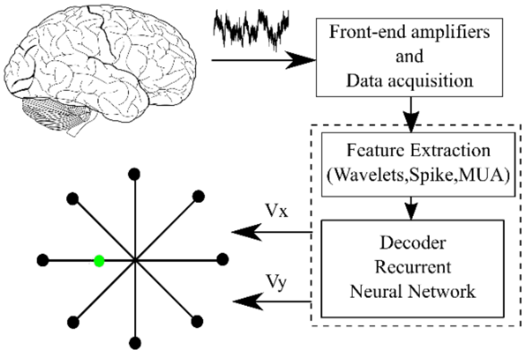
Front-end amplifiers and Data acquisition
Feature Extraction (Wavelets,Spike,MUA)
Decoder Recurrent Neural Network
Vx
Vy

Figure 1: Proposed Brain Machine Interface Architecture.

Introduction: The National Spinal Cord Injury Statistical Center reported a total of 19,105 new cases of spinal cord injuries with some amount of paralysis in 2019. Brain machine interfaces (BMIs) are a burgeoning technological solution to restore quality of life to these people through connecting brain signals to prosthetics. Current brain machine technology collects and transmits neural signals from implantable electrode arrays to be decoded using algorithms which are implemented on cumbersome and power inefficient computers. These systems also require daily calibrations by scientists and clinicians to maintain their usability. Herein lies an approach to address these problems through implementing specially designed algorithms that are memory and computationally efficient onto a low power application specific integrated circuit (ASIC) to decode patient intent using a fully implantable device. **The goal is to engineer hardware to implement intelligent decoding algorithms that will increase the reliability of neural decoding systems and decrease the physical size and power requirements by orders of magnitude through removing the need for transmission of unprocessed neural data**. I am working with professor Azita Emami at the Caltech Mixed-mode Integrated Circuits and Systems (MICS) laboratory to begin my work as an ASIC and algorithmic designer implementing these systems.

Previous Work: The promise of this project to produce reliable power efficient BMIs relies on the development of power efficient algorithms. The MICS lab has developed efficient algorithms that utilize deep multistate dynamic recurrent neural networks [1], as well as done an assay on decoding algorithms [2]. Furthermore, if energy efficient algorithms do not provide the power performance required, in memory processing architectures could be utilized to reduce power lost from transporting weights between memory banks and processing units [3].

Intellectual Merit: Current BMI technology largely utilizes hardware implementations which transmit digitized neural signals back to a compute station. There are few BMI ASICs that decode patient intent directly without off-the-shelf hardware implementations. Custom machine learning ASICs provide an opportunity to optimize for power and area efficient systems. BMI systems have unique signal processing constraints due to relevant information being mixed across time, frequency, and space in highly dimensional, redundant datasets. These constraints often require complex computational algorithms with high internal state complexity, that generally decreases power efficiency [2]. **This poses a particularly difficult engineering dilemma which requires a uniquely multidisciplinary approach to leverage knowledge of neuroscience with engineering expertise in ASIC design.** This dilemma is to fit the required computationally complex task of decoding patient intent from neural signals into a power and area efficient package.

Hypothesis: BMI ASICs implementing computationally, and memory efficient algorithms will be able to efficaciously ascertain patient intent while maintaining robust performance whilst still meeting power and area constraints requisite of fully implantable systems.

Research Plan: Several crucial prerequisite steps for this proposal are already underway in Azita Emami's lab including the evaluation of neural features measured from patient data for stability over time. Aforementioned specialized algorithms have been developed and evaluated for performance. While these prerequisite steps are crucial to the outcome of this project, the scope of this proposal is constrained to the implementation of these algorithms in hardware, optimizing for low power. Therefore, this proposal will only discuss the development, implementation, and testing of the algorithms into CMOS fabric. *Stage I- Algorithm CAD Design and Testing:* Digital and analog hardware will be described and laid out into Virtuoso Computer Assisted Design (CAD) program. The circuits will be fabricated with general power optimization techniques in mind such as clock and power gating portions of the circuit when they are not in use. Other techniques include using intentional specificity of transistor thresholds throughout the design so that leakage current is minimized, as well as a minimization of supply voltage levels. Furthermore, specific tradeoffs will be made between the bit precision of the algorithms and the resources required to run at those precisions. Significant energy is also lost by moving the algorithm's weights from memory to the processing hardware, so a layout will be designed such that the algorithm weights are physically closer to where computations are done. While designing the circuits, each component will be characterized at a unit level so that the performance of the entire circuit can be ensured. *Stage II-CAD Simulation and FPGA Testing:* After designing the circuit in Virtuoso, the entirety of the circuit will be tested with patient neural data collected over several weeks. The circuit simulation will give significant insight into the performance of the decoder system once fabricated. The simulation will also allow for design bugs to be caught and corrected before resources are spent fabrication of the design into silicon. The digital components of the circuit will also be implemented on a commercially available Field Programable Gate Array (FPGA) which will not only give physical proof that the logic and algorithms designed will work, but will give an upper bound on the power and area usage for the ASIC implementation. *Stage III-Hardware Fabrication:* Several test chips will be fabricated to investigate the performance of the decoding algorithms. This allows for verification and debugging of the major components of the algorithm, with the final system chip produced after all components are aptly constructed and verified. The circuits will be implemented into standard cell libraries for 65 nm CMOS technology using Design Vision. *Stage IIII-Hardware Testing:* The fabricated chip will be designed for testability using techniques such as built in scan chains to give access to the internal states of the circuit. Custom test systems will be built to feed neural signals to the device under test and validate the performance of the hardware. *Stage IV-Going Forward:* Once designed, fabricated, and verified, the integrated circuits could be tested *in vivo* using the same experimental therapy program from which the neural data was harvested.

Pitfalls and Alternatives: Due to variation in fabrication processes, the fabricated chips may exhibit characteristics and behavior that were not accounted for during simulation. If experienced, the test hardware designed into the chip will be utilized to identify and debug the flaw.

Timeline and Collaboration: The duration of the proposed project is three years and will be conducted under the supervision of Azita Emami. Azita Emami's MICS lab works in direct collaboration with Richard Anderson's neuroscience lab. This is a collaboration between two leading scientists in their respective fields. The MICS and Anderson lab collaboration has significant skill and expertise to confidently produce the anticipated results of this proposal. Emami's lab has adequate facilities and resources to fund the significant capital required to design and develop custom ASIC hardware. This project also has direct access to neural recordings of patients with implanted UTAH neural arrays which provides essential data for training the decoder algorithms.

Anticipated Results: Using specially designed decoding algorithms, significant reductions in power and area will be observed in the resultant neuro decoding system. It is important to note this project aims to maintain decoding accuracy and stability, despite using orders of magnitude less power and area.

Broader Impacts: Fully implantable neural intent decoders will not only greatly improve the quality of life of patients with paralysis, but also provide the basis for fully implantable ASIC chips designed for direct study of neural activity without the need to be linked to heavy, memory and power inefficient recording systems. The miniaturization of hardware and computational effort can further be generalized to many IOT or wearable systems which requires robust signal processing algorithms with limited power and area requirements. This will enable a variety of wearable devices to decode bio signals without the need to upload the signal data to an off-chip server, greatly improving security, power, and speed performance.

References: **(1)** B. Haghi, S. Kellis, M. Ashok, S. Shah, L. Bashford, D. Kramer, B. Lee, C. Liu, R. Andersen, A. Emami, **" Deep multi-state dynamic recurrent neural networks for robust brain-machine interfaces"**, *Program No. 406.04. 2019 Neuroscience Meeting Planner. Chicago, IL: Society for Neuroscience, 2019. Online.* **(2)** Mahsa Shoaran, Benyamin A. Haghi, Milad Taghavi, Masoud Farivar, Azita Emami, **"Energy-Efficient Classification for Resource-Constrained Biomedical Applications"** IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), 2018. **(3)** T.-J. Yang, V. Sze, "**Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators**," *IEEE International Electron Devices Meeting (IEDM), Invited Paper,* December 2019.