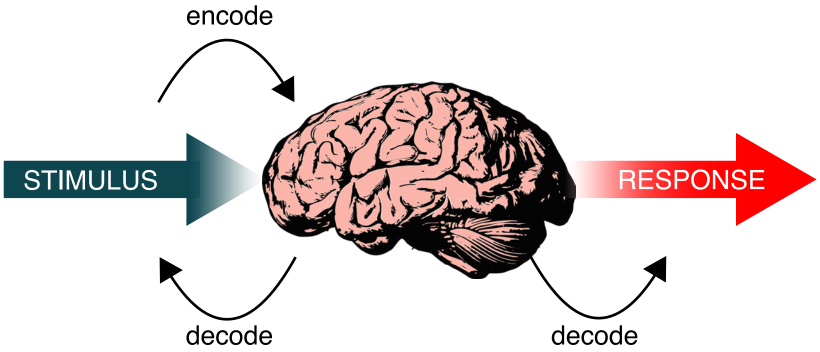


Figure 2.9: Encoding process attempts to predict brain response from the stimulus. Decoding process tries to predict participant's response based on the neural activity. Adapted from [45].

predict participant's response based on the measured, sub-sampled representation from the brain that is neural activity in this case (Figure 2.9).

Encoders effectively model the brain function and are much more challenging to build due to our sub-sampled way of collecting the brain data. Rather, it is often easier to model the outcome based on available data or to decode. To build true closed-loop connection with the brain both steps are necessary, but here we will focus on the state of the art decoding models.

2.6 This Work

The goal of this work is to develop a versatile, external, wearable neural interface that can be used in clinical and neuroscience research by allowing easy prototyping and testing of various paradigms in different setups that may require synchrony with other devices under research protocol. Because of this we will focus this work on improving capabilities of the research-oriented external systems. We will still make informational comparisons with implantable systems, but they are not the focus area of this work.

To address described problems with currently available external systems, this work utilized advanced, implantable neuromodulation platform, developed under DARPA SUBNETS program (Figure 2.10-left). This implantable system contains miniaturized and low-power ICs for sensing and stimulation, which we externalized and assembled small, wearable, and low-latency device, called Neuro-stack (Figure 2.10-right).

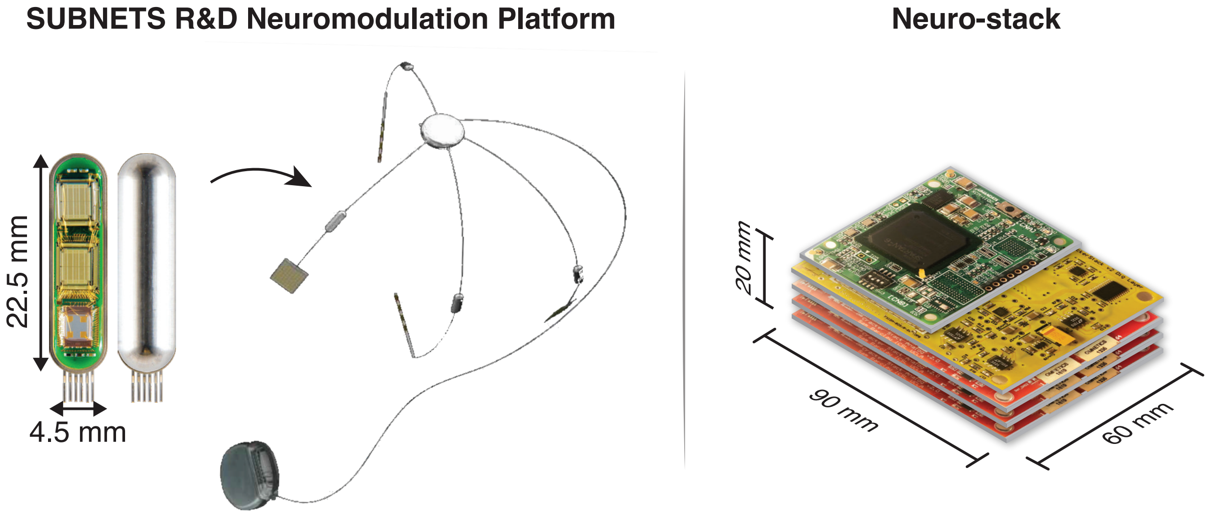


Figure 2.10: Leveraging the advanced implantable technology to build externalized, small neuromodulation device.

To briefly summarize key sensing and stimulation characteristics of a wearable device, such as the Neuro-stack, we compare it with existing systems (Figure 2.11). In the remaining text, we will detail how we achieved these numbers in the context of low-latency and wearable research experiments. External systems, Blackrock and Neuralynx, can support a large number (≤ 256) of both iEEG and single-unit/LFP electrodes at low input-referred noise levels ($\geq 1\ \mu\mathrm{V}_{\mathrm{rms}}$; Figure 2.11). Their stimulation engines can also support large number of channels (≤ 96) at sufficiently high maximum current (≤ 10 mA; Figure 2.11). Furthermore, levels of miniaturization of these devices ($\sim 10^{-3}\frac{\text{channels}}{\text{cm}^3}$; Figure 2.11) render them unusable in freely-moving experiments.

Ripple Nano2 has somewhat improved miniaturization ($\sim 10^{-1}\frac{\text{channels}}{\text{cm}^3}$ at the expense of number of supported channels (Figure 2.11). Ripple systems with its smaller size, compared to other external systems, is the only potential candidate for freely-moving experiments, but have not been used in these studies so far.

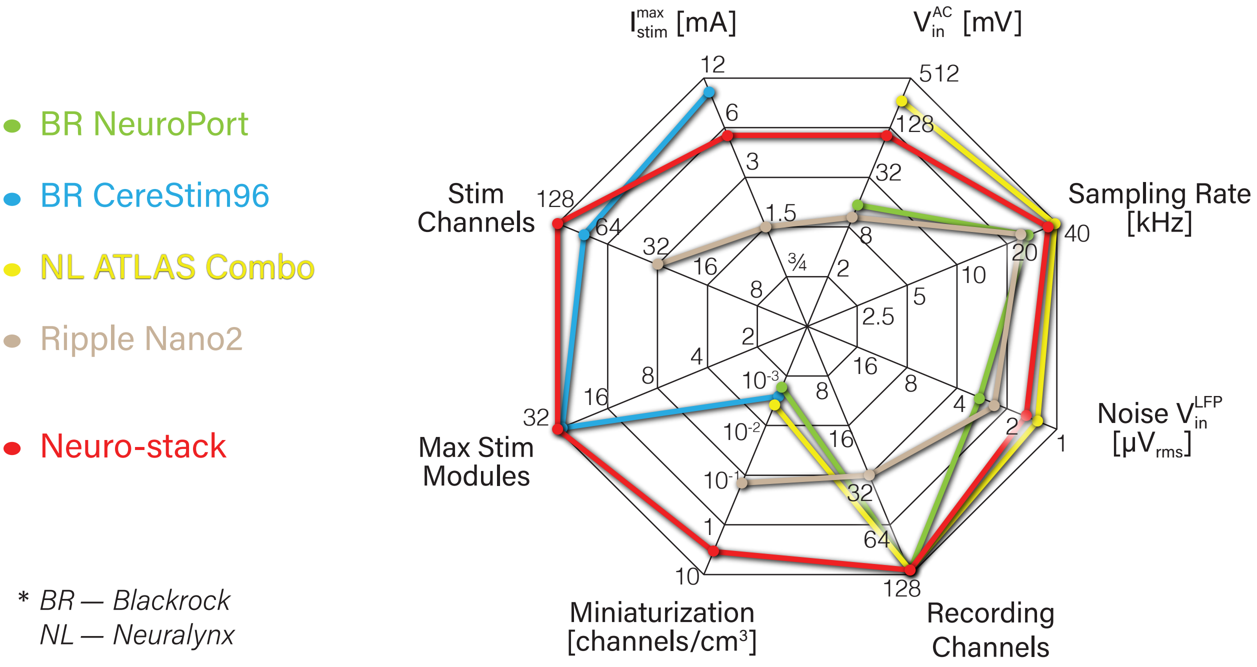


Figure 2.11: Comparison of current bedside intracranial recording and stimulation systems used in humans. Characteristics shown include the device sampling rate, noise of the input sensing front-end (Noise V_{in}^{LFP}), number of recording channels, linear input dynamic range (V_{in}^{AC}), maximum stimulation current (I_{stim}^{max}), number of stimulation channels (Stim channels), and maximum stimulation channels that could be used simultaneously (Max stim modules). BR – Blackrock; NL – Neuralynx.

2.7 Requirements

As previous review, key requirements for this work are centered around following: 1. Recording and stimulation capabilities; 2. Wearability; 3. Built-in computational hardware resoureces; 4. Online neuroscience application.

First requirement is to build an external system that can record large number of SEEG channels (100 or more), which means ability to acquire iEEG, LFP, and single-units. On the electronics level this translates to a requirement to record a signal that ranges from 10 µV to 1 mV at ≥30 kHz in a very noisy clinical environment. Further, the system needs programmable electrical stimulation with option to adjust current amplitude, frequency, pulse shape, duration, and all other timings including burst and multiburst protocols. Due to large stimulation artifacts that follow delivery, recording front-end needs high input dynamic range that exceeds neural signal levels (several tens of millivolts).

Second requirement is small dimensions and weight of the external device so that it could be comfortably carried on-body by the participants during walking and other physical movements, which are unavoidable part of naturalistic studies.

Third requirement is on-board resources for online neural processing, including artificial neural networks, and their training and inference.

Forth requirement is an in-vivo human application that utilizes developed hardware with state of the art neuroscientific results in the field of learning, memory, and spatial navigation that validate the necessity for such external systems.

3 Proposed Sensing and Stimulation Device

Here, we present the Neuro-stack (Figure 3.1), a bi-directional neuromodulation platform for wide-band sensing and stimulation of deep-brain areas for basic and clinical neuroscience studies.

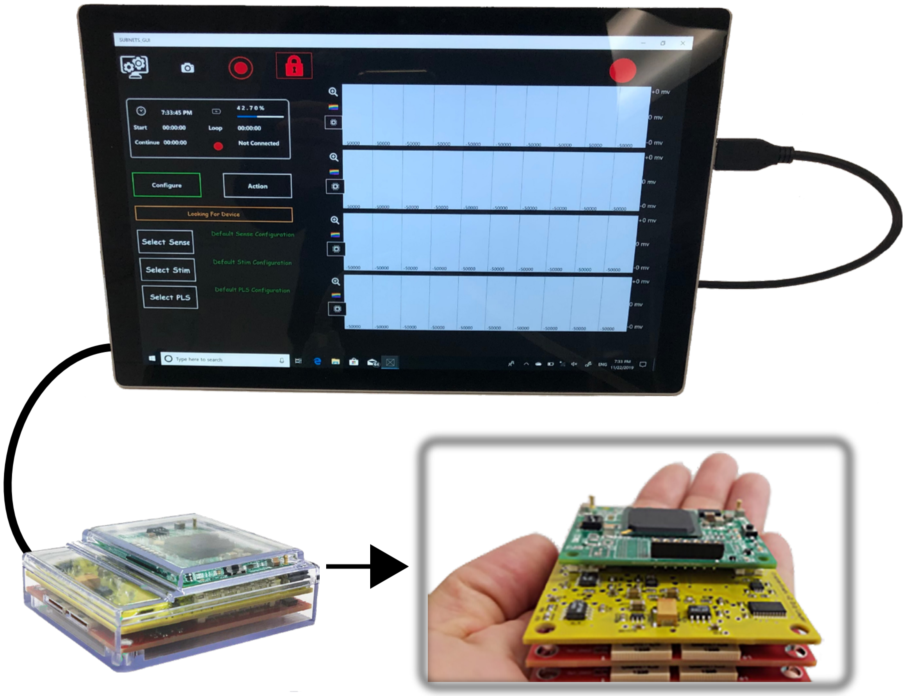


Figure 3.1: Base Neuro-stack platform, including a hand-held device and a GUI-based tablet for control, device configuration and data monitoring.

Compared to much larger existing devices (Figures 2.3 and 2.11) that are used bedside and carried on a cart, the Neuro-stack's small hand-held size enables concurrent stimulation and recording of real-time electrophysiology (single-unit and LFP activity) during freely-moving behavior by connecting to commonly used implanted macro- and micro-electrodes. Apart from its small form-factor and unique on-body wearability, the Neuro-stack can support:

1. Recording of up to 256 channels for a total of 128 monopolar or bipolar recordings with a sampling rate of up to 6,250 Hz. Further, wide-band sensing from up to 32 monopolar or bipolar recordings at up to 38.6 kHz allows for the recording of single-unit and LFP activity simultaneously.

2. Flexible and programmable stimulation allowing for delivery of bipolar/monopolar stimulation to any 32 out of 256 contacts simultaneously. Stimulation engines are current-controlled and allow the user to program current amplitude, frequency, timing, pulse shape, and other parameters.

3. Closed-loop neuromodulation. The Neuro-stack has built-in (hardware) oscillation power detection and thus the ability to trigger stimulation at a predefined phase of an oscillation (phase-locked stimulation [PLS] delivered at a particular phase of ongoing theta activity). Further, sensing of neural activity is concurrent with stimulation for true (full-duplex) closed-loop capabilities. Resources for designing custom closed-loop algorithms are available at both the embedded hardware and external software levels.

4. Software support that comes in two formats. First, a turnkey graphical user interface (GUI) running on a Windows-based tablet or laptop is available for research purposes (Figure 3.1). Second, a full-access application programming interface (API) library written in C++ allows the user to build custom research open- and closed-loop stimulation capabilities for research studies.

5. Tensor multiplication accelerator (Edge TPU) that is integrated with the Neuro-stack

device, enabling an extended range of applications such as real-time inference for neural decoding or closed-loop stimulation.

6. Wired or wireless mode. The Neuro-stack platform can be externally controlled and powered via a USB cable or remotely controlled through a secure local network using a battery-powered configuration. This flexibility allows researchers to perform wide-band recording and stimulation during either stationary or ambulatory (freely-moving) behavioral tasks.

3.1 Neuro-stack Hardware

3.1.1 Sensing and Stimulation

The central hardware component of the Neuro-stack platform (Figure 3.2) consists of three printed circuit board (PCB) layers: 1) analog, 2) digital, and 3) communication. Each layer is embedded with one or several dedicated integrated circuit (IC) chips. The analog layer (Figure 3.2-bottom) contains mixed-signal sensing IC (Sense IC and Spike IC) and stimulation IC (Stim IC) chips, which were previously developed as part of the DARPA SUBNETS program [37], [46]–[48]. The whole device is assembled by physically stacking the described layers (Figure 3.2). Furthermore, one Neuro-stack device supports up to four analog layers at the same time, for up to 256 micro-wire (LFP) electrode contacts (64 per layer) and up to 32 micro-wire (single-unit) electrode contacts (8 per layer). All analog and digital custom integrated circuits used in the Neuro-stack were fabricated using low-voltage 40 nm technology.

A single Sense IC (one per analog layer) accepts neural activity from up to 64 electrode contacts fed into voltage-controlled oscillators (VCO), which serve as analog-digital converters (ADC). Each VCO ADC supports 6,250/N Hz sampling frequencies, where N = 1,2,4,8, . . . , 128 and a 100 mVpp linear input dynamic range with 12/21 (macro/micro)

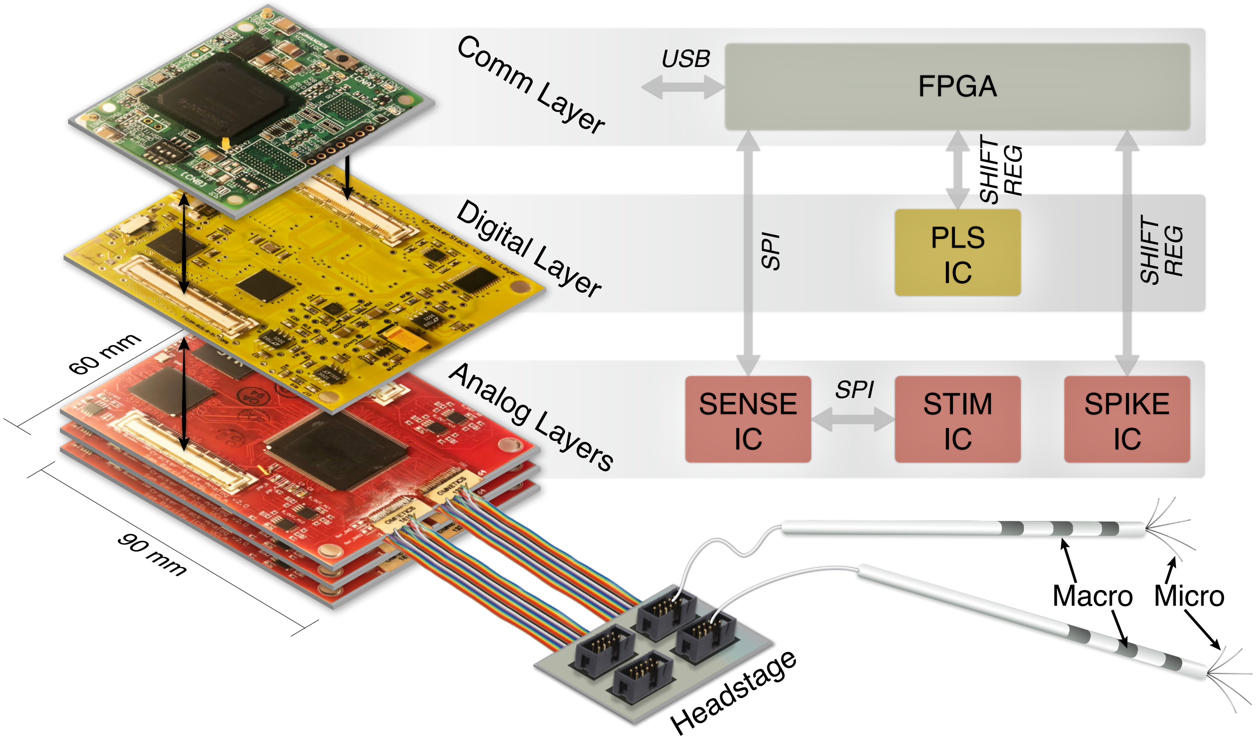


Figure 3.2: The Neuro-stack consists of three stacked layers: communication, digital, ana analog layer. Each layer carries dedicated ICs for sensing, open/closed-loop stimulation, and USB external communication. The Neuro-stack connects to commonly used neural electrodes.

bits of resolution, ensuring that the underlying neural signal is captured in the presence of large artifacts (e.g., from stimulation). The Sense IC contains digital nonlinearity correction to account for nonlinear amplification across the input range. Moreover, it also contains a digital logic for adaptive stimulation artifact rejection that subtracts a template stimulation artifact extracted from adjacent channels [48]. The total power consumption per channel is 8.2 µW. A single Spike IC (one per analog layer) accepts neural activity from up to 8 micro-wire contacts and supports sampling rates of up to 38.6 kHz [49] (Table 3.1).

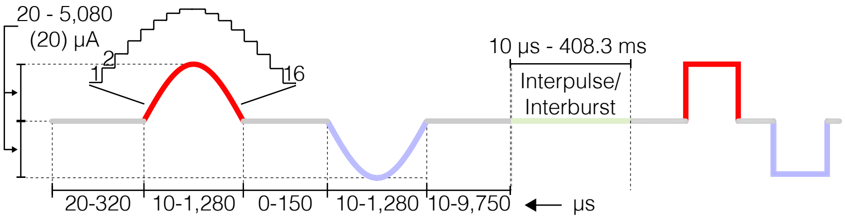
A single Stim IC contains eight engines that can, with the appropriate configuration,

Table 3.1: Neuro-stack sensing capabilities.

Sense IC (iEEG)	
Channels	128
Sample Rate	6.25 kHz
Input Range	± 50 mV
Noise	5.2 μV$_{rms}$
Sample Res.	16/21 bits

Spike IC (Single-units)	
Channels	32
Sample Rate	38.6 kHz
Input Range	± 20 mV
Noise	2 (7) μV$_{rms}$
Sample Res.	16 bits

drive current through any individual or combination of the connected 64 electrode contacts. Stimulation output current is highly configurable (Figure 3.3), including selection of amplitude, frequency, and multiple or custom waveform shapes. This flexible programmability allows for stimulation using previously used burst protocols as well as exploration of novel stimulation patterns for investigative research and therapy development. These capabilities also enable increased degrees of freedom (timing, amplitude parameters; Figure 3.3) compared to currently available intracranial neurostimulation systems.



Stim IC (Bipolar-Macro)	
Channels	32 out of 256
Control	Current
Amplitude	20 - 5,080 μA
Frequency	2.37 Hz - 16.67 kHz
Polarity	Anodic/Cathodic

Figure 3.3: Neuro-stack stimulation capabilities.

The Neuro-stack's digital layer (Figure 3.2-middle) routes signals between the analog and communication layers and contains a custom IC chip (PLS IC) for closed-loop stimulation based on the detected oscillatory (e.g., theta) phase in the recorded neural signal coming from the analog layer to enable PLS [50], [51]. A field-programmable gate array (FPGA, Xilinx Spartan 6 board) serves as a communication layer (Figure 3.1-top) between an external devices and custom ICs (Figure 3.1-right).

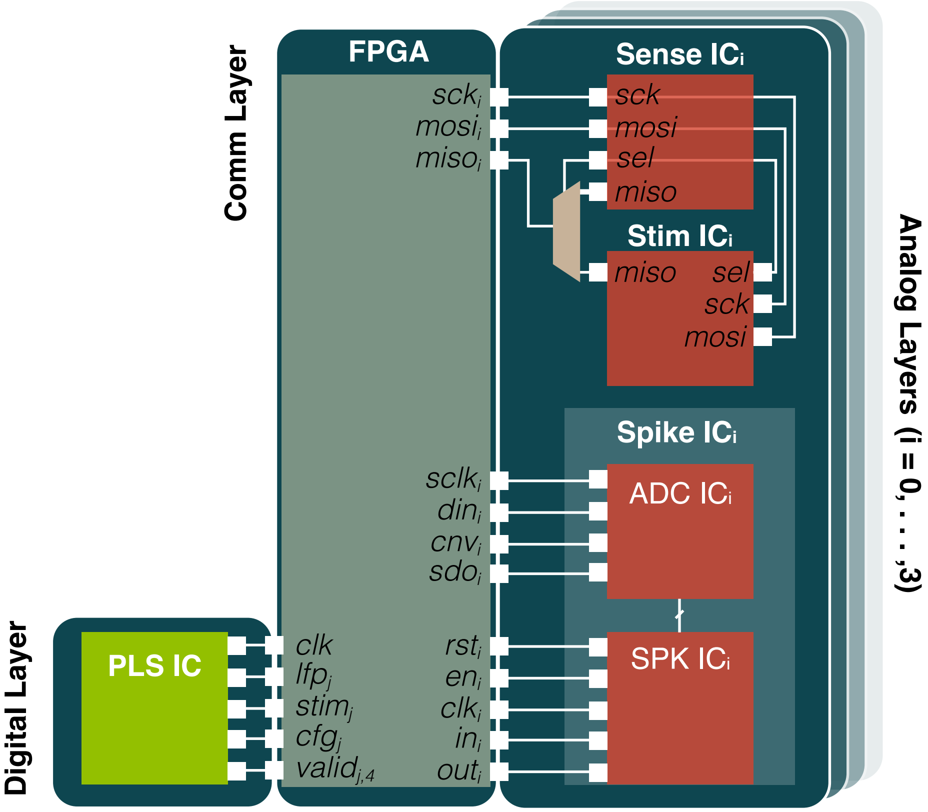


Figure 3.4: Neuro-stack high-level block diagram shows interfaces between assembled chips. Sense and Stim IC are part of one 3-wire SPI interface, while SPIKE and PLS IC use Shift Register transfer via data and valid lines.

The Neuro-stack uses the serial peripheral interface (SPI) at 12 MHz (Sense IC and Stim IC) and serial shift register (PLS IC and Spike IC) for internal communication between layers and IC chips and a USB interface for external communication and power supply (Figure 3.4). SPI interface between FPGA and Sense/Stim IC is specifically designed for lower area of the implantable solution as three input/output wires occupy less space. The chip select control is pushed through to the Sense IC, and can be accessed through packet communication protocol. Spike IC consists of two separate chips, amplifier (SPK IC) and analog to digital converter (ADC IC).

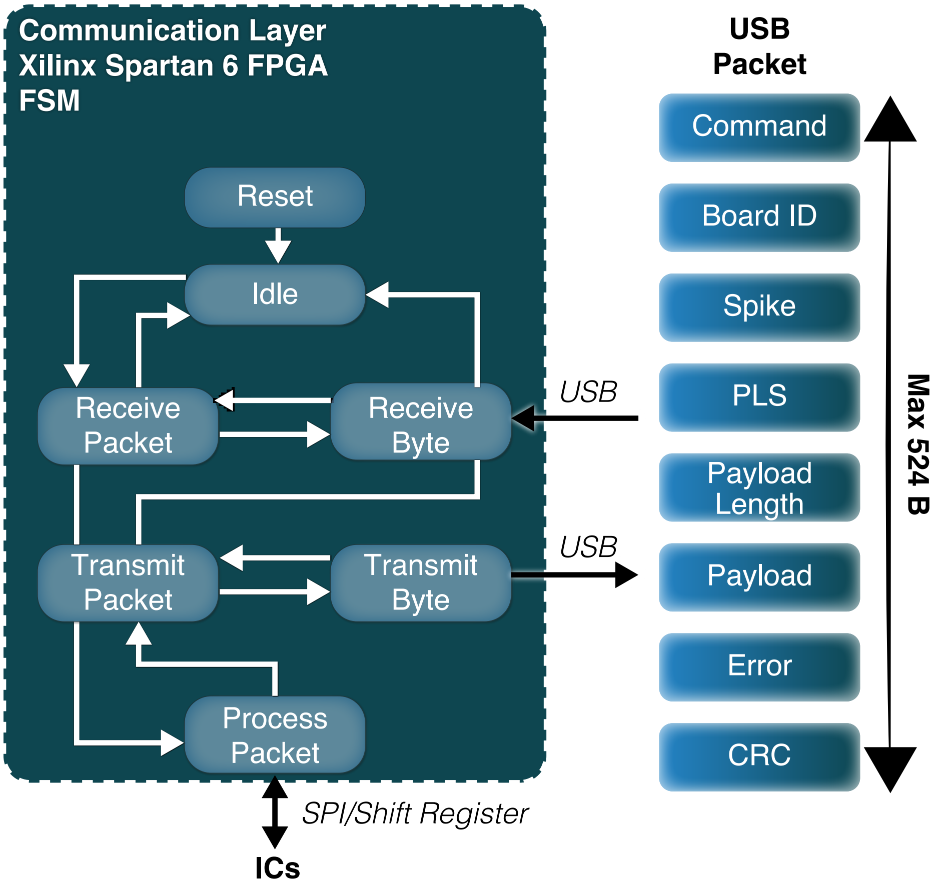


Figure 3.5: FPGA finite state machine and communication protocol to and from the Neuro-stack. USB packet structure (right) is used to address and control all Neuro-stack ICs.

The communication layer (FPGA) runs Mealy finite state machine (FSM) that is responsible for unpacking and rerouting USB packets to each IC addressed in the message, and vice versa (Figure 3.5). In the process, this also converts USB interface to SPI or to Shift

Register interfaces. This FSM is bi-directional and thus also processes SPI packets received from the Neuro-stack and converts them into USB packets, which are then transmitted to the external device. The FSM always begins with a Reset state after a reboot, and then enters an Idle state in which it waits for incoming packets. Once a packet is available, the FSM receives it byte by byte (Receive Byte) until the complete message is transferred (Receive Packet). The received packet is then being processed (Process Packet), converted into the appropriate interface (e.g., USB to SPI), and transmitted to the Neuro-stack ICs (via SPI or Shift Register). Similarly, after the processing is done, the response packet from the ICs enters a state during which it can transmit the packet (Transmit Packet) byte by byte (Transmit Byte) externally. Once the transmission is done, the FSM goes back to the Idle state and waits for new packets unless the streaming of the neural data is taking place, in which case the FSM enters Process Packet state indefinitely until the recording is stopped. The structure of the USB packets sent from external devices to the Neuro-stack Communication Layer, which contains up to 524 bytes that describe the type of Command, Board ID to address specific analog layer, Spike byte, phase-locked stimulation (PLS) byte, and Payload for additional information where its length (Payload Length) depends on the type of command. The packet also contains bytes for error codes (Error) and a cyclic redundancy check (CRC) to detect accidental changes in the raw packets.

This FSM can be controlled directly from a ready-to-use GUI, which allows real-time multi-channel monitoring and control of sensing and stimulation. The GUI application maps actions directly into the USB packet. However, in order to build versatile system for research applications, there was a need for an intermediate layer that can provide API functions as foundation on top of which user space can be built.

3.1.2 Wireless API and Online Inference Acceleration

So far, we have explained assembled Neuro-stack and its fundamental capabilities for sensing iEEG, LFP, and single-units, as well as to stimulate from large number of electrodes.

However, the focus of this work is real application in neuroscience research and before that could be possible a few other features needed to be developed.

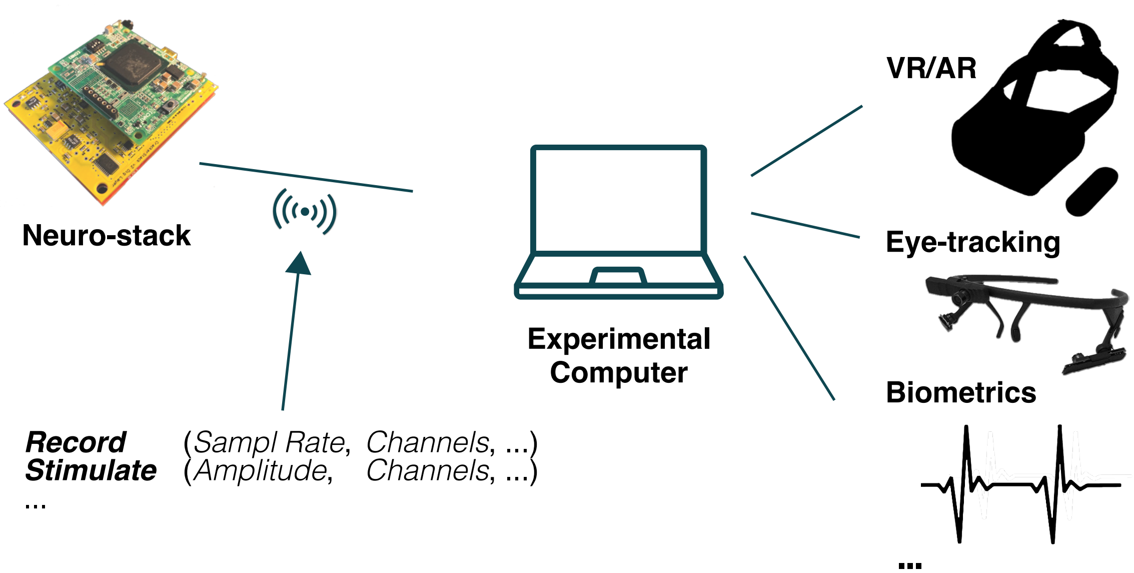

Neuro-stack
Record (Sampl Rate, Channels, ...)
Stimulate (Amplitude, Channels, ...)
...
Experimental Computer
VR/AR
Eye-tracking
Biometrics
...

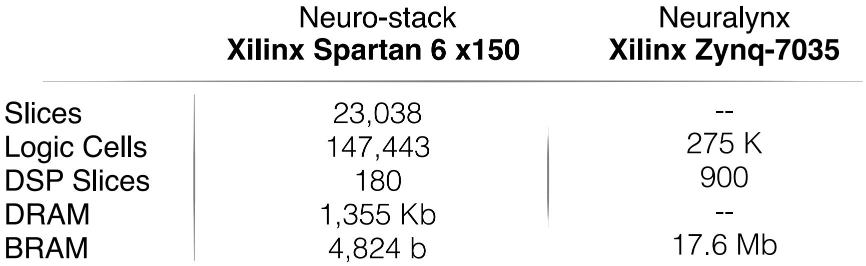
Figure 3.6: Human neuroscience research often involves many other devices for controlling the experiment, data collection synchronized and managed by a central experimental computer.

Human neuroscience research is often carried out from a single experimental computer, running a script that stimulates precisely defined behavior and synchronizes it with human brain signals. In order to establish greater control of the experiment as well as to collect additional data non-neural data that may be correlated with neural data and behavior, researchers are using increasing number of devices and wearable technology. For example, virtual/augmented reality (VR/AR) goggles give great control of what participants can see and thus, great control of the experiment that can be automated. On the other hand, eye-tracking and various biometrics data give further insight into participant's behavior especially one that involves cognitive and emotional functions. These different devices need to run in synchrony and be easily accessed and controlled from the experimental computer.

By allowing wireless API functions over local network, Neurostack could widen possible research setups used for testing human neuroscience hypotheses (Figure 3.6, [38]).

To enable wireless API and online processing, we needed to integrate additional hardware with Neuro-stack. The primary options for wireless communication suitable for research applications are Bluetooth or Wi-Fi peripherals. Implementation of both options is relatively straightforward by rerouting FPGA packets towards a dedicated chip or a wireless peripheral. We opted to go with Wi-Fi option due to its faster transmission, easier to implement security protocols, and because it is much easier to synchronize multi-device setups running on a same local network (Figure 3.6).

Table 3.2: FPGA boards used in Neurostack and Neuralynx recording system.



	Neuro-stack	Neuralynx
	Xilinx Spartan 6 x150	**Xilinx Zynq-7035**
Slices	23,038	--
Logic Cells	147,443	275 K
DSP Slices	180	900
DRAM	1,355 Kb	--
BRAM	4,824 b	17.6 Mb

Neuro-stack's FPGA board, Xilinx Spartan 6 x150, is an outdated board with only one role, a communication hub. The resources available on it were not sufficient for online neural data analysis other than applying basic spectral digital signal processing. Furthermore, it does not possess Wi-Fi peripheral. Newer FPGA boards, such as Xilinx Zynq-7000 series, which are used as part of Neuralynx online processing hardware, possess higher number of DSP slices and memory (Table 3.2). However, in order to perform ∼100M MAC operations per inference in a machine learning algorithm the numbers need to be higher to keep the inference latency low.

Instead, we opted for Edge Tensor Processing Unit (TPU), mounted on top of TPU Dev Board. It is an ARM-based single-board computer, running a Mendel Linux distribution.

The board has sufficient amount of on-chip and peripheral memory, an ARM Quad Cortex-A53 processor, Edge TPU co-processor, and a peripheral for Wi-Fi. The Edge TPU, designed by Google, can process 2TMAC/s at 2 W of power using a 64 × 64 MAC matrix specifically designed for machine learning matrix multiplication (Figure 3.7). The TPU Dev Board uses PCIe2 interface to the TPU co-processor ensuring high data bandwidth.

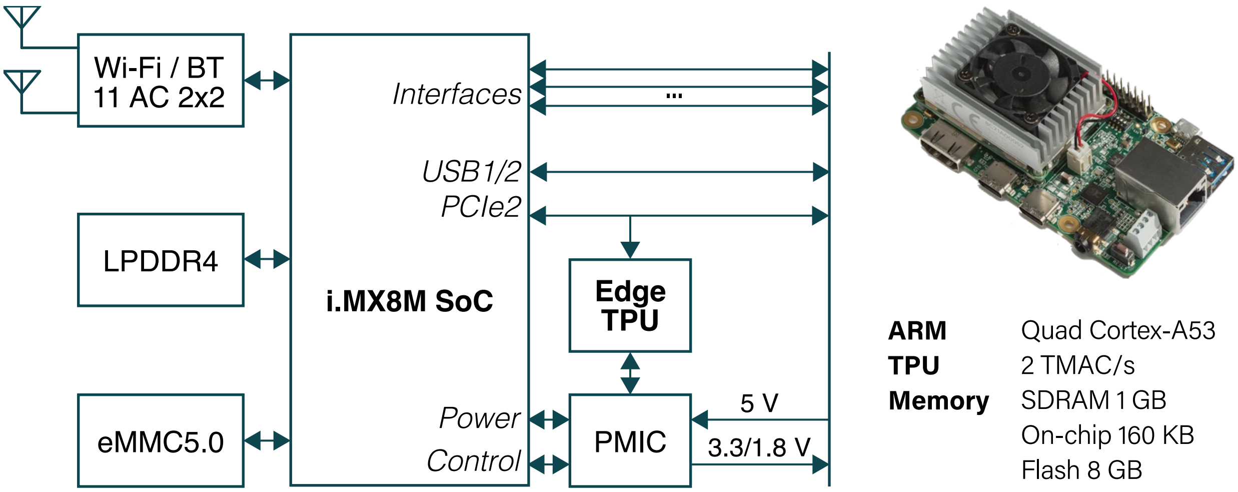


Figure 3.7: High-level block diagram of the TPU Dev Board. TPU Dev Board contains range of peripherals suitable for custom research applications such Edge TPU, ARM processor, Wi-Fi, and sufficient amount of memory.

The TPU Dev Board also supports external USB interface making it easy to prototype integration with Neuro-stack (Figure 3.8). External battery was used to power TPU Dev Board, which in turn also powered Neuro-stack. Battery power also helps reduce line noise from the system and recordings in extremely noisy environment such as hospital room. This setup can now offer wireless connection and local online processing using machine learning and artificial neural networks.

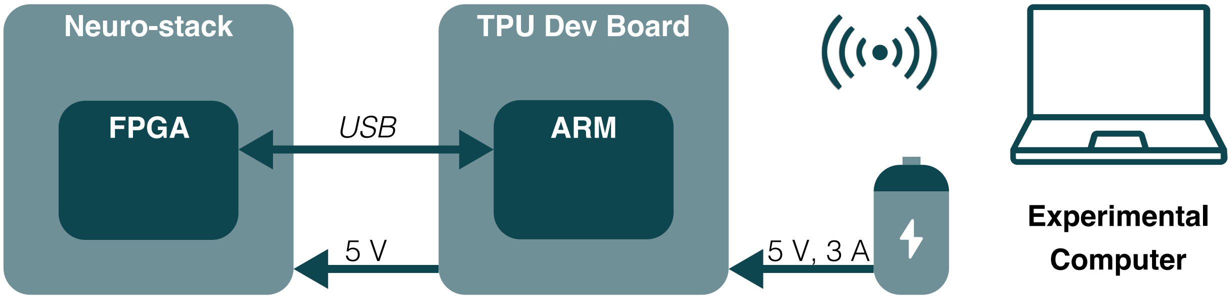


Figure 3.8: Neuro-stack and TPU Dev Board interface via USB.

3.2 Neuro-stack Firmware and Software

A large portion of this work went into the development of firmware and embedded software that can support Neuro-stack's data bandwidth. Recording from all Sense and Spike IC channels with all four analog layers means up to 30 MB/s of neural data. Other than keeping up with the pace of incoming data, the software also needed to support many features of the Neuro-stack, such as programmable open- and closed-loop stimulation, online processing through TPU Dev Board peripherals, etc.

For this purpose, we developed a speed-optimized, real-time pipeline in form of a C library. Through multi-threading operation, it supports all ICs with dedicated software process as part of a software clone model (Figure 3.9). Further, this library is compatible with all commonly used operating systems and platforms. To achieve the goal of wireless API used for easy access in research, we used this library as part of a general application built on top that can offer API functions externally using both wired (USB-C) and wireless (TCP/IP socket) connection. This library can also be used as a base for building new custom applications on top of it. Here, we show a modified version that other than key neuromodulation functions also offers TPU peripheral access through the same connection. This way researchers can easily connect to the Neuro-stack running client code within research

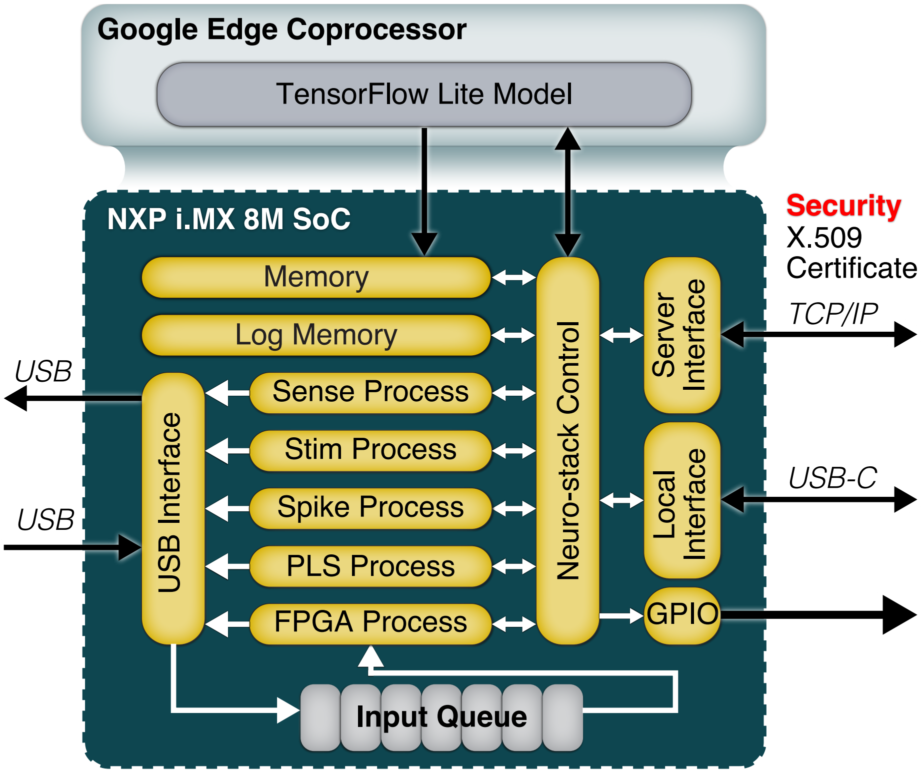


Figure 3.9: Block diagram of the Neuro-stack's wireless API server.

paradigm using any device (computers, phones, VR/AR goggles, etc.; Figure 3.9).

The Input Queue handles streams of both neural data and acknowledgment receipts from the Communication Layer and redirects them to the appropriate block on the TPU Dev Board responsible for each ICs (e.g., Sense, Stim, ... Process). The Neuro-stack Control block contains all of the API functions, which are then multiplexed to additional layers responsible for wireless (via Server Interface) or wired (via Local Interface) communication with the experimental computer. Additionally, the Neuro-stack Control block also contained functions for controlling the TPU, such as loading/saving the machine learning model (TensorFlow Lite Model) to/from the Memory block, redirecting the data streams directly towards the TPU, and receiving the TPU's output once it is ready (see Chapter 4.). The incoming neural data streams can also be stored locally in Log Memory or transferred to an external storage on the experimental computer through the Neuro-stack Control block. Furthermore, an LED light can be triggered (to turn on/off) through available general-purpose input/output (GPIO) pins for synchronization purposes. These triggered on/off events are internally temporally aligned with the incoming neural data in order to synchronize it with data from eye-tracking cameras. A local network can be created either by using a separate access point (e.g., router, hotspot, etc.), or by the TPU Dev Board, which contains a network controller that can support access point topology and thus can create its own local network. This wireless mode means that a server is created on the TPU Dev Board to allow for other devices, such as the experimental computer (e.g., to view neural data in real-time) to access the Neuro-stack API functions. Wired mode is also supported through Local Interface block (e.g., experimental computer connected via USB-C). All devices connected to the local network use Network Time Protocol (NTP; [52]) to log events with timestamps fetched from a common server in order to synchronize them.

Neural data acquisition and especially electrical stimulation from inside the brain are sensitive functions that require security. For that reason, we required X.509 certificate for a device to connect to the wireless API server running on TPU Dev Board. X.509 is a digital certificate that uses public key infrastructure. In this prototype version, we only used one experimental computer to connect to the TPU Dev Board and thus a self-signed certificate served the purpose.

The high-level structure of the code was shown (Figure 3.9, and now we will zoom in on the base functionality that allowed high-speed data acquisition from many channels (Figure 3.10). Here, we are describing the translation from API calls to USB packets and vice versa, handled by multithreading architecture (Sense, Stim, FPGA, ... Processes; Figure 3.10). The API calls were handled by a central thread (Main Process), which accepts calls and accordingly constructs USB packet for each command (Figure 3.5) and sends it immediatly to USB Controller within USB Interface block, which then leaves TPU Dev Board and goes to Neuro-stack communication layer. Each command was intended for specific IC, which was then extracted at the FPGA level. Likewise, for each command, dedicated IC will respond with specific acknowledgement or in the case of recording with a neural data stream (Sense and Spike IC). Packets received back from the communication layer were first accepted inside Input Queue (FIFO Queue) as soon as the packet was available at the USB inputs. Reader Process awaits an interrupt signal from the FIFO Queue, indicating the the new packets are available, and distributes the packets based on the IC that sent it. Each dedicated IC has a queue and a process that handles its incoming packets. All processes first extract the message and check for its correctness (Error and CRC codes). If there was an error or mismatch in CRC code, an interrupt is sent to the Main Process, which then halts the pipeline and notifies the external computer. If everything is in order, Sense and Spike Processes send data samples to a shared, mutex-protected, Data Process that is responsible for saving samples locally and transmitting it to TPU processing or towards Neuro-stack Control block, which then forwards the data to the experimental computer. In this prototype version, FPGA and Stim Processes only check for the correct acknowledgement from the FPGA and Stim IC and then go to idle mode until next packet is received. The PLS Process also receives the data from Sense IC and then notifies the Main Process when the stimulation should be triggered based on the theta oscillations.

The basic test software running on an experimental computer was a Python 3.6.9 script, which connects to the network socket connection and fetches incoming data samples and stores/plots them. Tests that required additional functions on the experimental computer software side will be described later in the text. For the remainder of the text, we will always consider that the base wireless API and experimental scripts are used for all in-vivo and in-vitro experiments and we will just focus on the modified additions for specific applications, if they were necessary.

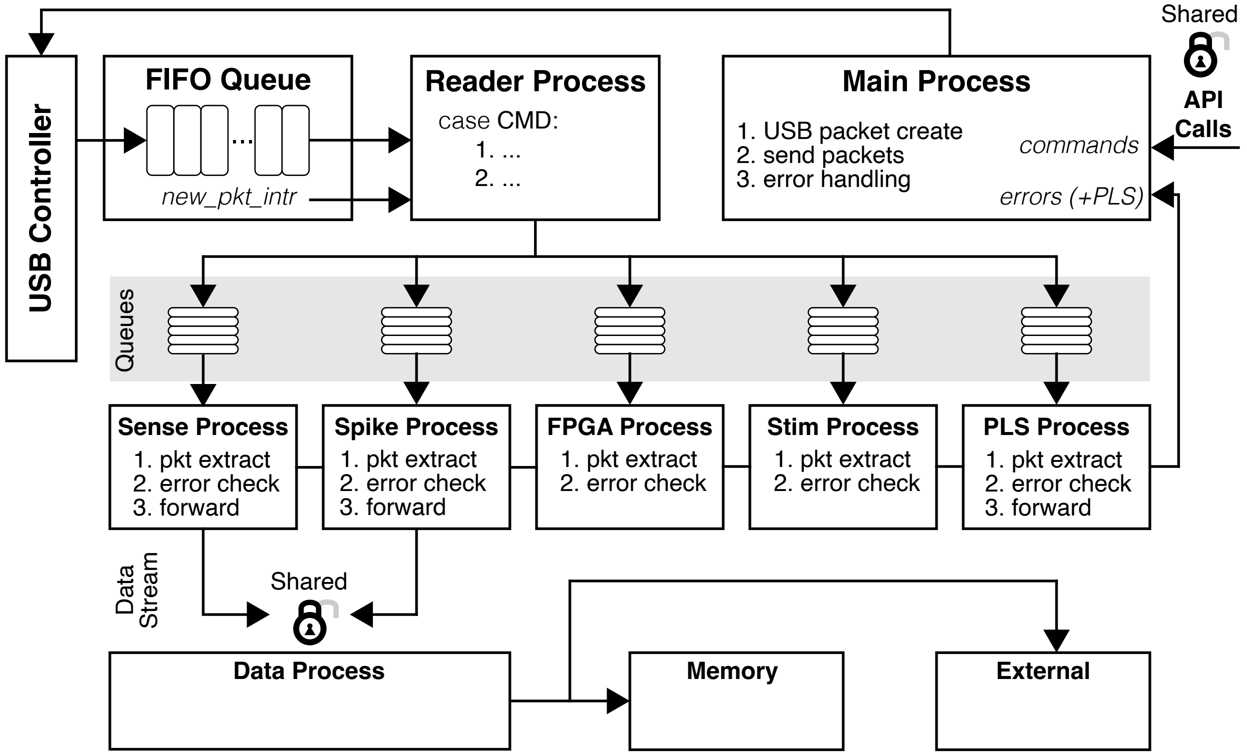


Figure 3.10: Simplified block diagram of the real-time pipeline that handles sending commands towards Neuro-stack ICs and receiving acknowledgements and data.

3.3 In-Vitro Validation

The Neuro-stack IC chips (i.e., Stim, Sense, Spike, PLS) were validated in-vitro separately [46]–[50] and some (Sense and Stim) as part of an implantable system [37]. Before moving to human in-vivo studies, in-vitro validation of all chips in the Neuro-stack was also completed. The setup for validating sensing capability included the feeding of pre-recorded analog neural data via an National Instruments (NI) PXI System (digital to analog converter) through a phosphate-buffered saline (PBS) solution, use of an oscilloscope to observe true signals at front-end inputs, and a computer to control and power the Neuro-stack (Figure 3.11).

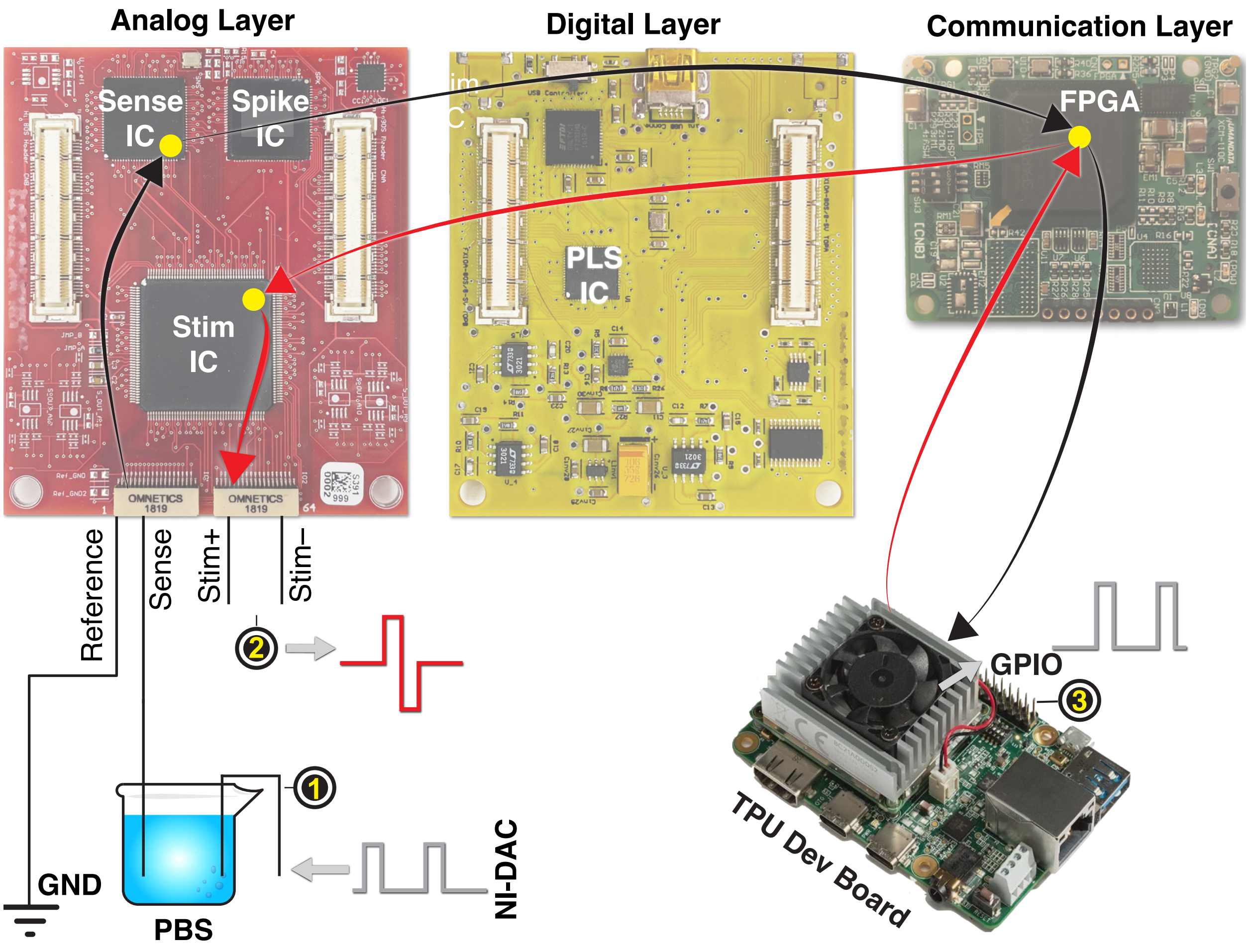


Figure 3.11: Neuro-stack in-vitro test setup.

Testing of the Sense and Spike ICs involved feeding 100 s of pre-recorded LFP/single-unit data through the NI-DAC. The analog signals were observed using an oscilloscope and recorded by a single channel using the Neuro-stack. For visualizing results, a time domain comparison was used for Sense IC and Spike IC (Figure 3.12). The Stim IC was tested as part of closed-loop delay measurements and in previous reports [37]. Delivered stimulation was captured by the oscilloscope and on one channel using the Neuro-stack (Figure 3.11). The PLS IC was tested in-vitro as part of a previous study [50], [51] and using the same in-vitro setup (Figure 3.11). For 300 s of LFP data, the results showed 400 detections within the theta band (3–8 Hz) and triggered stimulations with a circular variance of 0.3 [51].

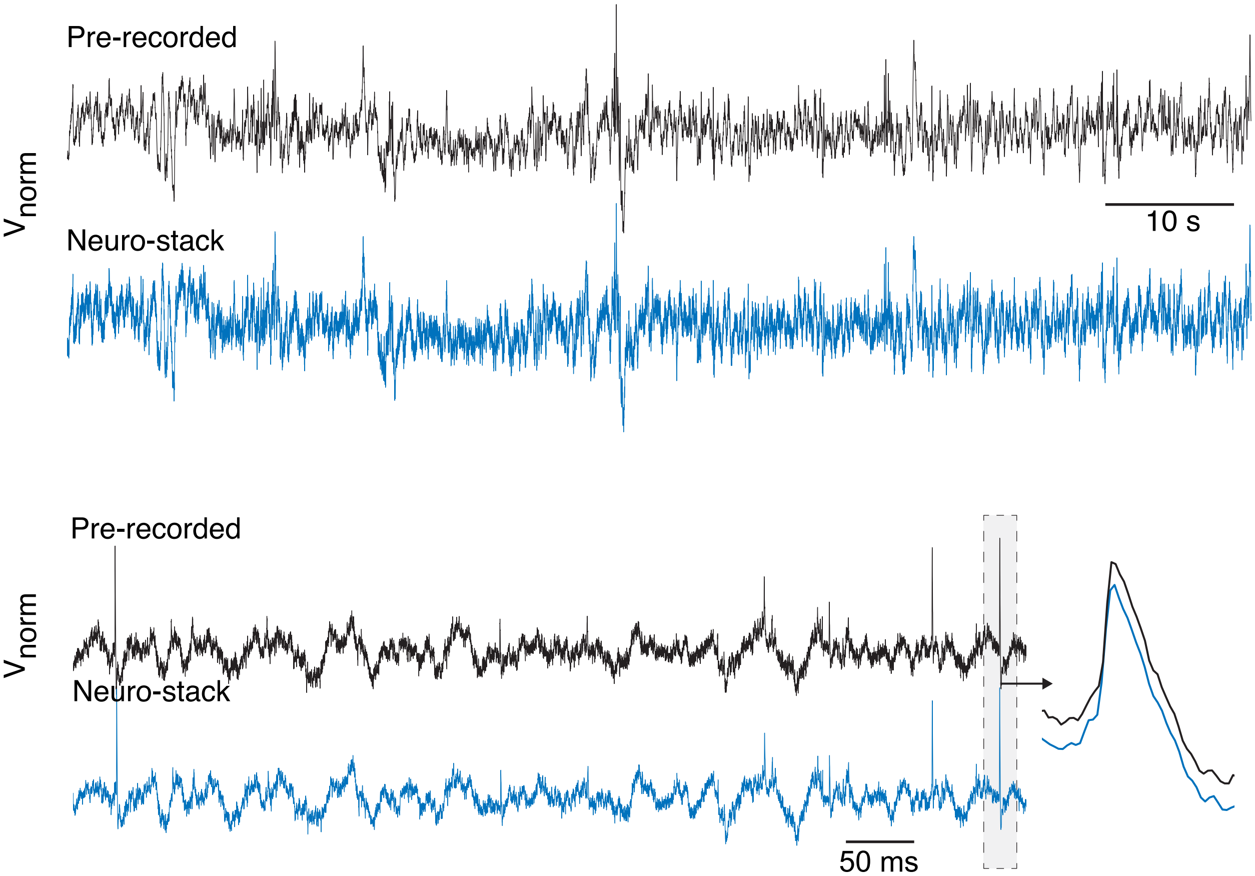


Figure 3.12: Neuro-stack in-vitro sensing validation.

Measurements of stimulation and synchronization delivery delays were also characterized for ensuring accurate closed-loop implementation as well as alignment between behavioral stimuli, neural data, and other devices that run in parallel.

First, the round-trip delay, important for closed-loop stimulation, was measured from sensed input to stimulation output by feeding a pulse train (50 pulses, 20 mV amplitude, 1 s pulse width, duty cycle 50%) from the NI-DAC to one channel recorded using the Neuro-stack. The modified software on the TPU Dev Board continuously pooled incoming samples and detected the increase from zero (rising edge) in these incoming values. Once detected the rising edge triggered one-pulse of stimulation. The delay (mean ± standard deviation [std] for 50 pulses) was measured on the oscilloscope by capturing both the recording input and stimulation output rising edges and their time difference (Figure 3.11). The pulse rising edge detection triggered stimulation on the TPU Dev Board software side (connected to the Neuro-stack via USB; Figure 3.11). Input/output observations by the oscilloscope showed a 1.57 ± 0.19 ms round-trip delay (Figure 3.13-left). This result was consistent with the PLS-based round-trip delay of 1.7 ± 0.3 ms measured from the sensed input to stimulation output [51].

Second, synchronization with external devices was done by timestamping neural samples using the TPU Dev Board; accuracy depended on the system latency through hardware and software. The Neuro-stack system and software latency from the recording input to the Sense Process thread on the TPU Dev Board was measured using the same pulse train process but instead of triggering stimulation, the detected rising edge triggers a 1 s pulse to the TPU Dev Board general-purpose input/output (GPIO) pin. We used the oscilloscope to observe the recording input and GPIO output, and measure the time difference between the rising edges, which was equivalent to the system latency (mean ± std for 50 pulses). Measured latency was 0.56 ± 0.07 ms (Figure 3.13).

To conclude this section, Neuro-stack can record iEEG, LFP, and single-units based on in-vitro tests. Also, it introduces millisecond delays into recording pipelines, which is crucial for online processing and closed-loop applications.

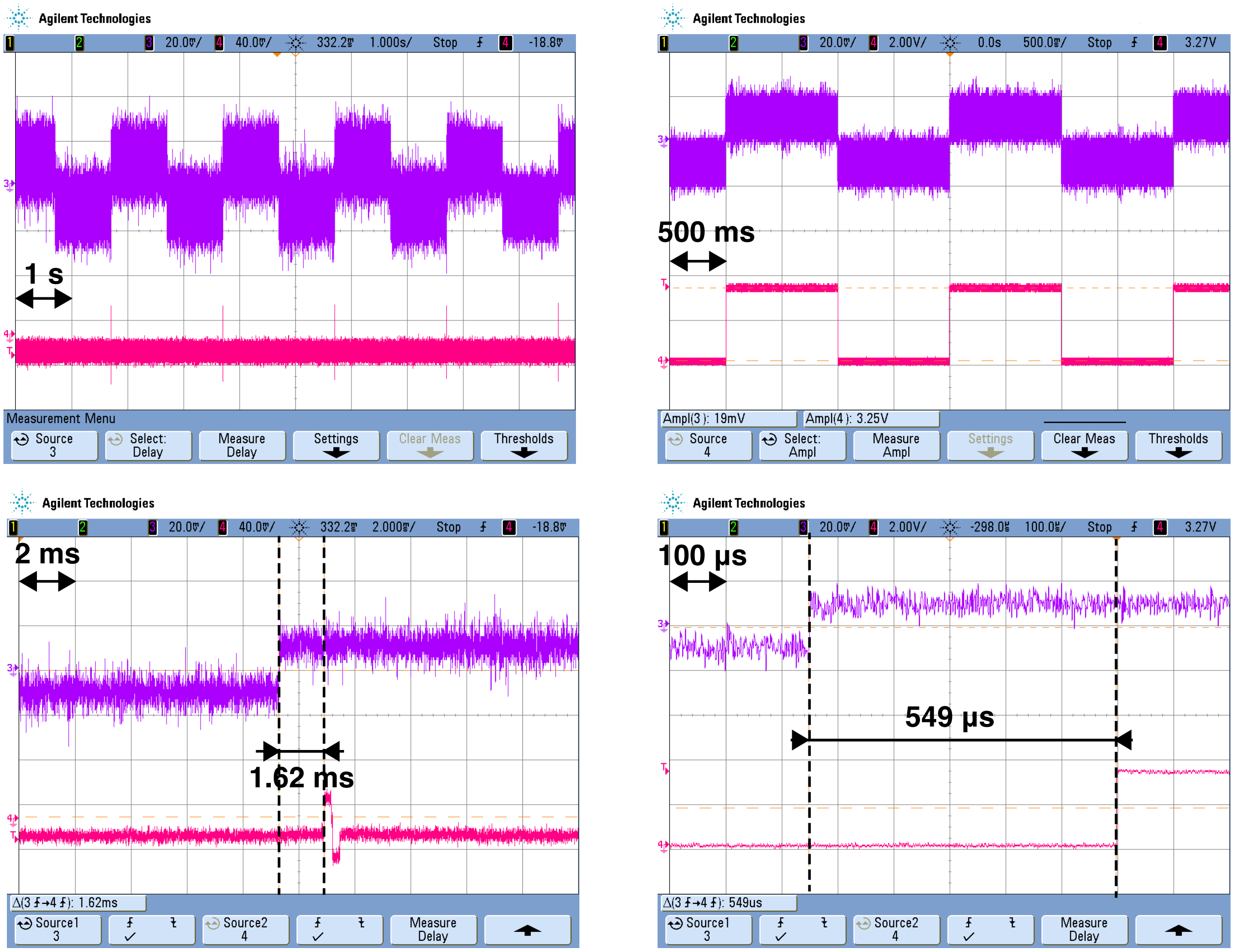


Figure 3.13: Neuro-stack sensing and closed-loop delays.

4 Proposed Neural Activity Decoder

This chapter outlines software and embedded implementation of a neural network model for decoding memory performance based on LFP activity from MTL. Before explaining technical implementation, we will describe a behavioral task that can trigger memory storage as well as methods from previous reports used to decode memory performance in similar studies.

4.1 Background

4.1.1 Verbal Memory Task

Verbal memory task tests participant's ability to successfully encode and recall words presented on a screen. The words appear serially, one after another, for 2 s with 4 s fixation cross between two words, which helps keep participants focused. There are usually from 8 to 10 of them, depending on participant's memory abilities. Words were drawn from clusters of six and seven of the word norms and were all 4-8 letter nouns that were rated as highly familiar (range 5.5-7 on a 1-7 scale), moderate to high on concreteness and imagery (range 4.5-6 on a 1-7 scale), and moderate in pleasantness (range 2.5-5.5 on a 1-7 scale) [53]. The word presentation phase is called encoding. To trigger long-term storage of the words, participants are then asked to perform mathematical operations and answer a question whether a sum of two numbers is odd or even. This phase is called distraction and lasts usually for 30 s. The last phase is called retrieval, and participants have 30 s to list out loud all the words that they remembered from encoding phase. One encoding, distraction, and retrieval

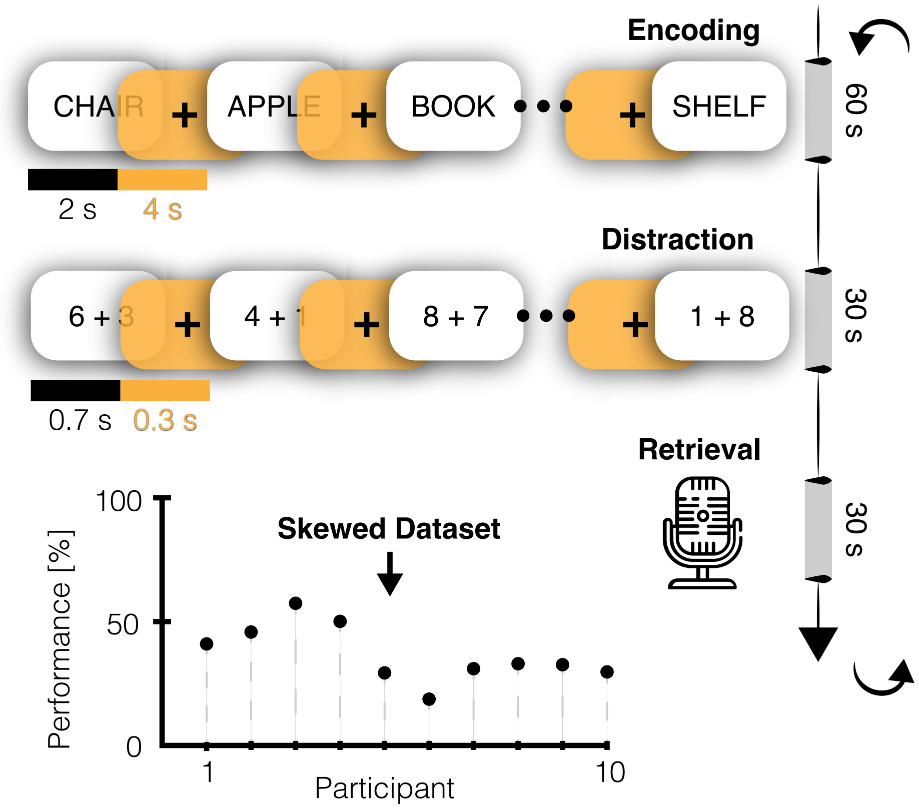


Figure 4.1: Verbal memory task and memory performance metrics in 10 participants.

cycle is called a trial, and during one experiment, participants often complete 5 - 10 trials (Figure 4.1).

This behavioral task was chosen for two reasons. First, it is easy to implement and run with participants. Second, we were in possession of a dataset from 10 participants recorded using Blackrock/Neuralynx during described task. The recordings were micro LFPs from hippocampal and entorhinal regions of the MTL. Doing the exact same task would give us an opportunity to compare analysis performance across systems or reuse the knowledge about memory encoding and recall in order to improve performance on the Neuro-stack. If we divide the data into two classes, remembered and forgotten words, the dataset itself is a skewed one as participants usually recalled less then 50 % of the words (Figure 4.1). Memory performance was calculated as the proportion of previously encoded verbalized words that were recalled.

4.1.2 Prior Work on Memory Classification

Previous reports have analyzed data from verbal memory task. In one prominent example, researchers have used logistic regression to perform binary classification to differentiate neural activity in the case of remembered and forgotten words [10]. The inputs to the classifier are sets of bandpower across different frequency bins and electrodes. Achieved average area under curve was 0.63 (Figure 4.2).

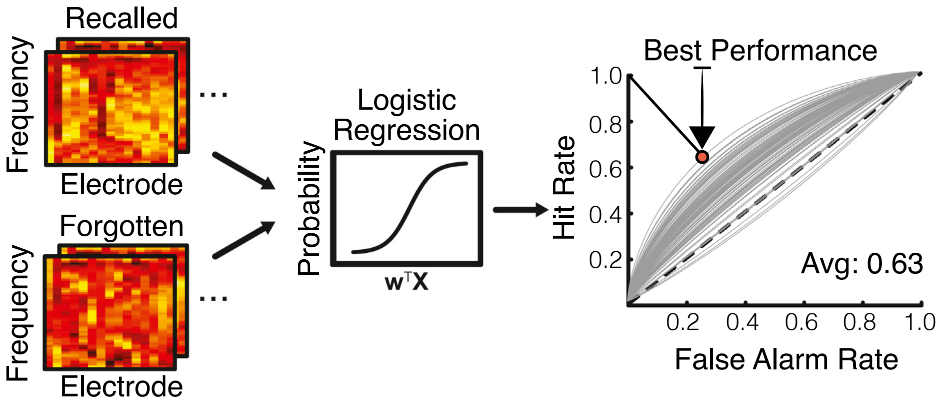


Figure 4.2: Verbal memory task binary classification using logistic regression. Adapted from [10]

.

The idea here was to use our dataset and surpass the performance of this reported classifier using artificial neural networks. Given that neural networks require a lot of data for proper training and testing, not many reports have used them on the data from the MTL because it is one of the hardest regions to access. Unlike MTL, neural networks have been successfully used for the data from cortical areas such as decoding speech from motor cortex [54].

4.2 Memory Decoder Software Implementation

4.2.1 Base Model

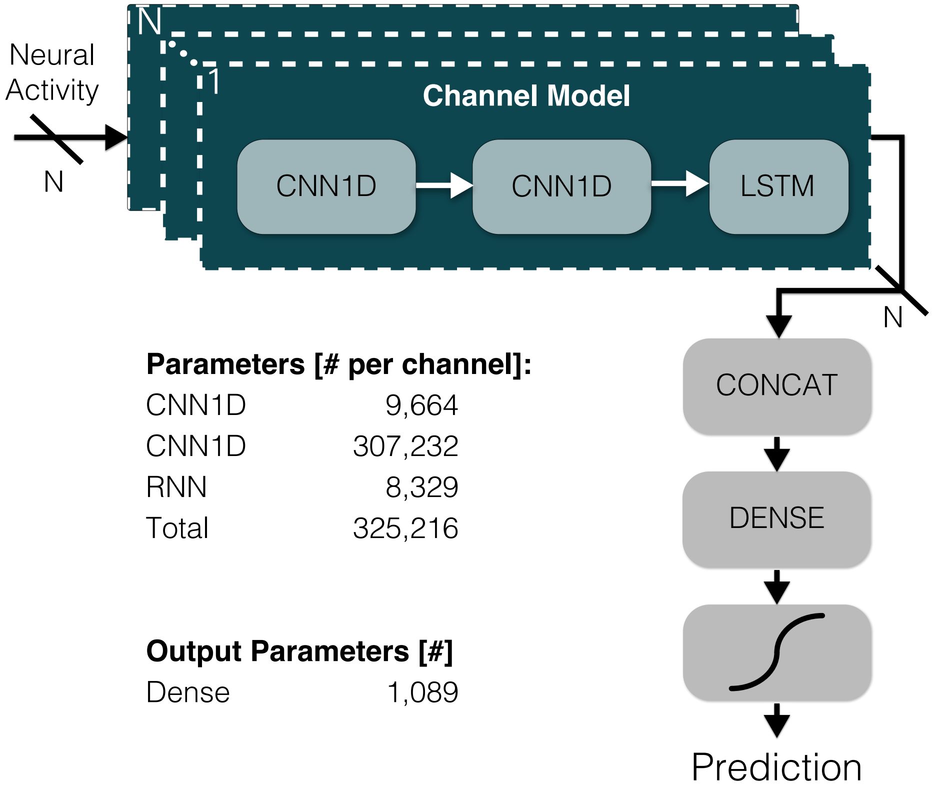
In order to perform binary classification and separate remembered/forgotten classes, we used an artificial neural network model (Figure 4.3). The model architecture included two one-dimensional convolutional neural network layers (CNN1D), first 32 nodes and second 64 nodes, and a long-short term memory (LSTM) recurrent layer with 64 nodes. We named each CNN1D+CNN1D+LSTM branch a channel model and used separate channel models for each brain region (N). After that, LSTM outputs from all channel models were concatenated and pushed through fully-connected Dense layers, and finally the classifier.

Separate channel models were implemented in anticipation of easier model explainability later and figuring out which regions contributed the most to model decision making. All models throughout this work were built using Keras with Tensorflow backend in Python 3.6.9.

4.2.2 Training and Testing

The model was trained offline using data from 6 MTL regions (left/right anterior hippocampus, left/right posterior hippocampus, left/right entorhinal cortex) from 10 participants (Figure 4.1).

LFP data (downsampled to 250 Hz, batch size 512) was extracted in chunks of 10 s (± 5 s) around the word onsets. Because word onsets were separated by 6 s, chucks had overlaps of 4 s, which were stored in separate windows as training samples. Furthermore, in order to give the model more information about the location of the word other than just time



Parameters [# per channel]:	
CNN1D	9,664
CNN1D	307,232
RNN	8,329
Total	325,216

Output Parameters [#]	
Dense	1,089

Figure 4.3: Neural network model includes multiple branches for each brain region and concatenated output.

positions, the 10 s chunks were multiplied by a Gaussian window function (mean:0, std: 2.5 s, cut-off: ±5 s), depending on the region. In order to augment dataset, we considered that all 10 s chunks coming from the same micro-wire bundle were just different samples from one input variable. Reasoning behind this was that LFP traces from 8 micro contacts are usually highly correlated (Figure 4.4).

To balance the dataset's two classes (positive – remembered word; negative – forgotten word), we shuffled the dataset and picked the randomly (uniform distribution) positive cases so that their number match the negative ones. Note that the number of cases for negative class was expectedly lower for all participants (Average memory performance: 35.86 ± 10.95

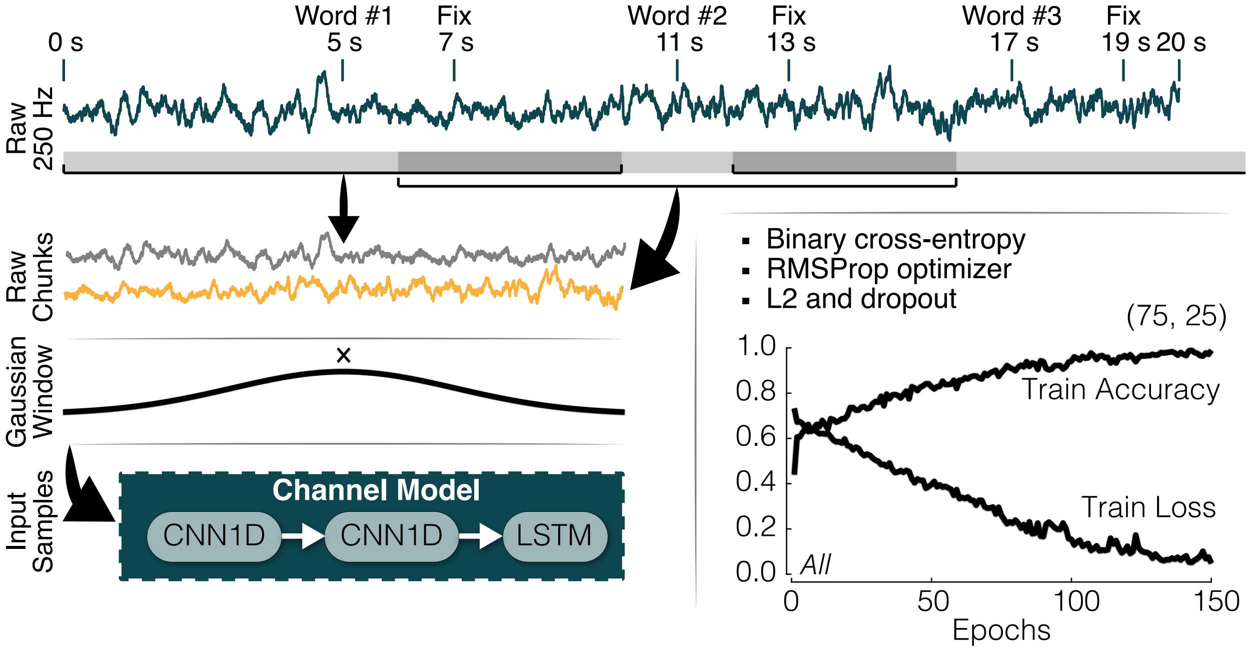
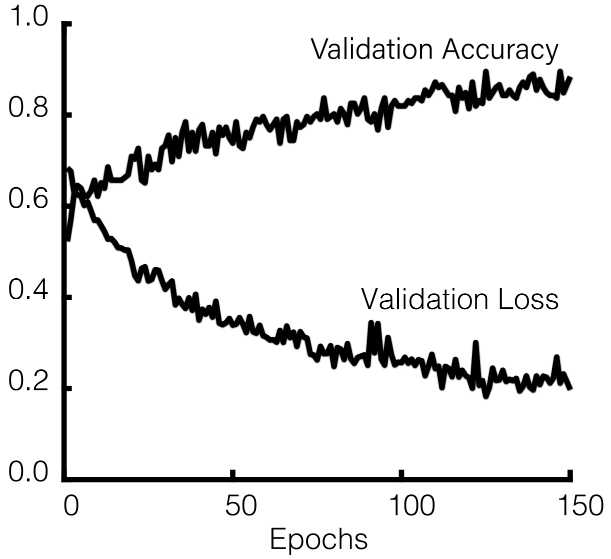


Figure 4.4: Neural network model training procedure with raw LFP data and training results.

%). The data from all participants was then divided into training (50%), validation (25%), and test (25%) sets. Then training and validation datasets were combined, shuffled, and used for training of the base model. Binary cross-entropy was used for the loss function, with root mean square propagation for the optimizer (learning rate of 0.001). The L2 regularization was used in the CNN1D and Dense layers and was proportional to the square of the weight coefficients' value. Moreover, to reduce overfitting further, the training dropout technique [55] was applied after each layer with a 0.2 rate, except for the LSTM, which used a 0.1 rate and a recurrent dropout (0.5 rate). Five-fold cross-validation (Figure 4.4 – training average across folds, Figure 4.5 - validation average across folds) was used for validation using the presented hyperparameters. Hyperparameter optimization was done during the validation phase and with respect to the F-1 score (0.5 threshold).



P	F1
1	85.79
2	81.48
3	82.59
4	83.34
5	92.94
6	87.15
7	96.11
8	95.34
9	85.78
10	93.32
μ	88.6

Figure 4.5: Neural network model validation and testing in 10 participants.

The final trained model was then used to make predictions for each participant's test set (4.5-right). Average F1-score across all participants was 88.6 ± 5.5 % (mean ± std).

The above-described neural network model was chosen after an extensive trial and error process during which multiple classification algorithms were tested on the same dataset. Specifically, before utilizing the neural network model, the data was classified using shallow methods such as Support Vector Machine (SVM). As part of the feature engineering process, we supplied SVM models with raw, power, and phase data in 0-250 Hz range chunks of 7 s (word onset at 3.5 s) or in a sequence of 1 s sliding time windows (with no overlap). Before choosing the final decoding model, we also tested several convolutional (CNN) and recurrent neural network (RNN) architectures. Summary of accuracies for each of these decoding methods is presented in Table 4.1. The final base model (Figure 4.3) that we chose based on the highest performance had inference latency of 2.8 ms, while performing ∼193 million MAC operations. This meant that with the closed-loop round-trip delay of 1.57 ms (See chapter

Proposed Sensing and Stimulation Device, section In-Vitro Validation), an estimated total delay would be ∼4.4 ms for a closed-loop that would involve neural network inference.

Table 4.1: Performance and latencies across platforms and models.

Classifier Algorithm	Accuracy Range	Input Type (Domain)
SVM	55 — 65	Power (Time-frequency)
SVM+PCA	65 — 90	Power/Phase (Time-frequency)
GoogleNet	69 — 96	Power/Phase (Time-frequency)
CNN2D	63 — 75	Power/Phase (Time-frequency)
LSTM	75 — 88	Raw (Time)
CNN1D	72 — 87	Raw (Time)

4.2.3 Explainability and Visualization

Other than being data hungry, one other challenge with neural networks is that they are hard to interpret. This especially holds true when neural networks are used for drawing neuroscience conclusions based on a trained and high-performing neural network decision making. Potential application in clinical treatments would be even more challenging and require full explainability of network's input to output mapping. It should be noted that the recurrent layers, such as LSTM layers, are far more challenging to interpret compared to convolutional layers. This is because LSTM output of a trained layer depends on the current state of the layer and not just the input, whereas convolutional layers input to output mapping can easily be extracted and visualized. Although further analysis is pending, we performed two methods to try to tackle explainability problem.

First, we tried to utilize initial decision to structure the input per brain region and we applied ”one-hot” encoding, where we fed the test data from only one region into corresponding channel model of the trained network, while keeping other inputs at zero. Example from one participant shows that certain hippocampal regions are contributing more to the decision than entorhinal regions (Figure 4.6). Results varied across participants, but the general trend was that the data from posterior and middle hippocampus contained the most information based on which the model successfully predicts the outcome of encoding phase alone with the accuracies of up to 88 % (Figure 4.6).

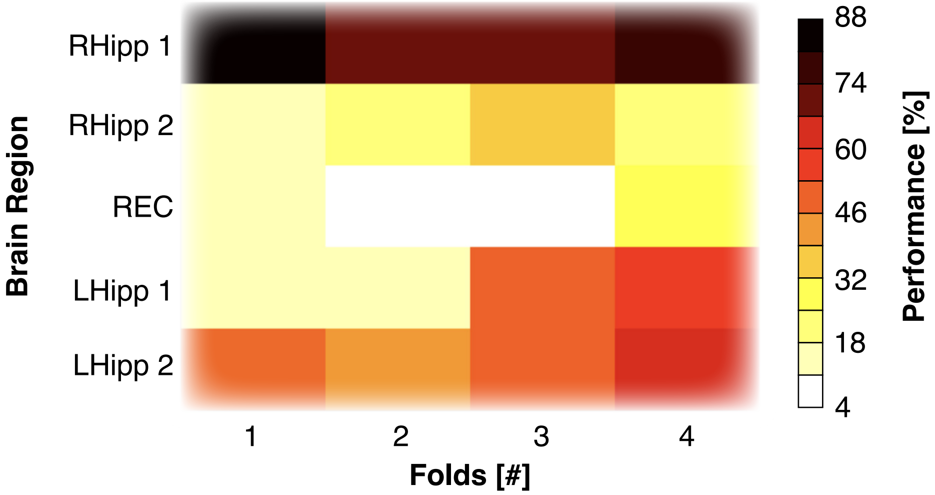


Figure 4.6: Neural network model partial explainability using 'one-hot' encoding input.

Second, one of the ways to check what trained convolutional part of the network does during the inference is to visualize its filter activations by displaying patterns that filters are meant to respond to. In order to do this, we applied gradient ascent at the input, that is to apply gradient descent to the input chunk values so as to maximize the response of a specific filter. The starting input chunk was 10 s with all samples being zero. Resulting chunk was the one that chosen filter is maximally responsive to. In our specific case, we performed this at the output of every filter in the second CNN1D layer for all channel models. The process was to build a loss function that maximizes the output of each filter and then to apply stochastic gradient descent, which adjusts the input chunk values so that filter output values is maximized. The loss function was average of the output for a given filter, and the gradient was with respect to the channel model input chunk. We also used L2 normalization

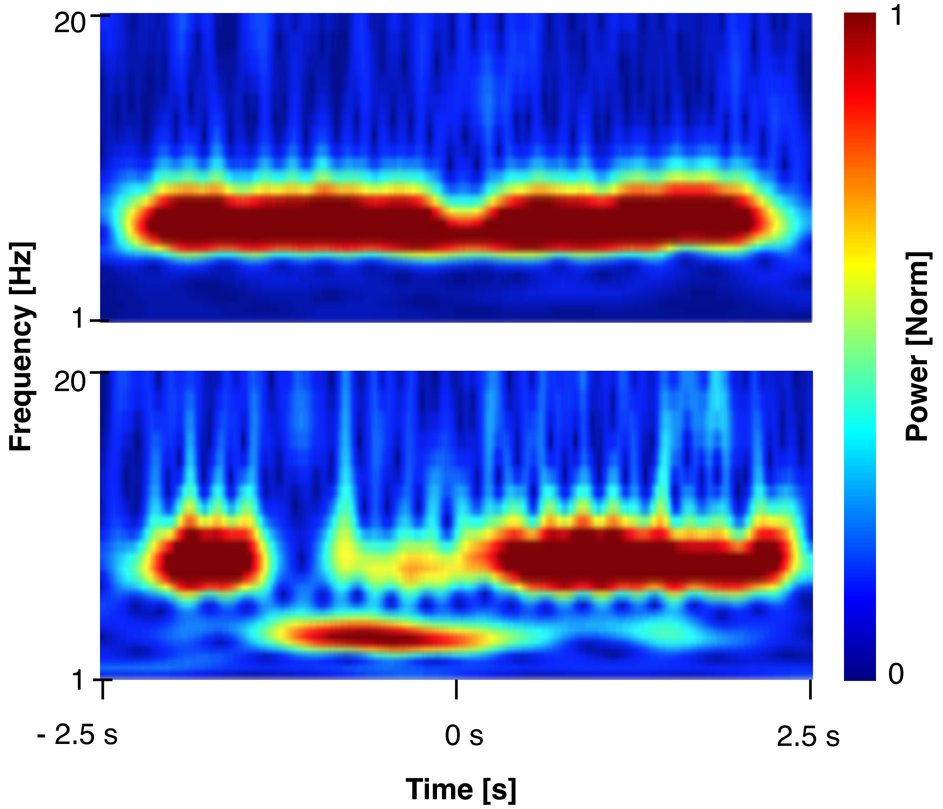


Figure 4.7: Neural network model partial explainability using CAM visualization presented as two CNN1D activation filters in time-frequency domain.

during gradient descent. Once completed the resulting input chunk was transformed into time-frequency domain using continuous wavelet transform with complex Morlet base in order to visualize whether the CNN1D layers were looking for specific oscillatory bands known to be signatures of verbal memory encoding. Although not all filters made sense in the context of neuroscientific knowledge, we are presenting the most interesting filters from the middle hippocampal branch, which maximally responded to theta bands (4 – 8 Hz) around the word onset (Figure 4.7). Association between MTL theta activity and memory functions is well established in literature. These results (Figure 4.7) should not be confused with a conclusion that neural 10 s chunk with strong theta power around the word onset is predictive of a successful encoding. Rather, this merely pointed out that filters with time-frequency transfer function that isolates theta bands (Figure 4.7) contributed to the model's final decision, ultimately made by layers that follow the second CNN1D layer, which could have been either a remembered or forgotten word. Prior reports suggested that successful verbal memory encoding is linked to lower theta band power ([10], [56]).

4.3 Embedded Implementation

4.3.1 Inference on the TPU

After successful offline decoding of the neural activity during verbal memory task, we wanted to explore online options that could utilize hardware resources of the Neuro-stack. To do that we needed first to convert trained model into the one that could be executed on the Edge TPU using the Neuro-stack's software infrastructure that was built for TPU Dev Board. Software neural network model is consisted of 32-bit floating point coefficient. For it to run on the TPU, those coefficients needed to be quantized to 8-bit fixed point and the model converted to TensorFlow Lite, before it could be compiled and transferred to TPU's memory.

Structure of the software responsible for the TPU inference is consisted of channel slices, which accept incoming neural data, preprocess them and then forwards to the Edge TPU once inference is externally triggered (Figure 4.8 – a zoomed-in portion of the Neuro-stack Control block from figure 3.9). Each channel slice was intended for one brain region and was used to restructure incoming data and preprocess it for the neural network. Every preprocess step involved downsampling to 250 Hz and z-score. Signal statistics were calculated during scanning phase (20 s) for each channel before every experiment and were kept inside channel slice memory for online normalization. Three 10 s FIFO (5 KB for 250 Hz data) were then continuously filled with incoming data. There were three FIFO blocks to account for three word onsets that had overlapping neural activity. Once the inference trigger was initiated,

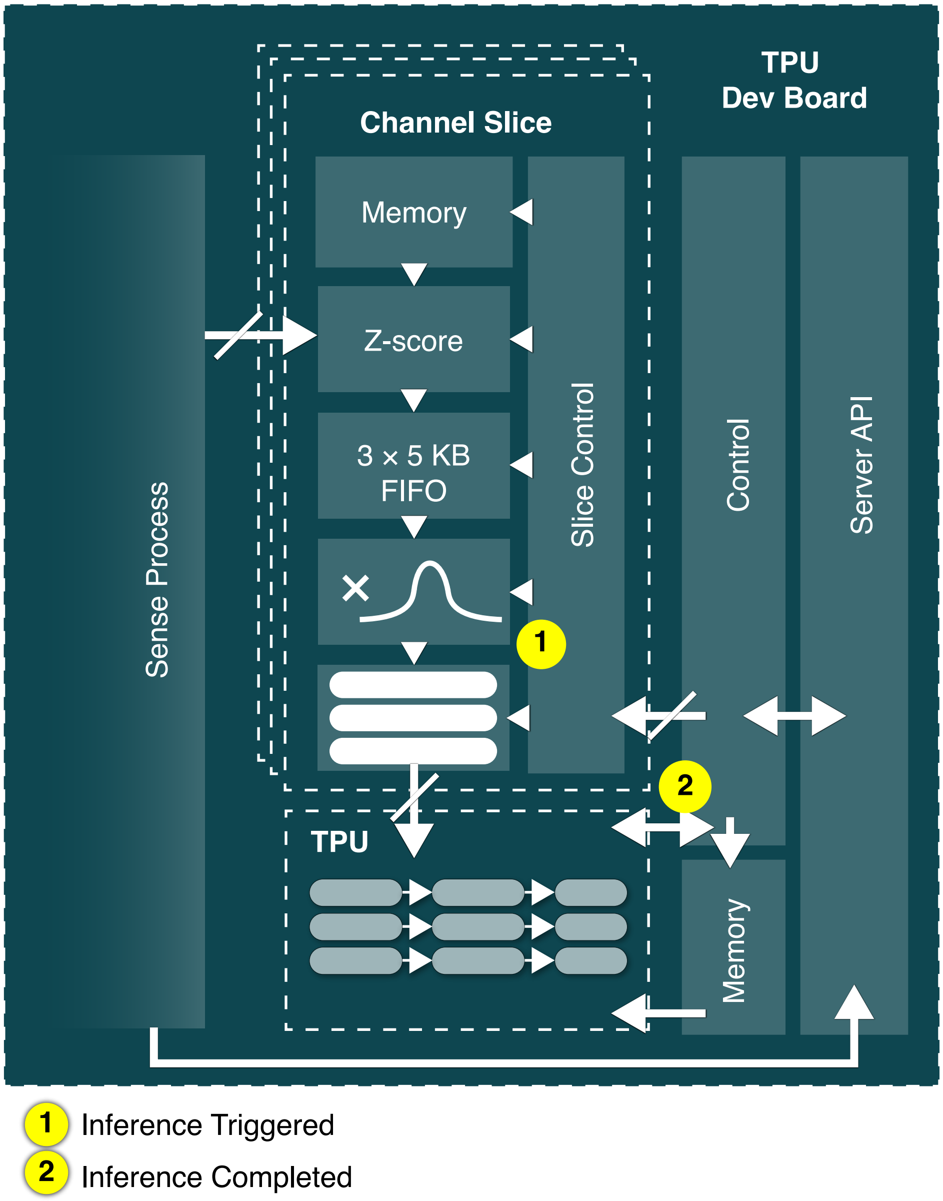


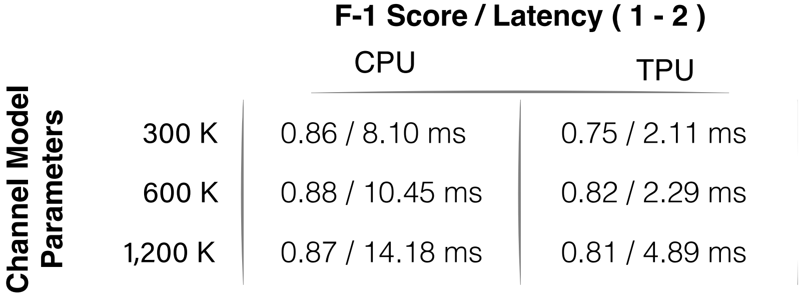
Figure 4.8: Neural network TPU embedded implementation.

either locally or externally through wireless server, the contents of the three FIFO buffers were multiplied with Gaussian window and transferred to the TPU from each of the channel slices. Inference predictions were then stored locally and/or transferred externally (Figure

4.8).

To test this operation, we converted software trained base model to TensorFlow Lite and transferred it to the TPU memory. Unless otherwise stated, all tests with this embedded neural network implementation involved emulation of the real-time data recordings by feeding prerecorded samples from the memory to channel slice inputs in time. Because the whole point of using the TPU was to shorten the latency in an online application, while retaining high decoding performance, we tested models with three different number of parameters and measured the inference latency and achieved performance accuracy (Table 4.2). The results justified the decision to move inference to the TPU as latencies decreased at the expense of reduced performances due to quantization compared to software implementation running on CPUs. In the case of 1.2 million parameters, the TPU implementation partially executed using ARM resources, which resulted in slightly higher latency, however our base implementation had lower number of parameters and worked entirely on the TPU.

Table 4.2: Performance and latencies across platforms and models.



	F-1 Score / Latency (1 - 2)	
Channel Model Parameters	CPU	TPU
300 K	0.86 / 8.10 ms	0.75 / 2.11 ms
600 K	0.88 / 10.45 ms	0.82 / 2.29 ms
1,200 K	0.87 / 14.18 ms	0.81 / 4.89 ms

4.3.2 Transfer Learning

Since the neural network model could execute on the TPU, the next question was whether it can operate in real-time and how can we retain the performance with new participant and previously unseen data. There were two challenges to achieve this. First, the TPU co-processor can only be used for inference and not for training. This was true when this work was under development, however at the time of this writing, partial retraining of the models on the TPUs was enabled. Second, working with participants in the hospital understandably comes with significant restrictions, especially in terms of time length allocated for experimental sessions. This meant that full training of the model on-site was not an option. Multiple separate sessions are most often not possible.

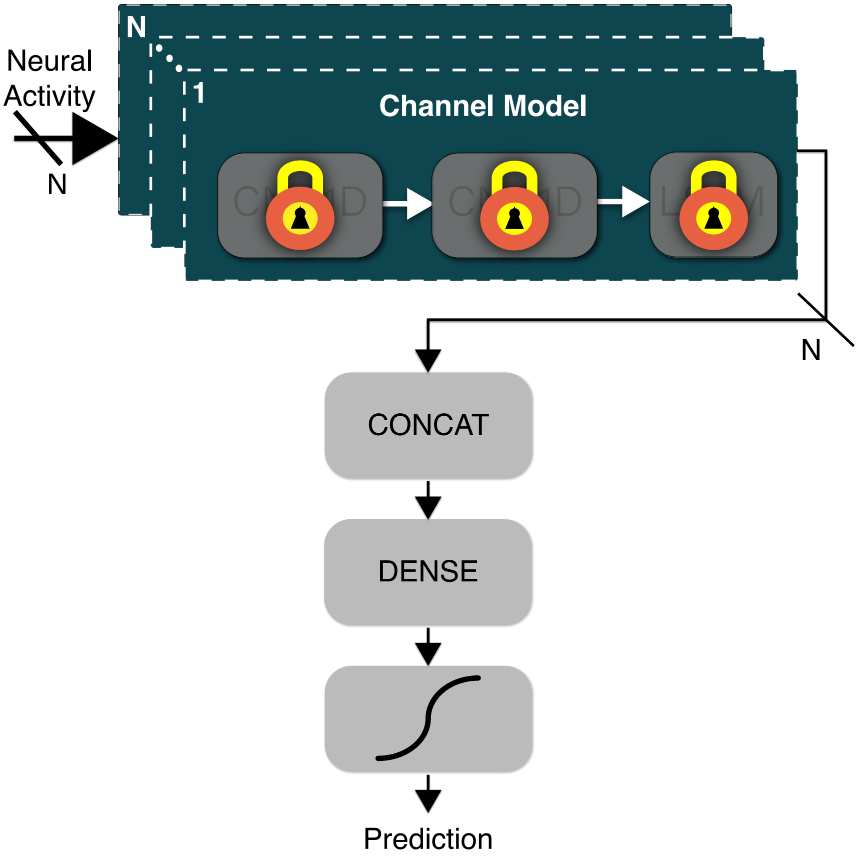


Figure 4.9: Partial retraining of the neural network model by locking channel model coefficients.

For these reasons, we opted for another technique common in deep learning field, transfer learning. The idea was to keep the trained base model mostly intact and just retrain output fully-connected layers with the new data. In our case that specifically meant keeping coefficients of the channel models fixed and retraining all layers coming after LSTM nodes (Figure 4.9). This way, time necessary for training phase would be significantly reduced, allowing a real-time operation during only one session with the participant, while still adjusting the model towards each participant's neural activity.

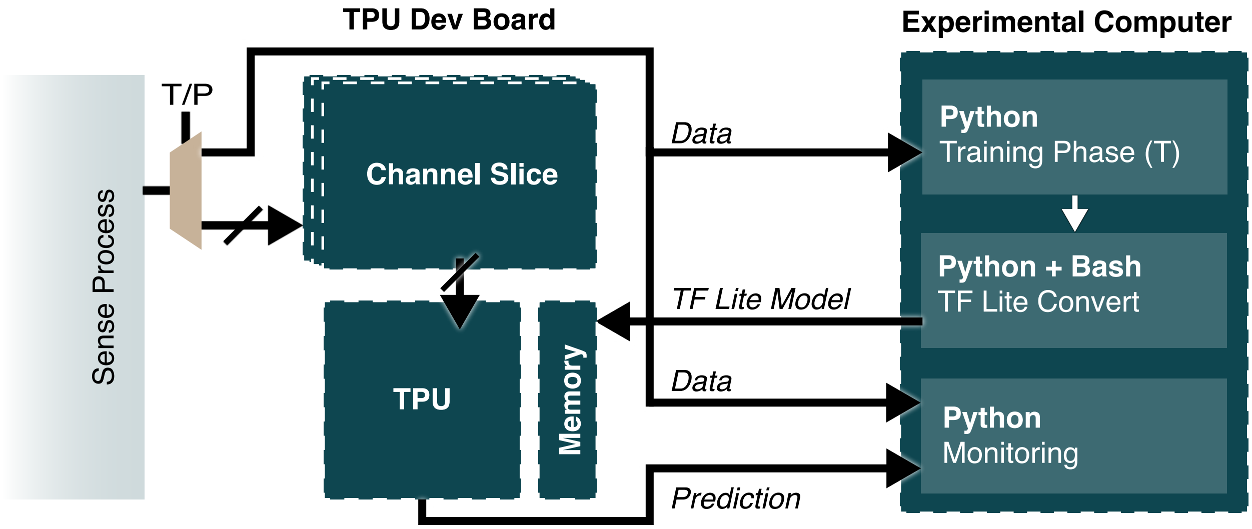


Figure 4.10: Transfer learning embedded implementation with partial external retraining and embedded inference.

Because the model cannot be retrained on the TPU model, we used the external computer to perform this phase, which was still reasonable given short time required for partial coefficient tuning. During retraining phase neural samples were directly forwarded over the wireless network to the experimental computer, where we ran Python and Bash scripts that automatically perform training and conversion to the TensorFlow Lita model. The model was then automatically transferred to the TPU's memory, ready to be triggered during prediction phase (Figure 4.10). Channel slices were equivalent to those described earlier (Figure 4.8), as transfer learning operation was just adjusting the model coefficient and not the way we perform inference. Neural samples were then streamed through channel slices and the TPU during inference. Samples and inference predictions were at the same time, once available, streamed to the external computer for monitoring.

As mentioned earlier, for in-vitro tests we emulated real-time neural streaming from the TPU Dev Board memory. The neural samples per brain regions were packaged in chunks of 10 s. During the training phase all incoming chunks were used for training as such that whenever we received new chunks, we shuffled them with the previous ones from the same session and used them for retraining (Figure 4.11). All model coefficients were kept and updated with the next iteration following the previous one. The dataset was not balanced in this case with the intention to use all data chunks from the new participant for retraining and testing regardless of the class. Furthermore, to reduce the burden on the TPU even further, we used the fact that only some channels significantly contribute to the final predictions (Figure 4.6) and ones all samples during training phase were used, we used 10% of the same data to run ”one-hot” encoding inference on the experimental computer. Three best performing regions were then selected and new model with only three channel branches was retrained again on the shuffled data. Final model was then ported to the TPU.

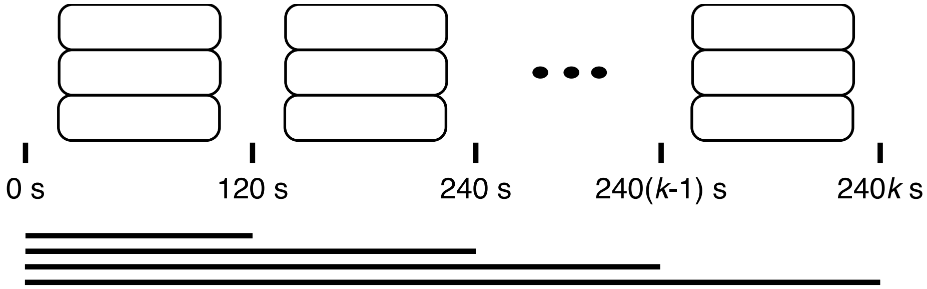


Figure 4.11: Real-time transfer learning retrains the model with shuffled old and new data.

4.4 In-Vitro Validation

4.4.1 Transfer Learning Compared to Software and Embedded Base Model

To test whether transfer learning operation could help adjust model's coefficient to the new participant we emulated real-time operation and timely forwarded pre-recorded LFP samples within TPU Dev Board. To make a correct inference on unseen data comparison, we also retrained the base model as well so that it contains information about 7 participants from our dataset and remaining 3 were used for testing on both CPU and TPU.

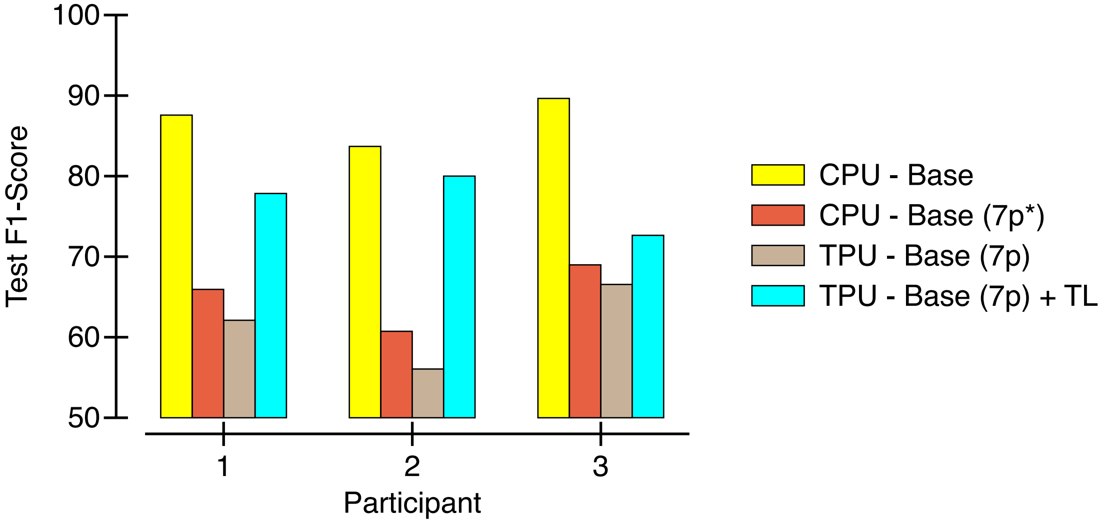


Figure 4.12: Transfer learning compared to software (CPU) and embedded (TPU) inference on unseen data.

Results clearly showed that transfer learning could quickly help readjust the model to work better on new participants (Figure 4.12). The best results were expectedly obtained using the software implementation, which possessed the information about the 3 participants. Once the data from these 3 participants was excluded from the training the F1-scores dropped on the CPU and even more so on the TPU due to quantization. However, when we used transfer learning retraining (Figure 4.10), F1-scores jumped by 15% on average. These results justified the use of the TPU to shorten the latency of online processing and transfer learning to make it perform better on previously unseen data.

4.4.2 Closed-loop Applications

Another question was whether the neural network decoders could be used in closed-loop stimulation applications to, in this case, enhance memory performance. Of course, this question can only be answered with comprehensive in-vivo closed-loop behavioral study, where we would trigger electrical stimulation whenever decoder predicts negative outcome, that is that the participant is likely to forget current word in this case. Even though successfulness of this cannot be tested in-vitro, we could still test technical feasibility of such experiment. In other words, we wanted to test how early with respect to a word onset could the classifier make most probable prediction. By testing this it could also give us insight about the round-trip delay, which would now additionally include the decoder in the loop compared to the first closed-loop test when we only performed pulse detection (Figure 3.11).

Given that closed-loop algorithms should operate independently, the inference had to be triggered periodically, and we needed a different way to provide task relevant information, such as timing of word onsets, to the model. A workaround that we used was to supply LFP data synchronized with a train of pulses (10 mV, 10 ms) via NI-DAC to sensing front-ends. Pulses were synchronized with the data by having rising edges happen simultaneously to the beginning of chunks (5 s before each word onset). For this experiment, we only chose one channel from one participant shown to significantly contribute to memory predictions. Then at a certain time t_0, representing the earliest point at which the decoder could already predict the negative outcome, we would deliver a stimulation burst with 10 pulses (Figure 4.13).

It should be noted that FIFO buffers on the TPU board were filled in serially, and when

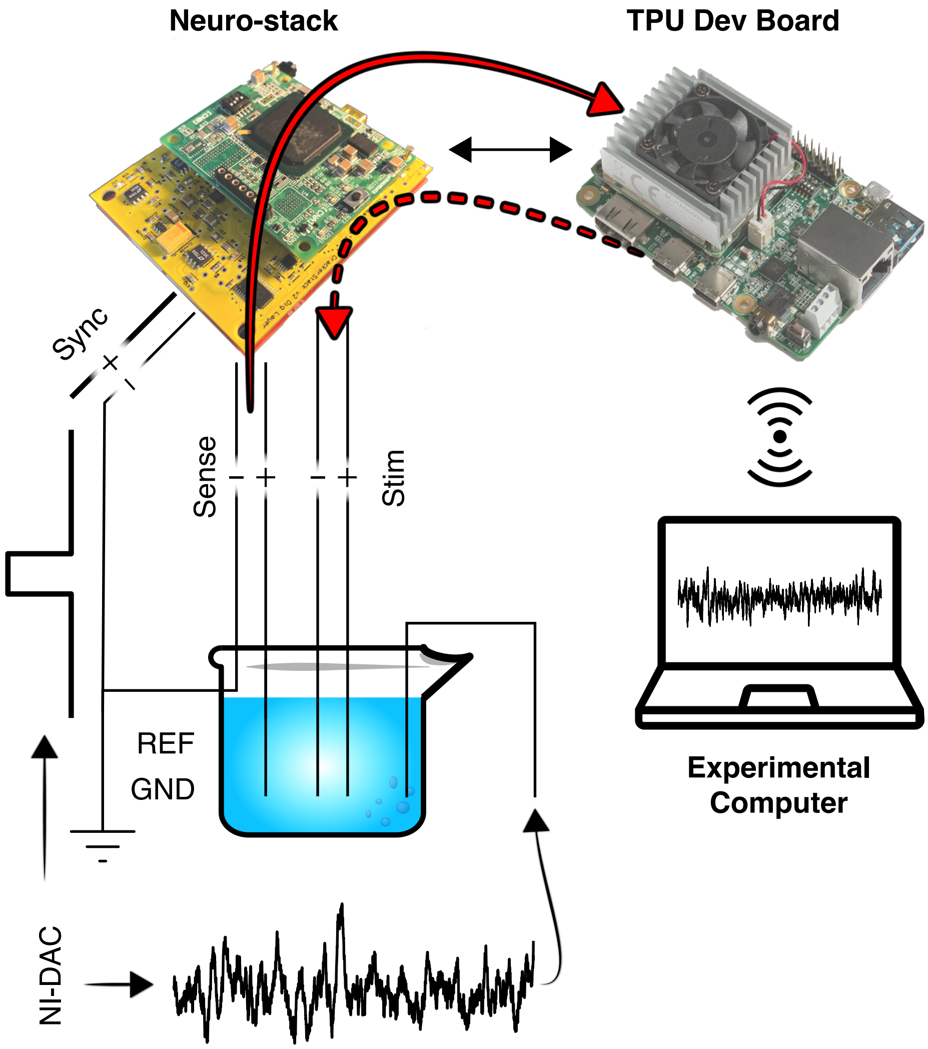


Figure 4.13: Neuro-stack decoder in-vitro validation setup for online inference and closed-loop triggers.

no data was present at the beginning of the experiment, missing samples would have zero value in the buffer (Figure 4.14). When we ran inference every 100 ms across 10 s chunk on the data from one participant, it was shown that in 83.1% of the words, the outcome turned out to be negative if at $t_0 = -100$ ms (where 0 s was word onset) decoder's output probability was less than 0.5. Inferences before -100 ms had success rate of $< 50\%$ and those after increased but not significantly. This meant that the outcome of the encoding process could be deciphered even before word was shown to the participant, which was in line with some previous reports.

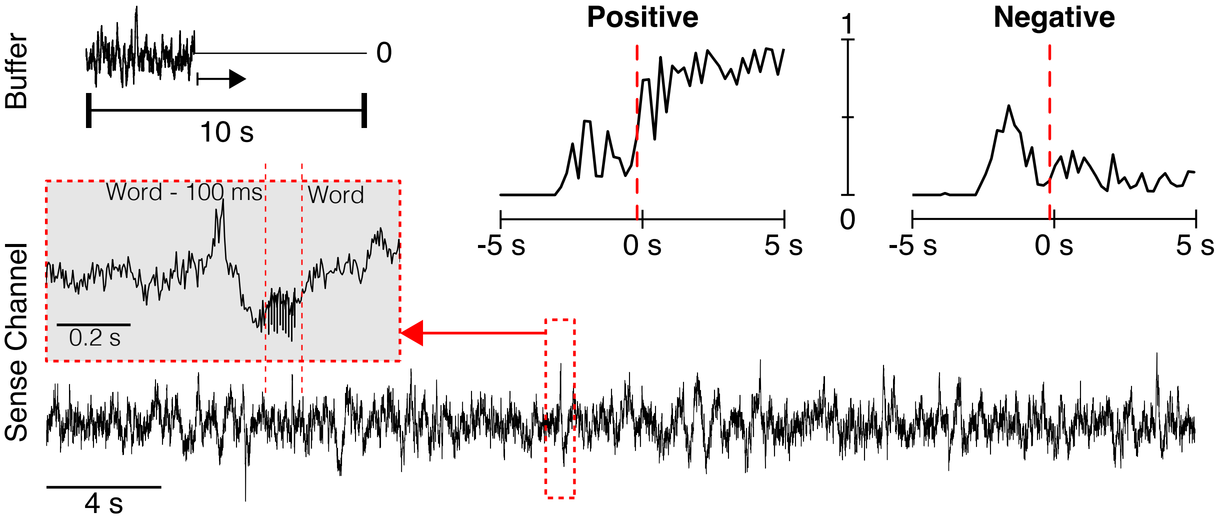


Figure 4.14: Neuro-stack decoder decisions can provide window during which stimulation could alter memory encoding.

In one example LFP trace, we shown the stimulation delivery by running inference at $t_0 = -100$ ms which resulted in negative prediction (Figure 4.14). This test also gave us insight about round trip latency because stimulation burst appeared one cycle after t_0. Plotted LFP had 250 Hz sampling rate, which meant that round trip latency was in the range of from 4 ms to 8 ms.

This test provided preliminary in-vitro proof that decoder can be used in a closed-loop in terms of its ability to process the raw data online and low round-trip delay. However, further work is required to establish meaningful hypothesis that could be tested through a separate study. Another point worth mentioning is that average probability plots over time for positive and negative classes (Figure 4.14-top) were generated prior stimulation attempt (Figure 4.14) on non-stimulated data to gather insights about how soon can the decoder make a decision. Further analysis is necessary to check how decoder's output change once stimulation artifacts are present in the data.

5 Human In-Vivo Validation

5.1 Experimental Equipment and Participants

One of the important aspects of this work was its incremental development informed at every step by human in-vivo experiments. This is often not the case when developing neural interface technology.

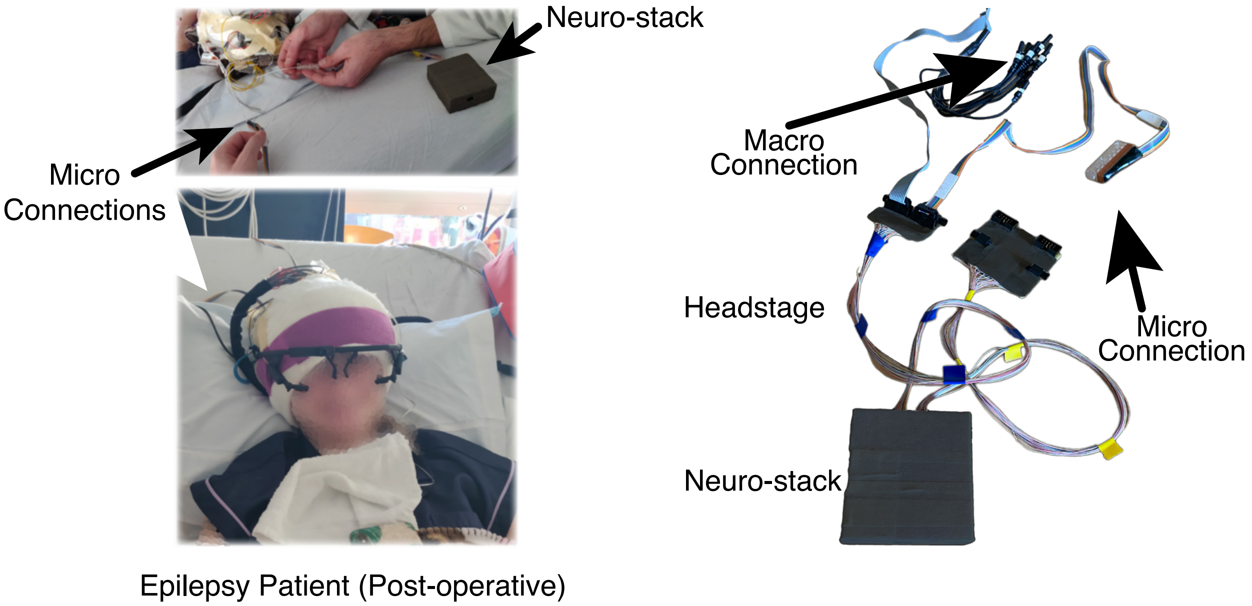
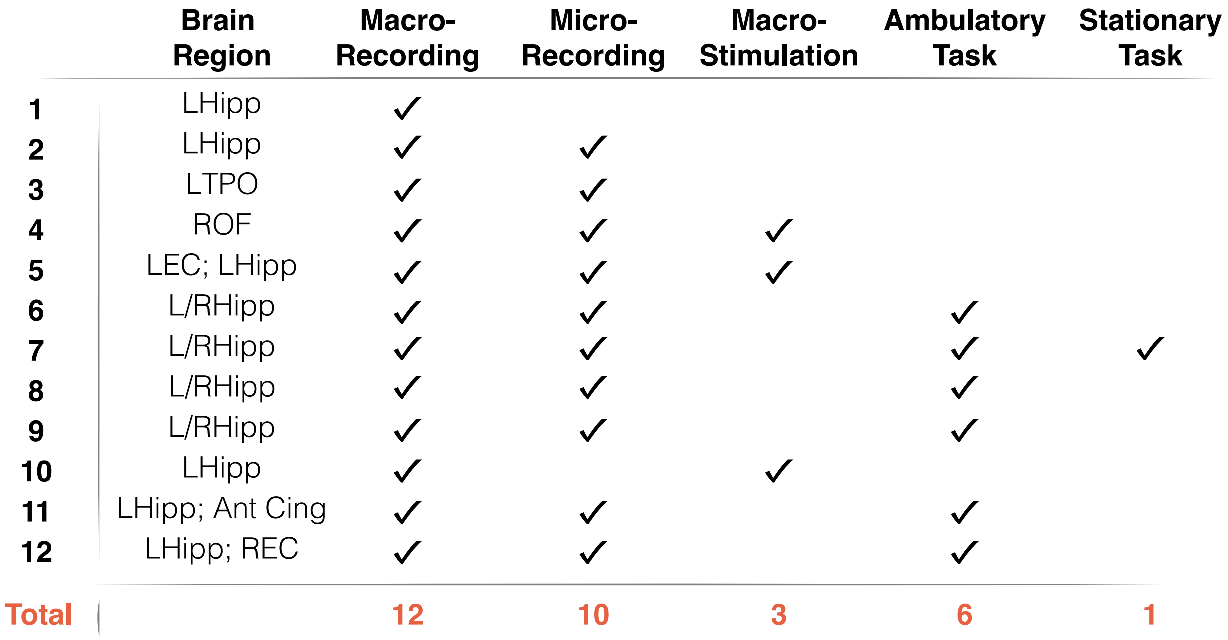


Figure 5.1: Epilepsy patient (post-operative) connected to Neuro-stack in the hospital (left). Neuro-stack headstages and connections to maco- and micro-electrodes (right).

5.1.1 Participants

Research participants (Figure 5.1, Table 5.1) were twelve patients (mean age 24.15 years, 9 females) with pharmacoresistant epilepsy who were previously implanted with acute stereo EEG depth macro- and micro-electrodes for seizure monitoring. Participants volunteered for the research study during their hospital stay by providing informed consent according to a research protocol approved by the University of California, Los Angeles (UCLA) medical institutional review board (IRB) approved protocol.

Table 5.1: Neuro-stack in-vivo validation with details about participants, brain regions, type of recording, stimulation, and behavioral task



	Brain Region	Macro-Recording	Micro-Recording	Macro-Stimulation	Ambulatory Task	Stationary Task
1	LHipp	✓				
2	LHipp	✓	✓			
3	LTPO	✓	✓			
4	ROF	✓	✓	✓		
5	LEC; LHipp	✓	✓	✓		
6	L/RHipp	✓	✓		✓	
7	L/RHipp	✓	✓		✓	✓
8	L/RHipp	✓	✓		✓	
9	L/RHipp	✓	✓		✓	
10	LHipp	✓		✓		
11	LHipp; Ant Cing	✓	✓		✓	
12	LHipp; REC	✓	✓		✓	
Total		**12**	**10**	**3**	**6**	**1**

Each Behnke-Fried macro-micro depth electrode (Ad-Tech Medical, Racine, WI) in every patient had 8-12 flexible polyurethane depth electrodes (1.25 mm diameter) and were implanted solely for clinical purposes and prior to completion of the research study. Each depth electrode contained 7-8 macro-contacts and terminated in a set of nine (8 recording, 1 reference) insulated 40 µm platinum-iridium microwires (impedances 200-500 kΩ) inserted through the macro-electrode's hollow lumen.

5.1.2 Neuro-stack Setup

Neural activity was recorded from macro- and micro-wire contacts using the Neuro-stack during wakeful rest, stationary, and ambulatory behavior from various brain regions (Table 5.1). The Neuro-stack setup was done bedside (Figure 5.1) or on-body during ambulatory movement, where the system was connected to implanted electrodes using a custom-built connector (i.e., touch-proof, Cabrio, and Tech-Attach connectors for commercial Behnke-Fried macro- and micro-electrodes, respectively; Figure 5.1). The main objective of the in-vivo validation studies was to test recording of iEEG, single-unit, and LFP activity and macro-stimulation during rest and behavioral tasks. The PLS closed-loop functionality has been tested in-vitro [51] with expected in-vivo validation to be a part of future behavioral studies.

For all in-vivo validation sessions, a Neuro-stack with two analog layers was used, which allowed for up to two micro-electrode bundles (16 channels) and eight macro-electrodes (16 bipolar channels). All micro- and macro-electrode recording sessions were sampled at 38.6 kHz and 6250 Hz, respectively. Base recordings were done without hardware decimation, non-linear correction, and artifact rejection on the Sense IC.

Stacked layers were placed inside a plastic enclosure (Figure 5.1) and wrapped from the inside with copper foil shielding tape to reduce the impact of the noise. Custom headstages (Figure 5.1) were built on a protoboard by placing two 5 × 2 connectors on each, which were internally routed to the Omnetics connector.

5.1.3 Electrode Localization

Electrodes were localized to specific brain regions using methods that have been previously used [57]. Briefly, a high-resolution post-operative CT scan was co-registered to a pre-operative whole brain MRI and high-resolution MRI using BrainLab stereotactic localization software (www.brainlab.com and FSL FLIRT (FMRIB's Linear Registration Tool [58]). Medial temporal lobe (MTL) regions, including the hippocampus and entorhinal cortex, were delineated using the Automatic Segmentation of Hippocampal Subfields (ASHS [59]) software using boundaries determined from MRI visible landmarks that correlate with underlying cellular histology. White matter and cerebral spinal fluid areas were outlined using FSL FAST software [60]. Macro- and micro-electrode contacts were identified and outlined on the post- operative CT. For a list of localized brain regions in all participants see Table 5.1.

5.2 Resting-State Validation

5.2.1 iEEG Sensing

First test was to perform a concordance study, where we recorded iEEG activity from macro-electrodes using both the Neuro-stack and commercially available electrophysiological recording systems (i.e., Nihon Kohden) for comparison purposes. We performed monopolar recordings (scalp reference) from anterior hippocampus and y-split the connections to record with the Neuro-stack at 6.25 kHz and with the Nihon Kohden system at 2 kHz (Figure 5.2). We also performed bipolar recordings from the same hippocampal region as well as the Neuro-stack recordings only without Nihon Kohden. We used audio pulses sent to both systems from the experimental computer for synchronization purposes.

Raw recordings from both systems were filtered and downsampled to 0 Hz – 250 Hz band, which we used for comparison in time domain (Figure 5.3-top). Then we calculated

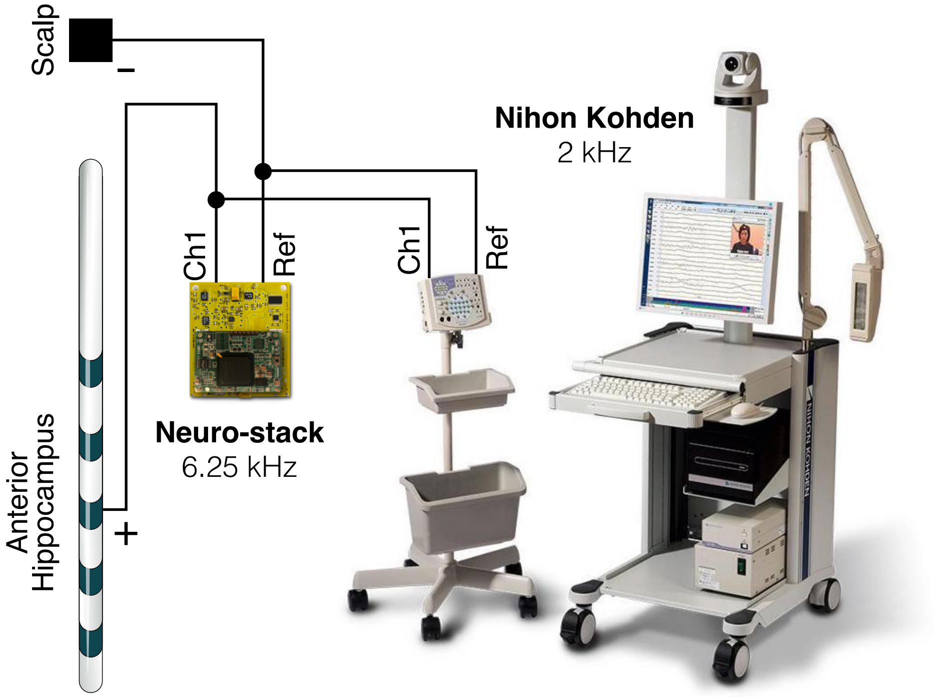


Figure 5.2: Concordance iEEG test with the Neuro-stack and clinical Nihon Kohden monitoring system.

continuous wavelet transform (CWT) using complex Morlet with 3 cycles as base wavelet in 77 logarithmic frequency steps (Figure 5.3-bottom/left). Finally, we calculated power spectral density (PSD) using FFT (Fast-Fourier Transform). The FFT length chosen was the largest power of 2, less than the length of the observed iEEG trace. The coefficients were then normalized with the trace length. Finally, the squared absolute value of the spectral coefficients multiplied by 2 (one-sided FFT) resulted in the PSD (Figure 5.3-bottom/right). All frequency and time-frequency domain plots correspond the time trace from one channel presented (Figure 5.3-top).

Obtained results showed nearly identical recordings from the two systems except for the slightly higher noise floor of the Neuro-stack at frequencies above 100 Hz (~ -40 dB). The y-

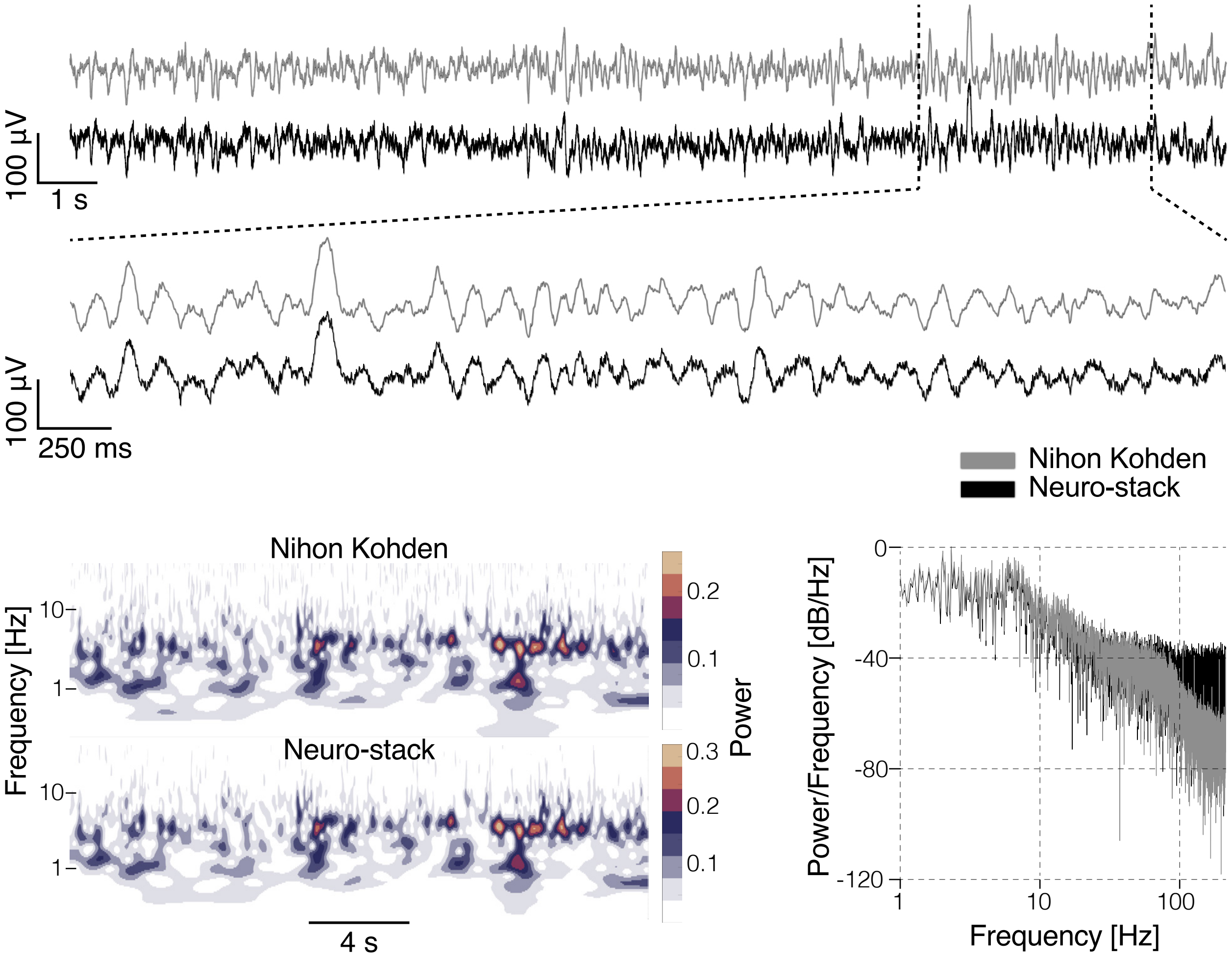


Figure 5.3: Concordance iEEG results in time, frequency, and time-frequency domain.

splitter connection between the systems affected the Neuro-stack recordings to higher degree as once we performed the Neuro-stack recordings alone the noise floor dropped by a ∼25 dB.

5.2.2 Macro-Stimulation

Stimulation was performed in three participants to test stimulation artifact propagation across channels and assess associated statistics with varying parameters. In the first two participants, bipolar macro-stimulation was applied to the left hippocampus (amplitude:

0.5 mA; Pulses/burst: 11; waveform shape: rectangular; pulse width: 1ms; frequency: 100 Hz). After successful delivery was observed in surrounding channels, a series of bipolar macro-stimulation bursts with varying parameters was delivered in a third participant. We recorded bipolar from eight macro-electrodes, one of which was used to deliver bipolar macro-stimulation (Figure 5.4).

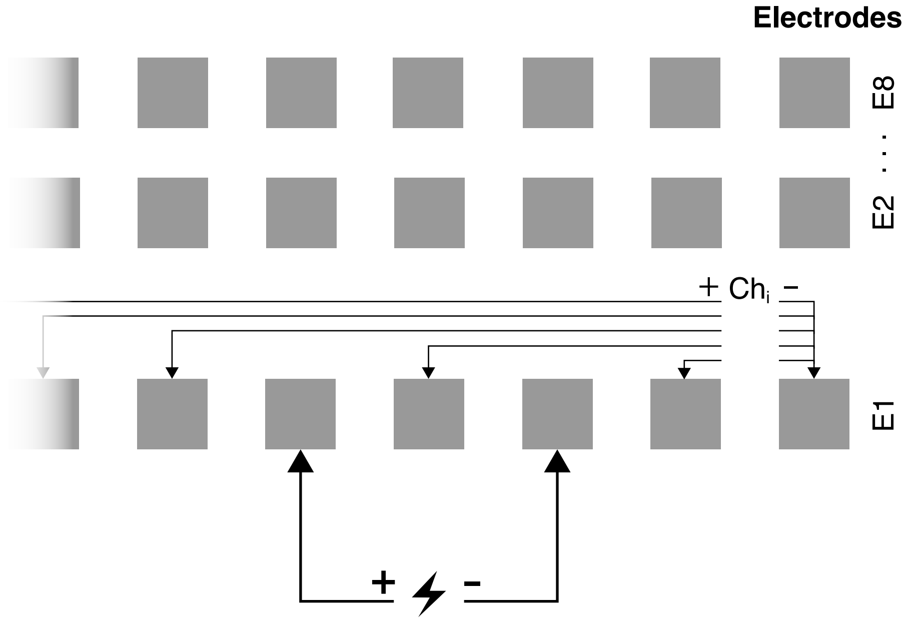


Figure 5.4: Neuro-stack bipolar macro-stimulation montage for in-vivo validation.

The parameter test space included [amplitude, frequency] combinations of [0.25, 0.50, 0.75, 1.00, 1.25] mA × [60, 80, 100, 120, 140] Hz where every combination was repeated four times for a total of 100 macro-stimulation bursts (Figure 5.5) with the following parameters (pulse width: 1.28 ms, interphase width: 150 s, rectangular pulse shape, interburst delay: 16.67 s). Stimulation delivery (Figure 5.5: top – entire session; middle – multi burst; bottom – single burst level) was observed on 40 nearby recording channels, obtained using the Sense IC (sampling rate: 6250 Hz). Overlayed pulses from one of the bursts with the same parameters (1.25 mA, 60 Hz) showed successful delivery across all channels (Figure 5.5-right – upsampled to 25 kHz and interpolated). This test showed ability of our sensing front-ends to capture stimulation artifacts without saturation or distortions. It also provided a large dataset for artifact propagation analysis based on stimulation parameters.

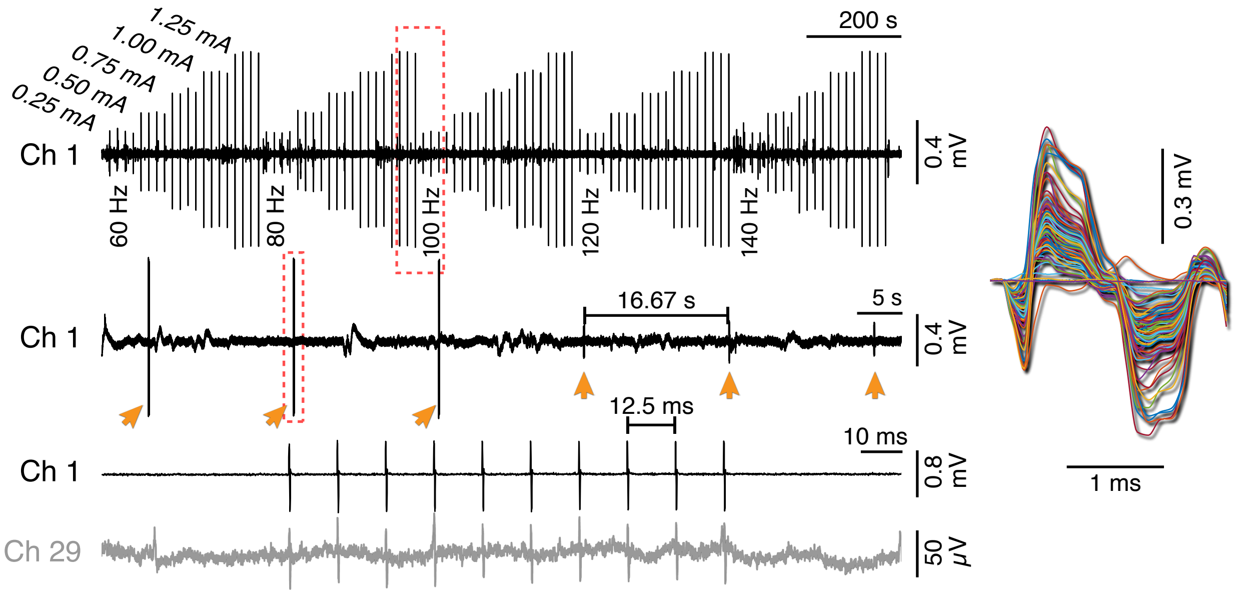


Figure 5.5: In-vivo macro-stimulation: 100 bursts, [60, 80, 100, 120, 140] Hz × [0.25, 0.50, 0.75, 1.00, 1.25] mA.

5.3 Ambulatory Behavioral Task

One of the most unique features of the Neuro-stack system is its ability to record mobile single-units, while participants are freely walking and wearing the system on-body. To our knowledge this kind of test has never been performed before.

Single-unit data was recorded in six participants during an ambulatory walking, while they carried a backpack with the Neuro-stack, TPU Dev Board, and an external Voltaic V75

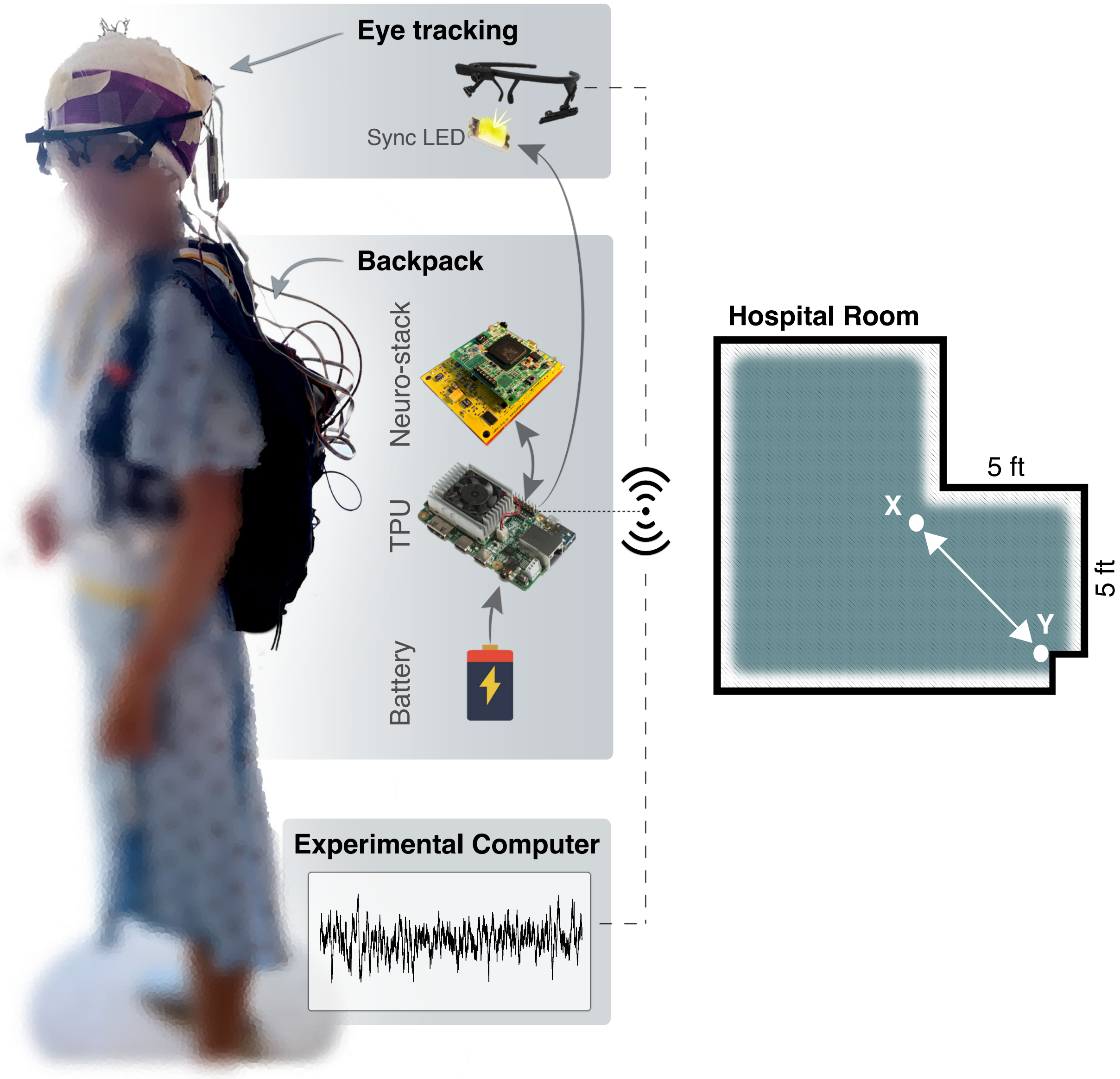


Figure 5.6: Neuro-stack in-vivo ambulatory setup. Participants were asked to walk repeatedly from one point of the room to another.

USB battery pack (Figure 5.6-left). Two of the participants were instructed to walk around their hospital room freely and visit prominent 'landmarks' such as locations near windows, doors, tables, etc. A separate group of four participants was instructed to walk repeatedly

(10 times) from one position to another position in the room using a linear path (Figure 5.6-right). The ambulatory movement was tracked using an eye-tracking headset (Pupil Labs Core device; [61]) which contained inward-facing eye cameras (sampling rate: 200 frames per s) and an outward-facing world-view camera (sampling rate: 120 frames per s). Neuro-stack was connected to two micro-wire electrode bundles (Behnke-Fried, Ad-Tech) to record from 18 micro-wire contacts (16 recorded single-unit activity and 2 served as reference contacts). Recordings with respect to local references (same bundle) were recorded at a sampling rate of 38.6 kHz.

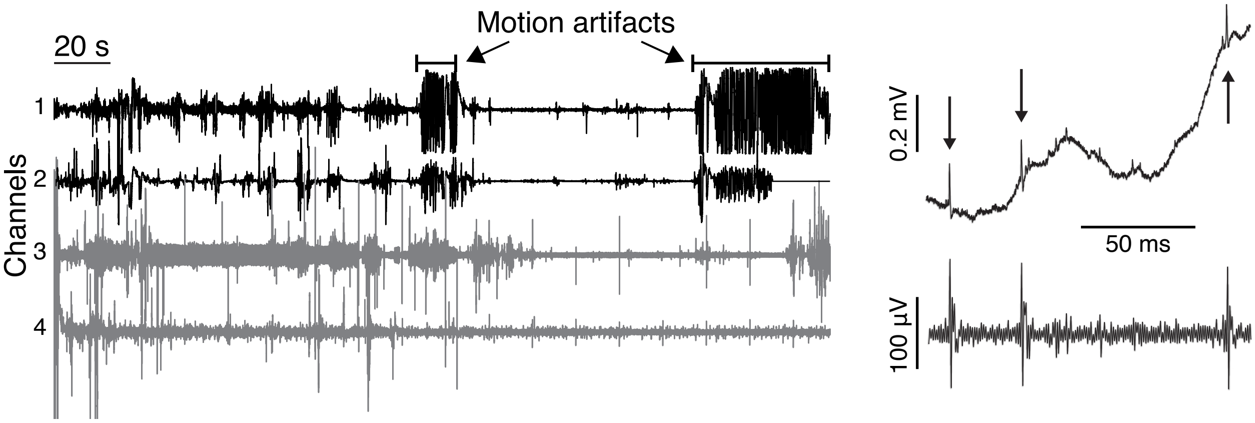


Figure 5.7: Motion artifacts and preservation of single-units.

The researcher used an experimental computer running an application (Python) to start/stop recordings and view in real-time the neural data. Both the Neuro-stack and eye-tracker were connected to the same local network from which the NTP timestamps were fetched. For a redundant method of synchronization, a miniature LED was attached to the corner of the world-view camera on the eye-tracking headset (Figure 5.6). The LED was programmed to turn on for 50 ms every 20 s during the experimental walking task, which was not visible by the participant and was also NTP-timestamped.

The first four walks were used to assess motion artifacts in recordings. Motion artifacts

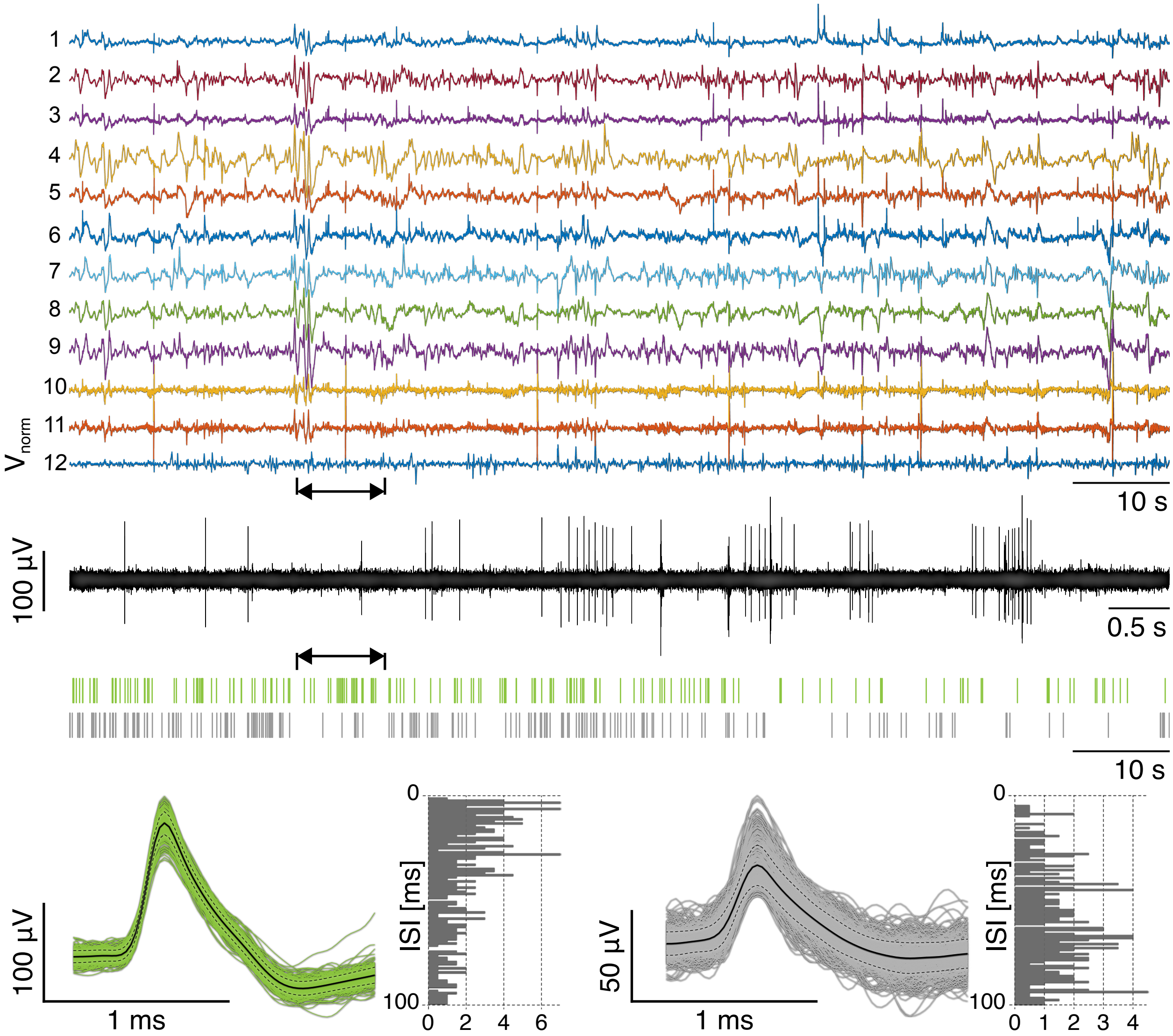


Figure 5.8: First recording of human ambulatory single-units. Raw, filtered (300 – 3000 Hz), and extracted units.

were present in the recordings, but the use of nearby electrodes (same bundle) as a reference resulted in reduced common noise artifacts using the front-end amplifiers (Figure 5.7). Raw (line noise removed) 12-channel neural activity recorded from one participant during walking from X to Y is shown in Figure 5.8. Although motion artifacts were reduced, slow voltage transients during movement were still present (Figure 5.7). Nonetheless, single-unit spikes were preserved and detected using a bandpass filter [300 – 3000 Hz] (Figure 5.8). After spike sorting [62] of the data, single-unit clusters were successfully isolated (Figure 5.8). We performed spike sorting using Wave_clus 3 [62]. Preprocessing included the use of a notch-filter to remove 60 Hz noise. Selected clusters were chosen so that more than 250 spikes were identified and that out of these, 1% or less had inter-spike-intervals (ISI) of less than 3 ms (Figure 5.8-bottom).

5.4 Stationary Behavioral Task

Neuro-stack's ability to record neural data in real-time and decode behavioral performance was tested bedside in a participant with indwelling micro-wire electrodes while they completed a verbal memory task (Figure 4.1). During the task, the participant was instructed to learn (encoding phase) a list of ten words that were presented on an iPad screen and then verbally recall as many words as possible after a distraction phase. Encoding, distraction, and recall blocks were repeated nine times during the experimental paradigm while the Neuro-stack recorded LFP activity from sixteen micro-wire channels, which was used to decode memory performance in real-time using artificial neural networks. During the verbal memory task, we used the Neuro-stack in a wireless configuration together with both the experimental computer and Stimulus Presentation device (iPad). Stimulus presentation on the iPad was implemented as a game using Xcode 11.2.1 and Swift 5.0.1 programming languages. For network communication, we used two TCP (transmission control protocol) channels (Figure 5.9; 1. Experimental computer – TPU Dev Board, 2. Experimental computer – iPad). During recall phase words were automatically converted to targets using built-in features of Swift language for iOS applications. Targets as well as event timestamps were transferred to the experimental computer as soon as decoded. NTP timestamping was used to log every event on the experimental computer, iPad, and TPU Dev Board, as well as every neural sample, which was later used for synchronization.

This was the setup for verbal memory task and in-vivo test of the transfer learning process described in the previous chapter. Out of nine trials, five were used for training phase and remaining four for making predictions. During prediction phase the data chunks were fed into the model for inference as they came without shuffling.

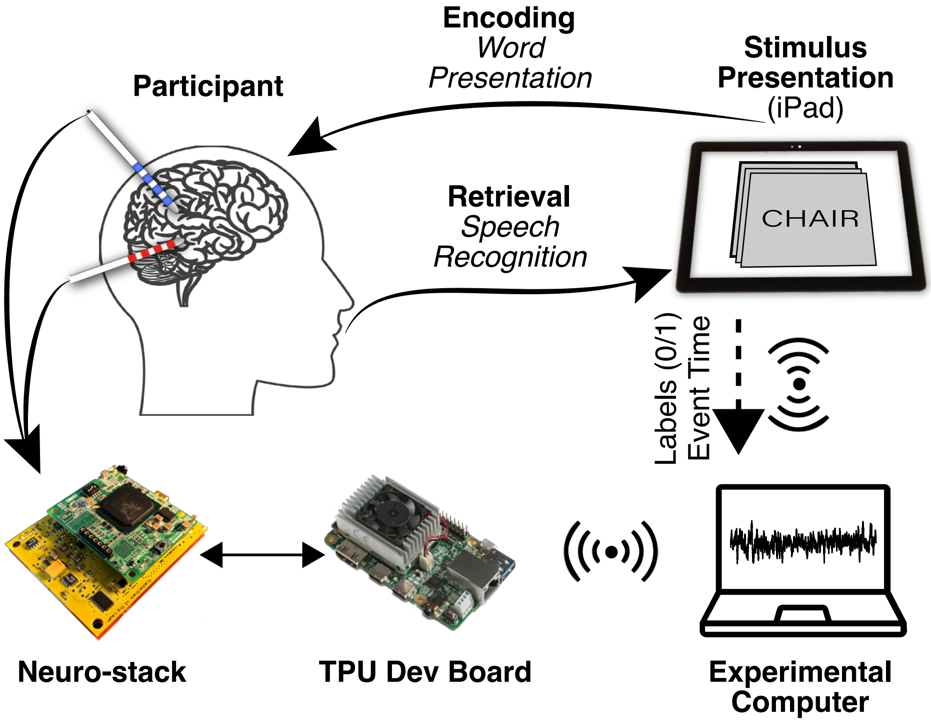


Figure 5.9: Neuro-stack in-vivo setup for performing stationary verbal memory task, recording, and online decoding.

After five trials of retraining on the experimental computer, achieved training F1-score was 90.5%. Final test F1-score after four prediction trials was 69.0% (72.5% accuracy; Figure 5.10). While results were not as good as some of the in-vitro or offline ones these were still well beyond the chance levels. Moreover, we only had opportunity to do this task in one participant, which resulted in completed verbal memory task and neural decoding using fairly complex system, which was a positive result. Experiments with more participants will really show the effectiveness of this real-time transfer learning operation as more and more Neuro-stack specific data is being incorporated into the base model.

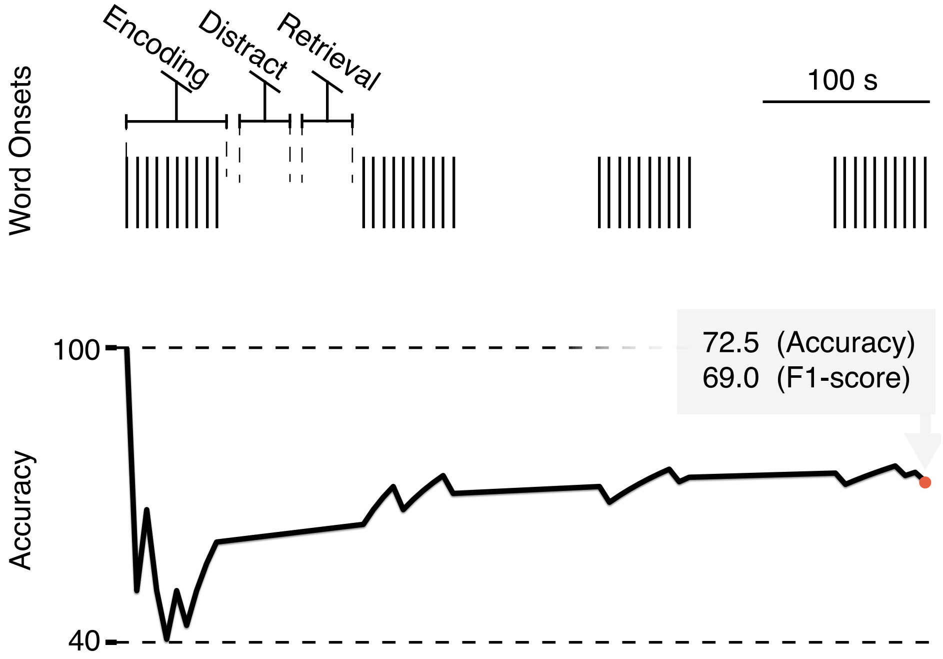


Figure 5.10: Prediction phase of the verbal memory task showing inference results synchronized with stimulus onsets.

6 Discussion

In this chapter, we are going to discuss open general topics, unanswered in previous section. The goal is to provide further clarification about the need for the Neuro-stack and its recording/stimulation advantages compared to available devices. Also, we provide a general description of neural interface research and development process in an academic setting as compared to industry, where all other existing neural devices are manufactured.

6.1 LFP/iEEG vs Single-unit

Both single-neuron and local fields have been used in human neuroscience research to map behavioral functions to a neuron or a neuron population activity. On the other hand, clinical treatments for certain disorders have mostly relied on local fields activity. In essence, the single-unit and LFP/iEEG signals are inter-dependent as such that neural oscillations in LFP/iEEG arise from collective synchronized activations within one population of neurons. Reversly, generated multi-component field can also influence firing of a single-neuron together with action potentials coming from a previous neuron on the signal pathway. It should be noted that this is a simplified description of how neural networks actually work. The primary goal of the Neuro-stack's development was a standalone, external, research device that was also envisioned as a potential pathway towards fully implantable system for single-unit/LFP/iEEG recordings and stimulation. In order to justify the need for both types of signals, in this section, we are discussing their difference and a need for both in the context of clinical applications.

A conclusion from decades of brain research has been that for any given observed behavioral function only a very small portion of neurons gets activated, that is, the brain employs sparse coding. For that reason, in order to explain complex human behavior or disorder, researchers sometimes need access to hundreds, if not thousands, of single-neurons to accurately describe the underlying mechanisms. On the other hand, a single channel of LFP or iEEG covers a population of tens to thousands of neurons. Even though the signal acquired that way is a low-passed superposition of network activity, neural oscillations or its transformations provide information about various levels of network synchronicity, which can, if significant for observed functions, be a reliable biomarker. So even though the brain uses action potentials for its function and communication, indirect measures such as brain oscillations also contain useful information. From engineering standpoint, it is much easier to record LFP/iEEG given that the analog front-ends require much lower operating frequency and that lower number of electrodes is necessary to cover observed brain region. Thus, depending on the observed function there is a tradeoff between technical feasibility and spatial coverage that one must consider before conducting research. To this date, there is no technical solution for implantable system that can record a large number (thousands) of single-unit channels and so there is also no approved treatment for any disorder based on action potentials signals. Approved treatments for some of disorders (epilepsy, Parkinson's disease) are exclusively based on iEEG biomarkers recorded from implanted macro-electrodes. The Neuro-stack cannot solve this and cannot be used for single-unit based treatments, however, it can be used as a research testing platform to provide insights into how can more naturalistic and ecologically valid single-unit research drive development of future implantable neuromodulation systems.

Although research and clinical applications based on iEEG have been successful so far it should be noted that the iEEG signal can carry information about multiple functions represented in form of neural oscillations that may overlap in time and frequency domains.

Similarly, macro-stimulation can affect more than just roughly localized centers of certain behavioral function or disorder. Further, a time-frequency uncertainty of the oscillation analysis, especially in real-time closed-loop applications, cannot be fully solved and will always limit ability for instantaneous action based on iEEG signal. Motivated by successes of single-unit motor cortex BCI interfaces, it feels natural that there will always be a need to build a closed-loop layer that can 'talk' the same language as the observed neural network. For the time being, research and clinical treatments for deep brain structures are more likely to continue its dependance on iEEG alone, but technology of the future may change that driven by the current investigational devices such as the Neuro-stack.

6.2 Bi-directional Full-Duplex Sensing and Stimulation

Concurrent sensing and stimulation have always been one of the key challenges in implantable neuromodulation technology development due to small margins in their input dynamic range compared to external systems. Given the implantable design of the Neuro-stack's ICs and a potential pathway to a fully implantable system that can record single-units, we discuss further in this section Neuro-stack's full-duplex capability.

Stimulation artifacts can be several orders larger in magnitude than underlying neural activity. This means that an artifact can overlap with the neural activity in time and frequency domain. Even more damaging is the amplifier saturation caused by a large stimulation artifact, which means complete loss of signal. Complete artifact rejection remains unsolved in general and is mostly handled on a case-by-case basis. Nonetheless, designers of neuromodulation devices have used various methods in order to allow bi-directional interface and prevent amplifier saturation, ranging from blanking the amplifier input during stimulation to allowing a certain montage, such as monopolar stimulation and symmetrical differential recording around the stimulation contact that would eliminate an artifact via amplifier's common-mode rejection, while leaving differential neural recording intact.

The Neuro-stack's sensing front-end in most cases does not require mitigation in order to record during stimulation. A large dynamic input range and digital non-linear correction (NLC) of the amplifier allows for capturing stimulation artifacts while maintaining high recording resolution for lower voltages, thus preserving neural activity during stimulation. While this does not remove artifacts it allows unrestricted full-duplex interface with no amplifier saturation nor neural signal degradation. Afterwards, artifacts may be removed separately during postprocessing stage or in an online manner, if necessary.

The Neuro-stack hardware sensing pipeline also includes digital adaptive stimulation artifact rejection integrated circuit (ASAR; [48]) that uses adjacent channels to adaptively learn the shape of the artifact, which is then subtracted from neural signal. However, ASAR was not validated as part of the Neuro-stack and is out of the scope of this work.

In conclusion, unlike existing devices, the Neuro-stack offers bi-directional full-duplex interface with plenty of margin to capture neural activity in presence of artifact regardless of the montage used.

6.3 Hardware and Software Development in Academy

All current devices described in introductory sections and compared with the Neuro-stack were developed in commercial environments (e.g., NeuroPace, Medtronic, Blackrock Microsystems, etc.). without constant interaction with clinicians nor human participants. To our knowledge, the Neuro-stack is the first neural interface completely developed, tested and used in human participants in an academic environment. Some of the development was even gradually informed by in-vivo human testing (e.g., wearability, stimulation protocols, and online processing). This was made possible by close collaboration between engineering, neuroscience, and neurosurgery departments at University of California, Los Angeles, as well as the ability of each department to perform cutting-edge research in their respective areas.

As mentioned earlier, implantable SUBNETS system with its integrated circuits was a multi-institutional effort initiated, supported, and funded by DARPA. The technology itself was developed in two fabrication cycles before meeting the set requirements through processes of verification and in-vitro and animal in-vivo validation. Next steps for evolving such technology are commercialization and acquiring of regulatory FDA (Food and Drug Administration) approval for human use. Both steps require years of work. The Neuro-stack, although consisted of the exact same integrated circuits with intention for human use, is an external, non-commercial, and research device, intended for human validation, was not subject to as strict oversight as implantable systems up to the stage presented in this work. Neuro-stack commercialization or use by other institutions would require large amounts of additional work in documentation, testing, and validation.

SUBNETS implantable Hardware and firmware were designed and tested in-vitro with an external oversight following FDA regulatory guidelines for verification and validation. The Neuro-stack assembly and software development were internally verified and validated in order to meet safety requirements set by UCLA IRB. This meant that every hardware and software functionality presented in this work was thoroughly tested and documented before obtaining the IRB approval. This included testing of recording functionalities as such that set parameters provide recordings from set channel, at set sampling frequency, under set configuration of the amplifier and not others. Likewise, stimulation tests meant that software control and trigger of the stimulation delivered current with exact preconfigured parameters and not others. Given that stimulation requires absolute safety for the purposes of this work a separate flag needs to be checked at firmware level to ensure that delivery can only happen during stimulation trials and not others. This ensured a redundant check in case an altered command is read by the firmware as a stimulation command. Further, all commands sent to the firmware contain a 8-bit CRC code to reduce probability of incorrect command delivery. Throughout the work only one password protected and encrypted experimental computer containing a signed certificate was used only by the author of this work to control the Neuro-stack, thus simplifying the necessary security infrastructure required for any commercial medical device.

Leakage currents of all channels were verified independently of software and hardware designers by a clinical engineer in idle, active recording, and active stimulation mode of operation. Further, all hardware and software documentation, including but not limited to design history, schematics, code, in-vitro tests, were reviewed and approved by the clinical engineer at Ronald Reagan Hospital at UCLA as part of the IRB review process.

6.4 Reproducibility

The Neuro-stack and SUBNETS implantable neuromodulation system were not a typical engineering research project. They were completed as part of a multi-institutional effort involving a large group of people over several years. This work presented a use of already developed custom integrated circuits to assemble a new device and the development of firmware and software to enable practical research applications with human participants. Thus, the reproduction of the whole system would first require replication of individual application-specific integrated circuits (ASIC) reported in prior work ([37], [46]–[50]). Only after that step can the full reproduction of the Neuro-stack be achieved based on this work.

7 Conclusion

7.1 Research Contributions

We present the Neuro-stack, a novel miniaturized recording and stimulation system that can interface with implanted electrodes in humans during stationary (bedside) or ambulatory behaviors. The Neuro-stack can record up to 256 channels of LFP/iEEG activity and 32 channels of single-/multi-unit activity. Macro-stimulation can also be delivered through any of the channels (up to 32 channels simultaneously) during recording, allowing for bi-directional full-duplex capability. This is a significant advantage over existing systems in that it allows for characterization of ongoing neural consequences of stimulation as well as precisely timed closed-loop stimulation.

A second major advantage of the Neuro-stack over existing systems is its smaller hand-held size that enables it to be carried on-body and be wirelessly controlled. These features allowed us to record single-neuron waveforms (spikes) during walking, which to our knowledge are the first recordings of their kind in humans. Future studies using the Neuro-stack could determine the neural mechanisms underlying human freely-moving behaviors (e.g., spatial navigation) to identify, for example, spatially selective neurons and their modulation by cognition (e.g., hippocampal place or entorhinal grid cells [41]) that have been previously discovered in freely-moving animals. Doing so would bridge decades of findings between animals and humans and potentially lead the way towards scientifically informed therapies for hippocampal-entorhinal-related dysfunctions (such as Alzheimer's disease). While we did not identify any spatially selective single-units in the current study, possibly due to the restricted spatial environment in which walking took place, further analysis from our ambulatory task and other future studies using the Neuro-stack over longer distances (e.g., hallways) may be able to identify these neurons in humans.

A third advantage of the Neuro-stack is its API that allows fast and flexible prototyping of the experiments with range of backend functions that accurately align behavioral and neural events (i.e., spikes). We demonstrated how the Neuro-stack's API integrated with a TPU can, in real-time, decode verbal memory performance in a single participant with accuracy levels that are comparable to previous reports [10]. Specifically, we used neural network models applied to hippocampal recordings to predict whether a previously learned item would be remembered, with offline results exceeding those previously reported [10], when equivalent metrics (F1-scores at the optimal thresholds) are compared. Future studies with larger sample sizes will confirm whether reported decoding accuracy can be generalized across subjects. It should be noted that we tested the decoding algorithm in one participant using the model pretrained with recordings from a different device with different noise levels (Figure 5.3), hence it is reasonable to assume that performance could go up as more Neuro-stack data are incorporated into the pretrained model. Given the increasing benefit of using machine learning approaches [54], [63] in neuroscience studies, the Neuro-stack could be useful for validating decoding models and testing novel closed-loop stimulation therapies (e.g., to improve memory in patients with severe memory impairments).

7.2 Future Work

There are numerous options for future work involving Neuro-stack system both on engineering and neuroscience side.

In short-term, large datasets collected during stimulation, ambulatory, and stationary tasks need further analysis. Stimulation data with varying output parameters could be used to model artifact propagation, depending on the parameters, which could be useful in a challenging problem of stimulation artifact rejection. The data from ambulatory tasks need further analysis to find correlates of behavior if present in the recordings as well as to further suppress motion artifacts. Finally, in order to continue upgrading the base neural network model, the data from a single Neuro-stack participant could provide more insights into how necessary adjustments before another attempt is carried out.

While the Neuro-stack offers several advantages over currently available systems, there are limitations that warrant discussion. First, this Neuro-stack prototype can only support a maximum of 32 wide-band single-unit recording channels. While it can also simultaneously record up to 256 LFP recording channels (using four analog layers), other existing bedside systems can allocate more than 256 channels solely for unit recordings. The use of multiple Neuro-stack devices, however, would address this issue and increase single-unit channel count substantially. Second, although the Neuro-stack is small enough to be carried on-body and thus, allow for full mobility, its connection with implanted electrodes is still wired, similar to other bedside systems. Thus, significant movements can result in motion artifacts. However, single-unit spike waveforms can still be detected and isolated during walking behavior as we show using techniques such as differential recordings between nearby contacts, as well as proper wire isolation and fixation. Lastly, the Neuro-stack currently can only be used in research studies with patients who have externalized electrodes implanted during clinical (e.g., epilepsy) monitoring. Since these patients need to be continuously tethered to bedside intracranial recording systems to assess for symptomatic episodes (e.g., seizures), this limits the amount of time a patient can be freely-moving. However, future studies can complete ambulatory studies after clinical data has been captured as was done in the current study, on the last day of the patient's hospital stay prior to electrode de-plantation surgery, or during circumstances where continuous monitoring may not be necessary (e.g., depression or chronic pain studies [64]). Furthermore, proper precautions and safety measures should be implemented, such as waiting to complete studies until epilepsy patients are back on antiepileptic medications to minimize risks associated with seizures during ambulatory tasks.

Future studies can determine which stimulation parameters are most beneficial for restoring cognitive or behavioral functions given the Neuro-stack's highly flexible programmability compared to existing human-approved stimulators. For example, continuous adjustments of custom pulse shapes, timing of complex burst patterns, and/or timing of stimulation relative to ongoing neural activity events could allow for the development of more effective stimulation therapies. Given the wireless and wearable nature of the Neuro-stack, studies could also determine whether closed-loop stimulation protocols effectively translate to more naturalistic behaviors during everyday experiences that occur during mobility.

Although Neuro-stack is much smaller than other external systems, an even smaller version could be tested in future in-vivo studies since its IC chips are all implantable by design [37], [46]–[50] and require a combined area of just 113 mm2 (4 analog layers). An implantable version of the Neuro-stack [37] but with its added single-neuron and closed-loop stimulation capabilities thus presents an exciting avenue towards a completely wireless intracranial single-unit and LFP recording system that would not be susceptible to motion artifacts. Very recently, there has been an emerging trend from neural interface manufacturers to develop digital headstages, which are fixed right at the head and are used to digitize signal, thus shortening the analog wire paths susceptible to noise. Given that Neuro-stack's ICs are currently the smallest ones, they could be perfect for similar short-term solution. Both implantable and digital headstage type of system would present a significant advancement over current FDA-approved chronic neurostimulation devices in that it would allow for single-neuron and multi-channel (current state-of-the-art is 4 channels; Neuropace RNS) recordings, bi-directional recording and stimulation (full-duplex) capability, and the ability to use advanced strategies for decoding (e.g., neural network models for inference) behavior or disease-related states. Altogether, these novel capabilities would provide cognitive and clinical neuroscience studies with a promising future pathway towards determining the deep-brain mechanisms of naturalistic behavior in humans and developing more effective closed-loop intracranial neuromodulation strategies for individuals with debilitating brain disorders.

References

[1] U. Topalovic, S. Barclay, C. Ling, *et al.*, "A wearable platform for closed-loop stimulation and recording of single-neuron and local field potential activity in freely-moving humans," *bioRxiv*, 2022.

[2] R. J. Gardner, E. Hermansen, M. Pachitariu, *et al.*, "Toroidal topology of population activity in grid cells," *Nature*, vol. 602, 2022.

[3] S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland, "Single-trial neural dynamics are dominated by richly varied movements," *Nature Neuroscience*, vol. 22, 2019.

[4] A. Schulze-Bonhage, "Brain stimulation as a neuromodulatory epilepsy therapy," *Seizure*, vol. 44, pp. 169–175, 2017.

[5] A. L. Benabid, P. Pollak, A. Louveau, S. Henry, and J. de Rougemont, "Combined (thalamotomy and stimulation) stereotactic surgery of the VIM thalamic nucleus for bilateral Parkinson disease," *Applied neurophysiology*, vol. 50, pp. 344–6, 1987.

[6] M. Vidailhet, L. Vercueil, J.-L. Houeto, *et al.*, "Bilateral Deep-Brain Stimulation of the Globus Pallidus in Primary Generalized Dystonia," *New England Journal of Medicine*, vol. 352, pp. 459–467, 2005.

[7] A. M. Lozano, L. Fosdick, M. M. Chakravarty, *et al.*, "A Phase II Study of Fornix Deep Brain Stimulation in Mild Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 54, pp. 777–787, 2016.

[8] R. F. Thompson and J. J. Kim, "Memory systems in the brain and localization of a memory," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, 1996.

[9] F. Bartolomei, P. Chauvel, and F. Wendling, "Epileptogenicity of brain structures in human temporal lobe epilepsy: A quantified study from intracerebral EEG," *Brain*, vol. 131, 2008.

[10] Y. Ezzyat, J. E. Kragel, J. F. Burke, *et al.*, "Direct Brain Stimulation Modulates Encoding States and Memory Performance in Humans," *Current Biology*, vol. 27, pp. 1251–1258, 2017.

[11] N. Suthana, Z. Haneef, J. Stern, *et al.*, "Memory enhancement and deep-brain stimulation of the entorhinal area," *New England Journal of Medicine*, 2012.

[12] Y. Ezzyat, P. A. Wanda, D. F. Levy, *et al.*, "Closed-loop stimulation of temporal cortex rescues functional networks and improves memory," *Nature Communications*, vol. 9, p. 365, 2018.

[13] R. Zelmann, A. C. Paulk, I. Basu, *et al.*, "CLoSES: A platform for closed-loop intracranial stimulation in humans," *NeuroImage*, vol. 223, 2020.

[14] I. Fried, U. Rutishauser, M. Cerf, and G. Kreiman, *Single Neuron Studies of the Human Brain.* MIT Press, 2015.

[15] I. Fried, C. L. Wilson, N. T. Maidment, *et al.*, "Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients: Technical note," *Journal of Neurosurgery*, vol. 91, 1999.

[16] K. Lehongre, V. Lambrecq, S. Whitmarsh, *et al.*, "Long-term deep intracerebral microelectrode recordings in patients with drug-resistant epilepsy: proposed guidelines based on 10-year experience," *NeuroImage*, p. 119 116, 2022.

[17] M. J. Nelson, P. Pouget, E. A. Nilsen, C. D. Patten, and J. D. Schall, "Review of signal distortion through metal microelectrode recording circuits and filters," *Journal of Neuroscience Methods*, vol. 169, 2008.

[18] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes," *Nature Reviews Neuroscience*, vol. 13, 2012.

[19] *Blackrock Neurotech (research)*, `https://blackrockneurotech.com/research`, Accessed: December 3, 2021.

[20] *Neuralynx*, `https://neuralynx.com`, Accessed: December 3, 2021.

[21] *Nihon Kohden EEG system: EEG-1200*, `https://us.nihonkohden.com/products/eeg-1200`, Accessed: December 3, 2021.

[22] *Ripple Neuro, Custom Neuroscience Research Tools*, `https://rippleneuro.com`, Accessed: December 3, 2021.

[23] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, 2005.

[24] J. Zheng, A. G. P. Schjetnan, M. Yebra, *et al.*, "Neurons detect cognitive boundaries to structure episodic memories in humans," *Nature Neuroscience*, vol. 25, no. 3, pp. 358–368, 2022.

[25] C. S. Inman, J. R. Manns, K. R. Bijanki, *et al.*, "Direct electrical stimulation of the amygdala enhances declarative memory in humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, 2018.

[26] A. S. Titiz, M. R. Hill, E. A. Mankin, *et al.*, "Theta-burst microstimulation in the human entorhinal area improves memory specificity," *eLife*, vol. 6, 2017.

[27] E. A. Mankin and I. Fried, "Modulation of Human Memory by Deep Brain Stimulation of the Entorhinal-Hippocampal Circuitry," *Neuron*, vol. 106, 2020.

[28] E. A. Mankin, Z. M. Aghajan, P. Schuette, *et al.*, “Stimulation of the right entorhinal white matter enhances visual memory encoding in humans,” *Brain Stimulation*, vol. 14, 2021.

[29] M. T. Kucewicz, B. M. Berry, L. R. Miller, *et al.*, “Evidence for verbal memory enhancement with electrical brain stimulation in the lateral temporal cortex,” *Brain*, vol. 141, pp. 971–978, 2018.

[30] C.-H. Kuo, G. A. White-Dzuro, and A. L. Ko, “Approaches to closed-loop deep brain stimulation for movement disorders,” *Neurosurgical Focus*, vol. 45, E2, 2018.

[31] N. C. Swann, C. de Hemptinne, M. C. Thompson, *et al.*, “Adaptive deep brain stimulation for Parkinson’s disease using motor cortex sensing,” *Journal of Neural Engineering*, vol. 15, p. 046 006, 2018.

[32] F. T. Sun and M. J. Morrell, “The RNS System: Responsive cortical stimulation for the treatment of refractory partial epilepsy,” *Expert Review of Medical Devices*, vol. 11, 2014.

[33] D. D. Cummins, R. B. Kochanski, R. Gilron, *et al.*, “Chronic Sensing of Subthalamic Local Field Potentials: Comparison of First and Second Generation Implantable Bidirectional Systems Within a Single Subject,” *Frontiers in Neuroscience*, vol. 15, 2021.

[34] S. Stanslaski, J. Herron, T. Chouinard, *et al.*, “A Chronically Implantable Neural Coprocessor for Investigating the Treatment of Neurological Disorders,” *IEEE Transactions on Biomedical Circuits and Systems*, 2018.

[35] V. Kremen, B. H. Brinkmann, I. Kim, *et al.*, “Integrating brain implants with local and distributed computing devices: A next generation epilepsy management system,” *IEEE Journal of Translational Engineering in Health and Medicine*, 2018.

[36] R. Gilron, S. Little, R. Perrone, *et al.*, “Long-term wireless streaming of neural recordings for circuit discovery and adaptive stimulation in individuals with Parkinson’s disease,” *Nature Biotechnology*, vol. 39, 2021.

[37] D. Rozgic, V. Hokhikyan, W. Jiang, *et al.*, “A 0.338 cm3, Artifact-Free, 64-Contact Neuromodulation Platform for Simultaneous Stimulation and Sensing,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, 2019.

[38] U. Topalovic, Z. M. Aghajan, D. Villaroman, *et al.*, “Wireless Programmable Recording and Stimulation of Deep Brain Activity in Freely Moving Humans,” *Neuron*, vol. 108, 2020.

[39] M. Stangl, U. Topalovic, C. S. Inman, *et al.*, “Boundary-anchored neural mechanisms of location-encoding for self and others,” *Nature*, vol. 589, 2021.

[40] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, “Speech synthesis from neural decoding of spoken sentences,” *Nature*, vol. 568, 2019.

[41] M. B. Moser, D. C. Rowland, and E. I. Moser, “Place cells, grid cells, and memory,” *Cold Spring Harbor Perspectives in Biology*, vol. 7, 2015.

[42] Z. M. Aghajan, D. Villaroman, S. Hiller, *et al.*, “Modulation of human intracranial theta oscillations during freely moving spatial navigation and memory,” *bioRxiv*, 2019.

[43] J. Jacobs, C. T. Weidemann, J. F. Miller, *et al.*, “Direct recordings of grid-like neuronal activity in human spatial navigation,” *Nature Neuroscience*, vol. 16, 2013.

[44] H. Wu, K. J. Miller, Z. Blumenfeld, *et al.*, “Closing the loop on impulsivity via nucleus accumbens delta-band activity in mice and man,” *Proceedings of the National Academy of Sciences of the United States of America*, 2018.

[45] P. K. Douglas and A. Anderson, *Feature Fallacy: Complications with Interpreting Linear Decoding Weights in fMRI*, 2019.

[46] W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, and D. Marković, “A ±50-mV Linear-Input-Range VCO-Based Neural-Recording Front-End With Digital Nonlinearity Correction,” *IEEE Journal of Solid-State Circuits*, vol. 52, 2017.

[47] D. Rozgic, V. Hokhikyan, W. Jiang, *et al.*, “A true full-duplex 32-channel 0.135cm3 neural interface,” IEEE, 2017, pp. 1–4.

[48] S. Basir-Kazeruni, S. Vlaski, H. Salami, A. H. Sayed, and D. Markovic, “A blind Adaptive Stimulation Artifact Rejection (ASAR) engine for closed-loop implantable neuromodulation systems,” IEEE Computer Society, 2017, pp. 186–189.

[49] H. Chandrakumar and D. Markovic, “An 80-mVpp linear-input range, 1.6-G input impedance, low-power chopper amplifier for closed-loop neural recording that is tolerant to 650-mVpp common-mode interference,” *IEEE Journal of Solid-State Circuits*, vol. 52, 2017.

[50] A. Alzuhair and D. Marković, “A 216 nW/channel DSP engine for triggering theta phase-locked brain stimulation,” vol. 2018-Janua, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1–4.

[51] A. Alzuhair, “Theta Phase-Specific Closed-Loop Stimulation in Implantable Neuromodulation Devices,” Ph.D. dissertation, University of California, Los Angeles. ProQuest Dissertations Publishing, 2019.

[52] S. S. Sandha, J. Noor, F. M. Anwar, and M. Srivastava, “Time awareness in deep learning-based multimodal fusion across smartphone platforms,” 2020.

[53] B. J. Underwood, M. P. Toglia, and W. F. Battig, “Handbook of Semantic Word Norms,” *The American Journal of Psychology*, vol. 92, 1979.

[54] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, “Speech synthesis from neural decoding of spoken sentences,” *Nature*, vol. 568, 2019.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, 2014.

[56] E. A. Solomon, J. M. Stein, S. Das, *et al.*, "Dynamic Theta Networks in the Human Medial Temporal Lobe Support Episodic Memory," *Current Biology*, vol. 29, 2019.

[57] N. A. Suthana, N. N. Parikshak, A. D. Ekstrom, *et al.*, "Specific responses of human hippocampal neurons are associated with better memory," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 10 503–10 508, 2015.

[58] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, 2002.

[59] P. A. Yushkevich, J. B. Pluta, H. Wang, *et al.*, "Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment," *Human Brain Mapping*, vol. 36, 2015.

[60] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, 2001.

[61] M. Kassner, W. Patera, and A. Bulling, "Pupil," ACM Press, 2014, pp. 1151–1160.

[62] F. J. Chaure, H. G. Rey, and R. Q. Quiroga, "A novel and fully automatic spike-sorting implementation with variable number of features," *Journal of Neurophysiology*, vol. 120, 2018.

[63] J. A. Livezey and J. I. Glaser, *Deep learning approaches for neural decoding across architectures and recording modalities*, 2021.

[64] K. W. Scangos, G. S. Makhoul, L. P. Sugrue, E. F. Chang, and A. D. Krystal, "State-dependent responses to intracranial brain stimulation in a patient with depression," *Nature Medicine*, vol. 27, 2021.