Ensembles of Implantable Microdevices as a Multi-node Network for Neural Sensing and Stimulation

Jihun Lee
B.S., Seoul National University, 2012
M.S., Seoul National University, 2016

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Biomedical Engineering at Brown University

Providence, Rhode Island
May 2021

This dissertation by Jihun Lee is accepted in its present form by the Division of Biology and Medicine and the School of Engineering as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Dr. Arto V. Nurmikko, Advisor

Recommended to the Graduate Council

Date_____ _____

Dr. David A. Borton, Reader

Date_____ _____

Dr. Wilson Truccolo, Reader

Date_____ _____

Dr. Lawrence Larson, Reader

Approved by the Graduate Council

Date_____ _____

Dr. Andrew G. Campbell,
Dean of the Graduate School

Abstract of "Ensembles of Implantable Microdevices as a Multi-node Network for Neural Sensing and Stimulation" by Jihun Lee, Ph. D. Brown University, May 2021

This study focuses on the development of implantable biosensor microdevices for future brain-machine interfaces. Ensembles of spatially distributed sub-mm size devices, "neurograins, were designed to operate as a wireless sensor network. The core of each device is an ultralow-power silicon integrated circuit fabricated at the TSMC 65 nm MS/RF CMOS process. To enable wireless operation involving potentially large populations of miniaturized implants, an efficient microantenna transceiver scheme was developed, allowing for simultaneous powering and bidirectional data communication at near 1 GHz to/from an external RF hub. Each neurograin (650 μm × 650 μm in the area), one node houses an on-chip coil, an RF energy harvesting circuit, custom circuits for either electrophysiological recording or electrical microstimulation, device identifier, plus specialized circuits for telemetry. Implementing a binary phase-shift keying modulation (BPSK) scheme on-chip was demonstrated to ensure a 10 Mbps data uplink using RF backscattering from each neurograin. An amplitude shift keying and pulse width modulation demodulation (ASK-PWM) scheme was advanced to achieve sufficient network robustness under asynchronous clock conditions across the neurograin population at 1 Mbps downlink rate. Based on simulations and experiments, we have demonstrated how a bidirectional uplink/ downlink can communicate with the external telecom hub for up to 770 neurograins under a customized time division multiple access (TDMA) protocol operating in a call-and-response manner. Extensive characterization of the neurograin ecosystem including wireless networking, wireless powering, and electrophysiological recording or stimulation was performed both on benchtop and in-vivo. Finally, we have demonstrated a multinodal network by implanting populations of neurograins into the rat model for wireless recording and microstimulation of cortical dynamics, with the animal head size limiting the scale of the implant to 48 neurograins. This thesis also describes neural population recording and decoding from the auditory cortex of non-human primates, which employed a platform, called "Dockex" previously developed in our group, for scalable machine learning experiments. The success in these scalable data acquisition and decoding approaches is encouraging further development of the concept of large-scale cortical interfaces in primates.

Curriculum Vitae

Address: 184 Hope St, Box D, Brown University, Providence, Rhode Island, 02904, USA
Phone #: +1-346-212-9617
Email address: jihun_lee@brown.edu, jihun2412@gmail.com

Education

Brown University School of Engineering, USA Ph.D. candidate in Biomedical Engineering	Sep 2016-Apr 2021 (expected) Cumulative GPA: 4/4
Seoul National University, Graduate School of Convergence Science and Technology, Seoul, Korea Master of Science in Applied Bioengineering	Mar 2014-Aug 2016 Cumulative GPA: 3.86/4.3
Seoul National University, Seoul, Korea, Bachelor of Science in Applied Biology	Mar 2008-Feb 2012 Cumulative GPA: 3.73/4.3

Publications

Paper

1. **Lee, J.**, Leung, V., Lee, A. H., Huang, J., Asbeck, P., Mercier, P. P., ... & Nurmikko, A. (2021). Wireless Ensembles of Sub-mm Microimplants Communicating as a Network near 1 GHz in a Neural Application. *bioRxiv*. (revised version pending in Nature Electronics) (citations 4)

2. Lee, A. H., **Lee, J.**, Laiwalla, F., Leung, V., Huang, J., Nurmikko, A., & Song, Y. K. (2020). A scalable and low stress post-CMOS processing technique for implantable microsensors. *Micromachines*, 11(10), 925. (citations 3)

2. Heelan, C.*, **Lee, J.***, O'Shea, R., Lynch, L., Brandman, D. M., Truccolo, W., & Nurmikko, A. V. (2019). Decoding speech from spike-based neural population

recordings in secondary auditory cortex of non-human primates. *Nature Communications Biology*, 2(1), 1-12. (citations 7)

3. Jeong, J., Laiwalla, F., **Lee, J**., Ritasalo, R., Pudas, M., Larson, L., ... & Nurmikko, A. (2019). Conformal Hermetic Sealing of Wireless Microelectronic Implantable Chiplets by Multilayered Atomic Layer Deposition (ALD). *Advanced Functional Materials*, *29*(5), 1806440. (citations 30)

4. Huang, J., Laiwalla, F., **Lee, J.**, Cui, L., Leung, V., Nurmikko, A., & Mercier, P. P. (2018). A 0.01-mm 2 Mostly Digital Capacitor-Less AFE for Distributed Autonomous Neural Sensor Nodes. *IEEE Solid-State Circuits Letters*, *1*(7), 162-165. (citations 8)

5. **Lee, J.**, Jang, J., & Song, Y. K. (2016). A review on wireless powering schemes for implantable microsystems in neural engineering applications. *Biomedical Engineering Letters*, *6*(4), 205-215. (citations 15)

6. Lee A.H.*, **Lee J.***, Jang, J., Song, Y. K. & Nurmikko, A. V. Synchronous Network of Cortical Microstimulators Powered by Near-infrared light (under preparation).

7. Heelan C.*, **Lee J.***, Nurmikko, A. V. & Truccolo, W. Single and multiunit population activity allow comparable decoding of speech sounds from nonhuman primate secondary auditory cortex (under preparation).

Conference paper

1. Lee A.H.*, **Lee, J.***, Choquette, K., Song, Y. K. & Nurmikko, A. V. (2021, May). A Distributed Ensemble of Wireless Intracortical Microdevices for Charge-balanced Photovoltaic Current Stimulation. *In The 10th International IEEE EMBS Conference on Neural Engineering (NER)*. IEEE. (accepted)

2. Laiwalla, F.*, **Lee, J.***, Lee, A. H., Mok, E., Leung, V., Shellhammer, S., ... & Nurmikko, A. (2019, July). A Distributed Wireless Network of Implantable Sub-mm Cortical Microstimulators for Brain-Computer Interfaces. *In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6876-6879). IEEE. (citations 5)

3. **Lee, J.**, Mok, E., Huang, J., Cui, L., Lee, AH., Leung, V., Mercier, P., Shellhammer S., Larson, L., Asbeck, P., Rao, R., Song, YK., Nurmikko A., and Laiwalla, F. (2019 March) An Implantable Wireless Network of Distributed Microscale Sensors for Neural

Applications. *In The 9th International IEEE EMBS Conference on Neural Engineering (NER)*. IEEE. (citations 15)

4. **Lee, J.**, & Nurmikko, A. V. (2018, October). Multi-coil High Efficiency Wireless Charger System for Hermetically Sealed Biomedical Implants. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1-4). IEEE. (citations 1)

5. **Lee, J.**, Laiwalla, F., Jeong, J., Kilfoyle, C., Larson, L., Nurmikko, A., ... & Leung, V. W. (2018, October). Wireless Power and Data Link for Ensembles of Sub-mm scale Implantable Sensors near 1GHz. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (pp. 1-4). IEEE. (citations 16)

6. Leung, V. W., **Lee, J.**, Li, S., Yu, S., Kilfovle, C., Larson, L., ... & Laiwalla, F. (2018, September). A CMOS Distributed Sensor System for High-Density Wireless Neural Implants for Brain-Machine Interfaces. In ESSCIRC 2018-IEEE 44th European Solid-State Circuits Conference (ESSCIRC) (pp. 230-233). IEEE. (citations 15)

7. Song, Y. K., **Lee, J.**, Jang, J., Lee, C. E., & Lee, A. H. (2018, January). A neural recording microimplants with wireless data and energy transfer link. *In 2018 6th International Conference on Brain-Computer Interface (BCI)* (pp. 1-5). IEEE. (citations 1)

8. Seo, J., Choi, G. J., Park, S., **Lee, J.**, Baek, C., Jang, J., ... & Kim, S. J. (2017, July). Wireless navigation of pigeons using polymer-based fully implantable stimulator: a pilot study using depth electrodes. *In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 917-920). IEEE. (citations 8)

Honors and Awards

- Awarded for excellent master's thesis by Seoul National University (2016)
- Awarded for the best research posters in neuroengineering during Brown University biomedical engineering annual retreat (2017)
- ETH-Wyss-Brown Workshop on Neurophysiology for Neural and Biomedical Engineering (Zermatt, Switzerland, 2017) Travel Awards from Brown Institute for Brain Science

Experience

Research

Ph.D. Thesis, Brown University, Providence, Rhode Island — Sep 2016- Apr 2021 (expected)
Neuroengineering and Nanophotonics Laboratory, Advised by Prof. Arto V. Nurmikko

- 4-coil resonant wireless charging system design for a subcutaneous brain implant
- Development of wireless powering and network communication for multinodal sub-mm wireless implants and in-vivo demonstration through recording/ stimulation experiment in the rat model
- NHP auditory cortex recording with multi-microwire array and decoding of neural signal for the reconstruction of high-dimensional auditory spectrum using machine learning

Visiting student, Brown University, Providence, Rhode Island — Jan 2015
Neuroengineering and Nanophotonics Laboratory, Advised by Prof. Arto V. Nurmikko

- Characterized deep tissue powering with a mid-field wireless antenna in tissue-simulating phantom

Master's Thesis, SNU GSCS, Seoul, Korea — Mar 2014-Aug 2016
Nano-Neuro Photonics Laboratory, Advised by Prof. Yoon-Kyu Song

- Developed a dual-band mid-field wireless antenna (2.45 GHz and 4.9 GHz) for deep tissue powering using an electromagnetic simulator (HFSS) and RFID sensor for neural recording
- Conducted deep brain stimulation experiments in a pigeon for 3D-movement control

Presentation

- **2021.** Society for Neuroscience Annual Meeting, Chicago, USA. *Decoding Speech from Spike-based Neural Population Recordings in the NHP Auditory Cortex* (accepted for Mini-symposium)

- **2020.** IEEE Brain Workshop on Advanced Neurotechnologies, online. *A Wireless Network of Microimplants For Neural Recording and Microstimulation* (Poster)

- **2019.** Society for Neuroscience Annual Meeting, Chicago, USA. *Decoding Complex Sounds Using Broadband Population Recordings from Secondary Auditory Cortex of Macaques* (Poster)

- **2019.** Society for Neuroscience Annual Meeting, Chicago, USA. *Pattered Cortical Microstimulation by Wireless Ensembles of Chronically Implanted Microelectronic Chiplets in Freely Moving Rat Model* (Poster)

- **2019.** 9th International IEEE EMBS Conference on Neural Engineering (NER), San Francisco, USA. An Implantable Wireless Network of Distributed Microscale Sensors for Neural Applications (Poster)

- **2018.** IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, USA. *Wireless Power and Data Link for Ensembles of Sub-mm scale Implantable Sensors near 1GHz* (Poster)

- **2018.** IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, USA. *Multi-coil High Efficiency Wireless Charger System for Hermetically Sealed Biomedical Implants* (Poster)

- **2014.** International Biomedical Engineering Conference (IBEC), Gwangju, Korea. *Optimal Antenna Design for a Brain Implantable Microsystem Powered via Mid-field Energy Transfer*

Reference

Arto Nurmikko, Ph.D.

- Professor at Department of Physics, Brown University, 184 Hope st, providence, Rhode Island, 02912, Arto_Nurmikko@brown.edu

Yoon-Kyu Song, Ph.D.

- Professor at Graduate School of Convergence Science and Technology, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea, songyk@snu.ac.kr

Acknowledgments

It has been such a privilege to me pursuing my thesis work at Brown University that has great strength in biomedical engineering and support collaborations. First of all, I would like to express my sincere appreciation to my advisor, Dr. Arto Nurmikko, for his guidance and all his support, which has been critical in the completion of this work. In this multi-disciplinary project, our team would not be able to be tightly interconnected as of right now without his academic leadership. I am really grateful to have studied under his direction, and it has been truly inspiring to me. I would also like to thank the members of my thesis committee, Drs. David Borton, Wilson Truccolo, Lawrence Larson for their availability and feedback on this dissertation.

It has been an exciting journey to work with a team of competent experts from various disciplines at Brown University, UCSD, Baylor University, Seoul National University, and Qualcomm. Particularly, I like to thank Dr. Farah Laiwalla, Dr. Vincent Leung, Dr. Peter Asbeck, Dr. Patrick P. Mercier, and Jiannan Huang, who contributed significantly to this work by designing the ASICs described in this work. Also, I would like to also thank Dr. Lawrence Larson and Stephen Shellhammer for helping in RF communication for this work. Dr. Y.K Song, Dr. Joonsoo Jeong and Ahhyoung Lee have been very helpful in navigating issues in microfabrication in our system. I want to thank Dr. Jacob Rosenstein as well for allowing our team to use his equipment for laser ablation.I would like to express my appreciation to Ethan Mok and Chester Kilfoyle for their contribution to the RF demodulation process. I also like to thank Dr. Chris Heelan

for providing great help to me in auditory decoding work and Laurie Lynch for helping me in NHP training. I would like to thank the former and the present members of our group who helped me in various ways: Dr. Ilker Ozden, Dr. Zeyang Yu, Dr. Songtao Chen, Dr. Joonhee Lee, and Dr. Stefan Sigurdsson.

Finally, I want to thank my parents and wife for their love and support throughout my graduate study. This work is dedicated to my family with great respect and appreciation.

CONTENTS

List of Figures

CHAPTER 1. INTRODUCTION

1.1. Sensors for Brain-Machine Interface (BMI)

Detecting and stimulating physiological electrical activity at multiple points helps to gain insight into the operation of targeted biological circuits such as those in the brain cortex. Current brain-machine cortical interfaces are based on the micromachined silicon-based multi-electrode array (MEA), one example of such device in Figure 1-1a, connected to active external electronics for neural signal processing, and a single monolithic MEA offers multiple tissue contact points numbering on a scale of a hundred [1- 4]. For example, the so-called popular "Utah" MEA, which is commercially produced nowadays, has been used in a vast number of studies in non-human primates and clinical trials in humans to reveal principles underlying cortical circuits [1-5]. This Utah array consists of pyramidal-shaped 1-2 mm long p-doped silicon shank electrodes with Pt or Pt/Ir coats and parylene-C insulation, and electrodes are connected to an adapter to an external electronic through the wire-bonded bundle, as shown in Figure 1-1b. A functional cortical motor prosthesis has been firstly demonstrated using the multichannel Utah array a decade ago [1, 6]. In this demonstration, the 96 channel Utah array was implanted in the primary motor cortex of a

human with tetraplegia after a spinal cord injury, as shown in Figure 1-1c. Neural signals from MEA provided information about an intended movement and, after decoding relevant neural signals, the patient was able to open an e-mail and operate devices, such as a multi-jointed robotic arm through this cortical interface. So far, one or two MEA (96 or 192 channels) has typically been used for studies [7-11] and, to achieve a higher number of channels, multiple MEAs can be used to access broader cortical areas. However, utilizing multiple and distributed MEAs presents physical and technical challenges in tethering architecture, data transfer, and implantability, which limits its scalability allowing only up to few thousand at most.

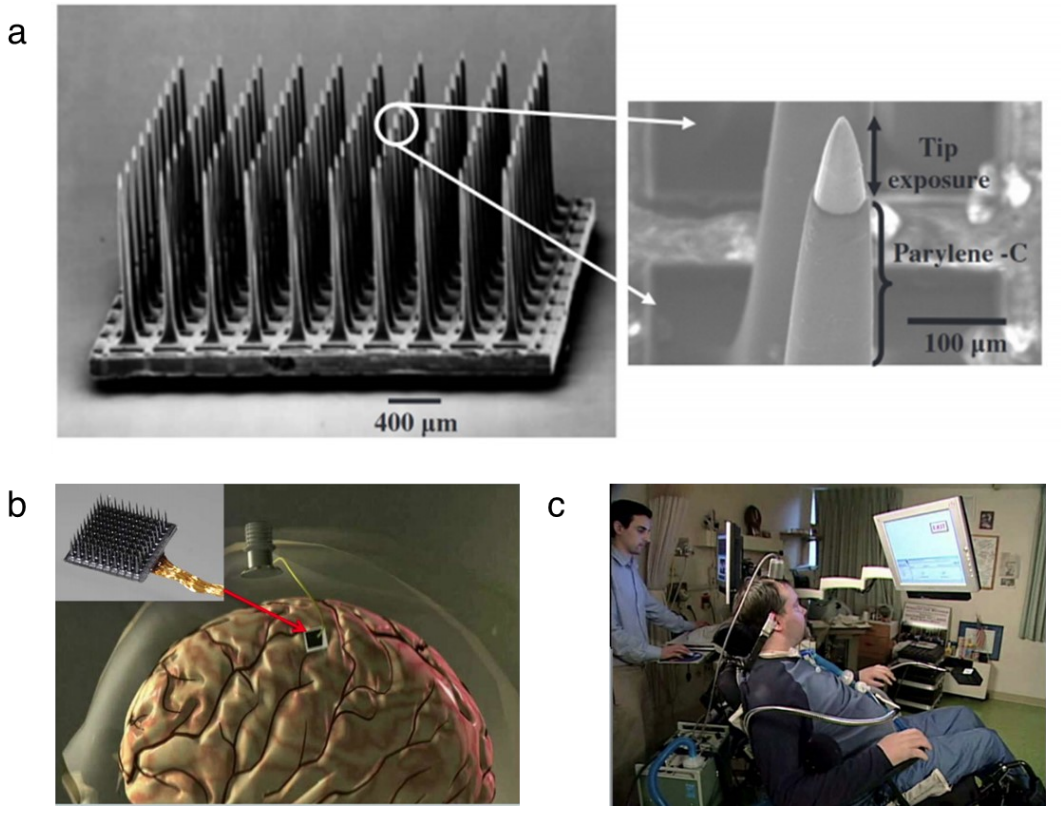

Figure 1-1. The monolithic Utah array for the multipoint neural interface. a. Scanning electron microscopy images on the Utah array and its tip. **b.** The Utah array and a gold wire bundle connecting the array to a connector for external electronic equipment. **c**. A picture for the human trial of BMI using the Utah array. Adapted from [1, 4].

1.2. Recent advances in BMI

1.2.1. Wireless neural implants

Tethering between the MEA and active external electronics is one of the major hurdles in BMI that limits system scalability and restraints the movement of the subject. In addition, the transcutaneous connection from the array to an external device increases the risk of infection and requires routine sterilizations. Wireless methods to entirely eliminate tethering to external electronics have been explored [12-19]. Figure 1-2a shows one of these systems, which uses a wireless neural processing module for multichannel recording in freely moving animals. In this work, the wireless system consists of a battery and an amplifier, multiplexer and digitizer, and wireless transceiver circuits, which is equivalent to the equipment for neural processing but a portable version. This device can simultaneously record neural signals from up to 512 channels and transmit them through RF communication. However, this wireless neural processing unit still relies on the transcutaneous connection between electrodes to the device, and the hardware size is significantly large. Previous members in our group pursued another effort using a subcutaneous neural wireless implant shown in Figure 1-2b [13]. This implant houses a wireless charging circuit, battery and a neural amplifier, multiplexer, digitizer, and transceiver circuits, which is also equivalent to neural processing equipment but is much more compact. In this system, transcutaneous tethering to the external device was eliminated and 96 channel wireless recording was achieved in freely moving animals. One of the major features in this implant is that the entire system except for recording electrodes is packaged inside a hermetic titanium enclosure for chronic biocompatibility, which has

been used in many other biomedical implants, such as deep brain stimulators or pacemakers and proved its safety for an extended period even more than ten years.

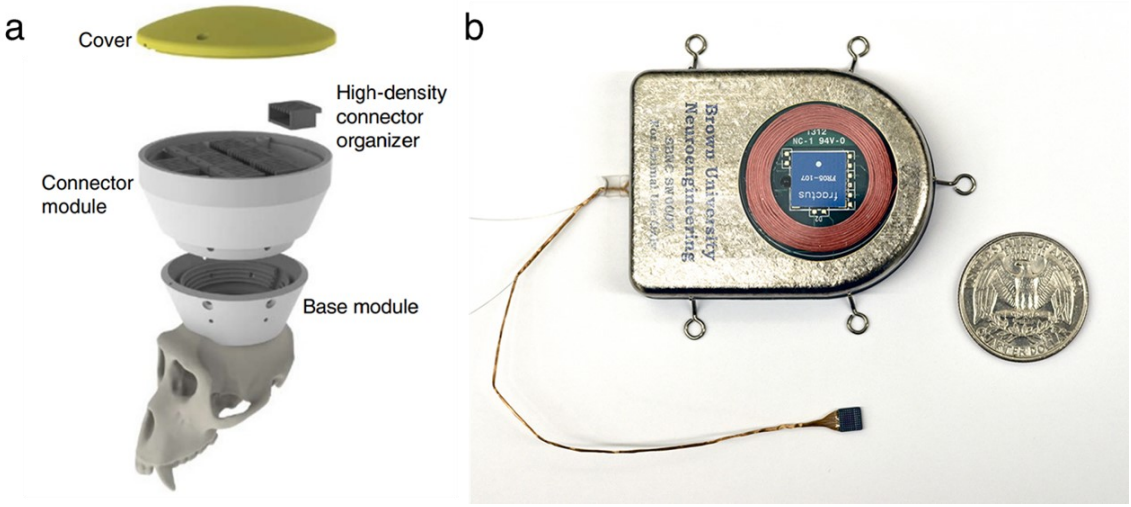

Figure 1-2. Wireless implants for neural recording. a. Large-scale portable hardware for wireless neural recording. **b.** A subcutaneous neural implant encapsulated by the hermetic packaging. Adapted from [13, 20].

1.2.2. The Neuropixel probe for recording across cortical layers

As an alternative to the MEA, several new approaches have recently been proposed leveraging recent advances in semiconductor microfabrication processes and implantation methods to scale up the channel size beyond thousands. One approach to scaling makes use of the vertical dimension with multiple multiplexed probe sites microfabricated along rigid penetrating silicon shanks, called Neuropixel [21]. The Neuropixel implant is a research device produced by IMEC (Leuven Belgium) and has vertically distributed 960 recording sites on a single, 1-cm long, shank with 70 μm × 20 μm cross-section, which extends from the cortex to deep subcortical structures to allow simultaneous recording of neural activities from multiple brain regions. Figure 1-3a shows electrodes, the probe, and

the entire system. The monolithic silicon shank and the base of the Neuropixel were fabricated under a custom 130-nm Complementary metal-oxide-semiconductor (CMOS) process, which housed amplifier, multiplexer, and digitizer circuits. Digitized neural data was directly transmitted to a head stage through a flexible connector. This system can simultaneously monitor 384 channels, while all 960 channels can be accessed through switching circuits. Using these dense recording sites, the Neuropixel can achieve a significantly high spatial resolution in recording, which can yield well-isolated activities from hundreds of neurons per probe.

Compared to the Utah array, the Neuropixel probe has significantly dense recording sites for recording along the single cortical column and subcortical areas. Therefore, this probe provides a detailed depth profile of the neural activities which can capture a layer-to-layer interaction within the column [22, 23]. The long shank structure can also in principle access subcortical areas, such as the hippocampus and the amygdala, thus providing information on an interaction between the cortex and subcortical structures, which has not been explored extensively. However, in contrast to the Utah array, the Neuropixel provides limited horizontal access due to its single shank configuration. Researchers have proposed that neural processing can occur in a horizontally distributed manner involving even multiple cortical areas and columns [24, 25]. For instance, the studies suggest that the representation of the human body is broadly distributed in the sensory-motor cortex [26, 27] and the visual field map can be found across the visual cortex [28]. Thus, the absence of the ability to access multiple cortical areas may limit the use of the locally probing Neuropixel in some applications.

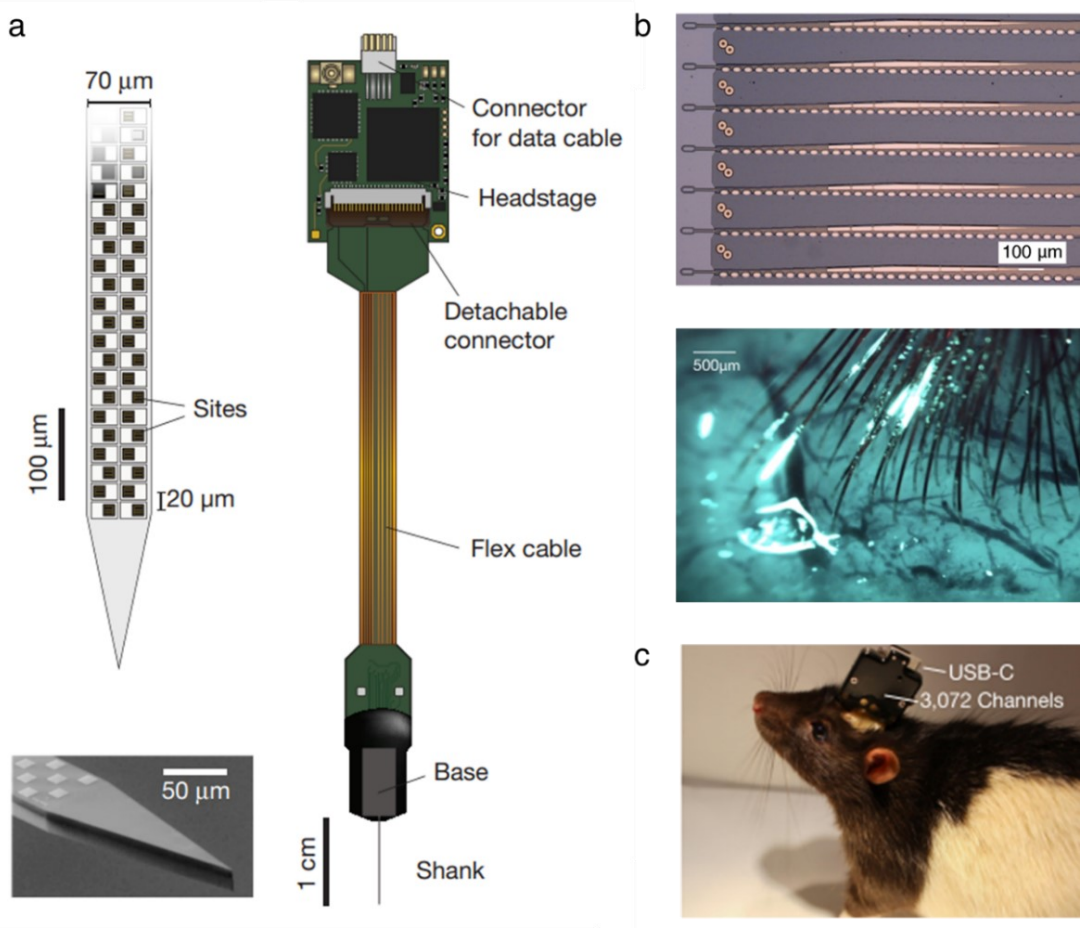

Figure 1-3. Advanced large-channel recording devices. a. The illustration of the Neuropixel system showing the probe tip, packaging, the flex cable, and the headstage. **b.** A picture for shanks with 32 electrode sites per each (top) and implanted shanks (bottom) in the Neuralink system. **c.** The Neuralink spike detection device for 3072 channels. Adapted from [21, 29].

1.2.3. The 'Neuralink' system for a high-bandwidth neural interface

Elsewhere, Neuralink Inc. has presented techniques for high-throughput electrode implantation and neural processing brain implants toward a high-bandwidth brain-machine [29]. This system leverages the most advanced implantation method using an automated machine for the insertion of multiple flexible electrodes. Figure 1-3b shows those passive flexible threads with 32 electrode sites for each on parylene-c substrate, which is not rigid

enough for the cortical tissue penetration, thus carried by a shuttle for the insertion. The robotic electrode inserter developed uses light modules with multiple independent wavelengths and four cameras to avoid vasculatures during insertions. The reported rate of electrode implantation is up to 29.6 electrodes per minute. Recording electronics was also presented, which is compatible with the implanted flexible electrode described above. The recording device consists of multiple ASICs developed by Neuralink, and two versions of the system have been proposed: 1) a 1536-channel device for broadband recording and 2) a 3072-channel device for the spike detection (Figure 1-3c). Compared to other systems, it has a remarkably small volume (24.5 mm×20 mm×1.6 mm) for the given channel size. However, as in the case of the Neuropixel, each shank carries 32 channels, and the total number of shanks is 96 for their current system with the horizontal access comparable to the 96 channel Utah array.

1.3. Wireless microscale devices for neural interface

An entirely new approach to fully 3-dimensional neural sensors, proposed by several groups, which envisions multichannel implants comprising ensembles of individual stand-alone wireless microscale sensors. These microimplants harvest wireless energy for their autonomous operation and often incorporate an ASIC to perform electrophysiological recording or cortical microstimulation across a targeted cortical volume.

1.3.1. Neural microimplants using an ultrasonic transcutaneous link

In one such approach, piezoelectric materials were used for energy harvesting and data transmission using ultrasound as the intermediary modality [30-32]. An ultrasonically powered wireless neural recording device with a single custom transistor and a piezoelectric material has been proposed, shown in Figure 1-4a [31]. In this device,

extracellular voltage changes across recording electrodes modulate the gate of the transistor, which subsequently changes the amount of current flow between the transistor terminals. By doing so, the load impedance on the piezoelectrical crystal changes altering the intensity of reflected ultrasound energy. Therefore, the electrophysiological voltage signal can be recovered on an external device by decoding the ultrasonic backscattering envelop, as illustrated in Figure 1-4a left. Recording of this wireless device has been demonstrated in vivo while collecting signals from the sciatic nerve of an anesthetized rat. This device, which is one of the ultrasmall implants, has a volume of 2.4 $mm^{3,}$ while the noise floor of the sensor was relatively high compared to other systems. Another version of an ultrasonically powered implant was proposed recently, scaling the implant volume down to 1.7 mm^3 [32]. The microdevice, designed for the electrical microstimulation for the neural tissue, consists of the piezoelectric transducer, capacitor, and ASIC, as shown in Figure 1-4b. The stimulator ASIC was able to harvest energy from the output from the transducer, deliver current-controlled stimulation pulses and communicate bidirectionally with an external transducer under the ultrasonic link. The downlink from the external transceiver set stimulation parameters on-chip to deliver programmable microstimulation and the uplink from the implant reported whether the stimulation occurred through backscattering modulation as depicted in Figure 1-4b. This system was also demonstrated in the sciatic nerve of anesthetized rats.

The ultrasonic transcutaneous link for these kinds of neural implants can be efficient compared to other modalities since the attenuation of ultrasound can be significantly lower in soft tissue [33, 34]. However, since the reflection on the bone tissue attenuates ultrasonic waves dramatically [35], such an ultrasonic powering of the neural

implants may not be suitable for the cortex below the skull, limiting their usage only to the peripheral nervous system.

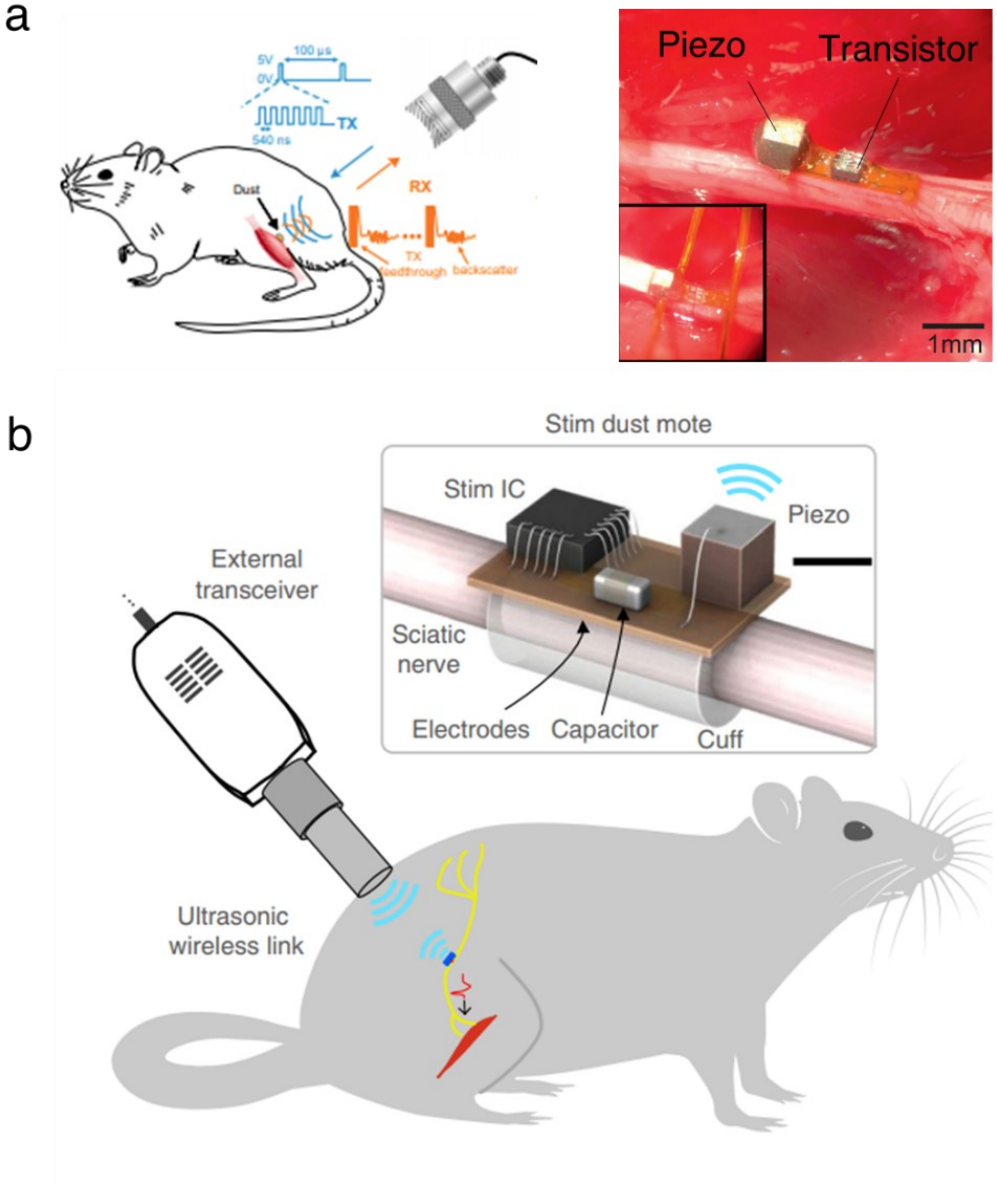

Figure 1-4. Neural implants with piezoelectric transducers. **a.** The schematic for the neural recording implant with a piezoelectric material and a transistor, which utilizes ultrasonic backscattering for data communication. **b.** The illustration for a neural stimulator powered by the ultrasonic wireless link. Adapted from [31, 32].

1.3.2. Neural microimplants powered by RF wireless energy harvesting

Purely electromagnetic wireless power transfer (WPT) and communication in the near-field (inductive) have been studied, e.g., by Ghovanloo and co-workers for mm-sized devices [36, 37, 38]. Since achieving good electromagnetic coupling with a miniature coil is inherently challenging, exploratory studies have been done for these microimplants [39, 40]. When properly optimized, RF (Radio Frequency) inductive coupling can provide an efficient path for wireless energy harvesting and telecommunication, which can also benefit from mature RF ASIC technologies.

One example of such a wireless implantable neural sensor was proposed by Ghovanloo's group, shown in Figure 1-5 [41]. This sensor has an ASIC architecture that houses a rectifier, an uplink RF communication circuit for Impulse-Radio Ultra WideBand (IR-UWB) telemetry, an amplifier for broadband neural recording, and an analog-to-digital converter. Even with this multifunctionality, the volume of the sensor was 1 mm^3 and the noise floor of broadband neural recording was 3.78 $\mu V_{rms.}$ However, this implant relies on off-chip components, including a 6-turns bonding-wire wound coil and surface-mount capacitors, which increase the overall size of the implant and also require extensive post-CMOS processing.

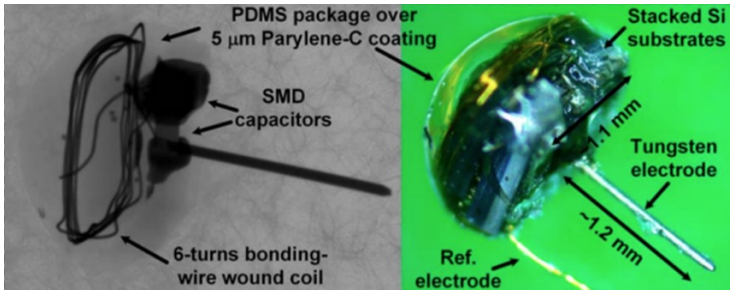

Figure 1-5. One mm-scale wireless microimplant. This implant uses inductive coupling for wireless powering and data communication [41].

The RF microrantenna coil and the matching capacitor can be directly integrated into the CMOS circuit utilizing its metal layer and on-chip capacitors so that a monolithic chip can perform wireless energy harvesting without off-chip components. But it requires a high-frequency operation of chips to reduce the matching capacitance for the coil resonance. One such implant has been proposed in [42, 43], which is called 'microbead'. The overall volume of the microbead is only 0.009 mm^3, which was achieved by integrating the on-chip coil directly above the circuit layers and custom-designed electrodes on the side of the CMOS chip. However, in this implant, the functionality of ASIC is limited for passive stimulation under the rectified voltage, and a high RF power transmission, up to 4 W, is required for the chip operation. Researchers have strived to reduce the size of the wireless neural implant, including our work described in this thesis [44-47]. Further miniaturization may allow the implantation of these devices to even an intracortical space allowing the direct interface between the device to layer 4 or 5 target neurons. So far, neural microimplants with the on-chip coil and RF energy harvesting circuits, as in 'microbead', have achieved the most miniaturized system since the monolithic integration can be easily achieved with CMOS fabrication processes.

However, to the best of our knowledge, what is missing from all prior work is the crucial aspect of the networking of multiple microimplants. Previous studies described above have demonstrated their implants on a benchtop or in-vivo, but with only a single node since their telemetry systems are not compatible with multimodal communication. We believe that one of the major motivations in developing such microimplants should be to build a spatially distributed network of microdevices to get access to multiple cortical

areas. This kind of networked microimplants can be potentially scalable beyond thousands of channels due to their wireless configuration. For that, a communication scheme, that is scalable to the desired large numbers, needs to be incorporated in neural microdevices. Recently, we have proposed particular solutions to wireless powering and communication as well as mixed-signal integrated circuits for neural signal recording and electrical microstimulation in [48-51] as building blocks towards the development of wirelessly networked microchip ensembles. In this thesis work, we report the comprehensive system-level description and characterization of such, which we term "neurograins", including in vivo demonstration of neural recording and stimulation in the rat cortex.

1.3.3. Overview of the networking scheme for wireless microsensors presented in this thesis

Figure 1-6 summarizes the salient features of the neurograin system, reported in this thesis e.g., for epicortical ECoG recording or intracortical microstimulation, where wireless communication-compatible microcircuits (ASICs) are realized in a 65 nm RF-CMOS process, each chiplet occupying a total volume of ~0.1 mm^3 (Figure 1-6c). The electromagnetic interface for communication and near field power harvesting between an external hub and the ensemble of neurograins is conceptualized in Figure 1-6a,b, which also shows the additional subcutaneous resonant co-planar relay coil designed to improve the WPT efficiency.

Standard Amplitude-Shift Keying (ASK), Frequency-Shift Keying (FSK), Load Shift Keying (LSK), or IR-UWB modulation schemes have been used to communicate with microimplants, demonstrated so far in limited channel bandwidth systems, typically below 1 Mbps [41, 52, 53]. For large ensembles of spatially distributed implants, a more versatile

networking protocol with larger data bandwidth is needed. Multichannel techniques such as frequency-division multiple access (FDMA) and code-division multiple access (CDMA) are well-established in cellular networks [54]. However, when considering their adaptation to microimplants, the FDMA approach can encounter challenges in the clock frequency tuning (described in detail in Chapter 3). CDMA implementation requires fine timing calibration and complex circuit configuration and thus has been tested only as a two-channel exploratory system so far [55].

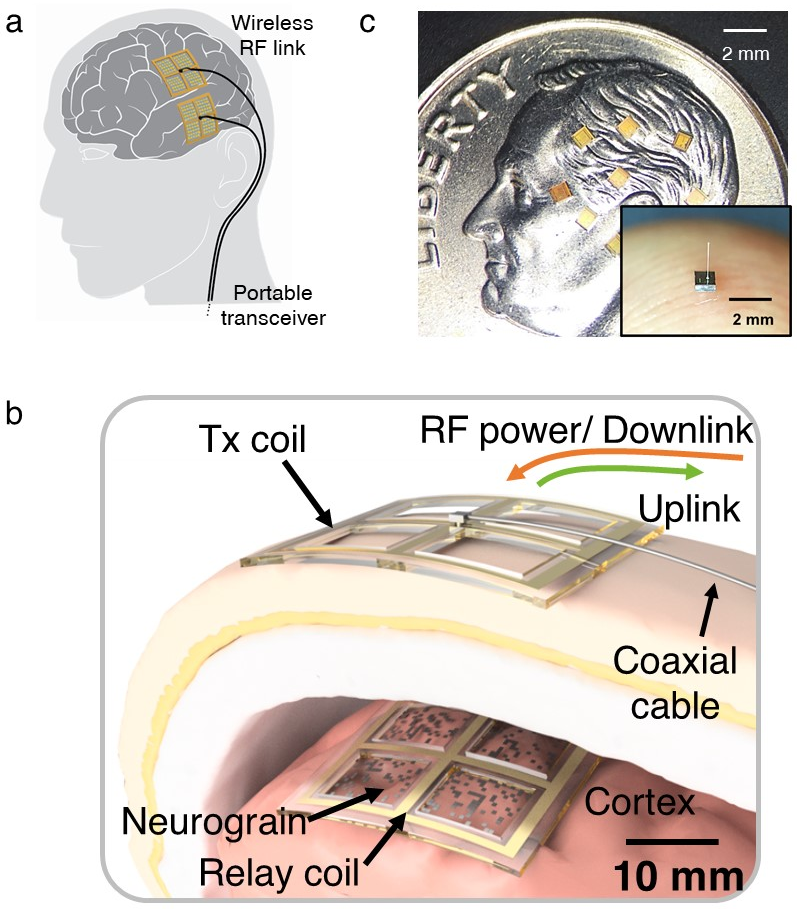

Figure 1-6. The wireless neurograin system of this thesis for distributed autonomous networking. a,b. Concept of a transcutaneous RF power and data link for a neurograin array. **c.** Handful of present 650 µm × 650 µm × 250µm sized chiplets on a U.S. dime. Adapted from [56].

In light of these considerations, we have opted to implement in our neural implants a specific power-efficient time-division multiple access (TDMA) network strategy for communication and active management of the system latency, which provides the necessary scalability and power and size efficiency in our neurograin multi-user network. The key accomplishment described in this thesis is the integration of ECoG-type neural recording, intracortical stimulation, and telecommunication functions on a single sub-mm microchip and demonstration of a multi-nodal wireless microsensor system, including initial application in-vivo. In this approach, each recording or stimulation site operates as a microminiaturized autonomous sensor node with its own independent wireless telemetry channel.

1.4. Introduction to the work presented in this thesis

The thesis research in this document involves two research projects. The main text below describes the author's progress toward building wireless networks of implantable neural microsensors and electrical microstimulators. The Appendix summarizes the outcome of a parallel research project where conventional intracortical microelectrode arrays were used to record circuit activity from the auditory cortex of non-human primates. The work benefited from neural population decoding techniques deploying machine learning techniques by the author's collaborators and provides further impetus for the need and potential benefits of future large-scale electronic cortical interfaces such as by the neurograin system.

Through the following first five chapters of the main text, this thesis will describe the progress made toward the scalable network of wireless implantable neurograin microdevices:

1. Development of the microminiaturized integrated circuits, the system-on-chip ASIC elements, for the multiple functions of wireless RF energy harvesting, neural signal amplification and digitization, and means for intracortical microstimulation, all enabled by and the accessible bidirectional communication through RF telemetry (Chapter 2)
2. Characterizing the RF networking protocols for a multichip ensemble and analyzing communication bit-error rates (Chapter 3)
3. Development of a 3-coil wireless microantenna system for optimal neurograin wireless powering (Chapter 4)
4. Application of custom microfabrication techniques to make microelectrodes suitable for neural recording and microstimulation and packaging neurograins with biocompatible materials. This chapter also describes coil fabrication and package methods (Chapter 5)
5. Demonstration of the neurograin system on benchtop and in-vivo (Chapter 6)

Chapter 2 begins with a detailed description of the circuit architecture the author, with crucial discoveries by expert collaborators, has made toward the autonomous operation of neurograin microsensors at near 1 GHz RF frequencies. The development path involved several milestone test ASICs to validate specific functional purposes. On a practical note is the long lead time typical in innovative ASIC research, between a circuit design, its tapeout and the final emergence of a semiconductor foundry. The goal was to

develop a neurograin ASIC as autonomous systems-on-chip (SoC), leveraging the TSMC 65 nm MS/RF LP CMOS process, with no off-chip components. This chip houses an on-chip microantenna coil, an RF energy harvesting circuit, electrophysiological recording or stimulation circuits, a device identifier (address), and the relevant network telemetry circuits. Each circuit block was extremely constrained in both area and power and the entire circuit architecture was realized within 500 μm × 500 μm area. The achievements made in this chapter have been published in several peer-reviewed conference proceedings. The subsequent chapters are dedicated to the characterization of each aspect of neurograin and the demonstration of neurograin interface in air, tissue-mimicking phantom, and in-vivo:

Chapter 3 concentrates on the one-to-all telecommunication challenge of an IoT-like wireless network of distributed microsensors. We introduce two networking protocols customized for neurograins 1) autonomous TDMA and 2) call-and-response TDMA under RF data telemetry. Autonomous TDMA offers a simplistic method to communicate with multiple chips only through the uplink but with the risk of packet collisions. We simulate and analyze packet collisions to find the effective channel bandwidth of this protocol. In call-and-response TDMA, the downlink triggers neurograin to backscatter the unlink in a given time using its unique device identifier to avoid packet collisions so that the network can achieve a low bit-error. While this protocol requires a bidirectional telemetry link, our experiments and simulation showed that it can be scaled up for up to 770 nodes. An RF hub, which consists of software-defined radio, RF amplifier, and SAW duplexer, is also described in this chapter, which is required for the management of RF wireless power transmission and bidirectional data communication for neurograins.

Chapter 4 focuses on the near-field electromagnetics and describes a 3-coil microantenna wireless system where power and data are communicated from an external central hub to a neurograin ensemble, developed particularly for wide-field neurograin access. By leveraging an additional epicortical resonator coil, more efficient RF wireless powering was achieved in the 3-coil system compared to the 2-coil system. The efficiency of the 3-coil system was evaluated using the neurograin clock frequency as an RF power level indicator.

Chapter 5 summarizes microfabrication techniques for neurograins, describing the multiple post-processing steps which are necessary to convert monolithic silicon CMOS chips from fab into fully functional neural implants. Here, we also describe a method to fabricate and package the microantenna coils as well.

Finally, chapter 6 presents the demonstration of the entire neurograin interface employing both recording and stimulation versions of neurograins, the customized 3-coil system, and the RF hub described in previous chapters. Up to 48 nodes of recording neurograins were tested in the saline and a small animal (rat) model to record and report the distribution of field potentials under autonomous TDMA. As in call-and-response chips, stimulation chip injected pulse-width programmable bipolar stimulation following command sequences carried in the downlink and successfully evoked changes in the field potentials and firing rates in the rat cortex. Chapter 7 will briefly compare neurograin with other state-of-art microsensors and microstimulators and describe the future goal of the neurograin project.

As noted, and in parallel to the neurograin project, I also undertook a series of studies in non-human primates using conventional microelectrode arrays to investigate

how sounds are encoded in the auditory cortex. This project, summarized in the Appendix, was motivated in anticipation of the application of the neurograin concept to study multiple and diverse cortical areas in primates. The work was published in a high-impact journal and has been invited to be presented at the Annual Meeting of the Society for Neuroscience in 2021.

1.5. Author's contributions to a neurograin team project

The neurograin project discussed in this thesis is a synthesis by a group of researchers from various disciplines to complete who have expertise in RF communication, wireless powering, circuit design, and in-vivo experiments, respectively. The purpose of this section is to define my role in the cross-disciplinary neurograin project as the lead graduate student researcher. To summarize, neurograin SoC was designed by collaborators, with my input, who are circuit experts in the University of California San Diego (UCSD), Baylor University, and Brown University. While I worked very closely with circuit designers, my contribution in this project was mainly focused on the assessment of the neurograin ASICs, including the relatively complex chip post-processing, the design and development of the resonant electromagnetic microantenna scheme including the external RF and the software-defined-radio (SDR) demodulation tools, and designing and conducting both the benchtop and in-vivo experiments. Below, I dissect my role according to the structure of this thesis. I much hope that my independent ideas, while having the privilege to engage with my senior scientist colleagues, have had a broader impact on this ambitious and complex project.

CHAPTER 2. ULTRALOW-POWER MICROSCALE INTEGRATED CIRCUITS FOR THE WIRELESS NEUROGRAINS: A SYSTEM-ON-CHIP (SOC) DESIGN

Neurograin SoC ASIC design

The circuit design for neurograin SoC was mainly implemented by team collaborators who are circuit experts at the University of California San Diego (UCSD). The circuits for the RF backbone for wireless power harvesting and data telemetry were developed under collaboration between the UCSD, Baylor University, and Brown University teams. The specific analog front-end neural amplifier (AFE for ECoG recording) and the biphasic current stimulator were co-designed by researchers at UCSD and Brown University, respectively. Note that for none of the details of the circuit blocks I claim to be the principal circuit designer.

Neurograin ASIC assessment and early implementation

I designed and configured a set of neurograin test modules consisting of wireless powering PCB antenna coils, integrating wire-bonded chips and matched capacitors. I performed multiple analyses such as that for the neural recording AFE power spectral density to obtain the chip gain spectra. As for the neural microstimulation chip, I characterized the multiple biphasic stimulation waveforms and evaluated the power-dependent current output in representative (saline) proxy neural media.

CHAPTER 3. NETWORK COMMUNICATION IN A WIRELESS NEUROGRAIN ENSEMBLE

Characterization of neurograin wireless communication on a benchtop

Using several versions of neurograin ASICs, examining the efficacy of one-to-all network communication, I performed all experiments and simulations for autonomous TDMA and organized data. For the call-and-response TDMA chip, a collaborator in Baylor University collected data and I analyzed it. I performed the demodulation experiments on the chip uplink data using the SDR configuration which I set up (below), and using detailed electromagnetic simulations designed the wireless powering antenna coils for these benchtop experiments. Subsequently, I analyzed and modeled simulated packet collision rates and associated bit-error rates for the different telecommunication protocols.

The RF wireless hub for data communication

I programmed the SDR and reconfigured its continuous memory allocator to collect continuous RF uplink data. I programmed the custom C-codes using the Linux interfacing libiio (Analog Devices) to store data from the RF front end in the dedicated RAM temporally. I performed the digital (pulse) detection referring to a code designed by a previous undergraduate student at Brown University, Ethan Mok. I also built the rest of the RF hub using a commercial RF amplifier and SAW filter.

Demodulation for bit recovery by Binary Phase Shift Keying (BPSK)

For demodulation (recovery) of the backscattered neurograin neural data collected from the SDR, I designed MATLAB codes by importing SIMULINK codes developed by Chester Kilfoyle, a former undergraduate student at Brown University. By importing codes into MATLAB, I created the demodulation process flow to be parallelly executable and much faster. I performed all demodulation processes for data in this thesis.

CHAPTER 4. WIRELESS POWERING OF NEUROGRAIN ENSEMBLES

The resonant 3-coil wireless electromagnetic link

Guided by my electromagnetic modeling and simulations, I designed the resonant 3-coil system, fabricated a series of micro and macrocoils, and performed multiple experiments to achieve impedance matching of the coils in the ASIC load environment. The modeling included realistic parameters for the scalp/skull tissue and coil geometries. I performed a grid search of the coil parameters to reach optimal wireless powering efficiency. I built PCB-based testbeds, and hermetically encapsulated these for immersion to brain phantoms.

Characterizing wireless power transfer

I realized that by using a neurograin as a power detector for received energy, I measured wireless throughput efficiency in coils for both a future primate and present rodent model. I simulated the spatial dependence of the transfer efficiency through an electromagnetic simulator and compared the result with that of the measurement. I built a dedicated RF measurement testbed using a liquid phantom to mimic dielectrics of body tissues.

CHAPTER 5. POST-CMOS MICROFABRICATION PHASE

Post-process Neurograin microfabrication

I invented the polymer-assisted microfabrication technique. My collaborator, Ahhyoung Lee at Seoul National University, performed most of the microfabrication and measured the impedance and alignment accuracy of microelectrodes on the chip.

Antenna Coil Fabrication and hermetic packaging

Polyimide PCB coils were made by a commercial service. A former postdoctoral researcher, Dr. Joonsoo Jeong in the Nurmikko lab, made the liquid crystal polymer (LCP) coil. I performed impedance matching on all coils and packaged them with polydimethylsiloxane (PDMS) or LCP.

CHAPTER 6. DEMONSTRATION OF NEURAL RECORDING AND ELECTRICAL MICROSTIMULATION BY ENSEMBLES OF NEUROGRAINS ON BENCHTOP AND IN-VIVO RAT MODEL

Benchtop validation of neurograin ensembles

I designed and performed all benchtop experiments and analyzed data both for recording and stimulation for ensembles of neurograins.

In-vivo experiments

I performed all the rat surgeries, neurograin implantations, and operated the neurograin interface to record neural signals or stimulate the cortex. I designed the in-vivo experiments and analyzed data.

APPENDIX: NEURAL POPULATION RECORDING AND DECODING FROM THE AUDITORY CORTEX OF NON-HUMAN PRIMATES

Animal surgeries and training (two NHPs)

As part of preparations for the implantation of commercial microelectrode arrays, I made an MRI-guided 3D model for the skull and the cortex of my two monkeys along with the metallic structure to guide the surgery. And I participated as an assistant during

the surgeries helping the primary surgeon. I trained the animal and collected neural data during passive listening with Laure Lynch, a former lab animal technician in Nurmikko lab.

Design and Perform Experiments to Record Population Dynamics from the Auditory Cortex.

Following the implantation of the MEAs into my two monkeys, I performed experiments over nearly three years in an acoustically isolated laboratory which I designed and constructed. The recording experiments made use of the head-mounted wireless transmitters developed in Nurmikko lab, which I used to acquire data from thousands of trials while animals listened to complex sounds.

Decoding Using Machine learning Approaches

A former postdoc, Dr. Chris Heelan in Nurmikko lab, developed a platform for scalable machine learning, called "Dockex", and he performed part of decoding experiments using Mel-spectrogram. I used Dockex to decode neural signals targeting the cochlear spectrogram.

CHAPTER 2. ULTRALOW-POWER MICROSCALE INTEGRATED CIRCUITS FOR THE WIRELESS NEUROGRAINS: A SYSTEM-ON-CHIP (SOC) DESIGN

2.1. Introduction

The material in this chapter is adapted from papers 'A 0.01-mm^2 mostly digital capacitor-less AFE for distributed autonomous neural sensor nodes.', IEEE Solid-State Circuits Letters 1.7, 2018, 'A CMOS distributed sensor system for high-density wireless neural implants for brain-machine interfaces.', In *ESSCIRC 2018-IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, 2018, 'Distributed microscale brain implants with wireless power transfer and Mbps bi-directional networked communications', In *2019 IEEE Custom Integrated Circuits Conference (CICC),* 2019, and "A distributed wireless network of implantable sub-mm cortical microstimulators for brain-computer interfaces", *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.

Wireless free-floating microdevices in the brain designed for neural recording or stimulation can be deployed in a distributed manner across a wide cortical area. High-resolution recording or stimulation necessitates the miniaturization of devices in the sub-millimeter range. Studies have also suggested that a small and untethered implant induces only modest tissue immune response [57, 58], thereby improving the implant system's chronic longevity. However, scaling neural microimplants down in size to sub-mm range brings significant physical and technological challenges to be overcome since individual microchips must incorporate multiple functions for their autonomous operation, such as wireless energy harvesting, data communication, neural recording, or stimulation.

Under these challenges, we successfully designed sub-mm size systems-on-chip (SoC) as application-specific integrated circuits (ASIC) under tightly constrained area and power while taking advantage of highly integrated TSMC 65 nm MS/RF LP CMOS foundry process. The block diagram in Figure 2-1 presents the overall electronic system consisting of the external RF hub and the neurograins, which shows their key features. Below, we describe in more detail the functional blocks in neurograin and subsequent versions of chips with different functional purposes. Note that this ASIC is designed by a team of circuit experts in the University of California San Diego (UCSD), Baylor University, and Brown University.

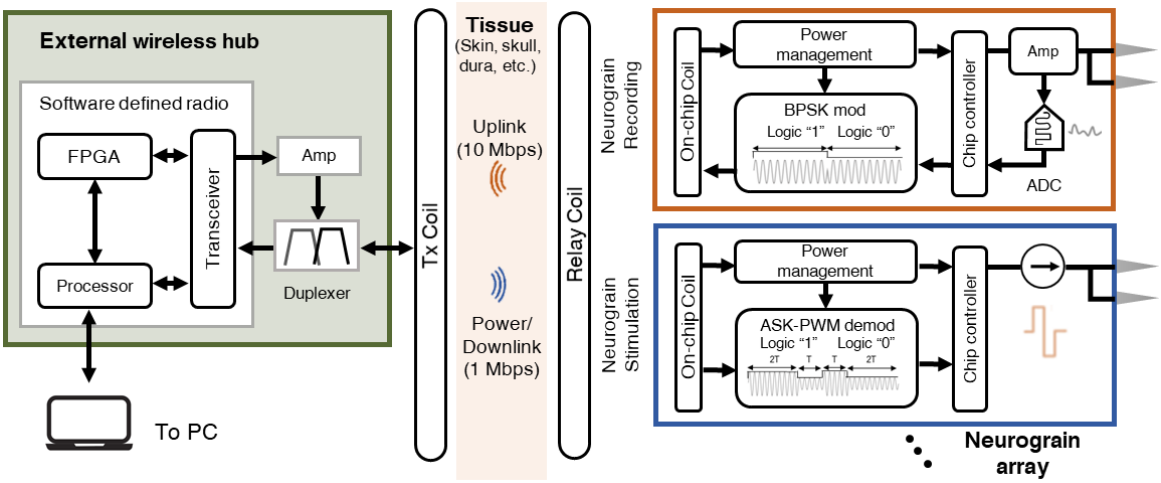

Figure 2-1. The block diagram of the overall electronic system. The system consists of the external RF hub, inductive coupling link, and key features of the neurograins for recording and stimulation, respectively.

2.2. Circuit architecture of a neurograin chip

2.2.1. RF wireless power harvesting

Since incoming RF energy is limited by a small-cross sectional neurograin on-chip microcoil, a highly efficient RF energy harvesting scheme is a critical feature for the practical feasibility of neurograin microimplants. Both the recording and the stimulating neurograin chiplets share a common "RF interface" (microcircuits indicated by the blue arrow in Figure 2-2, while the red arrow indicates the purpose-specific sub-circuit, here for microstimulation). This common RF interface includes a rectifier circuit that harvests RF energy, feeding off a three-turn on-chip coil magnetically coupled with the transmitting coil. The rectifier consists of three-stage, low turn-on, deep n-well NMOS transistors [59] to achieve up to 48.5 % rectification efficiency at -10 dBm input through avoidance of body effects in the NMOS transistor [49]. The on-chip rectifier is tuned at 915 MHz which is the transmitting "tone" in neurograin wireless communication.

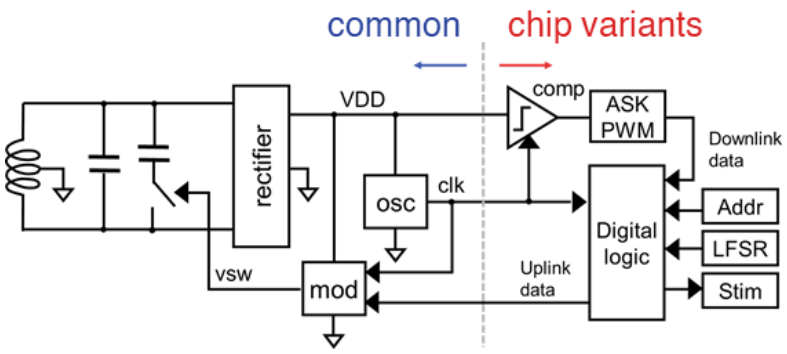

Figure 2-2. The stimulation ASIC circuit block diagram. The blue arrow highlights the RF interface (circuits common to neurograins of the different modality), while the red arrow points to blocks dedicated to particular physiologic interfaces (recording, stimulation).

On the other hand, the magnetic (inductive) coupling between neurograin coil (Rx) to the transmitting coil (Tx) via RF backscattering can vary depending on the location of a neurograin, a factor which subsequently affects the chip's rectifier output level. Therefore, in a higher magnetic field, the voltage supply (V_{DD}) on neurograin increases, which subsequently affects the following circuits. For instance, Figure 2-3 shows a clock frequency roughly proportional to the received RF power in decibel-milliwatts, which shows that our neurograin can operate at an RF incoming power as small as -10 dBm (100 μW). In the case of a higher B-field region, the voltage supply can exceed the chip's operating range and even damage the following circuit permanently. To accommodate this RF power distribution, we have employed a shunt diode that protects the chips from unintended excessive voltages. From another point of view, the chip has a clock sensitive to incoming RF energy so that the frequency of the clock can be used as an indicator for RF. For wireless efficiency calibration described in Chapter 4, we actually used this frequency dependent clock as an RF power detector to quantitively analyze features of our wireless link.

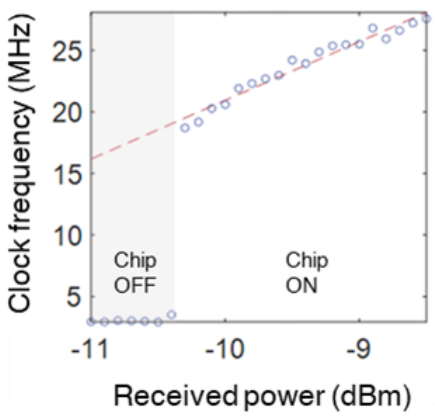

Figure 2-3. The relationship between the received power at the Rx coil and its impact on the on-chip relaxation oscillator's (clock) frequency.

2.2.2. RF uplink telemetry and on-chip modulation circuit for data transmission by RF backscattering

Neurograin microimplant transmits its data, such as recorded neural signal and device address, through an RF backscattering by modulating the circuit impedance ("mod" block in Fig 2-2). The RF backscattering data uplink channel operates at a rate of 10 Mbps near the ~1 baseband GHz wireless powering link. For this uplink communication, we use a Binary Phase Shift Keying (BPSK) modulation. The modulator ("mod" in Figure 2-2) operates by toggling a tuning capacitor to enable RF backscattering. While ensuring that the magnitude of its input impedance is maintained at a constant value, the phase modulation is maximized by selecting the value of toggling capacitor to maintain a stable power supply to the chip [49]. Data is upconverted prior to the backscatter using an on-chip, using a free-running relaxation oscillator with an operational frequency near 30 MHz resulting in the backscattering centered at 945 MHz. This conversion enables spectral isolation of the BPSK-modulated signal from the transmitted 915 MHz RF powering (Tx) tone. High fidelity isolation of the backscattered data facilitates its low-noise demodulation and recovery for online or offline analysis. An example of the demodulation of the BPSK backscattered signal is shown in Figure 2-4, which shows zero-bit errors over 77500 bits using a networking prototype chip that transmitting 31 repeating bits continuously.

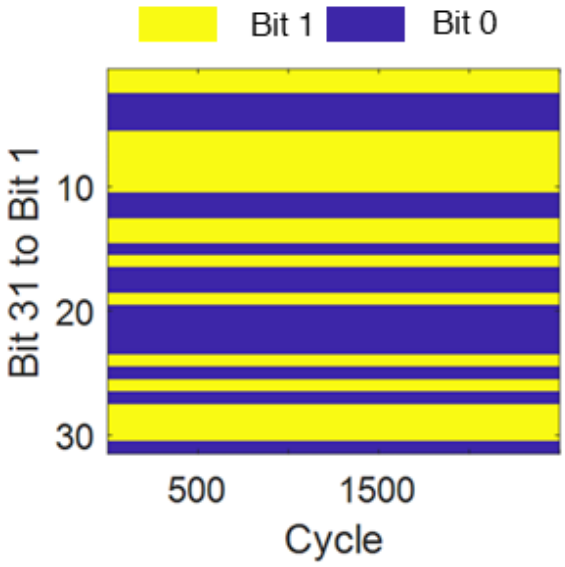

Figure 2-4. Continuous BPSK data from the test chip with 31 repeating LFSR bits over 2500 cycles showing zero-bit error.

2.2.3. RF Downlink telemetry and on-chip demodulation

The network architecture and band separation in neurograin allow the incorporation of downlink communication around the 915 MHz Tx baseband tone to complete a real-time bidirectional communication loop. For neural stimulation applications, we choose a 1 Mbps downlink rate, sufficient to include a range of command signals for commonly used electrical microstimulation protocols for the mammalian brain.

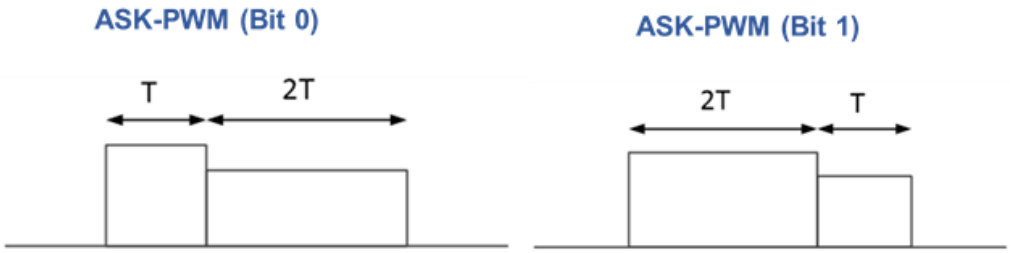

Figure 2-5. Sample ASK-PWM bits for data and frame synchronization; a "short high" represents bit 0 (left) while a "long high" is equivalent to bit 1 (right).

In configuring the downlink, we face a particular challenge for designing a communication protocol for neurograins due to the inherent variance in the on-chip generated clock frequency and phase which arises from the finite variation in the power harvested by each chip. This variation makes synchronization extremely difficult to achieve under usual protocols. We tackled this challenge by using a custom amplitude shift keying and pulse width modulation (ASK-PWM) scheme to implement the 1 Mbps downlink. ASK-PWM is the technique that encodes bits by amplitude and pulse width, as shown in Figure 2-5. Since the free-running oscillator on-chip generates a much faster clock frequency compared to the ASK-PWM data rate, the counters in the decoding circuit (Figure 2-6) can compare the relative pulse width and demodulate the bits. Since this process only compares relative pulse width, it can provide network synchronicity and be robust across the wide range of chip clock rates. This type of ASK-PWM is different from pulse width modulation (PWM). The low-state amplitude is only slightly smaller than that of the high-state, critical in the neurograin operation since by doing so, the power fluctuation during the downlink can be much reduced, and a stable wireless power supply achieved in the ASK-PWM scheme.

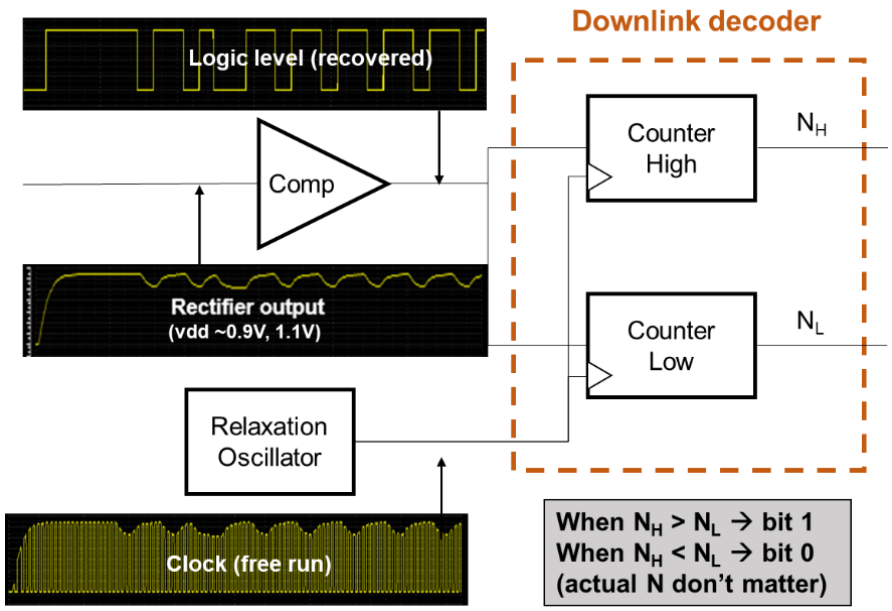

Figure 2-6. The logic decoder for relative pulse width comparison between the high and the low states. In the ASK-PWM protocol, the lower state (RF) energy total is smaller than that of the high state; this decreases the power level fluctuation during the downlink cycle.

2.2.4. Unique device identifier (neurograin address)

A unique device identification (ID) is imperative for the functional implementation of an addressable wireless network. For the neurograin microimplants, we have successfully tested two different ID approaches: (a) a physical digital address, engraved on-chip by laser ablation [60], and (b) an addressing scheme based on physically unclonable functions (PUF) [61, 62], the latter discussed below in a neural recording example. In the case of the laser ablation approach, the use of which is demonstrated below for microstimulation, a user-selectable 10-bit address is physically imprinted onto an array of ten metal interconnect traces on the chip top aluminum metal layer by selectively ablating a specific number of traces with a 532 nm pulsed laser. By contrast, the address circuits using the PUF approach take advantage of intrinsic 65 nm CMOS fab process

variations to generate unique randomized bit patterns. The number of bits in a PUF address has to be carefully considered to minimize the chance for address overlap for a given population of networked nodes. In our case, we have implemented 16 to 20-bit PUFs, anticipating an insignificant address overlap rate for ensembles of up to 1000 nodes. In general, the PUF approach offers distinct advantages in scalability due to the compact footprint and absence of labor-intensive post-processing; however, the unique address needs to be "discovered" for every neurograin. To discover the PUF address, each neurograin has to be tested individually before the experiments as they transmit their PUF address through the uplink.

To test whether a physically unclonable function (PUF) can provide 1000 unique device IDs, we analyzed the number of PUF bits and the probability of device ID overlap while assuming PUF bits are randomly determined having an equal chance to be either 0 and 1. Here the target number of nodes was 1000, and, as shown in Figure 2-7, the possibility of device ID overlap decreases for the higher number of PUF bit content. For the 16-bits and 20-bits PUF adopted in our system, we calculate a less than 1.5 % and 0.1 % chance of any address overlap between nodes, respectively. This simulation shows that PUF-based device ID can provide all 1000 nodes unique addresses without any post-processing labeling after the CMOS fabrication.

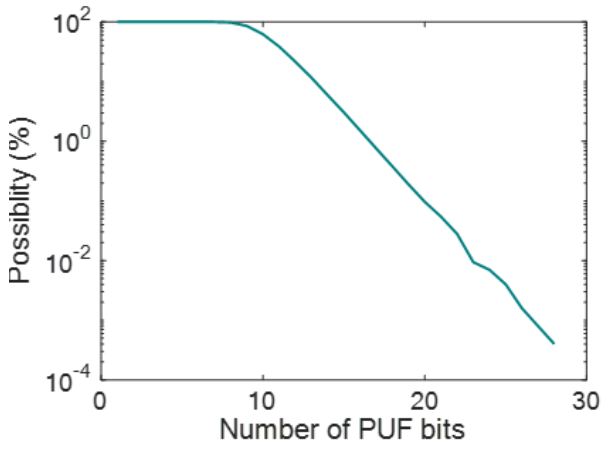

Figure 2-7. Analysis of the uniqueness of PUF addresses and the corresponding packet success rate in the autonomous TDMA scheme. The simulated possibility that any two PUF address among 1000 of them have the same bit sequences plotted under the assumption that the PUF bits are decided randomly.

2.2.5. Capacitorless analog-front end for ECoG recording

The front end of a neurograin SoC integrates high-fidelity physiologic sensing (recording) and actuating (stimulation) circuits with a particular focus on ultra-miniaturization and extremely low-power operation. To this end, we have designed an analog front end (AFE) for recording, a version of which is shown here for the acquisition of electrocorticographic (ECoG) signals using a DC-coupled V/I converter merged with an analog-to-digital converter (ADC) via an area-efficient capacitor-less approach [63]. The AFE directly senses neural potentials (without the need for bulky AC-coupling capacitors) with input-referred noise of 2.2 μV and 500 Hz bandwidth, and occupies an overall footprint of 100 μm × 100 μm while recording at a 1 kHz sampling rate and 8-bit resolution. This DC-amplifier allowed us to realize the recording circuit in such a small area but required proper DC-biasing for the circuit operation. We chose pair of Au electrodes and fabricated them on the chip (detailed in Chapter 4).

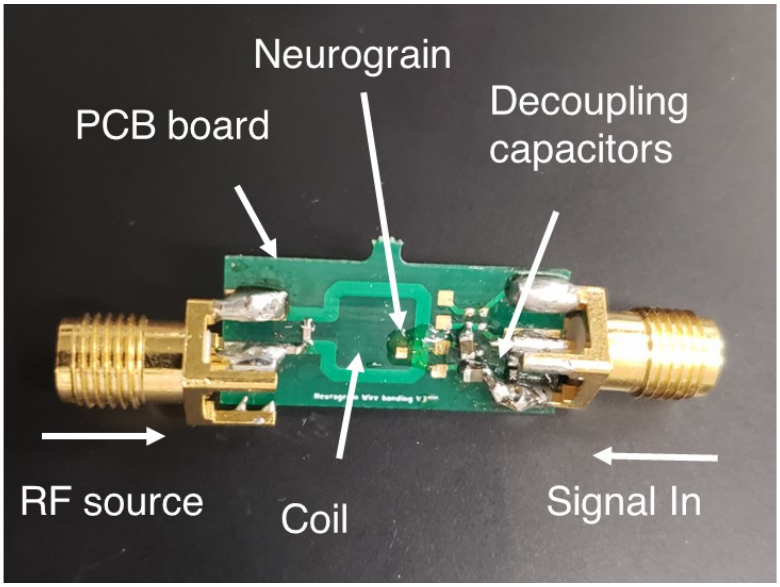

Figure 2-8. A testbench composed of a PCB board housing a receiving microcoil and a wire-bonded neurograin for the analysis of the performance of the AFE circuit.

To analyze the performance of AFE in the wirelessly powered chip, we built a test PCB structure with a coil shown in Figure 2-8. Here, decoupling capacitors are used to simulate the high impedance of the saline in the low frequency regime (<< 1 KHz) and to prevent the external signal source from affecting internal DC bias in AFE. For the input signal, we used an arbitrary function generator. Figure 2-9 (left) shows the result on AFE recording for 40 Hz sinusoidal wave according to its power spectral density, which shows the capability for low noise recording of AFE even with floating ground on-chip and presence of the wireless RF transmission. We also tested signals at different frequencies (Figure 2-9 right), where the large attenuation in the low frequency results from the decoupling capacitors. Even so, the AFE was able to capture the sub-Hz low-frequency signal as well as high-frequency components up to several hundred Hz, compatible with e.g. ECoG recording requirements.

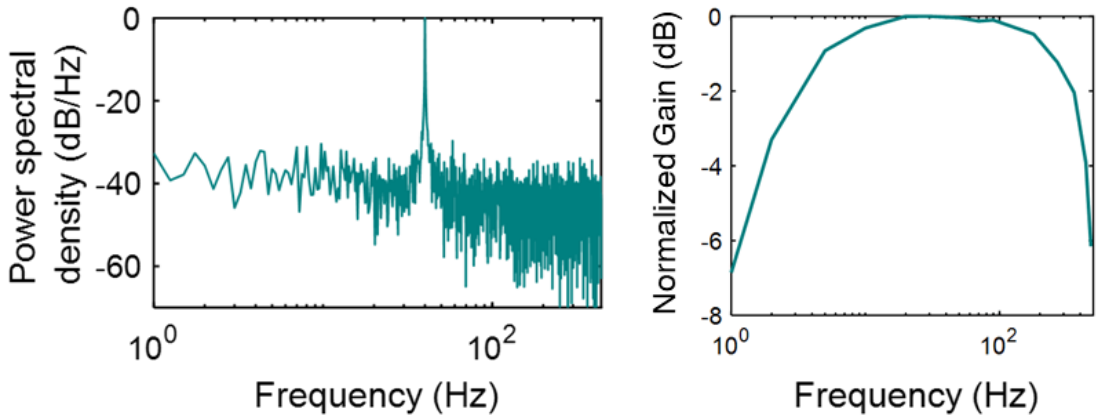

Figure 2-9. The normalized power spectral density (PSD) and the normalized gain spectrum of the recording neurograin. Neurograin detected a 40 Hz, 2 mV_{ppk} sinusoids. The low cutoff frequency is determined by the decoupling input capacitors in the test bench which mimics the high DC impedance in the electrode-electrolyte interface

2.2.6. Circuit delivering biphasic current injection for intracortical stimulation

Electrical microstimulation is facilitated through circuits designed to deliver biphasic current pulses with programmable pulse widths (here 100, 200, and 400 μs) as commanded by the external wireless hub through the downlink. "Stimulation neurograins" were likewise tested after wire-bonded on PCB structure to measure its biphasic stimulation. Figure 2-10 left shows stimulation output from the chip, which successfully generates biphasic shape and follows the downlink command for pulse widths. At present, the maximum level of current stimulation depends on the chip power level. It can range up to 25 μA as shown in Figure 2-10 right, and the level of stimulation current is sensitive to the received RF energy. Future work will incorporate stimulation amplitude regulation across the network.

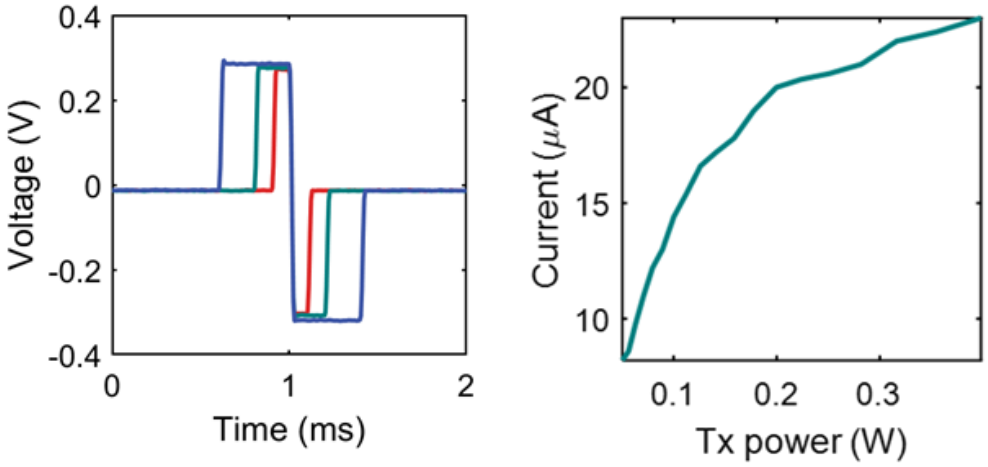

Figure 2-10. Biphasic current stimulation (voltage) waveforms generated by the microstimulating neurograin and corresponding injected currents as a function of the external Tx delivered power in a 2-coil test setup. Three different pulse widths per phase (100, 200, 400 μs) are shown and a 20 kOhm load was used.

2.3. Neurograin chip types for specific functional purposes

Introduction

Using blocks previously described, we have built three types of neurograin for different functional purposes as part of the journey towards the full SoC implementation of the wireless microsensors. A "Communication neurograin" was designed to quantify the network bit-error, thereby only carries networking circuits and transmits predefined bit-sequence used later for bit-error calculation. The "recording neurograin" is targeting epicortical ECoG recording with 1 kHz data sampling rate and 8-bit ADC resolution and transmits the uplink every 100 ms. "Stimulation neurograin" (mentioned above) performs biphasic current stimulation and ASKPWM demodulator for the downlink. The purpose of this series of chips was the goal to integrate the distinct functionalities on a single chip while recognizing the time-consuming logistics from the tapepout of any design to the

actual delivery of silicon die from the semiconductor foundry. The microphotograph of Figure 2-11 shows the overall footprint and main functional circuit areas of the different ASIC testbeds. The diced chips have a nominal areal size of 650 µm × 650 µm, including a 75 µm gap between the chip edge and the coil where the seal ring is placed at the foundry. Within the available 500 µm × 500 µm footprint, the 3-turn coil occupies 30 % of this area for RF harvesting, while the rectifier and uplink circuits employ around 23 % of the total area. The digital circuit, which handles data input/output and storage, regulator, and other circuits such as PUF, takes up most of the chip real estate in the recording chip.

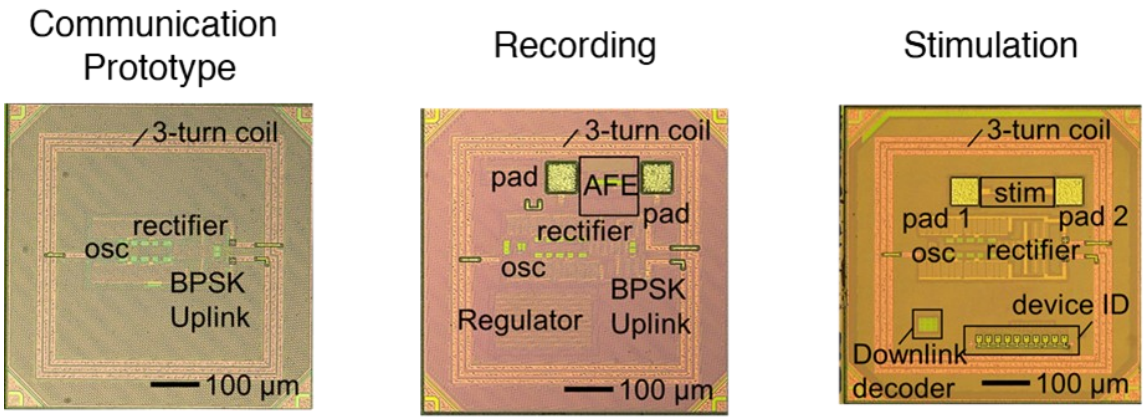

Figure 2-11. Photomicrographs of the principal ASIC test chips, active circuits in each design fitting inside the 500 µm × 500 µm 3-turn microcoil (identical for each design). (left) The networking test chip; (middle) A recording neurograin with the AFE and voltage regulator of VDD; (right) stimulator chip with downlink decoder and metal-fuse based device ID.

2.3.1. Designing circuits for network communication

Three types of circuits exploring the communication capability of a scalable neurograin network was developed. 1) continually backscattering chip, 2) autonomous TDMA (Time-domain multiple access) chip, 3) call-and-response TDMA chip. Chip type 1 was the first prototype chip which transmits repeating 31-bits from the Linear-feedback

shift register (LFSR). This chip operated as expected, validating the performance of the coil matching and RF harvesting circuit. We analyzed the 10 Mbps uplink path and its bit-error through this chip, as shown in Figure 2-4. Given that the clock frequency is proportional to the received power, shown above in Figure 2-3, we used this chip also as a wireless power detector and measured wireless efficiency by recovering its clock frequency from backscattered data. Chip type 2 shares most of its circuit with chip-type 1, but backscatters 32 cycles of 31-bits every 100 ms, a test for autonomous TDMA validation. Chip type 3 can perform bidirectional communication through the uplink BPSK modulator and downlink ASK-PWM demodulator circuits on the chip. The chip transmits 32 cycles of 31-bits only when the downlink calls out its unique PUF address. The type 2 and 3 chips were used in our full network communication validation, described in detail in Chapter 5.

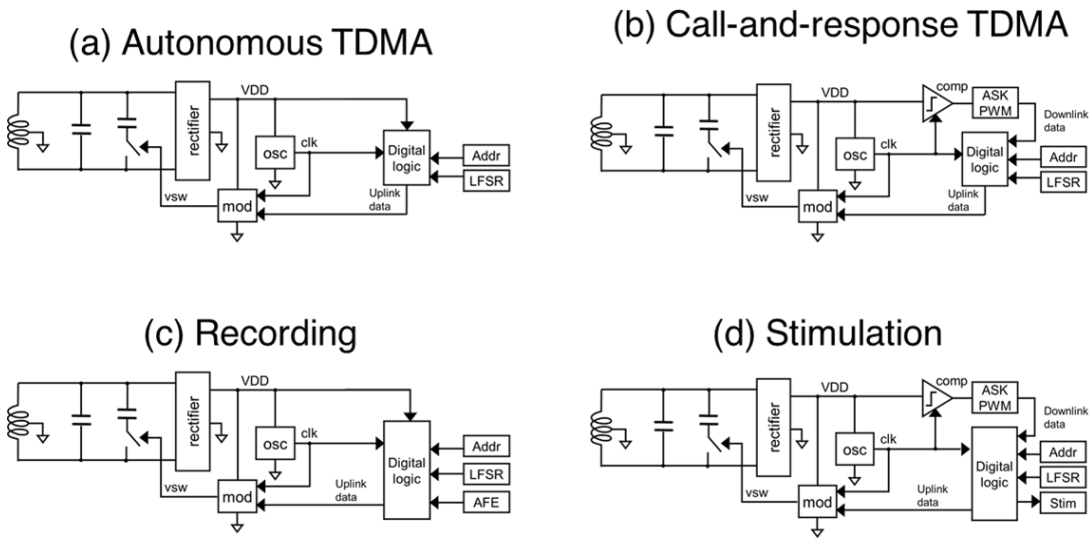

Figure 2-12. Neurograin chip block diagrams. (a) Autonomous TDMA networking chip (with PUF address). Using a BPSK modulator, each neurograin backscatters every 100 ms while its data packet length is 100 us. The data packets contain about 1000 predefined bits from a linear-feedback shift register (LFSR) for BER performance evaluation. (b) Call-and-response TDMA protocol-based chip, which shares RF energy harvesting circuits as in the autonomous version but adds an amplitude shift keying, pulse width modulation (ASK-PWM) demodulator for bidirectional communication. (c) Neural recording ASIC

with a PUF address and a capacitor-less analog-front-end (AFE). The recording chips have a voltage regulator to ensure stable recording. The chip backscatters in autonomous TDMA mode every 100 ms with the length of each data packet being 82 us. A packet contains predefined 32 bits from the LFSR, 20 bits dedicated to the PUF address, and 800 bits for the ADC output (8 bits/100 samples) (d) Stimulating Neurograin with an ASK-PWM demodulator and a biphasic current source stimulator. The chip generates current output for injection to neural tissue when the address code in the incoming downlink matches a 10-bit laser-written fuse address on-chip.

2.3.2. Design of the ECoG-type recording neurograins

The recording chip has a capacitorless AFE designed specifically for ECoG-type epicortical recording with corresponding on-chip data digitization. The uplink path uses autonomous TDMA networking while it transmits 800 bits (100 sample 8bits for 100 ms recording) of recording every 100 ms. Extensive analysis of recording capability has been performed particularly as this chip has a free-floating ground and records differential signal between two closely placed electrodes (Au-pads in Figure 2-11). Along with 800 bits of recorded data, it transmits 16 bits of PUF address and 32 bits of LFSR which allows for tracking the recorded data from each chip over time while also checking the bit-error encoded in those predefined 32 bits LFSR. The design calls for an on-chip regulator which provides the AFE a stable V_{DD}= 0.7 to stabilize the recording even during the ambient RF fluctuations. We will demonstrate the performance of the recording neurograins on the benchtop and in-vivo rat model below in Chapter 6. In the following chapters, we used the term, “recording neurograin” to refer to the version of the ECoG recording chip described here.

2.3.3. Design of neurograin for intracortical stimulation

As shown in Figure 2-12, a neurograin designed for electrical microstimulation also houses bidirectional communication blocks but need not operate in the call-and-response

TDMA mode as its main function is only to deliver electrical stimulation. The chip needs to transmit an uplink signal only when turned on. The uplink carries the on-chip 10-bit address, here a laser-written address, which enabled us to check the corresponding address and then deploy it for the downlink. Inclusion of the chip address into the downlink, allows the chip to recognize the address and command sequence to deliver intracortical stimulation on command. We implemented a means to program the temporal details of the stimulation to include both single shots as well as 100 Hz pulse train paradigms, the latter mitigating the data burden on the downlink. The single-shot mode can deliver burst stimulation at rates faster than 500 Hz using the recursive downlink, above the frequency commonly used for electrical microstimulation in the cortex. The stimulation pulse width was varied (100, 200, 400 μs per phase) to range across the commonly accepted strength-duration curve [64, 65]. Note that the present stimulation neurograin has a power-limited charge-injection capability allowing subthreshold stimulation by each device in an ensemble. Actual demonstration of the performance of the stimulation neurograins both on the benchtop and in-vivo rat model are described in Chapter 6. In the following chapters, the term, "stimulation neurograin" indicates the version of neurograin for the intracortical stimulation described here.

2.4. Conclusion

In this chapter, we have presented the analog and digital circuit designs and the RF telecommunication circuits as functional blocks within the overall SoC approach to neurograins and describe their separate functionalities. This system leverages recent advances in low-power microelectronics suitable for building miniaturized implants capable of on-chip housing of the multiple functional circuits. Under the collaboration of

multiple research teams, each block was designed as a low-power compact circuit using the TSMC RF LP 65 nm process node. We have described the details of each circuit block and the rationale in first designing and implementing a series of test chips for the different functional purposes of communication, neural recording, and neural stimulation. We have shown that the neurograin RF backbone design can efficiently harvest RF energy as well as communicate digitally at a low bit-error-rate, achieving 10 Mbps fast uplink and 1 Mbps downlink for the bidirectional communication system. Lastly, we have successfully demonstrated both the recording and stimulation capability of neurograins as the monolithic integrated wireless system validating their performance under wireless energy harvesting.

CHAPTER 3. NETWORK COMMUNICATION IN A WIRELESS NEUROGRAIN ENSEMBLE

3.1. Introduction

The material in this chapter is adapted from papers 'Distributed microscale brain implants with wireless power transfer and Mbps bi-directional networked communications', *IEEE Custom Integrated Circuits Conference (CICC),* 2019, and 'An implantable wireless network of distributed microscale sensors for neural applications', *9th International IEEE/EMBS Conference on Neural Engineering (NER),* 2019.

Networking of multiple wireless sensors such as being developed for IoT applications often presents a significant communication challenge due to the distribution of asynchronous clocks across wireless microdevices and the limited available data RF bandwidth. Our implantable neurograin system is no exception. We began by investigating networking protocols previously developed for telecommunication and compared those approaches in the context of the proposed neurograin networks. TDMA (Time-division multiple access), FDMA (Frequency-division multiple access), and CDMA (Code-division

multiple access) can all be employed for one-to-many communication of wireless neural microdevices, and a few studies explore those techniques but in a limited number of channels [55, 66].

As the major project goal, we aimed to communicate with up to 1000 neurograin nodes where each node transmits ECoG-type neural data collected at the 1 kHz sampling rate and digitized at 8-bit resolution (the total of 800 bits). To accommodate up to 1000-channel ECoG recording nodes, the networking protocol needs to achieve at least 8 Mbps data transfer rate in total.

In FDMA, each RF device uses a different frequency band to allow neural data sampling without the network latency while each node occupies a minimum bandwidth of 8 kHz for the current ECoG application. However, the channel bandwidth for each node needs to be increased for overhead to adapt to clock variances within the neurograin population. But at the same time, the RF harvesting approach here incorporates a 3-coil system with the intermediate (unloaded) resonant coil showing a high-Q factor and a narrow RF bandwidth. Therefore, the total group bandwidth available for the neurograin population is limited by the RF bandwidth of the resonant coil, which subsequently limits the total available channel size in the neurograin system. On top of that, the RF bandwidth allocation in FDMA, accomplished by fine-tuning the clock frequency of each neurograin, is formidably challenging in the large number, i.e., 1000 nodes. For that, each node needs to be tuned differently by the circuit design or post-processed after CMOS fabrication for selective frequency tuning, which can be labor-intensive.

For the synchronous CDMA application, the synchronization of the network plays a key role, and the clock frequency of all neurograin nodes needs to be precisely tuned as

well as in phase. Under the clock variance of neurograin, synchronous CDMA cannot operate as expected. On top of that, CDMA primarily operates by using mutually orthogonal digital signals (which are used as bits) but programming those codes for the individual chip can be challenging, though not impossible. Asynchronous CDMA is not applicable here as in the number of nodes over hundreds; the aggregated noise level will be too high. On the other hand, designing nodes with fixed clock frequency can provide a solution for FDMA and CDMA. However, this also has a significant technological challenge if one considers the circuit complexity, power, and area constraints of the microchip.

A practically useful TDMA protocol requires fast data transmission during a short period and has lower spectral efficiency. However, due to the usage of PUF (physical unclonable function)-based device ID, all chips can have the same fixed design and do not require individual post-processing for the channel separation. Also, when incorporating the ASK-PWM downlink as in the custom-designed call-and-response TDMA scheme (described below), the clock variance does not significantly affect the channel bandwidth. Under the neural interface perspective, call-and-response TDMA also allows channel selection for nodes with important neural information and more frequent checking to reduce the latency. Therefore, we employed a TDMA networking protocol in the neurograin system using a single carrier frequency at 915 MHz to enable periodic, scheduled communication to and from ensembles of implanted devices.

3.2. Neurograin TDMA network

Introduction

Neurograin telemetry circuits were designed to employ the TDMA protocol to accomplish one-to-many communication for ensembles of microsensors. We have investigated two versions of TDMA-based schemes, an "Autonomous" and "Call-and-Response" approaches, which are illustrated in Figure 3-1. The key tradeoff in large-scale network design balances hardware complexity versus networking efficiency. The former defines stringent power and area constraints for the microdevices, while the latter is critical for scalability. In recognition of these considerations, the neurograin SoC network was initially first realized using unidirectional "Autonomous" TDMA. This protocol is a simple, compact, and low-power approach for early ensemble testing at benchtop in rodents where channel counts are limited by anatomy. Benchtop experiments were performed to validate and compare the two communication protocols in the air using the 3-coil wireless system with 8 mm separation between the Tx to relay coil.

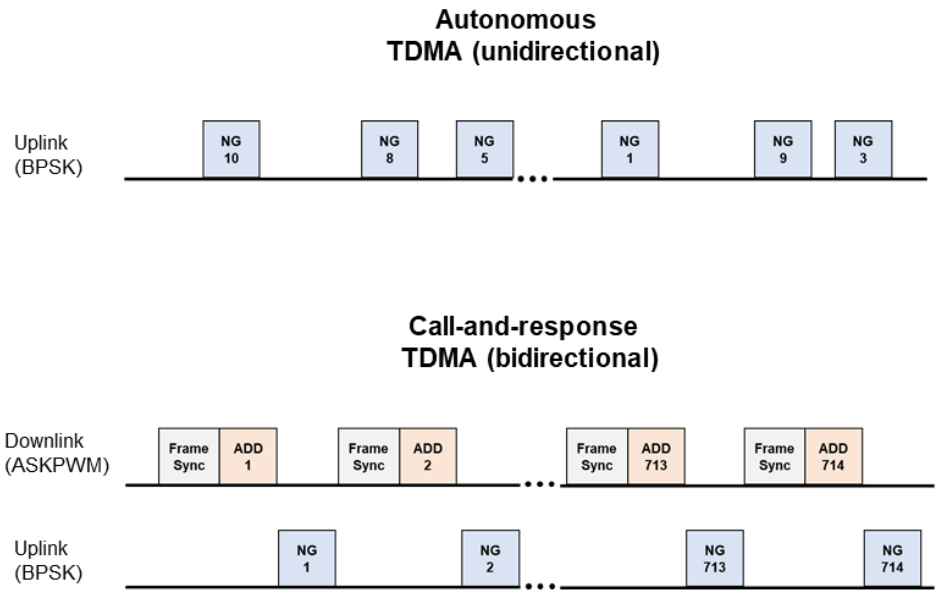

Figure 3-1. The schematic to illustrate the timing diagram for autonomous and call-and-response TDMA cases, respectively. Each contains a 100 ms frame and 82 us data packet length.

3.2.1. The autonomous TDMA scheme

The "autonomous TDMA" is a customized networking protocol that provides a simplistic way of multinodal communication, only incorporating the uplink stream from the chips. This protocol relies on the chip's unique device ID and this built-in prefabricated PUF ID probabilistically determines the device's initial communication timeslot in the network queue. Data is periodically backscattered at the predesigned network latency (nominally 100 ms). Each microdevice needs to buffer and store up to 100 ms of data (820 bits, PUF address, and recording); This data is packetized and transmitted over 82 µs (10 Mbps rate). The on-chip clocks are designed to be free-running at 30 MHz with +/-12.3%; this variation is due to the spatial dependence of the magnetic coupling strength and the variation in the received RF power. Figure 3-2 shows the measured transient data received at the external hub from a network of 64 chips. The "pulses" indicate periodic data packets sent by the nodes which present different backscattering amplitudes depending on the magnetic coupling efficiency.

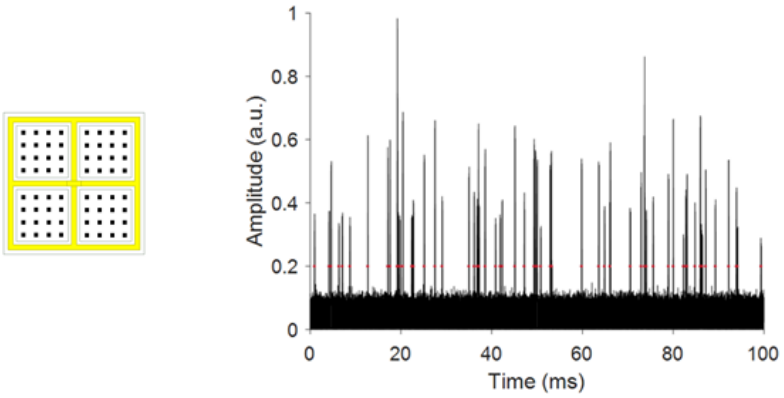

Figure 3-2. Demonstration of data communication on the benchtop with 8 mm separation between the Tx to relay coil in the air. (left) 64 autonomous TDMA chips, with placement within a relay coil area of 20.4 mm × 20.4 mm. (right) A transient recording on uplink data packets from multiple chips under "Autonomous" TDMA.

For each data packet, decoded bits are compared with the predefined sequence to determine the bit-error-rate (BER). Due to the randomized time slots, among 1160 data packets detected, 128 packets showed partial packet collisions, which checks with packet success rate simulation in Figure 3-6. Including the collided packets, we recovered data with an average BER of 1.82 %. The signal-to-noise ratio (SNR) and the BER for all 1160 packets are shown in Figure 3-3. As shown in the figure, the BER did not generally depend on the SNR, which is contradictory to the typical BPSK data link [67]. This is because the collision between packets was the dominant factor in the autonomous TDMA BER distribution, which resulted in the high BER in packets even with the high SNR.

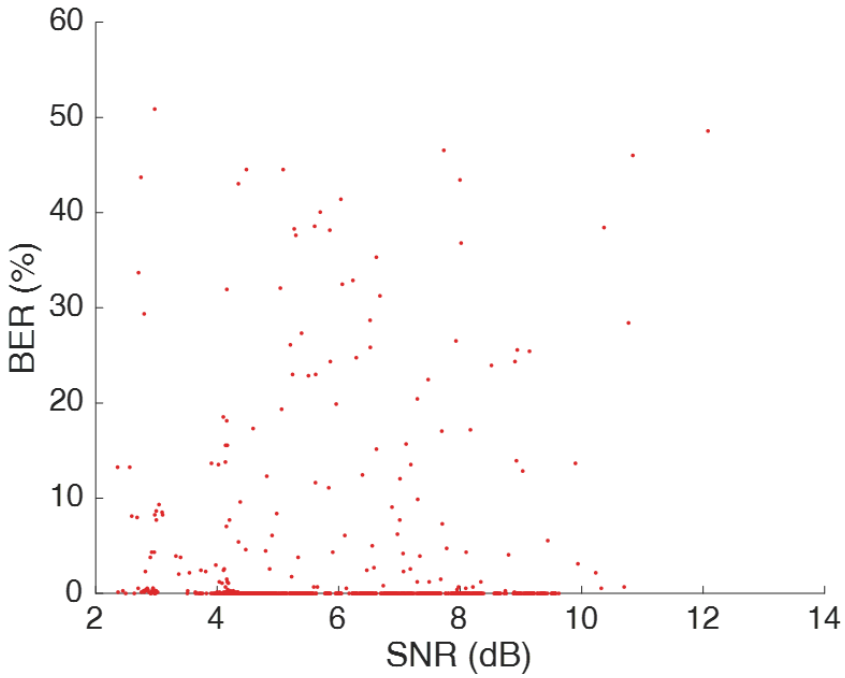

Figure 3-3. The Signal-to-Noise Ratio (SNR) and the Bit-Error-Rate (BER) of 1160 packets measured over 2 seconds from an ensemble of 64 autonomous TDMA chips. Partial packet collisions were observed from 128 packets. When including collided packets, we recovered data with an average BER of 1.82 %.

To examine the network's robustness over a long period, we collected data packets over 30 minutes from 64 chips. Since the current receiving path cannot support the continuous data sampling over 10 seconds, we subsampled 1800 backscattered packets and evaluated the packet BER and the clock frequency. The spread of their relative packet RF powers mapped the variable energy harvested as a function of the chip's spatial location. Packet RF power and recovered chip clock frequency are plotted in Figure 3-4, with each color representing an individual neurograin. These color labels were obtained by the k-mean clustering into 64 groups and showed a well-isolated, stable clock frequency over time. However, this hardware-efficient autonomous approach has network capacity for ~100 nodes, beyond which the possibility of packet collisions (as shown in Figure 3-6) increases for an ensemble of free-floating microimplants due to unscheduled TDMA. In the following section, we simulate the packet collisions in autonomous TDMA to analyze the effective bandwidth of the system.

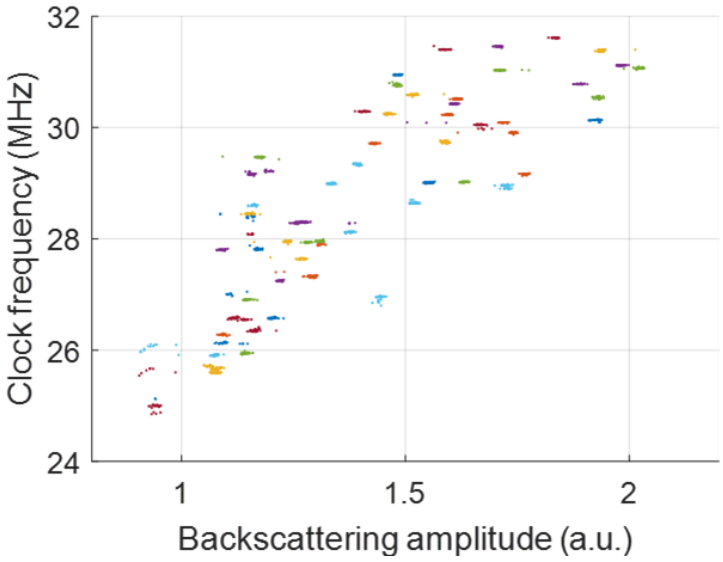

Figure 3-4. Recovered clock frequencies and backscattering amplitudes of 1800 data packets sampled for over 30 minutes from 64 chips.

3.2.2. Packet collisions in autonomous TDMA

Under autonomous TDMA, multinodal communication can be achieved while the "packet collisions", more than two chips transmitting packets simultaneously, can happen due to randomized time slots. One example of the packet collision is shown in Figure 3-5.

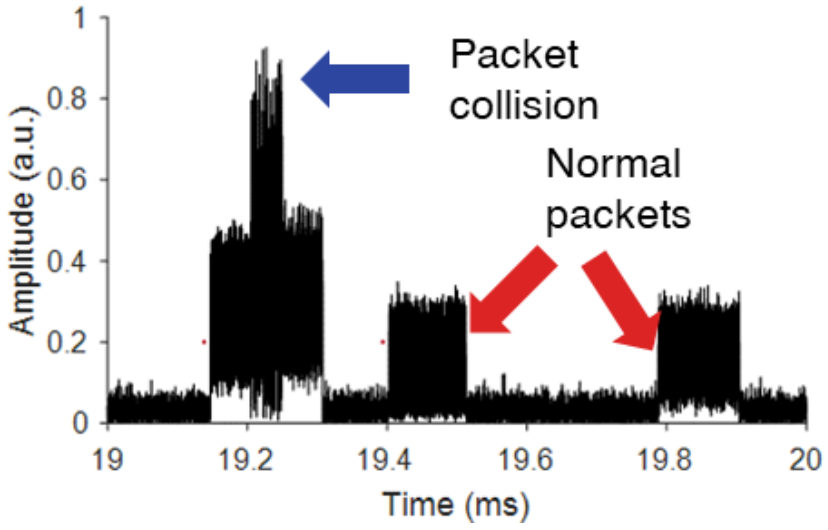

Figure 3-5. Appearance of a packet collision in the autonomous TDMA scheme. The transient waveform of received data showing uplink data packet collision from two neurograins operating in the autonomous TDMA mode. Collisions are caused by unscheduled backscattering events and chip clock frequency variations.

Autonomous TDMA chips backscatter in random time slots every 100 ms their unique 20-bits PUF address and the 800-bits neural data (total 820 bits). Given that the nominal backscattering time for 820 bits is 82 μs (10 Mbps), the chance of this data packet collision can be calculated as in Figure 3-6. The simulation shows that the packet success rate decreases as a function of the number of chips. In the case of 32 and 64 nodes operating in autonomous TDMA, we obtain a 95.28 % and 90.46 % packet success rate on average, respectively, when the clock variants among chips are included. Here, we note that any

partial overlap between packets has been counted as a packet failure. Thus the bit success rate would be expected to be much higher than the packet success rate. While autonomous TDMA circuits provide a compact and low-power approach for early ensemble testing in rodents, where available real estate for the implant is limited e.g., by a subjects' anatomy, it has limited scalability if one aspires for a higher number of nodes. For large-scale neurograin ensembles, we propose the "Call-and-Response" TDMA as a more sophisticated alternative for the communication protocol as follows.

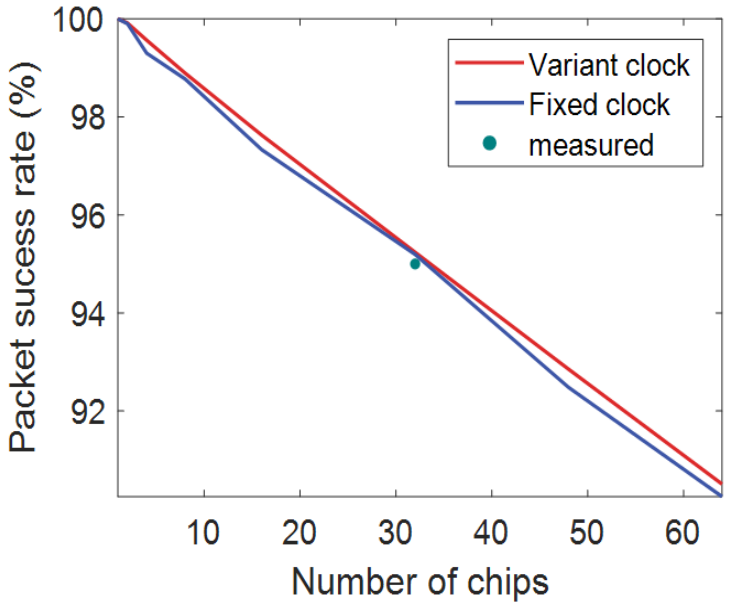

Figure 3-6. Simulation of the packet success rate assuming a randomized time slot for each chip. Here, any partial overlay between packets has been counted as a packet failure; a higher number of chips shows a lower packet success rate. For this simulation, 2 seconds of transient data was synthesized over 1000 repeats, and the averaged collision rate then calculated.

3.2.3. The call-and-response TDMA scheme

As noted, to avoid the inherent packet collision in the autonomous TDMA mode, we also investigated a "Call-and-Respond" TDMA scheme, where the ASK-PWM

protocol for downlink communication is integrated into the microdevices. The unique device identifiers (16-b PUF address), implemented on-chip, are discovered from the ensemble through an automatic start-up communication. The IDs are subsequently used to command individual devices through the ASK-PWM link. Each downlink command is received and downlink codes are deciphered across the ensemble in a synchronized manner. If the received downlink code matches a given device's address, it transmits a data packet ("respond").

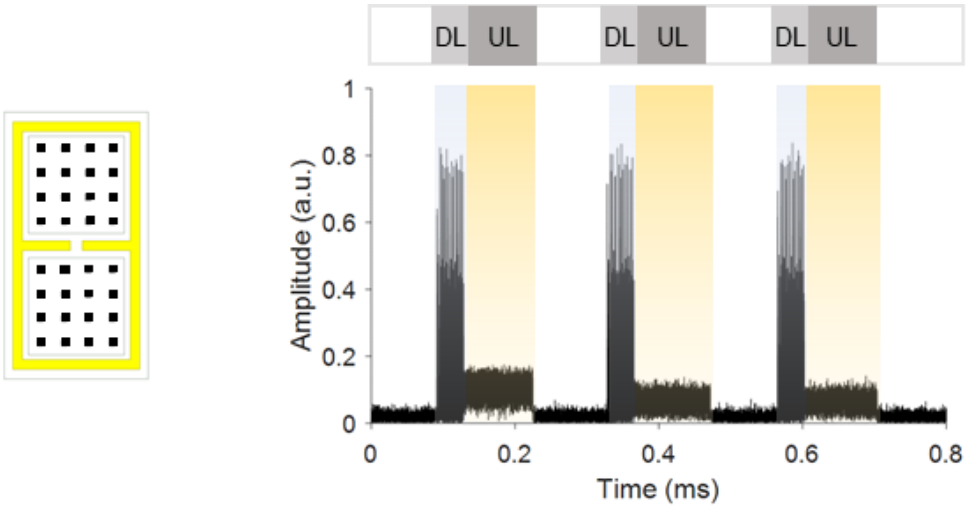

Figure 3-7. A 32-node call-and-respond networking within a relay coil. A transient recording on the downlink (DL) and uplink (UL) under "Call-and-Respond" TDMA.

Figure 3-7 shows the snapshot of the transient received signal at the external hub with the strong downlink call of 40 µs and the weak, uplink backscattered data of 100 µs for a network of 32 neurograins. Successful communications, measured by response rate to the downlink, occur with an average efficacy of 94.7% across this network. Three devices showed a response rate lower than 90% (Figure 3-8), possibly due to CMOS and dicing process variations. We tested all available chips arriving from the foundry and did not perform any prior screening/selection.

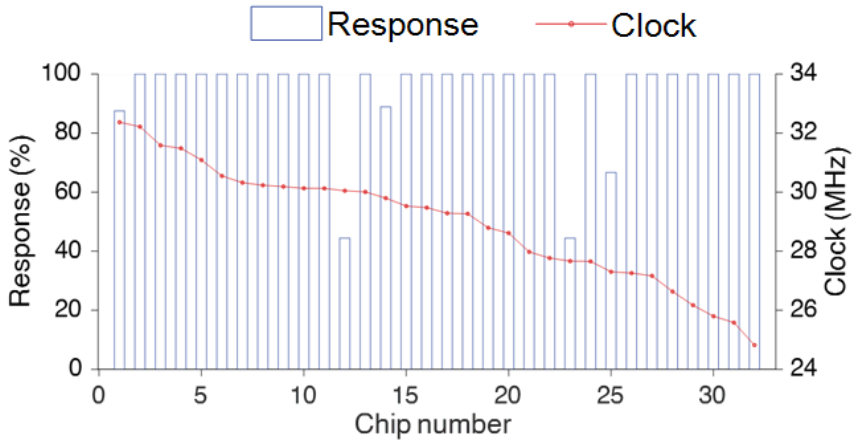

Figure 3-8. The ASK-PWM downlink response rate and the averaged clock frequencies of an ensemble of 32 call-and-respond TDMA chips.

The uplinked 967 data packets from these call-and-response chips, comprising 1000-bit predefined sequences. When comparing demodulated bit sequence with predefined bits, we could achieve an average BER of 0.004 % across all 32 chips in the total of 947 data packets and mean BERs for the individual chip are shown in Figure 3-9.

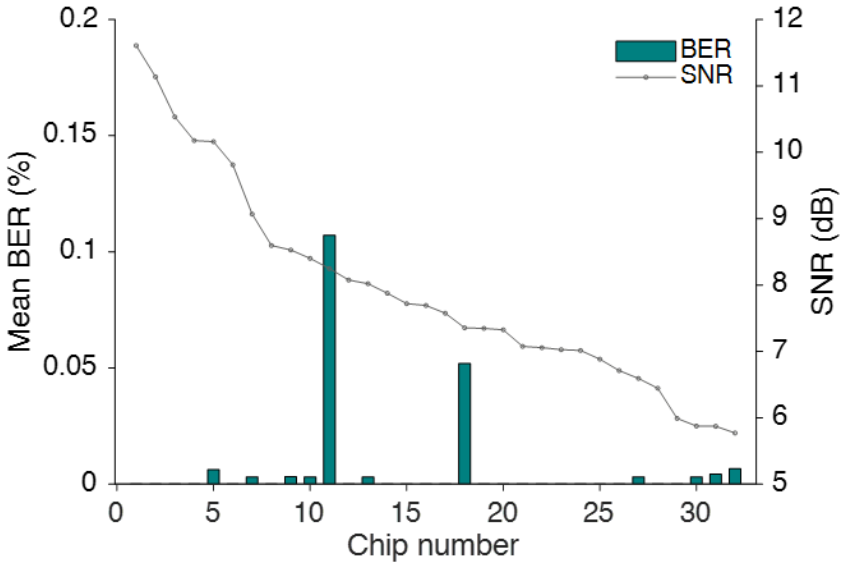

Figure 3-9. The simulated averaged BER from 1000 uplinked bits versus the SNR of 947 data packets from 32 chips each individually identified by its PUF address.

Figure 3-10 shows the BER and the clock frequency of 947 packets from 32 chips after sorting them by their PUF addresses. The on-chip clock did not show any significant changes over time and bit-errors occurred in a few data packets. This data suggests that the chip power level was stable during this period. Also, the bit-error was not observed across chips simultaneously, which suggests that the occasional bit-errors did not come from ambient transmitting energy fluctuations. However, bit-error mostly occurred in a single bit and an average BER of 0.004 %; thus, we concluded that this range of BER is acceptable for the neurograin interface.

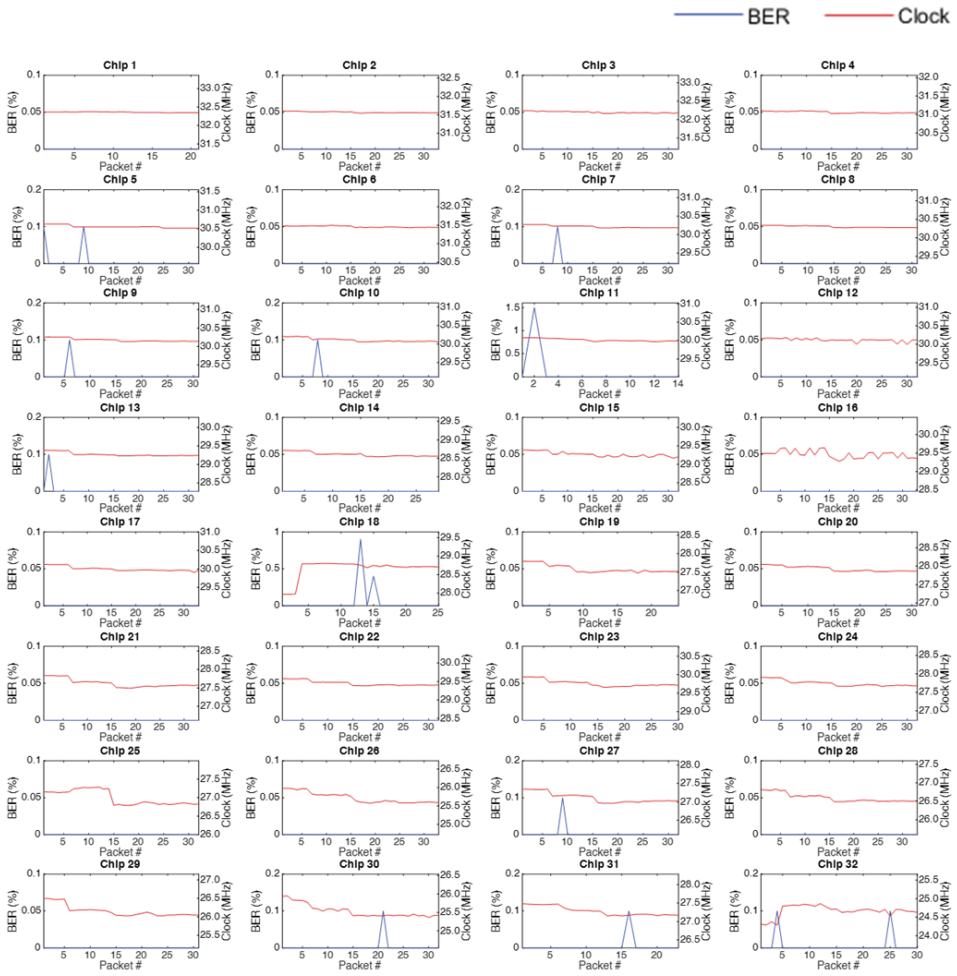

Figure 3-10. The BER and the clock frequency of 947 packets from 32 chips sorted by their PUF addresses and the predefined transmission sequence.

For the current exploratory version of call-and-response TDMA neurograin, with 16-bit PUF addresses, a downlink rate of 1 Mbps, and Manchester-coding to ensure a data stream, each downlink trigger frame takes up to 40 us. Figure 3-11a,b shows the transient RF amplitude on 32 chips call-and-response TDMA over 66 ms, which shows a strong downlink leakage and the following uplink packet from chips. To simulate a larger number of chips, the 32 sequences of the downlink were transmitted in a loop and the downlink assigned 281 timeslots during 66 ms.

This result shows that, under the call-and-response scheme, we can communicate with up to 425 neurograin within 100 ms, even with the present 230 μs gap between the downlinks. Figure 3-11c also shows the histogram for the duration of one cycle of the uplink and the downlink telemetry. According to this data, the downlink can be transmitted every 170 μs so as to communicate with up to 588 chips without any circuit modification. However, in the present version, one cycle of the downlink and uplink has a 145 μs duration, on average, given that the current version of ASIC transmits additional bit sequences (1024 bits) for the communication link test (such as repeating 31 LFSR sequence and zero paddings), which thus uses bandwidth less efficiently. For more efficient data communication, one might choose to transmit the 20-bit PUF address through the downlink (20 μs under the 1 Mbps) to enable 800 bits transmitted through uplink during 80 μs (10 Mbps). Even after adding a time gap between downlinks to accommodate the neurograin clock variance, the time slot for each chip can be reduced down to 130 us allowing communication with 770 chips in this call-and-response scheme. Note that the call-and-response approach also provides a way to down-select chip nodes which may contain the most meaningful neural information, paving a way to use this type of bidirectional

communication approach for future 'adaptive sensing' to reduce the burden on data processing and to reduce the system latency further.

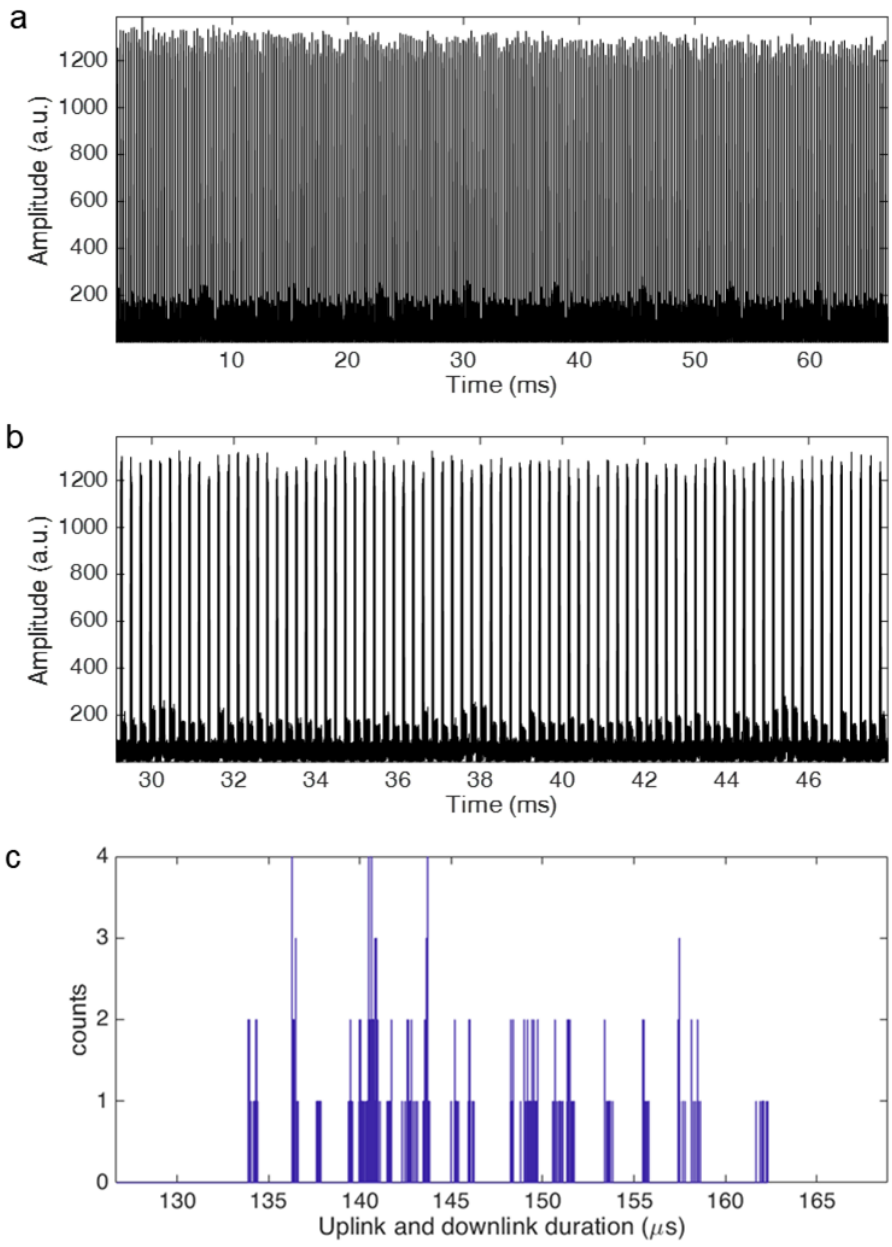

Figure 3-11. Demonstration of call-and-response TDMA. a & b. The transient waveform in a receiving channel with 32 neurograins shows a large amplitude leakage of downlink and the relatively weak uplink (data packet). Within a 66 ms duration, the equivalent of 281 cycles of the downlink and uplink were achieved, allowing in this example up to 425 neurograins to be multiplexed within 100 ms. **c**. The duration of the single downlink command and the immediately following uplink data packet.

3.3. The external RF wireless hub for data communication with the neurograin microimplants

3.3.1. Software-Defined Radio (SDR) in the RF hub

The RF wireless hub that communicates with neurograin needs to adapt to the customized uni- or bi-directional network communication for ensembles of neurograins [68]. Commercially available RF transceivers such as Wi-Fi, Bluetooth, Zigbee, cannot offer the required flexibility for our specialized custom RF data communication, which necessitates the use of a specialized software-defined radio (SDR) at the current development stage of the project. The SDRs used in this work (Raptor, Rincon Research Corporation or FMCOMMS3, Analog Devices) can operate across a wide range of frequencies (70 MHz- 6 GHz) and support up to 60 MSPS data sampling. Importantly, the SDRs have on-board processors (Quad-core ARM Cortex) and FPGA (Xilinx Zynq Ultra Scale+ XCZU9EG-1FFVC900E). By programing a C-code for the processer, we could use the available open library to interface with IIO devices (libiio) to handle the Rx data sampling and the Tx data transmission on the transceiver chip (AD9361, Analog Devices) on the board. We allocated most of the processor RAM space (a total of 4 GB) for the continuous memory allocator (CMA), in order to temporally store all sampled data from the RF front end. Since the chosen sampling rate is nominally 30-45 MSPS in the neurograin system, the RF hub can store data for 10 seconds or less. For the downlink, we can also pre-load the transmitting data on the CMA to send it at 1 MSPS through the 915 MHz transmitting path.

When sampling the 945 MHz-centered uplink packets retrieved from neurograins, the SDR receiving front-end performs the signal amplification, down-conversion (from 945 MHz to DC), and digitization. The digitized IQ data was then ported to a PC for offline demodulation processing (in MATLAB/Simulink) through LAN cable. Technically, one can import the demodulation codes directly onto the SDR board, which can accelerate the process by using the on-board FPGA for real-time demodulation.

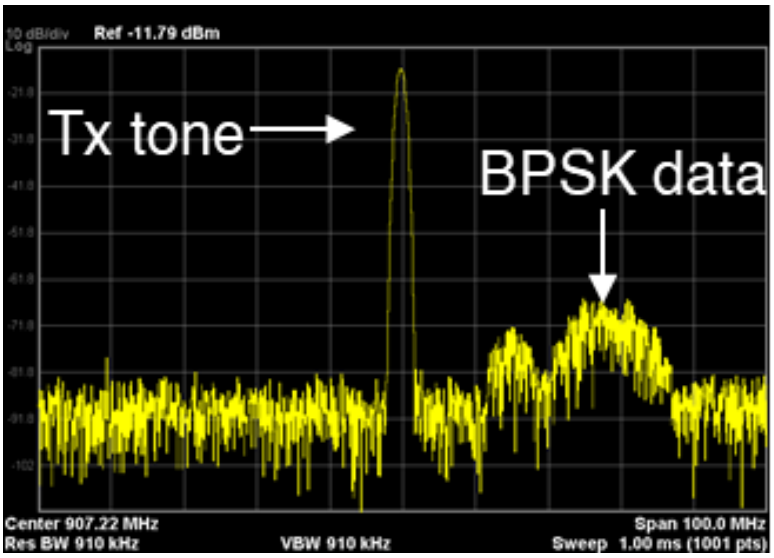

Figure 3-12. At the SDR: the full RF spectrum showing the baseband (Tx) tone, here near 907 MHz and the BPSK-modulated backscattered data lobe of the receiving channel, here centered around 937 MHz.

3.3.2. Building the external RF hub for completing the neurograin interface

The external RF hub also comprises other RF components, an amplifier and a duplexer for the RF signal processing in addition to SDR. The amplifier was used to boost the transmitting signal from the SDR and the duplexer isolated the BPSK uplink signals from the baseband Tx tone shown in Figure 3-12. Figure 3-13 shows the overall RF hub construct with these three principal components, including the RF amplifier, the SAW

duplexer and the SDR. Here, ADL5605-EVALZ (Analog Devices) board was used due to its high amplification gain (up to 23 dB) and low noise figure (4.8 dB). Also, P1dB (the 1 dB compression point) of this board was 30.9 dBm; thus, the amplifier could provide sufficient transmitting energy up to 1 W in all neurograin experiments. The RF SAW duplexer (D5DA942M5K2S2, Taiyo Yuden), used to isolate the receiving signals (945 MHz) from the transmitting tone (915 MHz), has acceptably high isolation of 58 dB between two bands. Since the average backscattered power from neurograin chips is only around -45 dBm whereas the transmitting base band tone can be high as 30 dBm, this degree of isolation is essential for viable neurograin communication. The resultant spectrum in Figure 3-12 is obtained within the receiving path immediately after the duplexer but still shows a strong leakage of baseband Tx tone. We also tried to replace the SAW filter by using the tunable RF notch filter and circulator for the uplink isolation. However, the noise level here was much higher and the rejection of the baseband Tx tone insufficient.

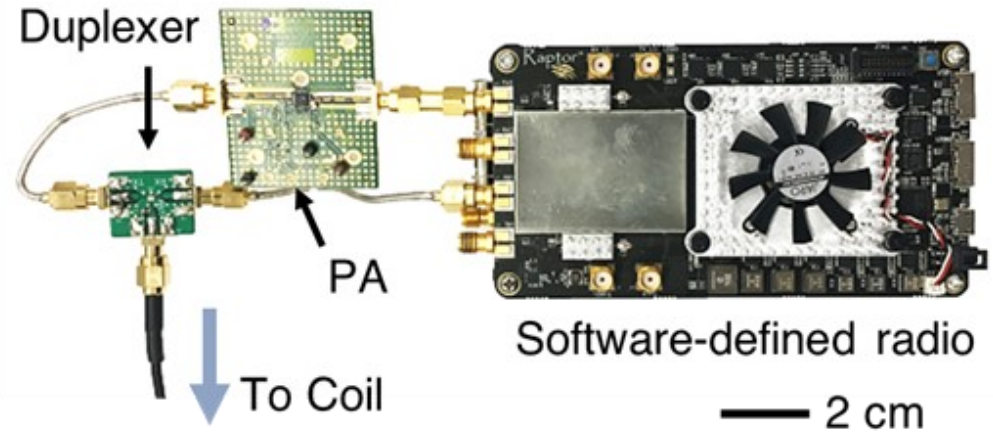

Figure 3-13. Photograph of the benchtop implementation of the external RF telecom hub consisting of software-defined radio (SDR), a power amplifier (PA), and a duplexer for isolation.

3.3.3. BPSK demodulation for digital bit recovery

BPSK uplink signals backscattered from neurograins were demodulated to recover the digital bit data on the chips. The BPSK bit demodulation process is depicted in Figure 3-14. The demodulation sequence starts with the data packet detection in MATLAB. For the packet detection, the digitized IQ data from the SDR was filtered by a window-based FIR filter with a 10 MHz high cutoff frequency so that the residual Tx tone can be filtered again in the digital domain. Then, we performed the automatic gain control to normalize the packet amplitude and coarse and fine frequency compensation to cancel the clock frequency offset. Since the clock frequency on-chip was used for the up-conversion of 915 MHz signals, the clock variance among the neurograin population subsequently resulted in the variant carrier frequency on backscattering. Thus, this carrier frequency offset in neurograin can range from -3 MHz to +3 MHz following the clock variation, which far exceeds the typical offset range in RF communication. Since the clock frequency also drives neurograin on-chip digital engine, it also determines bit data rate. Therefore, the recovered clock frequency needs to be utilized in the timing recovery process to improve the precision of bit recovery.

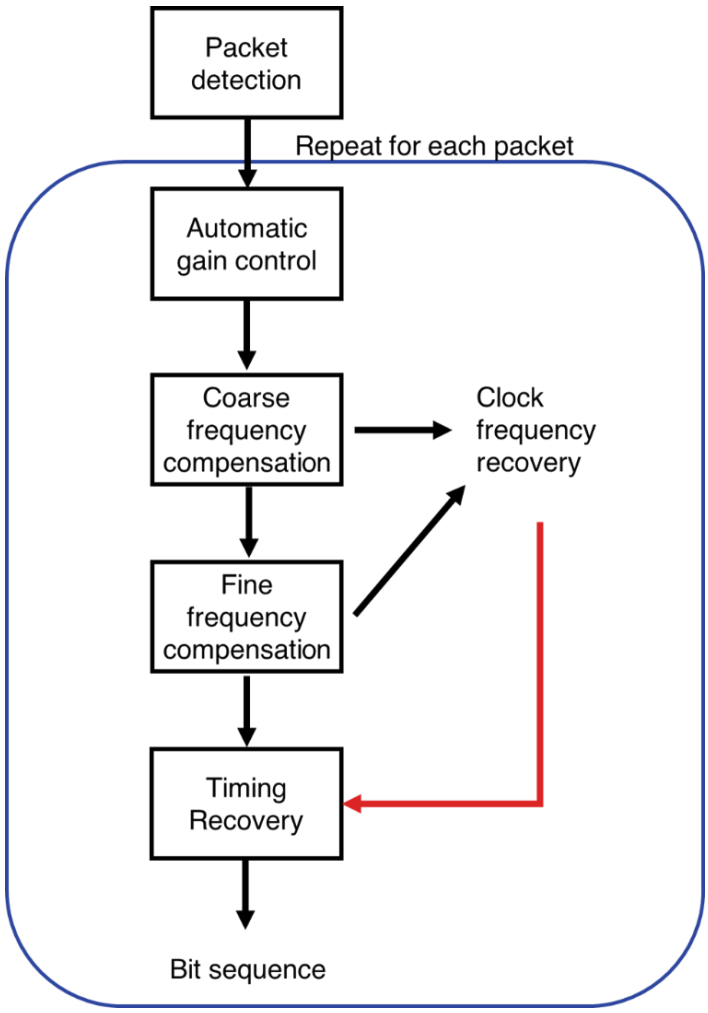

Figure 3-14. The BPSK Demodulation process implemented via MATLAB for bit recovery.

3.4. Conclusion

We have implemented and demonstrated one-to-many communication under two proposed TDMA schemes, autonomous and call-and-response TDMA, for the network of neurograin ensembles. Autonomous TDMA is a simpler and compact approach to achieve networking in a smaller number of nodes, but the packet collision probability resulted in high bit-errors for larger neurograin populations, probably limiting the adequate channel size in practice to ~100 nodes in a brain-machine application. The Call-and-response

TDMA requires a more complex circuit architecture for bidirectional communication but can offer more scalable data links. In our exploratory investigation on the call-and-response TDMA scheme, we have found that the RF hub can communicate with up to 425 microdevices within a 100 ms period. And, theoretically, this data link can be scaled furthermore to support up to 770 microdevices by incorporating a shorter "call" downlink (ongoing research), being the likely eventual choice for truly large-scale multi-node communication for distributed microsensor networks.

CHAPTER 4. WIRELESS POWERING OF NEUROGRAIN ENSEMBLES

4.1. Introduction

The material in this chapter is adapted from the paper, 'Wireless power and data link for ensembles of sub-mm scale implantable sensors near 1GHz', *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2018.

As noted in Chapter 1, wireless powering on active neural implants presents many obstacles regardless of the energy modality used (e.g., ultrasound, RF, optical, etc.). It is important to recognize that safety regulations impose strict limits to the maximum power to the external sources. The challenge the author faced was to design an inductively coupled (near-field) electromagnetic approach, beginning with the microantenna configuration integrated into the neurograin silicon chips. The design requires that a population of spatially distributed neurograins is inductively coupled with an external transmitting coil in the RF hub so that ensembles of microchips can be efficiently RF powered from the external power and signal source. Note that the wireless link serves the dual purpose of

simultaneously providing the means for bidirectional data transfer. In the present generation of sub-mm size neurograins the on-chip microcoil has a small cross-sectional area of 500 μm × 500 μm which intrinsically limits the transcutaneous wireless transfer efficiency (WPT) from a single transmitter antenna external to the body [40]. Aiming to develop an optimally efficient wireless powering system for neurograin, we began by comparing two electromagnetic schemes based on a 2-coil WPT and a 3-coil resonant WPT approaches and pursued a systematic analysis of the frequency-dependent characteristics of the system. Based on this analysis, we focused on the development of a 3-coil near field scheme, adding a relay antenna as the third coil to enhance the energy-transfer, similar to those explored in several other studies for miniature wireless implants [40, 69-71]. The one specific geometry and innovation reported here incorporates a characteristic "quadrant shape" WPT coil to maximize wireless transfer across a targeted area in the cortex. The following sections describe these design processes and how we quantified the 3-coil WPT system using functional neurograin chips.

4.2. Approach to wireless electromagnetics for powering and bidirectional data exchange at the neurograin interface

4.2.1. Comparison between the resonant 2-coil and 3-coil WPT configurations

In a 2-coil resonant inductively coupled WPT approach shown in Figure 4-1 left, the efficiency of wireless power delivery (η) between the Tx coil and Rx coil can be described by

$$\eta_{2-coil} = \frac{k_{12}^2 Q_1 Q_{2L}}{1 + k_{12}^2 Q_1 Q_{2L}}$$ Equation 4-1

where k_{12} is the coupling factor between the Tx and Rx coil and $Q_1 = \omega_0 L_1/R_1$ and $Q_{2L} = Q_2/(Q_2 Q_L + 1)$, where $Q_2 = \omega_0 L_2/R_2$ and $Q_L = R_L/\omega_0 L_2$ [72, 73]. Here, due to load transform from the resonant capacitor C_2, R_L can be solved as $\omega_0^2 L_2^2/R_c$ and R_c is the circuit equivalent impedance [40]. k_{12} depends on the size of the coils and the distance and arrangement between the two coils. In the case of on-chip micro-coils implanted, e.g., on the surface of the cortex, the size mismatch between Tx and Rx antennas is significant (where the latter are ~0.25 mm^2 in our system) so that, given the separation between two coils exceeding several millimeters, the maximal coupling k_{12} is severely limited.

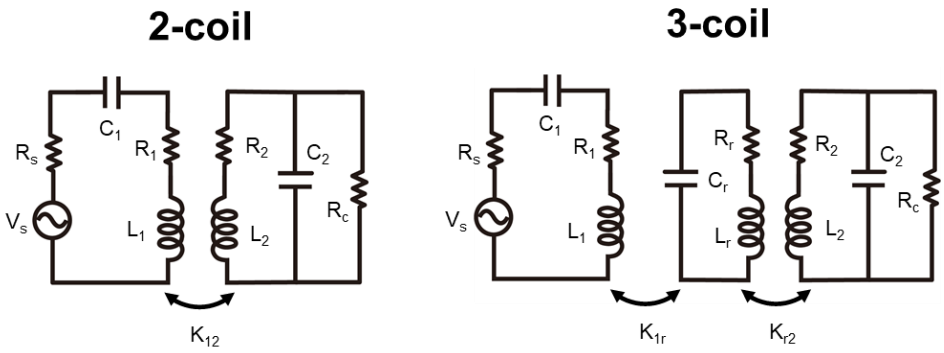

Figure 4-1. Comparison between the 2-coil vs. 3-coil WPT configurations. (left) Lumped circuit model for the 2-coil WPT system. (right) The 3-coil WPT model.

In the 3-coil WPT configuration, a secondary coil can be introduced to have moderate coupling both with Tx and Rx coils, thus increasing the overall efficiency of WPT (Figure 4-1 right). The coupling factor between the relay and Rx coil, k_{R2}, can be ten times larger than k_{12} in the co-planar arrangement and the efficiency in 3-coil WPT can be simply written as

$$\eta_{3-coil} = \frac{(k_{1R}^2 Q_1 Q_R)(k_{R2}^2 Q_R Q_{2L})}{(1+k_{1R}^2 Q_1 Q_R + k_{R2}^2 Q_R Q_{2L})(1+k_{R2}^2 Q_R Q_{2L})}$$

$$= \left[\frac{(k_{1R}^2 Q_1 Q_R)}{(1+k_{1R}^2 Q_1 Q_R + k_{R2}^2 Q_R Q_{2L})}\right] \cdot \left[\frac{(k_{R2}^2 Q_R Q_{2L})}{(1+k_{R2}^2 Q_R Q_{2L})}\right] = \eta_{1R} \cdot \eta_{R2}$$

Equation 4-2

where k_{1R} is the coupling factor between the Tx and the relay coil, and k_{R2} is the coupling factor between the relay and Rx coil [72]. Q_R is the unloaded quality factor of the relay coil. Using a secondary coil that has a size comparable to the Tx coil, k_{1R} can be increased significantly, and a high η_{1R} can be achieved. Also, co-planar placement of the relay and Rx coils can improve k_{R2}, which results in higher η_{R2}. Note that since the relay coil is not directly loaded by source or load impedance, η_{1R} and η_{R2} benefits from a high-quality factor for the relay coil Q_R. Consequently, the relay coil collects high magnetic energy to the target area, improving power transfer efficiency for the ensemble sub-mm Rx coils. Based on simulated coil parameters Table 4-1 and equation 4-1, 4-2, the maximum and minimum efficiency within the relay coil are calculated in the (proposed) primate and (implemented) rodent model coil for both cases of 2-coil WPT configuration and 3-coil WPT configuration as shown in Figure 4-2.

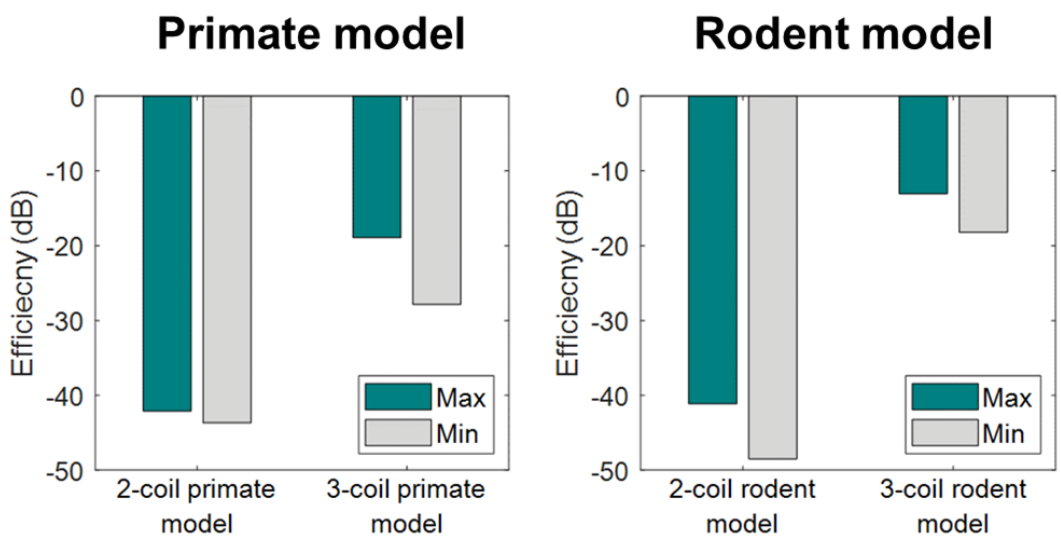

Figure 4-2. Efficiency comparison of the resonant 2-coil vs. 3-coil WPT configurations for the (proposed) primate and (implemented) rodent configurations, using Equation 4-1 and 4-2 with coil parameters acquired from Table 4-1. Vs: source voltage, Rs: source impedance, Rc: circuit impedance.

4.2.2. Rational for selecting the near 1 GHz baseband frequency

We chose the 1 GHz baseband RF range as optimal between several competing frequency-dependent factors. As the RF frequency is lowered, the achievable quality factor on the microchip Rx coil is limited, as shown in Figure 4-3, while the efficiency between Tx to relay coil can improve. At higher RF frequencies, tissue absorption losses increase, and the quality factors of the Tx and relay coils can decrease due to self-resonance effects [74, 75]. An HFSS simulation suggests an optimal frequency of the coil WPT system as 377 MHz assuming 1 kOhm circuit load impedance. However, in addition to the tight trade-off among the coil couplings, following RF circuit design can play a critical role in this type of wireless system. In a low-frequency RF band, a matching capacitor and BSPK modulator need to be scaled much larger in the footprint, which already occupies 23 % of the total at 915 MHz. Lumping the multiple factors together, we chose 915 MHz as the

baseband frequency, which could also benefit from the availability of mature wireless technologies. The near 1 GHz frequency region also allows the increase of available telecommunication bandwidth, which is important in scaling the system whether to, say, 770 neurograins (case example in this work) or beyond.

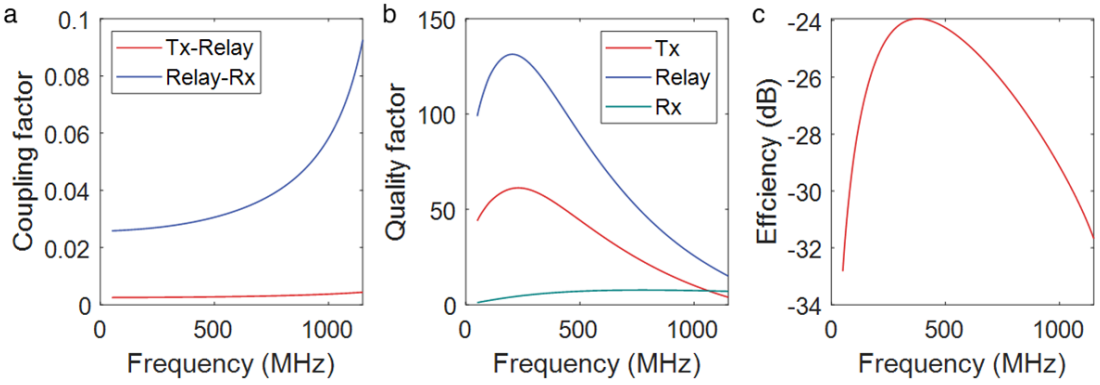

Figure 4-3. Simulated frequency dependence for the 3-coil WPT system for (proposed) primate use, with 8 mm transcutaneous coil separation. a. Coupling factors between Tx and relay coil and between the relay and Rx coil. The "Rx coil" refers to the miniature on-chip coil on each neurograin. **b**. Quality factors for the Tx, relay and Rx coils, respectively. **c**. Total efficiency calculated using simulated parameters and the equivalent circuit model described in the Method section. For this simulation, the skin and bone thickness were 4 mm for each.

4.2.3. Electromagnetic simulations

We utilized a three-coil WPT architecture for inductive coupling to optimize the efficiency of electromagnetic data and power transfer between the external hub and ensembles of spatially distributed neurograin implants. Design constraints for this coil WPT system, comprising the external Tx coil, the implanted Tx-size-matched relay coil and the on-chip receiver (Rx) microcoils, are informed by high-performance electromagnetic simulations. In the electromagnetic simulation, we modeled dielectric tissue layer, polyimide-based PCB coils, and packaging materials such as

Polydimethylsiloxane (PDMS) or liquid crystal polymer (LCP) since we aimed to simulate realistic coil parameters. The Tx and relay coils were both designed in a geometry that can be viewed as a de facto superposition of unit sub-coils, capable of area extension for tailored areal coverage while maintaining WPT efficiency. Leveraging the same underlying electromagnetic principles, we were able to customize the design of the Tx/Relay coil to tailor different versions of the system to optimally accommodate the anatomical structure of both future primate experiments and the present rodent model, as shown in Figure 4-4 (dimensions and simulation parameters shown in Table 4-1).

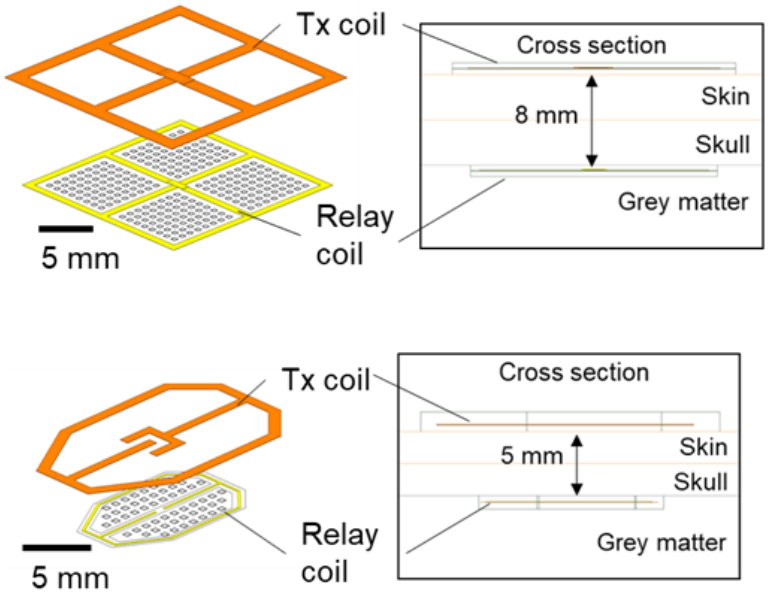

Figure 4-4. The 3-coil WPT system for neurograin. (top) The wireless power transfer (WPT) system proposed for a primate model consists of the transmitting (Tx) coil, the relay coil, and the receiver (Rx) micro coils targeting 8 mm separation (4 mm skin and 4 mm skull). (bottom) Coils for the rodent model targeting 5 mm separation (2.5 mm skin and 2.5 mm skull).

Based on the salient anatomical structures, the primate model coil was designed to transmit wireless energy to the relay coil and neurograin in 8 mm separation. The geometry

chosen here allows wide-field WPT, which can potentially transfer RF energy up to 1000 neurograins. In the rodent model coil, the target wireless powering distance was set to 5 mm considering the relatively thin scalp and skull of the rodent; thus, much higher efficiency could be achieved. But the rodent skull size limited the size of the overall WPT coil structure allowing only up to 64 neurograins placed within the rodent model coil peripheral. On the electromagnetic simulation, we tested 64 and 26 locations within the relay coil for the primate model and the rodent model, respectively, and the simulated result on mapped efficiency is shown in Figure 4-5.

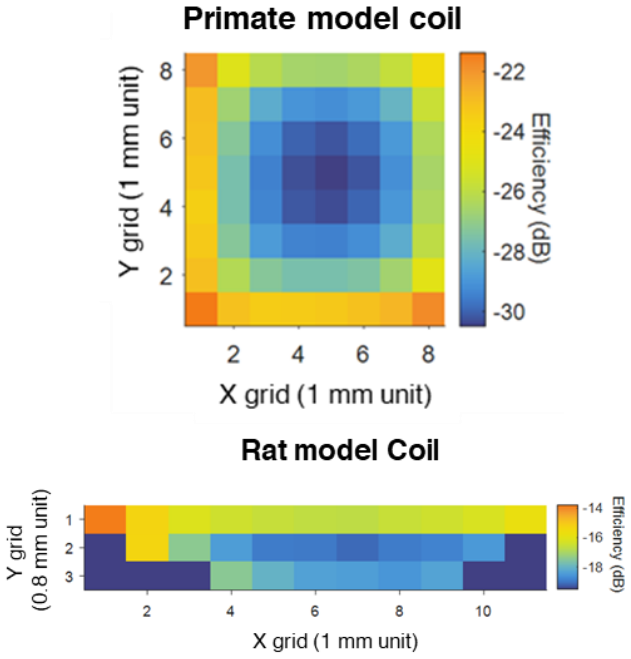

Figure 4-5. Simulated wireless efficiency in various Rx coil locations across the transverse plane in the primate model-based coil (top) and the rodent model coil (bottom).

In the electromagnetic simulation, the WPT was seen to be much more efficient around the boundaries of coils due to a stronger magnetic coupling between the relay coil and the Rx coil while, at the center of each sub-coil, ~ 10 dB less efficiency was achievable

compared to the maximum. These results show that the ASIC circuit on each neurograin chip needs to adapt to a 10 dB difference in wireless transfer efficiency within the coil in order to ensure its operation regardless of the position within the relay coil. The previous result shown above in Figure 2-3 presents that neurograins need to collect at least -10 dBm power for their operation. And 20 dBm Tx power should provide sufficient energy across the relay coil in the primate model coil. In the rodent model coil, much less transmitting power was needed due to the higher wireless transmission efficiency than the primate model coil.

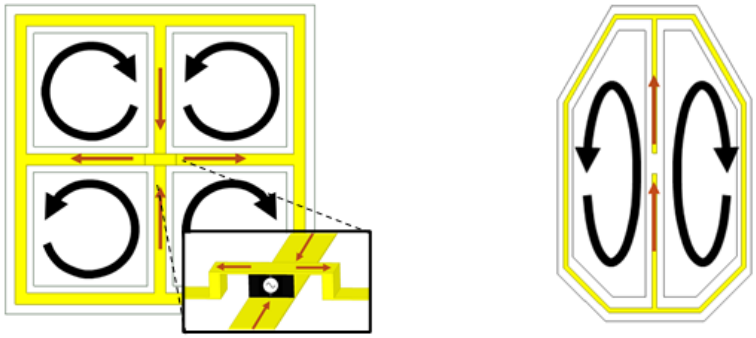

Figure 4-6. In this "window" relay coil geometry, four or two sub-coils are combined to realize specific current flow direction for the (proposed) primate and the (implemented) rodent models, respectively, to expand the effective power harvesting area for scalability to cover a wider cortical space.

The Tx and relay coils have a specialized geometry comprising two or four sub-coils connected in parallel. This parallel configuration of sub-coils decreases the overall inductance as well as the resistance of the coils, resulting in the same quality factor. On top of that, the overall coupling coefficient between the Tx and relay coils is equivalent to that between each sub-coil. As a result, this parallel configuration does not affect the system's overall efficiency while allowing the expansion of the wireless powering area. Figure 4-6

shows an example of simulated current flows in the Tx coil in our 3-coil WPT system. In the Tx coil for the primate model, the transmitting RF port can be attached in the middle of the coil, as shown in the expanded sublet plot of Figure 4-6. This 3D configuration drives the current in a specific direction so that the respective directions of magnetic flux of adjacent subunit coils are opposite to each other, as shown in Figure 4-7.

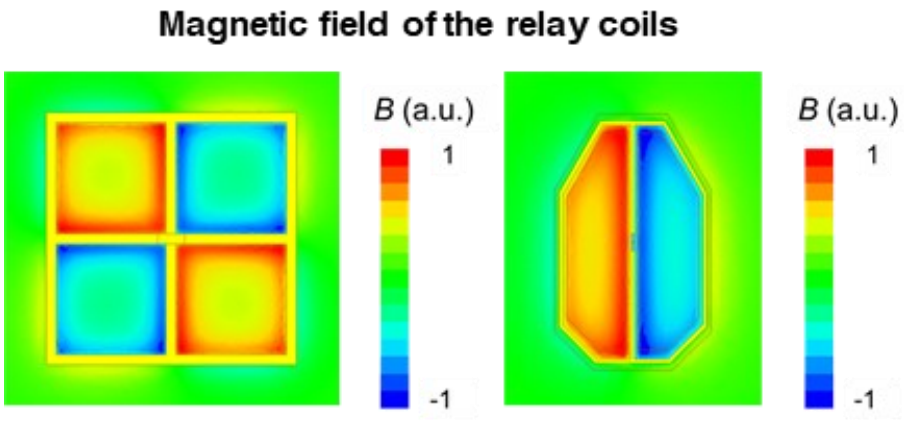

Figure 4-7. The magnetic flux generated by the specific relay coil designs shows the alternating direction of fluxes within the subunit coils. (the direction of the flux is shown as red or blue).

The on-chip micro antenna design (Rx) is optimized as a miniature Rx coil using the thick aluminum top metal layer within our semiconductor foundry process. The Rx coil design was optimized to be 3-turn since, in a lower number of turns, coupling to the high load impedance gets less efficient, while for the higher number of turns, the self-resonance decreases the quality factor of microantenna. The on-chip microcoil has an outer diameter of 500 μm and is able to accumulate magnetic flux with an efficiency compatible with powering on-chip operations of the high-load/low-power ASIC at 1 GHz. S-parameters from HFSS simulation results were imported on Cadence Design Systems Mixed-Signal Simulator to set up a full RF simulation test bench for wireless circuit design.

Type	Primate model coil	Rodent model coil
L_1 (nH)	12.2	20.9
L_R (nH)	11	16
L_2 (nH)	10.3	
Q_1	26.9	50
Q_R	30.8	109.8
Q_2	13.4	
k_{1R}	0.051	0.02
k_{R2}	0.0035-0.0099	0.008-0.015
k_{12}	0.0005-0.0006	0.0003-0.0007
Size Tx (mm)	22.5 x 22.5	20 x 12.3
Size Relay (mm)	20.4 x 20.4	14.2 x 8
Size Rx (mm)	0.5 x 0.5 (3 turns)	
Tissue	4 mm skin, 4 mm skull, >20 mm cortex	2.5 mm skin, 2.5 mm skull, >20 mm cortex

Table 4-1. Dimensions and (simulation) parameters of the resonant 3- coil system at 915 MHz

4.3. Measurements to characterize the Wireless Power Transfer for Optimal Performance for Neurograin Ensembles

4.3.1. Measurement of wireless power transfer

One way to quantitatively analyze the electromagnetics of the wireless link in-situ, without resorting to the confounding wired measurements due to external probe size, etc., is to measure the clock frequency of the chip, which serves as an estimate of the power level received by the chip, as per the simulated frequency-power relationship in Figure 2-3. We have used a liquid-form head phantom construct as a proxy for a physiologically relevant RF environment and analyzed the effect of tissue path loss and surrounding tissue permittivity on the WPT and data link [76]. We mapped the WPT efficiency at 64 spatial positions inside a primate model relay coil using a location sampling grid with an x-y pitch

of 1 mm (Figure 4-8 top) and in a rodent model coil at 26 positions with an x-y pitch of 1 mm and 0.8 mm, respectively (Figure 4-8 bottom). Based on the comparative anatomical characteristics, the separation between the Tx coil to relay coil was 8 and 5 mm in the primate and rodent model coil, respectively.

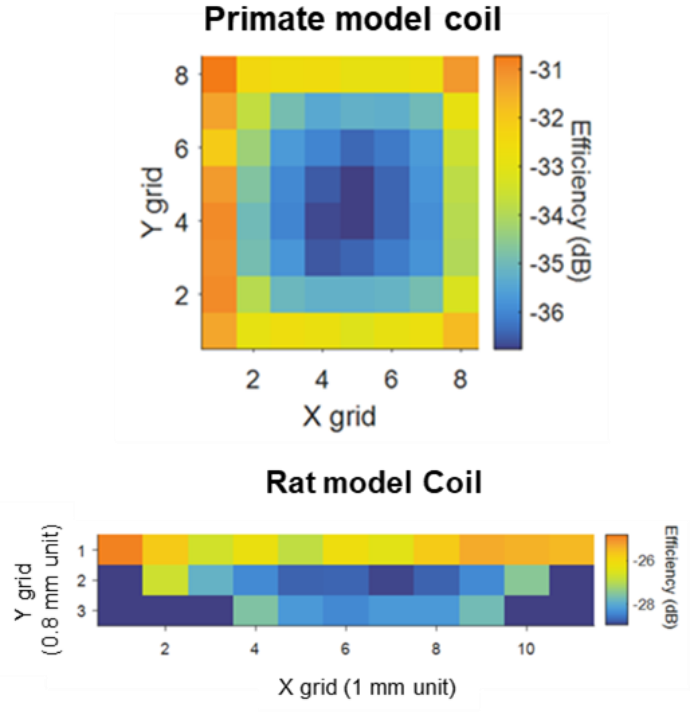

Figure 4-8. Wireless efficiency measured for 64 and 26 Rx coil locations in the primate model coil configuration (top) and the rodent model coil configuration (bottom), respectively.

As expected by the simulated results, the Rx coil achieved its highest efficiency when close spatial proximity to the relay coil perimeter and coupling decreased when moving toward the center of each subunit coil (Figure 4-8 top) in electromagnetic simulation. In the primate model coil, the WPT efficiency ranged from -36.77 dB to -30.73 dB (center to edge). In the rodent model coil, the maximum efficiency of -24.88 dB was located at the corners, decreasing to -28.90 dB at the center.

Received power at the microimplant also depended on the horizontal and angular alignment of Tx coils, as shown in Figure 4-9 a,b. For the prospective primate model, lateral movement (x-y plane) of the Tx coil resulted in a loss of 1.3 dB for a 3 mm horizontal displacement. For the asymmetric coil morphology used in the in-vivo rodent experiments, a y-directional misalignment of 3 mm resulted in a 0.6 dB penalty. In contrast, an x-directional misalignment resulted in a much larger 6.73 dB loss, which is expected given the coil aspect ratios. A 20-degree tilt resulted in 1.29 dB vs. 1.9 dB loss in WPT efficiency in the primate and rodent configurations, respectively. In both these simulations and measurements, a given neurograin was in the center of the subunit coil. In the actual operation of the neurograin system, we used the strength of the backscattered RF signal as the guide to optimally align the Tx coil with submillimeter accuracy.

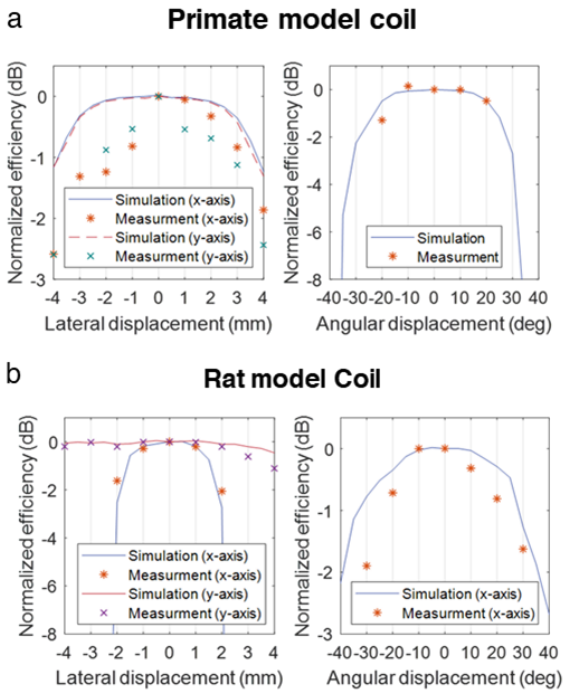

Figure 4-9. WPT efficiency as a function of the lateral and angular displacement of the external Tx power coil for the primate (a) and the rodent models (b), respectively. Here a neurograin was located in the center of the subunit coil.

Array efficiency changes in the lateral and angular displacement of the Tx coil is shown in Figure 4-10. Here, the wireless efficiencies in 36 neurograin were solved simultaneously in the primate model coil to simulate the multinodal neurograin network. The efficiency changes did not show a significant chip-to-chip variation across the neurograin population in the lateral and angular displacement of the Tx coil. This result shows that only the coupling between the Tx to relay coil was affected by the Tx displacement and the field distribution within the relay coil did not change considerably.

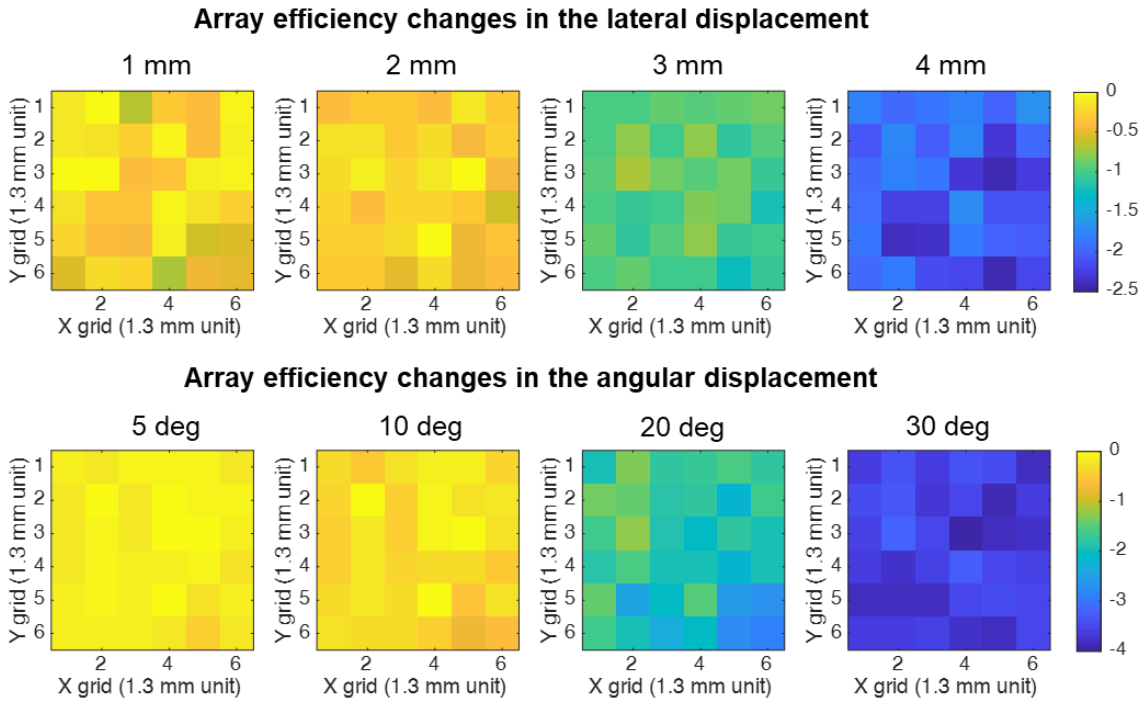

Figure 4-10. Simulation to show WPT efficiency changes across a 36 neurograin array and dependence on the lateral and angular displacement of the coil (primate coil model). Up to -2.38 dB and -3.86 dB loss is observed for 4 mm lateral and 30 deg angular displacement, respectively. The associated variance of efficiency on neurograin WPT was not significant as the displacement mainly affects the coupling between the Tx and the relay coil.

4.3.2. Extending the WPT concept to multiple Tx power sources

The work so far had emphasized wireless powering using a single Tx coil and single relay coil, similar to previous studies which deployed the 3-coil WPT system [41, 72]. Extending the concept to multiple pairs of Tx/relay coils can provide access to wider cortical coverage and multiple cortical areas. As such, the magnetic field by two adjacent coils can interact destructively, which can significantly impair wireless harvesting/communication. One remedy is to fine-tune the phase of coils to have a precisely 180-degree phase difference, thereby generating a magnetic field in the opposite direction. However, our wireless powering and data transmission scheme is intrinsically designed to be parallel-extendable to cover multiple cortical areas and increase channel capacity in the future.

In our coil WPT system, when the dual Tx coils are in phase, each subunit coil generates specific current flows so as to generates magnetic fields in Figure 4-11. Here their respective magnetic fluxes do not destructively co-interfere even between two Tx coils. Making them in-phase can be accomplished simply by connecting them into the same RF source, which is less challenging than making two coils have a 180-degree phase difference through a phase shifter.

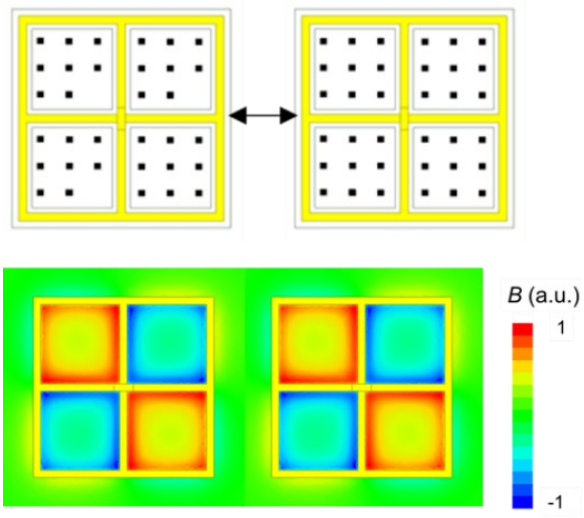

Figure 4-11. A pair of Tx coils for increasing the areal coverage, here for 69 neurograin nodes, displaying their simulated magnetic field distribution.

To demonstrate this type of dual Tx functionality, we have placed two Tx coils in 8 mm separation and place 69 autonomous TDMA neurograins and connect two RF hubs, respectively. First, we evaluated the data link when two Tx coils were in-phase with a shared RF source. After that, we also drove the Tx coils with separate RF sources to simulate the situation where the phase tuning is not ideal, which can be the case for two regular coils in a proximal distance. When two coils were in-phase, the data packet from neurograin showed the retained packet shapes in Figure 4-12 (left). However, when two coils were out-of-phase, the destructive interference on magnetic fields resulted in deformed packet shapes, as shown in Figure 4-12 (right).

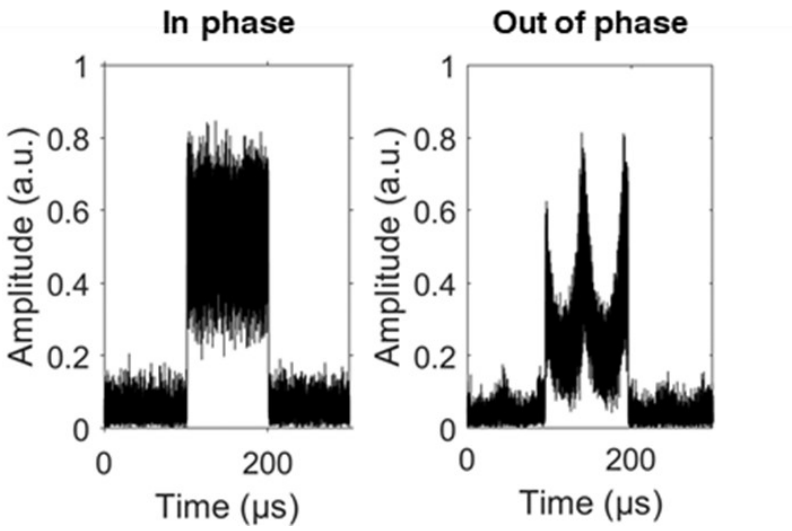

Figure 4-12. Time profiles of received RF data packet using one shared RF carrier source in-phase (left), or two separate RF carrier sources out of phase (right). The out-of-phase condition simulates the interference of two standard coils under non-ideal phase shifting.

When demodulating these packets, the averaged BER was 0.5% over 865 packets under in-phase Tx coils, as shown in Figure 4-13. However, in out-of-phase condition, the mean BER over 733 data packets then risen to 15.9%, which was much higher than in-phase condition. We note that, even with this dramatical difference in the BER, average clock frequencies across chips were similar in both conditions, suggesting that the WPT efficiencies were comparable in two cases. This result shows that the magnetic field interference between two Tx coils did not affect the wireless powering much, but it significantly increased the bit-errors.

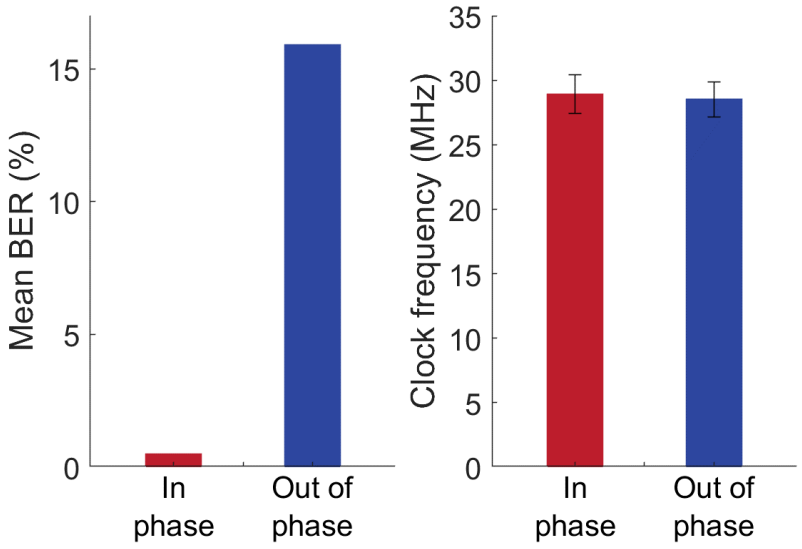

Figure 4-13. Averaged BER and clock frequency distribution for 865 and 733 data packets (in-phase and out of phase, respectively).

4.3.3. Assessment of specific absorption rate

To evaluate the RF tissue safety of the 3-coil WPT system, we solved the specific absorption rate (SAR) through electromagnetic simulation. In considering the current regulatory limits for RF power, Figure 4-14 shows computed SAR maps averaged over 10-g of tissue [76]. According to present IEEE guidance (and the related FDA dictum), adverse effects are not expected below the 10 W/kg SAR averaged over 10-g of tissue "threshold" value [77]. When applying this standard, our primate coil WPT system can transfer enough power for current ASICs to operate to 48.4% of the lateral relay coil area, including approximately 500 neurograins (Figure 4-14 left). For the rodent coil and values simulated for our experiments, an input power of 0.072 W (18.6 dBm) would provide sufficient RF energy to neurograins across the relay coil, at the equivalent peak spatial SAR of 3.69 W/kg (Figure 4-14 right). The actual measured WPT efficiency was, on average, 7.4 dB and 9.8

dB lower than the simulated values in primate and rat model coil, respectively, although the spatial dependence of the measured efficiency generally followed the simulated results.

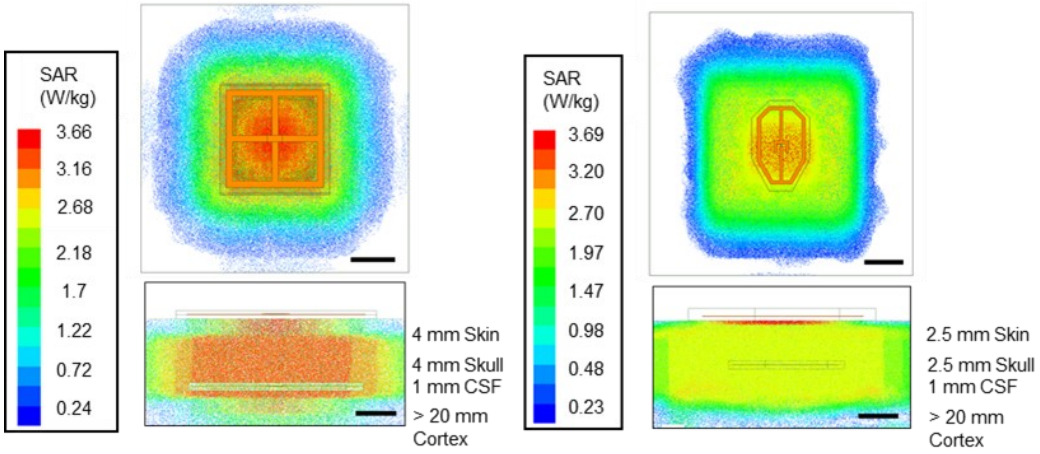

Figure 4-14. Specific absorption rate (SAR), simulated averaged over 10-g tissue. (left) SAR field pattern of the primate WPT system with 0.1 W Tx emission. (right) SAR field of the rodent model coil with 0.072 W Tx emission. The scale bar is 10 mm. SAR was averaged over 10-g tissue by following the IEC/IEEE 62704-4 [76].

We note that the power measurement includes losses at multiple stages across the link from the RF source to the on-chip circuit. The overall efficiency can be further increased, for example, by improving the on-chip rectifier efficiency or better tailoring the impedance match in the coil/RF circuitry. Due to the specifics of the TSMC 65nm foundry process, the embedded planarization-enabling filling structures across the multiple metal layers of the ASIC chip may have further limited the cross-section and coupling factor of the Rx coil [78]. Further improvement on peak SAR will be dependent on identifying the sources of the reported additional 7 dB measured loss in our system and mitigating these causes. When achieving the simulated maximal efficiency, 0.1 W (20 dBm) of Tx power in the primate model coil can provide adequate energy to a neurograin ASIC at any location

within the relay coil area, which corresponds to a peak spatial SAR of 3.66 W/kg, averaged over 10 g according to results.

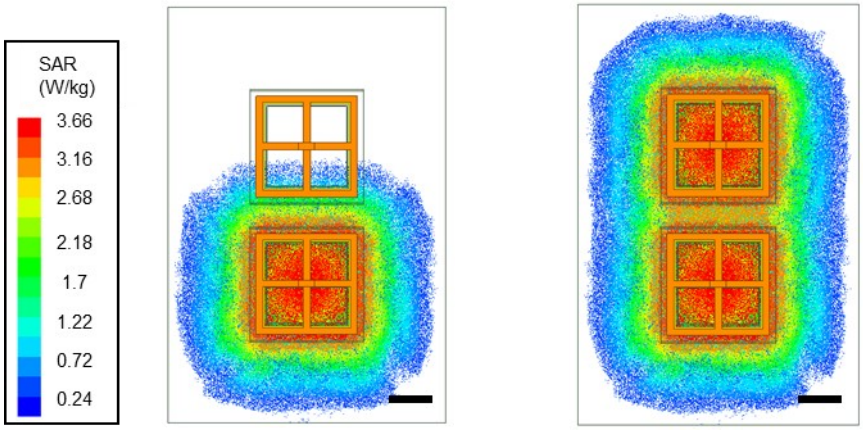

Figure 4-15. Comparison of "SAR fields" with one Tx channel enabled (left), and two adjacent Tx channels enabled (right) with 0.1 W emission for each. The composition of the tissue layer in this simulation was the same as in Figure 4-14. The scale bar is 10 mm

Lastly, Figure 4-15 shows the effect of an extra adjacent Tx/relay coil on the SAR value, demonstrating that adding another RF WPT/data path (e.g., for broader cortical coverage across multiple areas) does not affect the local SAR value. Thus, from the SAR point of view, our wireless system should be able to scale up using multiple RF paths (i.e., scaling for neurograin populations in the thousands) with minimal impact on the peak local SAR values while distributing the RF power.

4.3.4. The RF benchtop setup for conducting wireless efficiency measurements

For the benchtop test, we made an RF test setup consisting of a phantom bath and liquid head phantom to mimic the dielectric properties on surrounding tissues in the neurograin interface. Liquid head phantom with the permittivity of 41.8 and the conductivity of 0.97 S/m around 900 MHz was made in accordance with the IEEE standard

1528-2013 [76]. The liquid phantom consisted of 64.81% 1,2-Propanediol, 0.79% NaCl and 34.4% DI water. A phantom bath was designed for the wireless test using acrylic plates and a 0.1 mm thin plastic sheet to reduce RF interference (Figure 4-16). The separation between Tx and relay coils was 8 mm or 5 mm, and the gap was filled with the liquid phantom. Two plastic sheets supported the liquid phantom and the relay coil, respectively. For the efficiency measurement experiment, grid paper was attached inside the relay coil to guide the manual placement of the neurograin chip. For the Tx alignment, a 3D-printed Tx coil holder was placed below the bath and above the paper with grid marks which provides references for the alignment between the Tx and relay coil.

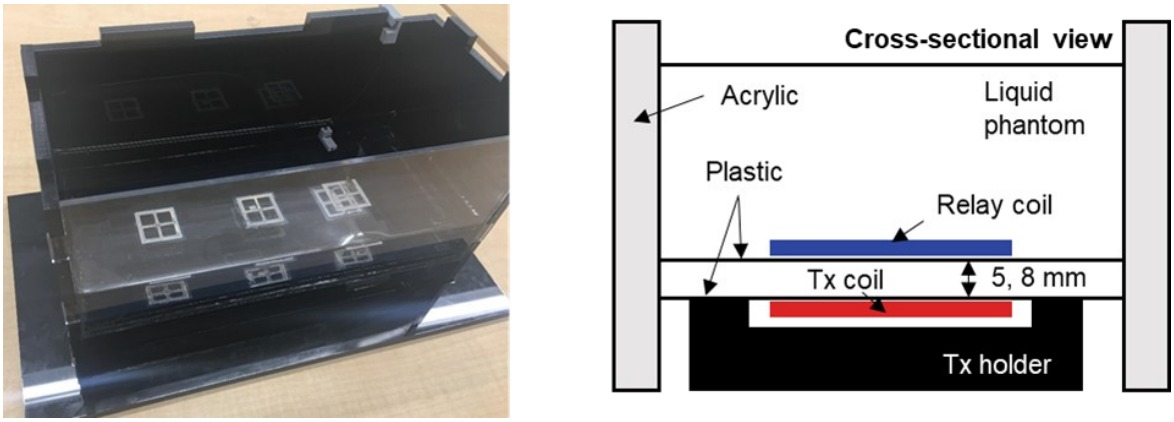

Figure 4-16. The acrylic-based benchtop setup designed to minimize ambient RF interference, including the liquid phantom to mimic dielectric properties of brain tissue [76].

4.4. Conclusion

Summarizing, the author has developed a 3-coil WPT system for neurograin considering multiple frequency-dependent competing factors which determine the overall WPT efficiency of the system. After comparing the 2-coil and 3-coil WPT systems, we have decided to use an additional resonator to focus the magnetic field in the target area

where the neurograins are implanted. We have designed a specialized custom RF coil through electromagnetic simulation, which comprises sub-coil structures to expand the wireless powering area. We measured the efficiency of the 3-coil WPT system using neurograin with a frequency-dependent clock as an RF power detector. However, compared to simulated values, the measured efficiency showed 7 dB additional losses, which may come from coil-to-coil or coil-to-circuit matching losses or the rectification loss. Considering that we need to transmit up to 27 dBm to provide sufficient RF energy to neurograin in all locations within the relay coil, further improvement on the impedance matching would be required to reduce the transmitting power and subsequently improves the tissue RF long-term safety of the system.

CHAPTER 5. POST-CMOS MICROFABRICATION PHASE

5.1. Introduction

The material in this chapter is adapted from the paper, 'A scalable and low stress post-CMOS processing technique for implantable microsensors', *Micromachines*, 2020.

The bare neurograin silicon CMOS die which arrives from the semiconductor fab must be further processed by a sequence of microfabrication steps to be a functional biomedical implant. This non-trivial post-CMOS processing phase includes building metal films for microelectrodes atop the on-chip Aluminum contact pads, applying biocompatible coatings on the chip, and so on. The processing requires microfabrication steps such as photolithography, metallization, etching, and bonding on the CMOS die. At the completion of the foundry CMOS process, the top metal layer typically consists of an aluminum layer (M9) which oxidizes over time and is unsuitable as the electrical recording or stimulation interface with biological tissue [79, 80]. Therefore, stable alternative materials such as gold (Au), platinum, or poly(3,4-ethylene dioxythiophene) polystyrene

sulfonate (PEDOT: PSS) need to be used for recording or stimulation to achieve long-term reliability of the electrode-tissue interface. Several studies suggested post-processing techniques using the sacrificial carrier substrate such as deep-etched silicon, silicon oxide substrate [81-83], or rigid SU-8 or epoxy polymer [84, 85]. However, the size variance of decided chips cannot be accounted for in the silicon-based carrier substrate, and the retrieval of CMOS can be significantly challenging with SU-8 or epoxy polymer carrier due to their high chemical resistance. On the other hand, CMOS chips are susceptible to chemical, mechanical, and temperature stress, limiting the choice of substrate materials and processing techniques.

The following section describes our novel chip processing technique using a polydimethylsiloxane (PDMS) substrate and demonstrates this on diced and thinned neurograin chips [60]. With this technique, neurograin chips were embedded into flexible PDMS in the form of a 2D grid and an aligned manner. Also, the chips and the PDMS formed a planar surface which allows for standard microfabrication techniques such as photolithography. We also note this process is also scalable, which means that we can simultaneously process a larger number of neurograins. Under this method, we were able to print pairs of Au and PEDOT: PSS microelectrodes on neurograins, and assessed the interface impedance before used on the benchtop or in-vivo for neural recording and stimulation (Chapter 6). The entire fabrication steps are designed as a dry assembly process that does not involve harsh chemicals, mechanical stress, and high temperatures to prevent damage to the CMOS device.

5.2. Post-process microfabrication process flow for neurograins

5.2.1. Planarization technique for a 2D neurograin array

To embed neurograin chips into flexible PDMS in the form of a 2D grid, we performed the PDMS-based planarization process illustrated in Figure 5-1. Firstly, to align multiple chips as an array, alignment marks were made on a glass slide with a photoresist. The glass slide was cleaned with acetone, IPA, and DI water and dried with nitrogen gas immediately. An adhesion promoter (hexamethyldisilazane, HMDS) was spin-coated on the glass at 2500 rpm for 15 sec, a thin photoresist layer (AZnLOF2035, MicroChemicals GmbH, Ulm, Germany) was spun at 3000 rpm for 60 sec, then was soft baked on a hotplate at 100 °C for 90 sec, and exposed to 80 mJ/cm² UV light at 375 nm. After post-baking (same conditions as the soft-bake), the photoresist was developed in AZ MIF-300 for 10 sec. Then, a homogenous mixture of PDMS and a curing agent (Sylgard 184, Dow Corning, Midland, MI, USA), at a ratio of 10:1, was directly spin-coated on the glass substrate with photoresist markers at 3000 rpm for 60 sec and cured for 7 min at 90 °C. The thin sacrificial PDMS layer protected photoresist markers from mechanical damage and provided an adhesive surface for picking up the chips.

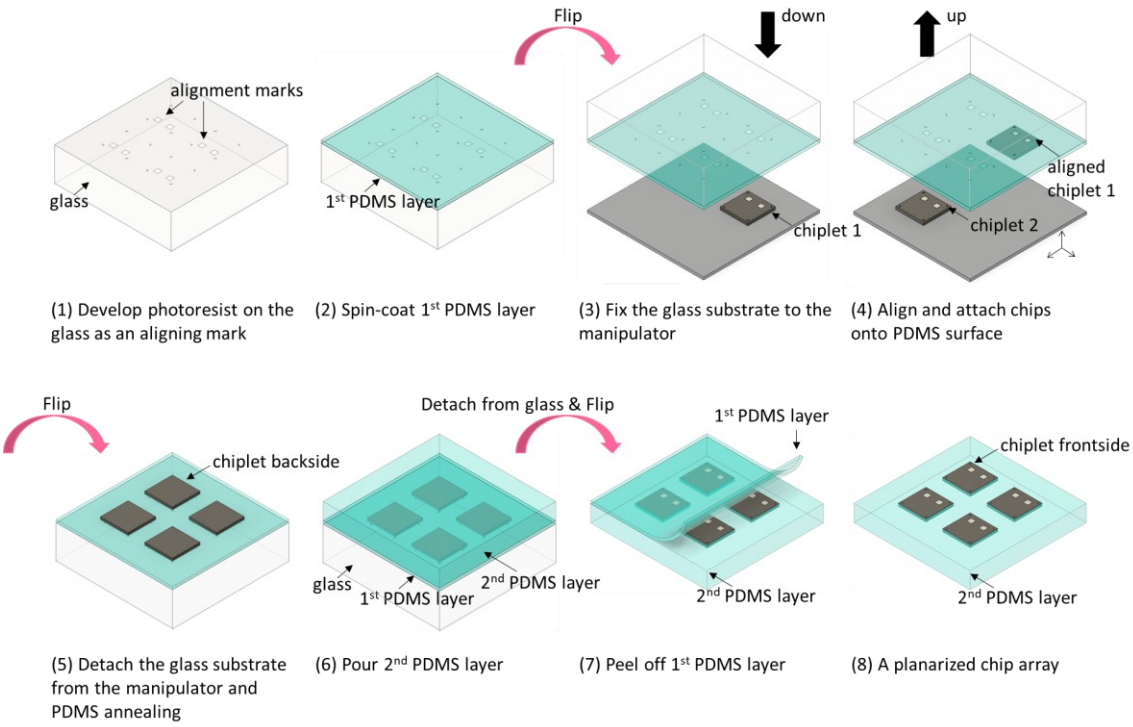

Figure 5-1. The polydimethylsiloxane (PDMS)-based fabrication procedure for post-processing of 2D-planarized microchip arrays: (1) A patterned array of photoresist aligning marks is created on a glass substrate. (2) The glass substrate is spin-coated with the first 65 µm-thick PDMS layer. (3) After baking, the glass substrate is mounted on a manipulator with the adhesive PDMS layer facing down. (4) The chips are aligned and picked up by the sticky PDMS surface, guided by the markers. (5) The glass substrate with the chips is detached from the manipulator and annealed. (6) With the glass substrate facing down, the second layer of planarizing PDMS is poured over the chips and cured overnight. (7) After detaching the glass substrate, the spin-coated PDMS layer is carefully peeled away. (8) The complementary metal-oxide-semiconductor (CMOS) chip array with active circuits exposed and flushed with PDMS carrier is ready for post-process fabrication.

Subsequently, the PDMS-coated glass carrier was fixed on a micromanipulator while facing CMOS chips at a bottom stage. The marker position was found manually in the XY-plane, and the PDMS-glass carrier was lowered to pick up the neurograin chip by the adhesive PDMS surface. By repeating these steps, multiple chips on PDMS could be precisely aligned as a 2D array. After the pick-and-place procedure, the glass substrate with neurograin was annealed at 130 °C on the hot plate for 1 h. The annealing process is crucial

since it prevents cross-linking between the sacrificial thin PDMS layer and a second PDMS carrier substrate. The PDMS mixture is then poured over the glass/PDMS/chip construct until the backside of the chips is submerged. The preparation was then cured overnight at 80 °C (> 8 h). After curing, the PDMS-chip-PDMS assembly was detached from the glass and attached to a silicon wafer substrate with the active chip surface facing upward. Finally, the thin sacrificial PDMS layer was gently peeled off from the second PDMS carrier substrate resulting in structure in Figure 5-2. We note that this process eliminated the use of excessive mechanical force/pressure, harsh chemicals, or heat/stress to ensure that the functional CMOS circuits were not damaged. For instance, in Figure 5-2, we could embed fragile neurograin thinned at 30 µm into PDMS without damaging it. On top of that, the PDMS-chip assembly could be disassembled since the PDMS maintains its flexibility throughout the process.

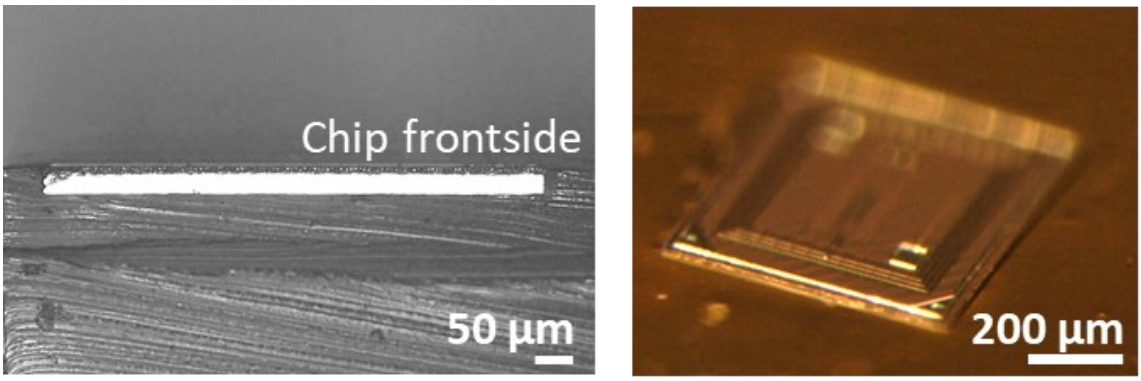

Figure 5-2. The final planarized chip/PDMS assembly is completed after the top PDMS layer is peeled off from the contacting chip/PDMS surface. Note the two small square Al-contact pads on the chip discussed in the text.

To analyze how effective this technique is, we compared the profile of planar photoresist spin-coated on neurograin with or without the carrier PDMS substrate, as shown in Figure 5-3. In the first case, neurograin was attached to silicon wafer only with

double-sided Kapton tape. In this experiment, 200 µm thick neurograin was used since most of the neurograin are not thinned, which results in a step difference of 200 µm at the edge of the chip. In the second case, we embedded neurograin into PDMS and attached it to the silicon wafer. When we spin-coated the photoresist and measure the profile, we could observe that 8.8 µm think photoresist in the middle of the chip in the chip-only case while it was much even in the chip-PDMS assembly condition. In this non-even photoresist layer, photolithography is extremely difficult since the required UV light exposure can change depending on the thickness of the photoresist. However, in chips with PDMS carrier substrate, planar photoresist layers were achieved, which are sufficiently flat for photolithography.

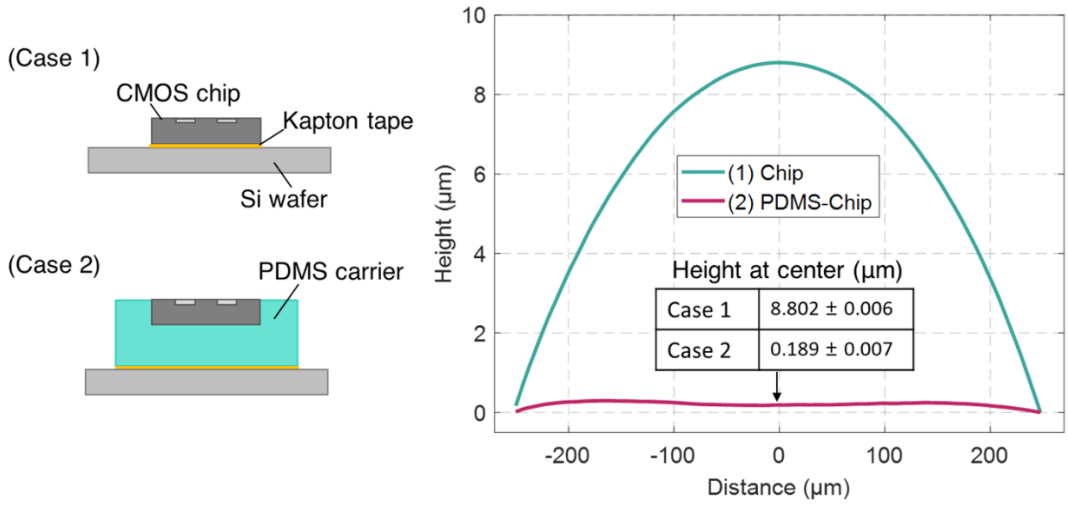

Figure 5-3. Comparing the surface profiles microchips coated by a 3.5 µm-thick photoresist film, without (case 1) and with the planarization process (case 2). Both samples were spin-coated at 3000 rpm for 60 sec and baked in the oven at 90 °C for 30 min. The unprocessed and PDMS-planarized chips (650 µm × 650 µm × 200 µm) were attached to the silicon wafer using a double-side Kapton tape.

5.2.2. Post-process fabrication of neurograin microelectrodes

Neurograin circuits were designed for electrophysiological sensing or stimulation by wirelessly networked ensembles of implantable microdevices. Following receipt from the foundry (TSMC; 65 nm RF low power LP process node), individual chips were diced to 650 μm× 650 μm afterward. Some of the dies were then mechanically ground to have a ~30 μm thickness by DISCO Hi-TEC (Munich, Germany). Each ASIC chip was configured for two or three 65 × 65 μm^2-sized aluminum (Al) pads, commonly available in the CMOS process, to provide direct electrical access to adjacent biological tissue. The impedance and longevity of implanted electrodes in the tissue call for careful consideration of the electrode material. We employed combinations of Au and PEDOT: PSS thin films as planar microelectrodes, coated on the Al-pads of the CMOS die to complete the contact structure. Unless otherwise specified, all contacts were approximately (70 μm^2) in surface area. For photolithography, we used both a mask aligner (Karl Suss MJB3, Karl Suss America, Waterbury, VT, USA) and a maskless aligner (MLA150, Heidelberg Instruments, Heidelberg, Germany) at the Brown Microelectronics facility.

The PDMS-planarized chips (Section 5.2.1) were first plasma-treated in a Fischione model 1070 plasma cleaning system in a 30 sccm of 25% O_2 and 75% N_2 gas mixture for 2 min. With the resulting hydrophilic surface modification on the PDMS, this process improves the bonding and printability of the subsequent photoresist and helps one achieve an even coating over the PDMS-chip assembly surface. Immediately after plasma treatment, AZnLOF2035 photoresist was first spin-coated on the PDMS-chip surface at 3000 rpm for 60 sec and soft-baked in the 92–94 °C oven for 30 min, then exposed with a wavelength of 365 nm at an irradiance intensity of 75 mJ/cm². After baking, patterns defining the desired

electrode openings were developed in AZ MIF-300 solution. An electron beam evaporator (Lesker Lab-18, Kurt J. Lesker Company, Jefferson Hills, PA, USA) was used for the deposition of Ti/Au (50/200 nm) thin films followed by the lift-off process using AZ 400T stripper on the 80 °C hot plate for 5 min. As a result, the deposited Au-patterns lined squarely on the top of the ASIC aluminum contact pads. Figure 5-4 demonstrates how the method is indeed scalable toward a batch fabrication process, including CMOS chips of different footprints. Figure 5-4 left shows an array of 650 µm × 650 µm × 200 µm sized chips, where a pair of 100 µm sized square Au microelectrode patterns are deposited on corresponding Al contact pads (as received from the fab). In Figure 5-4 right, 200 µm square Au patterns could be added around the proximate seal ring structure surrounding the edge side of the chip.

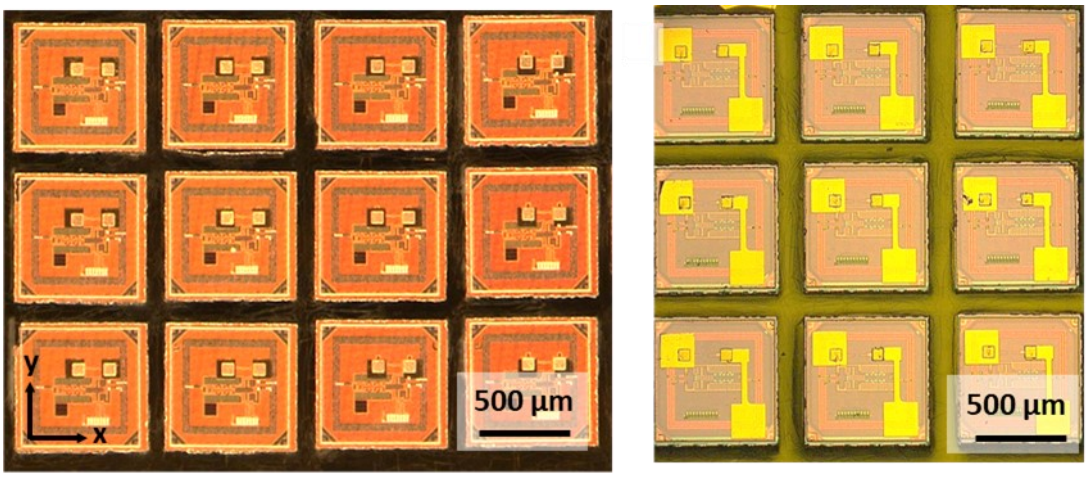

Figure 5-4. Examples of simultaneous post-processing of multiple CMOS chips by batch fabrication: Photo shows an array of 650 µm × 650 µm sized chips with 100 µm or 200 µm square Au patterns (electrodes) deposited on the two Al-pads within each chip.

Further thin-film PEDOT: PSS coating on Au patterns can be done to lower the impedance of neurograin. PEDOT:PSS is a conducting organic polymer which is seeing

increasing use in the electrical stimulation of brain tissue (see also below). Here, 20 mL aqueous dispersion of PEDOT: PSS (Clevios PH1000, Heraeus Epurio Clevios, Hanau, Germany) was mixed with 5 mL ethylene glycol, 50 μL dodecylbenzene sulfonic acid (DBSA), and 1 wt% of (3-Glycidyloxypropyl) trimethoxysilane (GOPS). Adding ethylene glycol enhances the conductivity of PEDOT: PSS, while DBSA helps to increase its coating properties by adjusting the surface tension. In addition, GOPS enhances the stability of PEDOT: PSS films in aqueous environments as a cross-linker [86]. Before applying PEDOT: PSS film on the Au/Al contact pads, AZ photoresist was spin-coated and exposed to UV aligning the PEDOT: PSS electrode sites in co-registry with the CMOS chip contact pads. The mixture of PEDOT: PSS was then evenly spin-coated onto the 3.5 μm-thick AZ-photoresist-patterned chips at 3000 rpm for 60 sec and soft-baked in the 130 °C oven for 20 min. The AZ photoresist was lifted off in all regions except PEDOT: PSS electrode sites using acetone under sonication for 1 min. Note that since PEDOT: PSS films are susceptible to a base solution, acetone was only used to dissolve AZ photoresist patterns in the lift-off process [87, 88]. The PEDOT: PSS film was subsequently cured on the 130 °C heat. As noted, a key goal of this work was to develop a method to enable high throughput batch post-processing of multiple sub-mm scale CMOS dies. Figure 5-5 presents the results of coating PEDOT:PSS selectively on patterned Au electrodes.

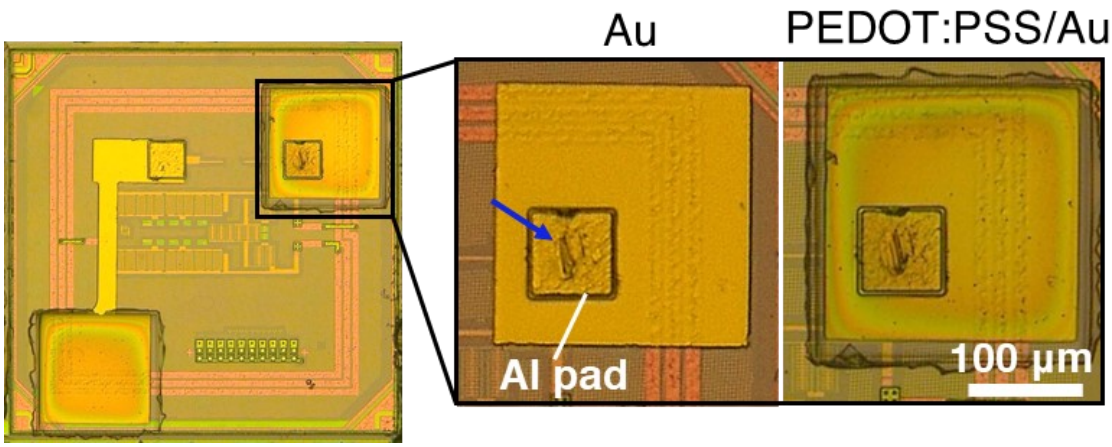

Figure 5-5 . A close-up view of micropatterns with both Au and PEDOT: PSS coating specifically on targeted planar microelectrode sites only. The blue arrow shows the indentation generated by the mechanical removal of the Al oxide layer.

To quantify our alignment accuracy, the angular deviation and center-to-center distance were measured by overlaying the photomask CAD drawing onto Figure 5-6 (left). During the calculation, the alignment offset of the mask image was canceled by subtracting the average of the measured values. The CAD measurement guide includes a grid of squares on pad sites and alignment marks at the four corners of each chip with a regular pitch and angular orientation. The result of the measured angular deviation is shown in Figure 5-6 (left). The achieved x–y precision is shown in Figure 5-6 (right). By measuring the center-to-center distance between the center of two pads on the CAD drawing and that of actual chips. The accuracy of the process was determined using the standard deviation, which was 0.3° for the angular orientation, and 8.5 µm and 6.4 µm for the lateral accuracy on the x-axis and y-axis, respectively.

The results show the alignment accuracy is well within the range of tolerance required for our implantable microchip fabrication processing. If it were necessary to further improve the accuracy and scalability of our technique to thousands of chips, we

envision an automated alignment procedure for large ensembles using a robotic microhandling tool guided by a computer vision system. This advancement would also improve the process yield by reducing the number of misaligned chips and the fabrication time required for chip placement.

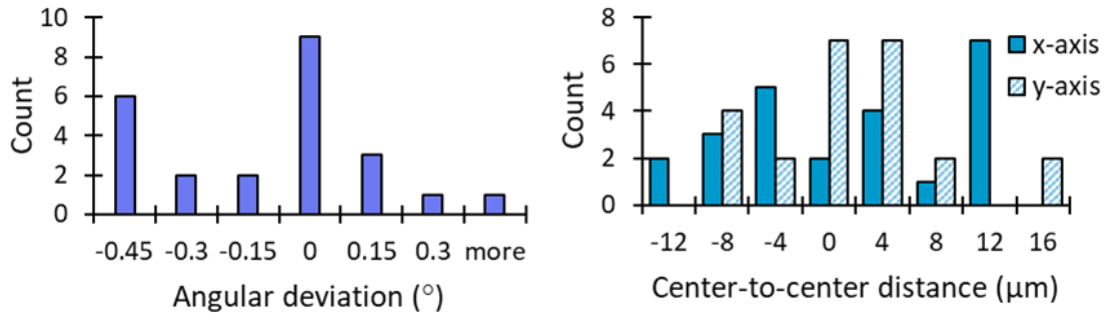

Figure 5-6. Histogram of the measured angular orientation (left) and center-to-center lateral distance (right)

5.2.3. Characterization of the post-processed microelectrode impedance

Following the microfabrication sequence above, we analyzed the impedance of microelectrodes fabricated on the chip to check whether they can provide an impedance suitable for the neural tissue interface. Since PEDOT: PSS offers biocompatible interfaces between the electrode and biological tissue and improves the efficiency of stimulation current delivery due to its low contact impedance, it has emerged as one of the promising electrically conductive polymers to be used as microelectrode material [89-91]. We fabricated 70 µm square Au and PEDOT: PSS-coated Au electrodes with extended structures, including additional pads on the chip surface for our impedance measurement. This chip was attached to a flexible liquid crystal polymer (LCP) film with patterned Au traces. Using a ball bonder outfitted with a 20 µm diameter Au bond wire (WestBond,

USA), two extended Au pads on the opposite electrodes were wire-bonded on the LCP substrate individually as shown in Figure 5-7, which helped to make the external connections with the tester during measurement. This structure is encapsulated using an epoxy resin except for the Au and PEDOT: PSS electrode sites exposed to saline solution. Among two exposed on-chip electrodes, one serves as the active electrode while the other serves as the reference electrode. The impedance measurement is performed at room temperature, and the frequency sweep is from 1 Hz to 2.04 kHz using a NanoZ impedance tester (White Matter LLC, Seattle, WA, USA).

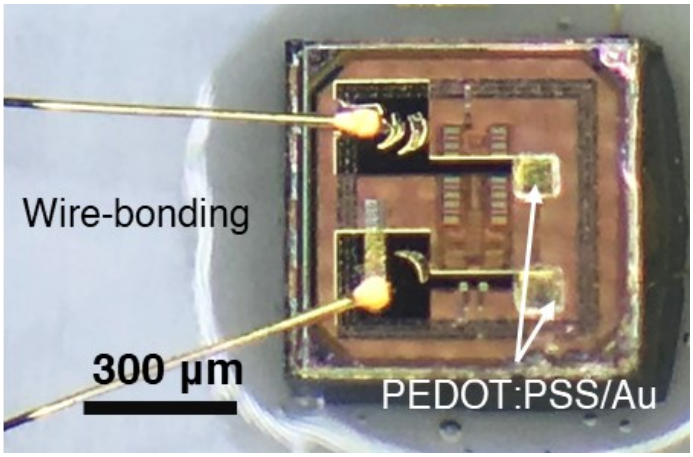

Figure 5-7. Test structure with an extended gold trace for auxiliary wire bonding to the ASIC circuit of ae neurograin.

To verify the electrical characteristics of the post-processed on-chip microelectrodes, either with Au or combinations of Au and PEDOT: PSS, we performed frequency-dependent impedance measurements in saline. Our measured impedance values in Figure 5-8 across the frequency range of interest were comparable with the previously reported impedance on conventional planar structures used in electrophysiology, both Au and PEDOT: PSS electrodes [92]. Likewise, we verified that the impedance of the organic

PEDOT: PSS was notably lower than that of Au electrodes across the overall frequency range (up to the approximate characteristic frequency of neural action potentials typically used as a reference point for impedance comparison). Here, the impedance of our 70 µm sized Au electrode (427 kOhm) was more than an order of magnitude larger than that of the 70 µm-sized PEDOT: PSS on Au (7 kOhm) at 1 kHz, consistent with previous findings [92]. This result supports that the quality of our post-processing for microfabricated electrodes on the sub-mm sized chip using the lift-off process is comparable with that of electrodes fabricated on the large planar substrate.

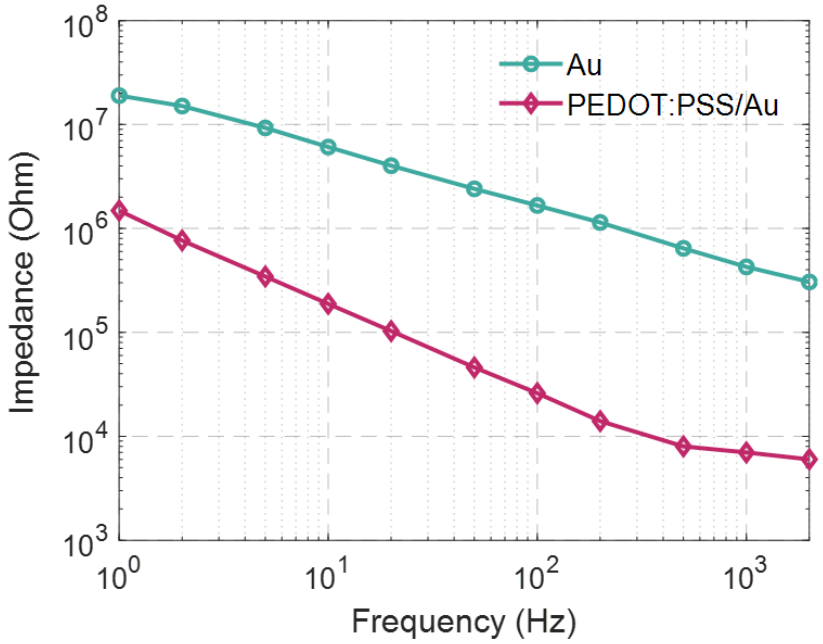

Figure 5-8. Comparison of the impedance spectra of PEDOT: PSS coated and uncoated Au microelectrodes of neurograins of the same size.

5.2.4. Microfabrication of intracortical electrodes for the electrical microstimulation

As described in Chapter 2, the "stimulation neurograin" was designed to deliver a biphasic current stimulation intracortically. For the present generation of our neurograins,

designed to float on the cortical surface, this requires adding penetrating electrodes on the die of 1~1.5 mm length. The current PDMS-based technique is not suitable for the fabrication strategy of adding long microwires on the chips. A post-process assembly approach was developed for intracortical electrodes on neurograin, as illustrated in Figure 5-9. We began by fabricating SU-8 cubic blocks (200 μm × 200 μm × 200 μm) on the glass and poured PDMS on the top of those patterns to transfer them to PDMS. After annealing PDMS at 90 ° C for 4 hours, one gets a PDMS structure with cubic holes which are then filled with a silver epoxy as shown in Figure 5-9b. We manually placed tungsten electrodes with 1.5~2 mm length for each hole, followed by thermal annealing at 130 ° C for 3 hours. The microwire electrodes within these silver epoxy cubic blocks, able to stand alone on the chip due to the base cubic structure, were dipped in the silver epoxy and manually attached to the neurograin. This electrode-neurograin structure was covered again with a non-conductive epoxy to improve the mechanical strength required for the penetration and also to provide the electrical encapsulation. Overall processes were performed manually and the electrodes were not strictly perpendicular to the surface of the neurograin. Yet the wire electrodes could still penetrate the cortex reliably and provided intracortical access to neurograins (Figure 5-9e). We have also explored alternative methods for further improvements, such as 3D microprinting of intracortical electrodes directly on neurograin using the Center for Nanoscale Systems at Harvard, a project presently under investigation in our group.

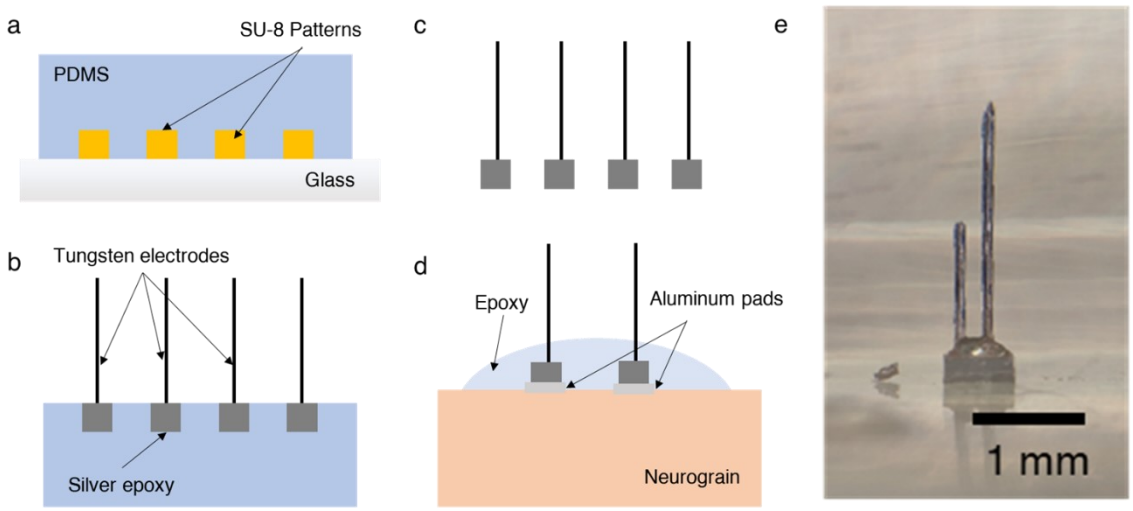

Figure 5-9. Assembly processes to attach intracortical penetrating tungsten electrodes on "stimulation neurograin" silicon die. a. Transferring the SU-8 cubic block structure on PDMS. **b**. Electrodes placed inside the PDMS cubic holes filled with the silver epoxy. **c**. Tungsten electrodes with silver epoxy blocks. **d**. Neurograin with intracortical electrodes attached to aluminum pads. **e**. Photograph for stimulation neurograin with pairs of intracortical electrodes.

5.2.5. Post-processing the identifier addresses onto chips by laser ablation

The wireless stimulation neurograins each haves unique RF device identification (digital address) for its downlink communication. As an example of hard encoding of the address on-chip, the 10-bit metal fuses shown in Figure 5-10 top need to be cut selectively to assign the different binary bits addresses to each chip. This fuse structure using a 3.4 μm-thick Al metal micropattern added to the top layer redistribution metal available in the 65 nm CMOS process. To conserve valuable silicon real estate, each fuse line was constrained to a 3 μm width and 6 μm length. In our design, ten metal lines of the 10-bit fuse structure are electrically connected to the ground within the internal circuitry underneath the dielectric passivation layer.

We used a commercial EzLaze 3 laser cutting system (New Wave Research, Sunnyvale, CA, USA; courtesy of Prof. Jacon Rosensteins' laboratory at Brown) to write the addresses as follows. Pulsed 355 nm UV laser radiation was first utilized to ablate a window in the dielectric passivation (scratch protect) bilayer (Si3N4/SiO2) on top of the chip's surface to access the Al metallization features in the designated RF address bus embedded within the CMOS circuit. Considering the fuse size, the laser target spot size for window opening was set to 2.5 µm × 15 µm defined using a 50x objective (7–8 single-shot exposures were sufficient to remove the dielectric layer). Subsequently, 532 nm green laser light was used to cut the Al-metal traces within the fuse. We chose a smaller aperture size (2.5 µm × 10 µm) to reduce physical damage on the proximate dielectric layer (3–4 times exposures single-shot exposures). Figure 5-10 (bottom) shows laser-ablated fuses under the scanning electron microscope (SEM), which has well-defined cutting edges of the fuse without excess damage on the surrounding structure.

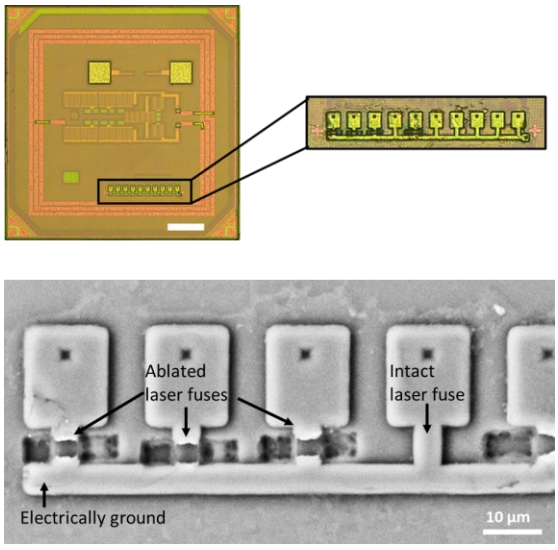

Figure 5-10. (top) A microscope image of a CMOS chip with 500 µm × 500 µm sized RF on-chip antenna coil and a close-up view of an embedded 10-bit fuse structure. Scale bar: 100 µm. (bottom) SEM image of the one fuse structure after the laser ablation process.

The laser-ablated fuse address was registered by stimulation neurograins internal circuitry and validated electronically through wireless radiofrequency (RF) power and data transmission. The laser-written 10-bit address and an interleaved header sequence were transmitted as a backscattering signal of stimulation neurograin, which happens only at the chip turn-on process [51]. Demodulated bits on this backscattered data from the chip in Figure 5-10 are shown in Figure 5-11, which indicate changes in 4 address sequence accordingly. These results proposed that our metal-fuse address and laser ablation technique can effectively provide a unique address for neurograin.

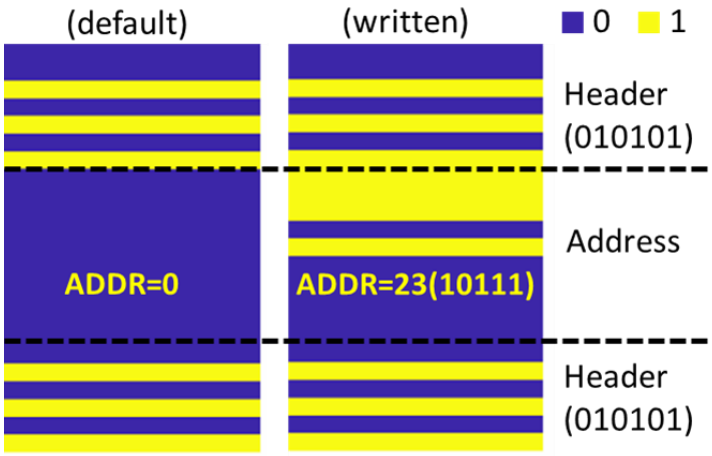

Figure 5-11. Validation of the hard encoded addressing using RF backscattered data for the 10-bit address and an interleaved header sequence (010101) [51].

5.3. Fabrication and packaging of the coils involved in RF power harvesting and bidirectional communication

5.3.1. Fabrication of polyimide coils

We fabricated the coil structures designed on the electromagnetic simulation to build the 3-coil WPT system for the neurograin interface. The Tx and relay coil were

fabricated on a polyimide substrate with a 0.1 mm thickness under a 2-layer PCB process using a 2 oz copper trace. For the Tx coil, trace width up to 1 mm was used to improve the quality factor while, in the relay coil, the trace width of 0.2 mm is used to reduce the effect of surrounding dielectric tissue. These design parameters were optimized in electromagnetic simulation before the fabrication. Coils were then assembled with matching capacitors and encapsulated with PDMS (Dow® SYLGARD™ 182) or 3M biomedical-grade epoxy resin. For Tx coils, 50-ohm mini-RF coaxial connector and cable are also attached, as shown in Figure 5-12.

Figure 5-12. (left) The Tx and relay coil for the (proposed) primate model packaged with PDMS or LCP. (right) The Tx and relay coil for the (implemented) rodent/rat model. The relay coil was packaged along with a set of 56 individually addressable neurograin chips.

5.3.2. Methods for LCP coil fabrication

We have mostly used coils fabricated on the polyimide PCB for our neurograin experiments, while long-term usage of these coils in the tissue can raise a safety issue due to the chronic toxicity of the copper trace. Thus, we developed a biocompatible version of a 3-coil WPT system that consists of the liquid crystal polymer (LCP) board and gold traces

to improve long-term safety. Several studies have shown that LCP has high biocompatibility allowing the usage of it within the tissue for an extended period [93], which also has a low dielectric permittivity making it suitable for the RF coil substrate. To build the LCP coil, we deposited 200 nm thin gold traces on LCP and performed electroplating of gold to make thick 2 μm gold traces as shown in Figure 5-13 left. We then attached the matching capacitors and encapsulated them with PDMS so that PDMS provides the electrical insulation (Figure 13 right). Figure 5-13 left also shows wirelessly powered neurograin-proxy devices with red LEDs light up in addition to the LCP coil. This neurograin-proxy device consists of a PCB Rx coil, a matching capacitor, a rectifier, and a red LED.

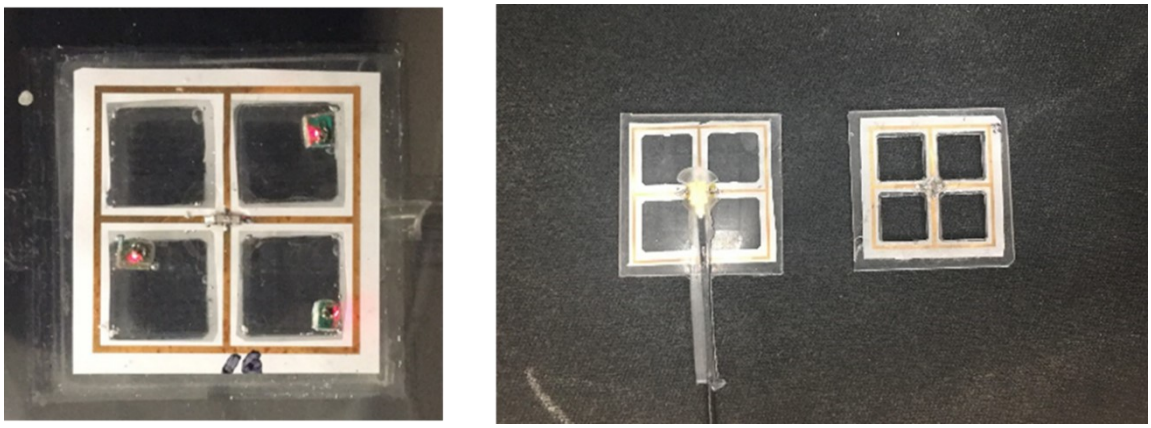

Figure 5-13. (left) The relay coil made with a gold trace and LCP showing wirelessly powered red LEDs within the coil. (right) The Tx and relay LCP coils encapsulated by PDMS.

5.3.3. PDMS packaging of coils

Polyimide or LCP coils need to be packaged with other materials such as PDMS for the electrical insulation. However, the electrical insulation itself is not sufficient for the relay coil operation since, if coils are packaged with thin layer materials, their frequency

resonance gets susceptible to surrounding biological tissue which mostly has a high dielectric permittivity. Notably, in the case of the relay coil, we found that the resonance shift within the tissue was significant. Thus, the coil needs to be packaged with sufficiently thick materials with a low permittivity to reduce the resonance shift in the tissue. In the neurograin 3-coil WPT system, coils were packaged with PDMS by using acrylic molds to make a 1-mm thick PDMS encapsulating layer. Before the packaging, coils were cleaned with IPA, acetone, and DI water, and molds were coated with an anti-adhesion spray. The base PDMS layer with a 0.5 mm thickness was cured for 3 hours at 100 °C and attached to the coil. The top PDMS layer covered the PDMS-coil again, resulting in an overall packaging thickness of 1 mm. PDMS-packaged coils are shown in Figure 5-12 and 5-13.

5.3.4. LCP packaging of coils

Our group tested LCP packaging on neurograin chips and analyzed their reliability over time, which shows that even the 25-50 μm thick LCP layer could provide hermetic sealing over 8 months [94]. LCP was also used for the package of the relay coils in the cases that longer-term hermetic encapsulation was needed. For LCP packaging, we used a metal jig and a high-temperature press machine. The total thickness of coil LCP packaging should be thick enough to provide a sufficient gap between the tissue to the coil as in PDMS packaging. So, we targeted the total thickness of 1 mm and we filled the gap with LCP powders.

Firstly, we placed two Teflon sheets on each side of the jig and press four layers of 25 μm LCP films at 230 °C for 1 min. After that, we pressed again at 284 deg for 10 min. We divided pressing into two steps since, in the case that the thermo-press was done in a single step, LCP sheets were often torn due to the high aspect ratio structure in our coil.

After cooling the metal jig down, we placed coils into the LCP socket made during the previous thermo-press shown in Figure 5-14 and filled the gap with the ground LCP powder [95]. The LCP sheet and coils were then covered again with additional two layers of 25 μm LCP sheets and pressed at 284 deg. The relay coil for the primate model in Figure 5-12 was encapsulated by the LCP through the procedure described here.

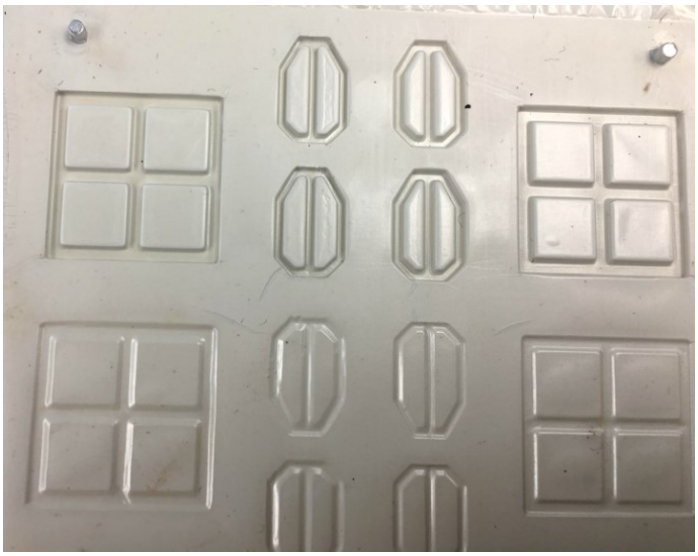

Figure 5-14. LCP sheets thermo-pressed by a metal jig used for the packaging of the relay coils.

5.4. Conclusion

In summary, considerable additional effort was required to transform the foundry delivered silicon chiplets to functionally useful neurograin microimplants. We developed novel post-process techniques for the sub-mm sized CMOS die, using PDMS to fabricate pairs of Au or PEDOT:PSS electrodes on neurograins, which showed an impedance suitable for the neural interface. Another microfabrication step employed for writing programmable addresses on the neurograins was developed using laser ablation in connection with the on-chip metal fuse. The metal fuses served as the device identifier, and

we verified that the correct laser-ablated on-chip addresses were successfully registered by the neurograin chip circuit. Additionally, for the use of the neurograin microdevices in intracortical stimulation, we developed a technique to attach intracortical electrodes to neurograin for the delivery of electrical microstimulation deeper into the cortex. Finally, fabrication and packaging methods for the 3-coil WPT system were developed using moderately biocompatible materials, PDMS, or materials suitable for even longer-term, such as LCP, for possible chronic use of the implants.

CHAPTER 6. DEMONSTRATION OF NEURAL RECORDING AND ELECTRICAL MICROSYIMULATION BY ENSEMBLES OF NEUROGRAINS ON BENCHTOP AND IN-VIVO RAT MODEL

6.1. Introduction

The material in this chapter is adapted from the paper, 'Wireless ensembles of sub-mm microimplants communicating as a network near 1 GHz in a neural application', *bioRxiv*, 2020.

While some groups have proposed ultra-miniaturized microimplants for targeting neural signals, such microdevices have been so far demonstrated only for a small handful of devices due to the challenges involved in the wireless networking of ensembles of microimplants [31, 41, 42]. In the work presented in this thesis, the neurograin SoC concept was built on the concept of a wirelessly powered chip that can communicate as a networked ensemble as described in previous chapters. This chapter demonstrates the performance of

multichannel neurograin implants, first, on the benchtop and the in-vivo rat model in demonstrating their functionality in neural recording and microstimulation, respectively, featuring ensemble networking, multichannel ECoG recording, and intracortical microstimulation [56].

While we aim to ultimately integrate each modality (recording, stimulation, and bidirectional TDMA link) into a single neurograin, we can currently leverage two types of chips to fulfill the specific needs of each separate function across neurograin ensembles. The recording chip reported here generates the uplink under autonomous TDMA. Since the head/skull size of a rat and associated craniotomy limits the number of practically implantable neurograins (here N=48), the probability of data packet collision is low. On the other hand, stimulation neurograin operates with the downlink in the call-and-response TDMA mode. Even in the current version of the stimulation chip, multiplexing between chips takes only about 40 μs, which is sufficiently fast for commonly used electrical microstimulation protocols for the mammalian brain. Both for the benchtop and in-vivo experiments, we obtained reference data employing a commercial neural signal processor instrument (Nomad, Ripple Inc), connected to physically wired electrodes. This reference data provides data for a comparison between the performance of neurograin with that of the current state-of-art commercial device where the former has a fully wireless configuration but with the tightly constrained area and power, and the latter does not.

6.2. Benchtop validation of the wireless operation of neurograin ensembles

6.2.1. Performance characterization of “recording neurograins”

We firstly analyzed the performance of the recording neurograins in the saline to characterize the AFE and the neurograin uplink data telemetry, respectively. The benchtop test on recording neurograin was performed in the setup shown in Figure 6-1 where the powering coil was also embedded in saline. The supporting layer of the setup consisted of PDMS and the PCB to form a planer surface for the saline and neurograin. The acrylic saline bath covered the top surface of the PDMS and PCB assembly and the saline was filled in the bath. The Tx coil on the PCB was connected to the RF hub for RF wireless powering and data communication of neurograin. Underneath the supporting PDMS- PCB assembly, we placed a 3D-printed holder, which fixed the saline benchtop setup on the metal board.

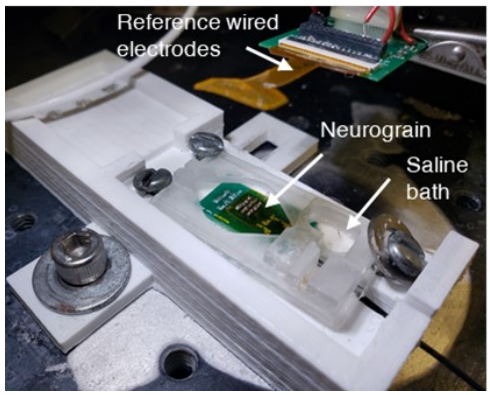

Figure 6-1. A neurograin saline benchtop test setup with an embedded Tx coil connected to the RF hub.

To obtain reference data for the neurograin system, we employed a commercial multichannel neural signal acquisition system (Nomad, Ripple Inc), which records signals from wired electrodes placed near the neurograins. In the version of neurograin for recording, we could track the recorded data from the target chip over time by using its

unique PUF device ID. For instance, Figure 6-1 top shows transient recordings on 200 Hz both from neurograin and the Nomad system. Clips of recorded data from neurograins were transmitted as multiple separated data packets, demodulated separately, and combined to achieve continuous waveform in this plot. The relative amplitude difference between recordings on the Nomad system and neurograin was observed possibly due to two factors, 1) the distance from the recording electrode to the signal source and 2) recording configurations. Note that while the neurograins measure voltage differentials between closely placed two electrodes on the chip, the wired Nomad system recorded voltages on a single recording electrode with respect to a distal ground electrode (Ag/AgCl in this case). In the case of the wired system, the differential voltage between recording and ground electrodes was much larger, resulting in a noticeable amplitude difference between recordings in the two cases.

Nonparametric power spectral density (PSD) estimates (200 Hz recordings) are shown in Figure 6-2 (bottom). While the recording signal quality from the neurograins appears comparable with that of the wired system such transient plots, the PSD estimates revealed that the noise floor in the case of the neurograins was generally higher. However, we note that in contrast with the rack-mounted commercial system, the wearable/portable wireless neurograins are tightly constrained in their area and power as well as data bandwidth.

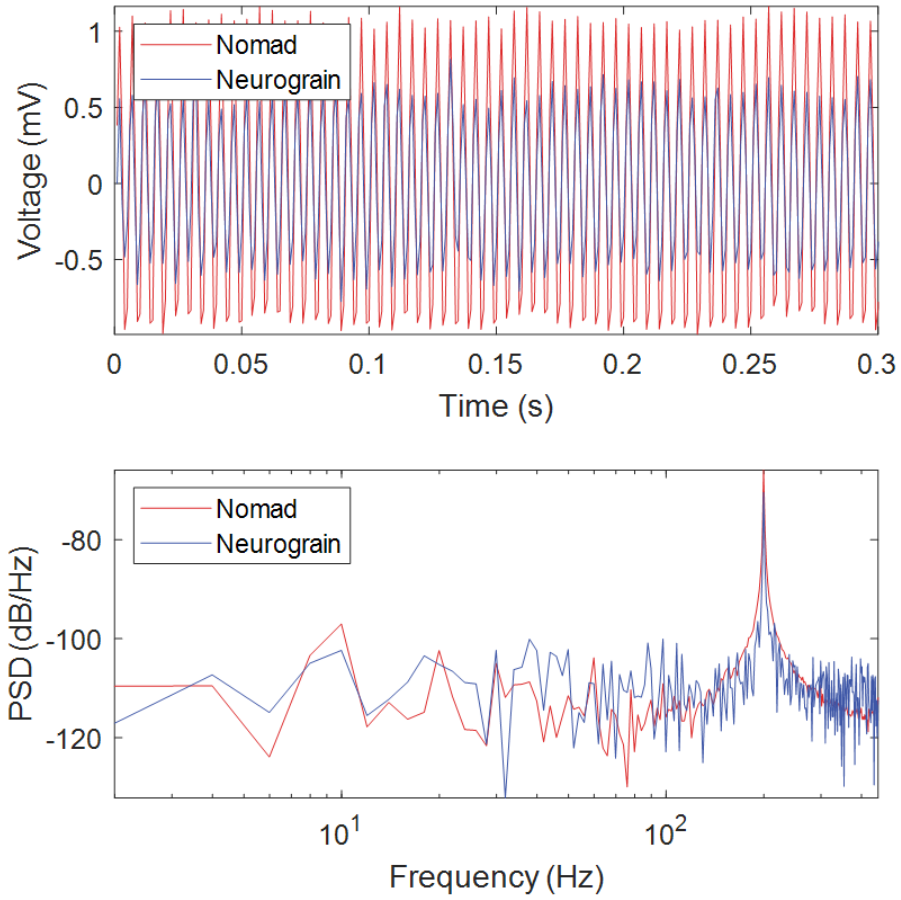

Figure 6-2. Response to a 200 Hz sinusoidal signal injected into the saline recorded by a neurograin and a wired commercial device (Nomad, Ripple Inc), respectively. Both the transient waveforms (top) and the nonparametric power spectral density (PSD) estimates (bottom) are shown. Compared to the commercial wired device, the recording quality of neurograin is understandable somewhat limited due to constraints on the circuit area and power. Still, the current version of recording neurograin can clearly capture low and high-frequency features of local field potentials.

To test whether neurograin can record physiologically relevant data beyond the sinusoidal waves, we performed another benchtop experiment and delivered previously found ECoG data into the saline through a function generator. The recording, shown in Figure 6-3, presents two waveforms from neurograin and the Nomad system, which successfully captured low and high-frequency components in ECoG data in both cases.

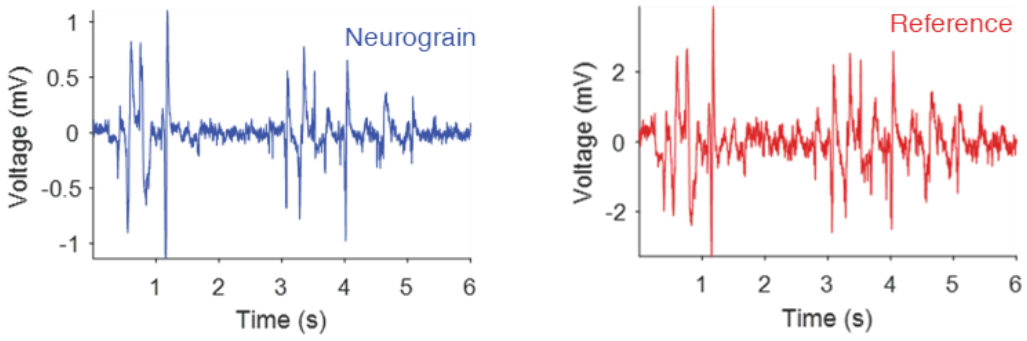

Figure 6-3. Recordings of ECoG-type signals injected into saline from neurograins (left) and the wired commercial system (Ripple Nomad) (right). The Nomad configuration used a distal ground electrode while neurograins measured a potential differential between the two on-chip gold electrodes.

Above, we mainly presented the recording from a single neurograin in the saline, whereas the networking of the neurograin is a critical requirement for the system. To demonstrate the networking performance, we built a 36 channel neurograin assembly testbed shown in Figure 6-4. Each neurograin had a pair of Au microelectrodes, post-process fabricated on the chips. In this experiment, we used a concentric tungsten electrode to generate a localized electric field above the array for each neurograin to register a signal, resulting in an amplitude distribution based on their location with respect to the current injection source. All of the 36 neurograins were embedded into PDMS along with the relay coil, magnetically coupled to the Tx coil underneath.

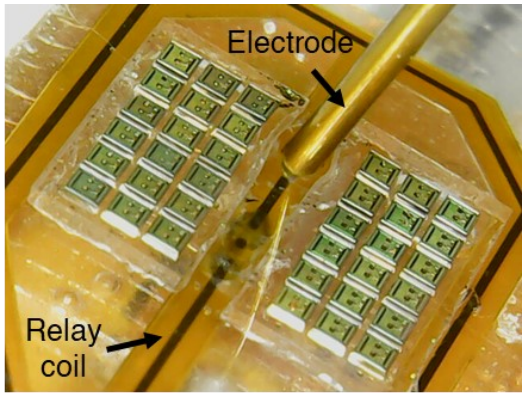

Figure 6-4. The photograph of the neurograin array in a saline testbench. The Tx and relay coils are located underneath the sensor array, for wireless powering and data communication.

Results from 24 channel recordings are shown in Figure 6-5 where the subplots are aligned based on their relative location in the 2D neurograin array. We placed the concentric electrode between the chip 9 and 13 and injected a 40 Hz sinusoidal wave through the saline media. Under autonomous TDMA, we could successfully collect recorded signals from 24 neurograin ensembles simultaneously and track the record data from each chip separately using their PUF ID. As expected, the amplitude of signals captured by neurograins was much higher around the signal source compared to that from neurograins far from the source electrode. However, even among the selected 24 channels, we observed that the noise levels on a few chips were unacceptably high. We carefully analyzed the noise source to find instability in the DC-coupled amplifier of neurograin. This instability may come from Au electrode impedance variations since the DC-coupling configuration makes the front-end amplifier susceptible to the impedance in the electrode-electrolyte interface. To improve the noise level, our group is now pursuing two approaches 1) developing an AC-coupling amplifier for neurograin and 2) improving post-processing by using multi-step electroless plating (not yet used for the recording neurograins).

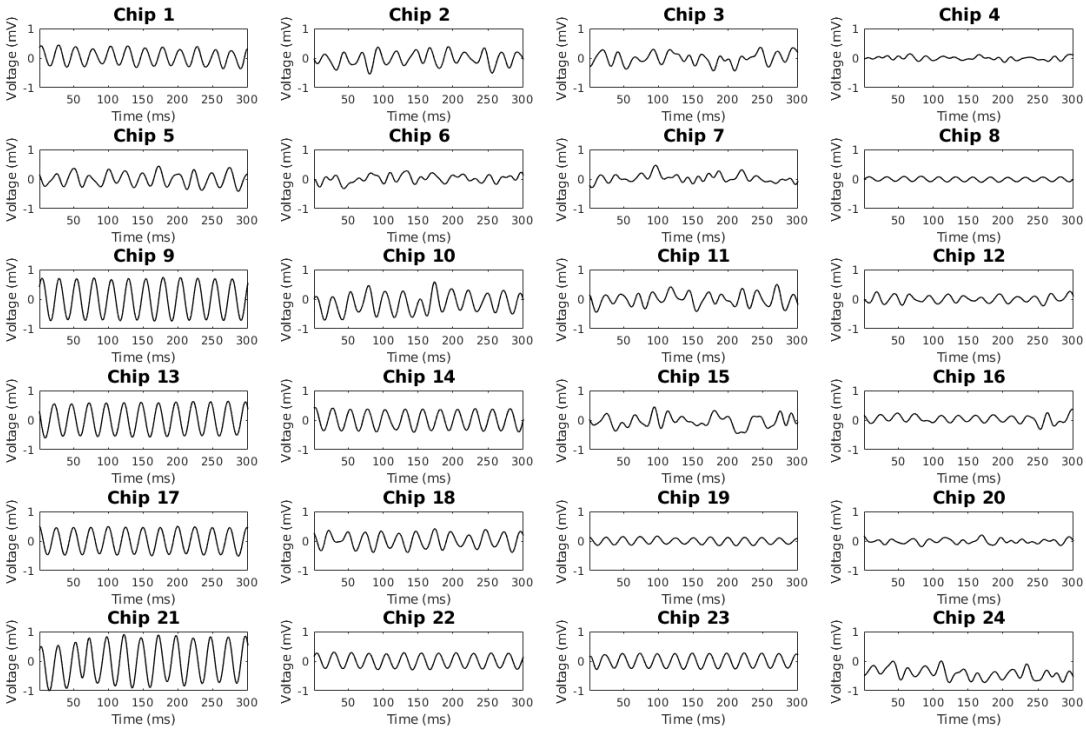

Figure 6-5. Recordings of 40 Hz sinusoidal waves by the distributed neurograin network in saline (only 24 channels are shown). The subplots are aligned based on their relative location in the 2D neurograin array. Here, a concentric electrode placed between chip 9 and chip 13 generated a localized 40 Hz signal, which has a higher magnitude around chip 9 and 13 and a lower magnitude around chip 4 and 8 due to the electrical field distribution.

Recordings for an extended period from 8 neurograins are shown in Figure 6-6 to give a sense of the stability of neurograin recording. Most of the time, the neurograins were able to report continuous waveforms so that we could identify chips by their address over 5 seconds. Occasionally. However, packets were missed at several points; some of these are labeled by the blue arrows in Figure 6-6. This type of discontinuity was observed because of packet collisions (detailed in Chapter 3), a feature of the current version of recording neurograin operating under the autonomous TDMA networking protocol.

Ongoing work involves implementing the full call-and-response TDMA approach (Chapter 3) to avoid these packet collisions and thus to reduce the data losses.

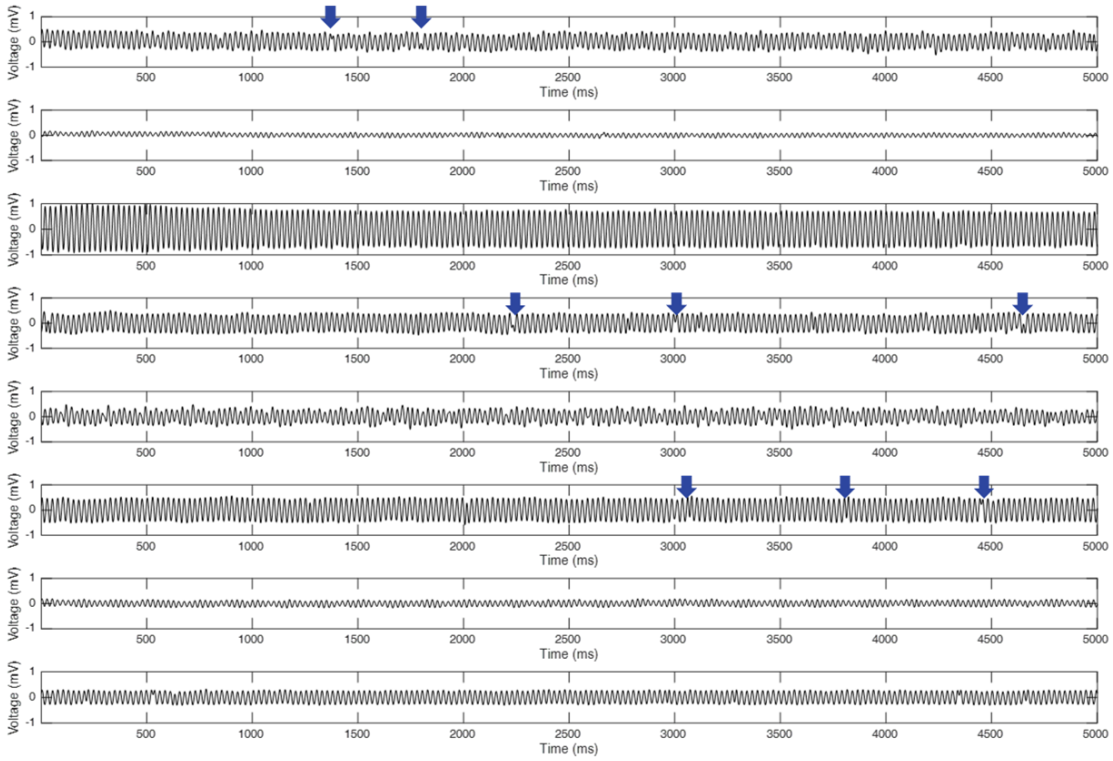

Figure 6-6. A sampling of 8 channel recordings from a small ensemble of neurograins over a 5 second period. Some of the missing data packets are labeled with blue arrows.

6.2.2. Demonstration of microstimulation by neurograins in saline

After attaching tungsten electrodes to the 'stimulation neurograins', we tested the chips in the saline benchtop setup shown above in Figure 6-1, to validate biphasic current delivery. The stimulation events from the chips were initiated by transmitting the previously discovered device ID via a downlink command. Simultaneously, we monitored the current output from neurograin chips by employing separate wired microwire electrodes and the Nomad commercial system. Figure 6-7 presents the recorded waveforms of electrical microstimulation from the chips in the saline, demonstrating that neurograins

successfully delivered reproducible stimulation into the saline following address-specific commands via the downlink. The actual stimulation waveforms were reshaped by filtering of the capacitive electrode-electrolyte interface so that the biphasic stimulation waveform in the media was "RC" distorted, as shown in Figure 6-7. Most importantly, however, the result validates the correct operation of our ASK-PWM downlink, the current injection circuit in the ASIC, and the pulse width remote programmability.

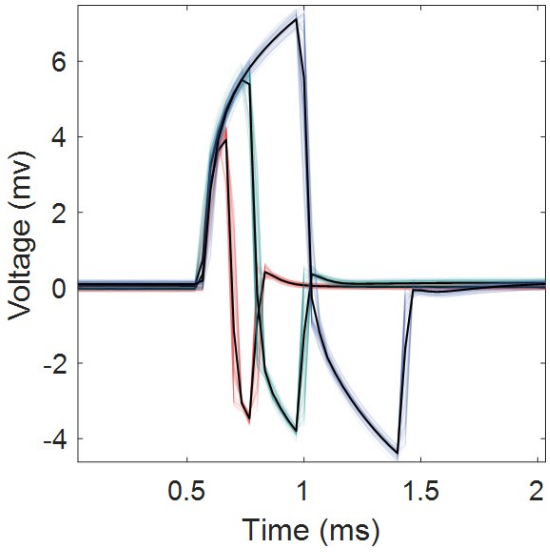

Figure 6-7 Superimposed biphasic stimulation pulses and their averaged waveforms (black) in saline as a function of the pulse width (red, green, and blue; 100, 200, 400 us per phase, respectively).

As already pointed out, the networking of the stimulation-type neurograins is critical for wireless delivery of spatially patterned excitation in the cortical tissue. We tested 12 neurograins in the saline bath while monitoring the remotely communicated stimulation patterns using an auxiliary array of recording electrodes connected by wire to the Nomad neural signal processor. A downlink sequence was transmitted every 100 ms, triggering a 100 Hz burst of continuous stimulation from each chip sequentially. In Figure

6-8, the green triangles indicate the downlink artifacts captured by the neural signal processor. The 10 stimulation artifact events between downlink commands correspond to the 100 Hz stimulation event by the neurograins. It was clear that only a selected neurograin injected the current-by-command into saline for a designated period since the electrical distribution of the evoked potential can be estimated from the distance between a neurograin and the recording electrodes. Upon repeating 12 downlink sequences in a loop, we could check that our neurograin network can reproducibly generate the stimulation pattern.

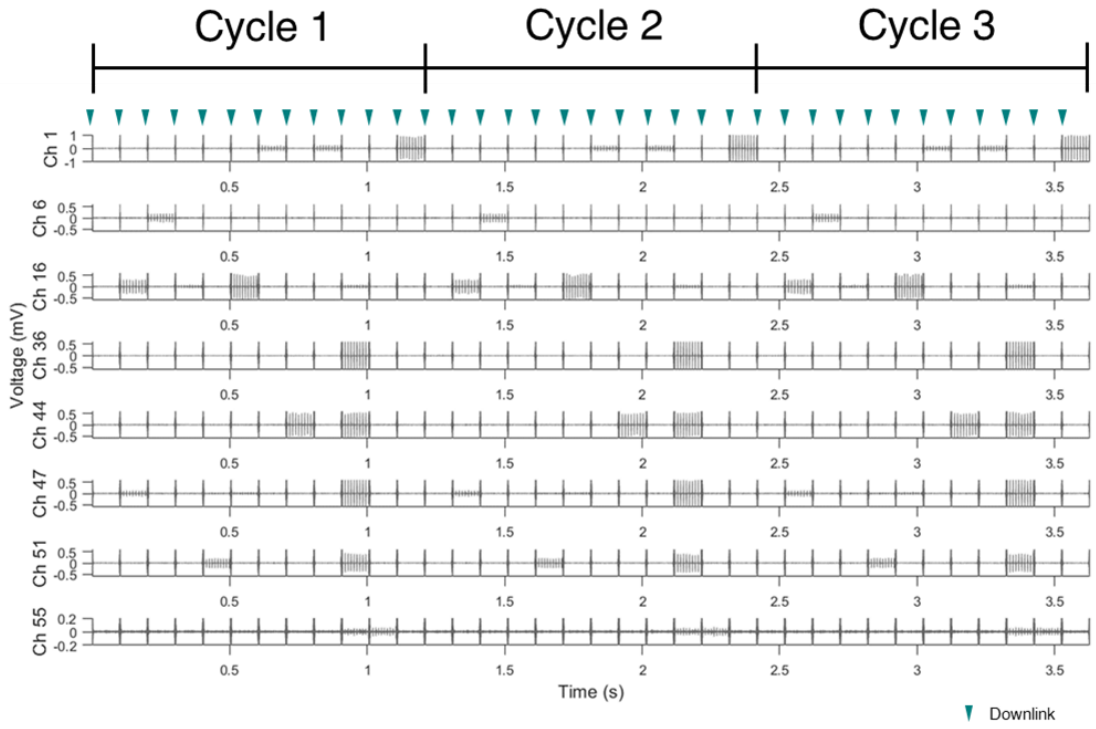

Figure 6-8. A spatial activation pattern of 12 Neurograins was recorded by 8 channels a conventional wired microelectrode sensor in saline. Each downlink triggered a specific neurograin to inject a current burst at 100 Hz. Accordingly, different stimulation patterns weres observed due to the distance dependence of the evoked potential by the stimulus to the recording electrodes. The figure demonstrates such wireless programming (indicated by downlink arrows) of stimulation patterns over 3 cycles.

6.3. In-vivo test of the neurograin ensembles in the rat model

Introduction

The validation of both recording and stimulation neurograins was performed separately using the in-vivo rat model. In experiments implanting recording neurograins, we deployed a 48 channel neurograin configuration with pairs of Au electrodes on chips. The ensemble of neurograins, encapsulated with PDMS along with the relay coil, was placed on the rat cortex to record epicortical signals, as shown in Figure 6-9 (left). The relay coil and the neurograin assembly were powered by the Tx coil shown in Figure 6-9 (right). The Tx coil was connected to the RF hub for wireless power (WPT) and data communication across the ensemble.

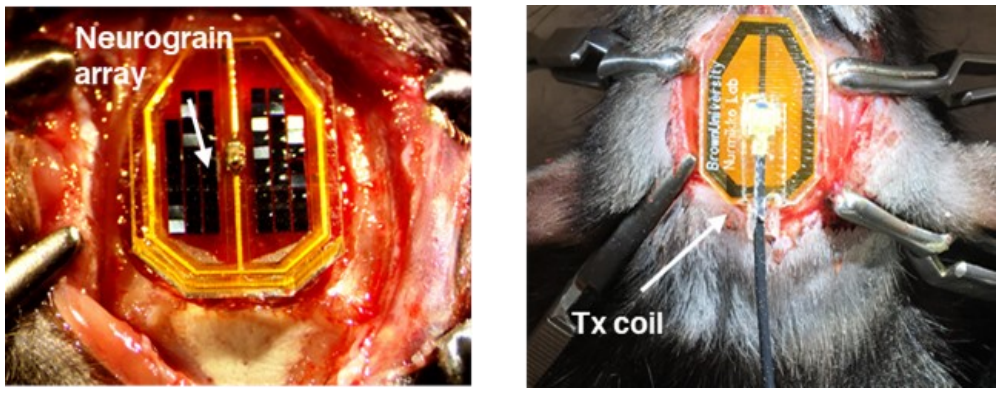

Figure 6-9. A photograph of the ensemble of recording neurograins encapsulated in Polydimethylsiloxane (PDMS) along with the relay coil mounted on a thin polyimide miniboard (left); the Tx coil was similarly mechanically supported by polyimide substrate (right).

In separate experiments for assessing the stimulation neurograin, we used a specially fabricated auxiliary test coupon assembly shown in Figure 6-10. The device consists of sets of wired recording electrodes, connected to the Nomad commercial neural signal processor, which were used to record the wirelessly evoked cortical stimulation by

the ensemble of neurograins. After assembling the recording test coupon onto a polyimide substrate, we placed the recording device in proximity of the stimulation neurograins. Since the modulation of neural circuits induced by subthreshold neurograin stimulation (I ~ 10 uA) was relatively weak if localized, we expected those recording electrodes needed to be placed near the stimulation sites to capture those localized responses. Penetrating intracortical microwires were fabricated on the neurograin ASICs as described in Chapter 5. The Nomad commercial neural signal processor, connected to wired electrodes, recorded intracortical signals to monitor the changes in the local field potential (LFP) and the firing rate in response to the neurograin stimulation.

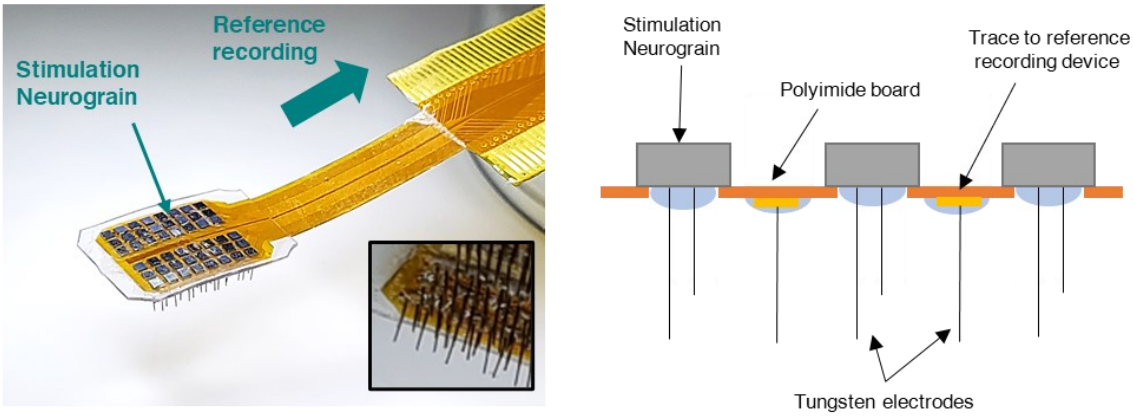

Figure 6-10. The ensemble of stimulation neurograins with post-processed penetrating intracortical electrodes and auxiliary wired recording electrodes for monitoring of evoked activity in the rat cortex. In this assembly, neurograins operate in a fully wireless manner, and the recording electrodes in proximity monitor spatial stimulation patterns and local neural activity to quantitively assess the effects of neurograin stimulation. Due to the manual electrode integration process, tungsten electrodes are not strictly perpendicular to the neurograin chiplet plane, as shown in the zoomed inset.

For the in-vivo experiments to date, the skull was not replaced, in part because the large craniotomy required to accommodate at least 48 neurograins made this impractical in the acute experiments. External wiring from auxiliary recording electrodes to the Nomad

commercial neural signal processor was also required to collect the reference data during the experiments. However, given the skull and the skin thickness of a rat (maximally 2.5 mm for each tissue layer), the associated RF attenuation is on the order of 2.5 dB, a minor effect on total losses when the skull and scalp are fully closed in the future experiments.

6.3.1. Animal surgery

We performed surgeries on wild-type Long-Evans adult male rats (500–700 g) in this study to test and demonstrate the neurograin system in-vivo. All research protocols were approved and monitored by the Brown University Institutional Animal Care and Use Committee (IACUC), and all research was performed in accordance with relevant NIH guidelines and regulations. Rats were anesthetized with 2–3% isoflurane throughout the surgery. After shaving the hair, the animal was fixed on a stereotaxic frame, and the head skin was sterilized. Up to 5 cm sagittal incision was made in the skin over the skull, and a bilateral craniotomy was performed using a precision surgery dental drill.

For neurograin implantation, up to a 4 mm × 10 mm craniotomy was made per each hemisphere, much larger than the typical craniotomy size. Drilling was done carefully, stepwise all around the target area until the thin layer of the skull left on all peripheral of the craniotomy. No vertical penetrating holes were made throughout the surgery. The skull was further thinned until it got removed, and the skull piece was lifted by a tweezer slowly. While lifting the skull piece, we used another tweezer to gently remove the connecting tissue below the skull to prevent damage to the dura. In case of a successful surgery, the dura was intact even with the large craniotomy, and could be cut the dura with a scissor.

The midline skull, between bilateral craniotomy, was thinned down to improve the contact of neurograin interface but not entirely removed since the large vein in the midline could easily be damaged. For implanting stimulation neurograins, insertion of the intracortical electrodes was done manually, and the ensemble of neurograins was fixed onto the skull with dental cement if needed. For interoperative recording and stimulation measurements, isoflurane was switched to 80 mg/kg Ketamine and 5 mg/kg Xylazine mixture, and the depth of anesthesia was monitored every 5 minutes by checking the breath rate and paw retraction responses. If needed, ¼ of the initial dosage was given to the animal to maintain a stable anesthesia level. A heating pad was placed below the animal to maintain the body temperature, and the saline was administrated on the cortex occasionally to prevent drying of tissue. Upon completion of the surgery, animals were euthanized with pentobarbital sodium.

6.3.2. Analysis of the data packets transmitted by neurograin ensembles in-vivo

In the initial assessment steps in-vivo, we transferred wireless energy to all deployed 48 neurograins and analyzed the received data packets after demodulation of the backscattered data. When demodulating packets, we could identify the 48 unique PUF addresses (Figure 6-11) being repeatedly transmitted. The unique address on each chip ensured continuously identifiable neural data flow from each recording neurograin, verifying that reliable communication with 48 channels was established.

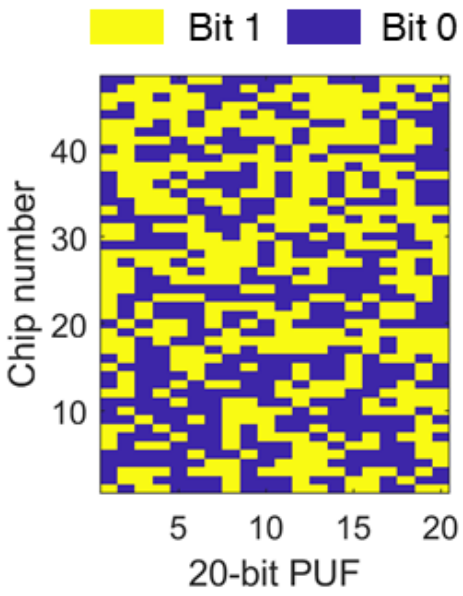

Figure 6-11. The retrieved PUF addresses, each 20-bit long, from an ensemble of 48 recording neurograins.

As one example, we detected 1697 data packets for 3.3 seconds, sorted them by their addresses, and solved the mean clock frequency of each chip as shown in Figure 6-12. Due to packet collisions and clock frequency variance, packet counts for each chip were not identical in all chips. Nevertheless, none of them showed packet counts less than half of the maximum possible value, consistent with all addresses being unique for each chip. Note that the clock frequency ranges from 28 to 32 MHz across this ensemble due to the dependence on received RF power (Chapter 3).

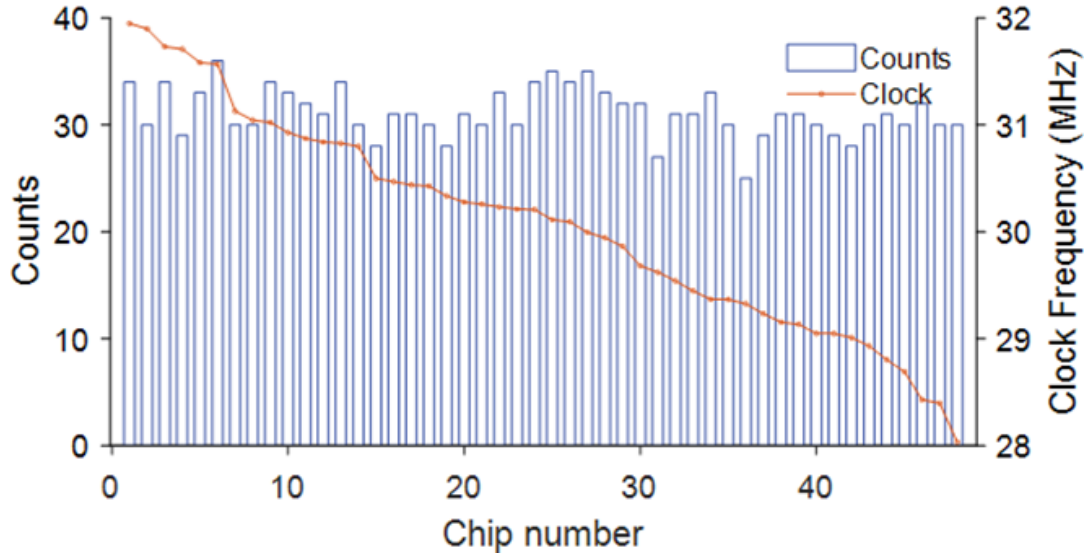

Figure 6-12. Packet counts and clock frequencies sorted by distinct PUF address across the 48 chips in the neurograin ensemble implanted in the rat cortex; measured over 3.3 seconds.

Figure 6-13 shows the histogram of the number of different PUF bits detected between any two chips. We used all 48 addresses in Figure 6-11, found the corresponding 1128 pairs of addresses, and calculated bit differences between each of them. If PUF addresses were determined randomly during the CMOS process and had an equal chance to be 1 or 0, one would see a normal distribution and an average of 10 bits, half of the 20 PUF bits on the recording chip. Figure 6-13 reveals that this was the case since it reveals the normal distribution with an average of 10 bits. While the distribution does not strictly follow the normal distribution, we expect that one will see a normal distribution if a larger number of PUF addresses is tested. Based on such observations, we conclude that as expected from basic principles, the 20-bit PUF address was random across the ensemble of neurograins, yet stable over time.

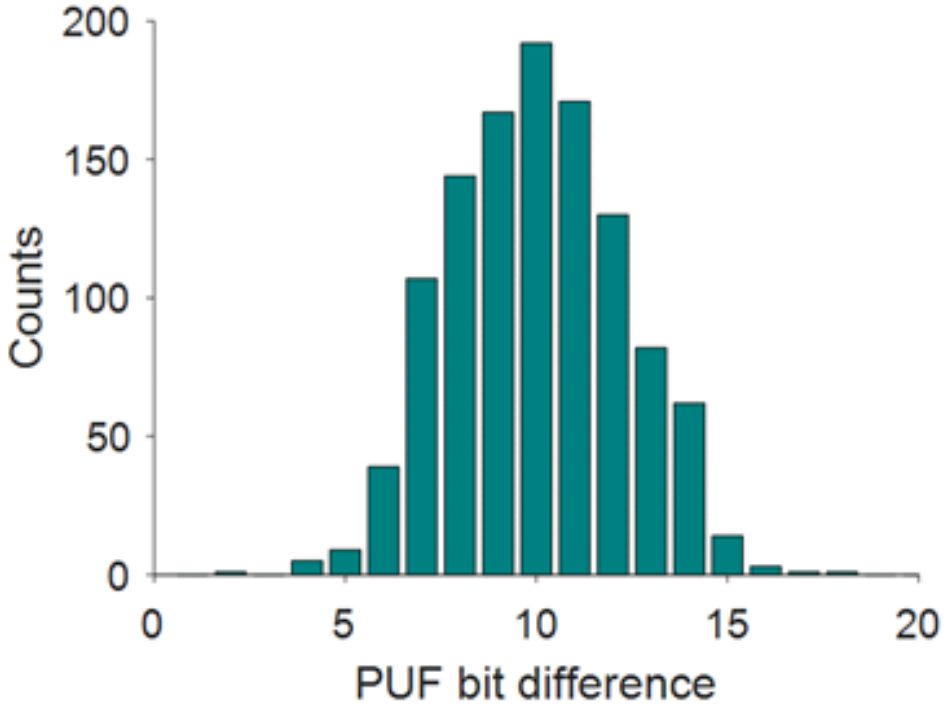

Figure 6-13. The histogram describing the number of bit differences calculated in the corresponding 1128 ($_{48}C_2$) pairs of addresses (using all 48 PUF address)

6.3.3. Recording cortical signals by ensembles of neurograins in-vivo rat experiments

The ensembles of recording neurograins were tested on the rat cortex to record epicortical signals and to validate its recording capability as a distributed network. The Nomad commercial neural signal processor was also configured to record the corresponding ECoG signals. After the implantation, the animal was anesthetized under ketamine, and isoflurane was discontinued. In the absence of external stimulus, the neurograin devices were able to record spontaneous ECoG signals, prominently capturing the sub-Hz slow-wave oscillations characteristic of ketamine anesthesia (Figure 6-14). The characteristic low-frequency oscillation was also observed by the commercial neural signal processor.

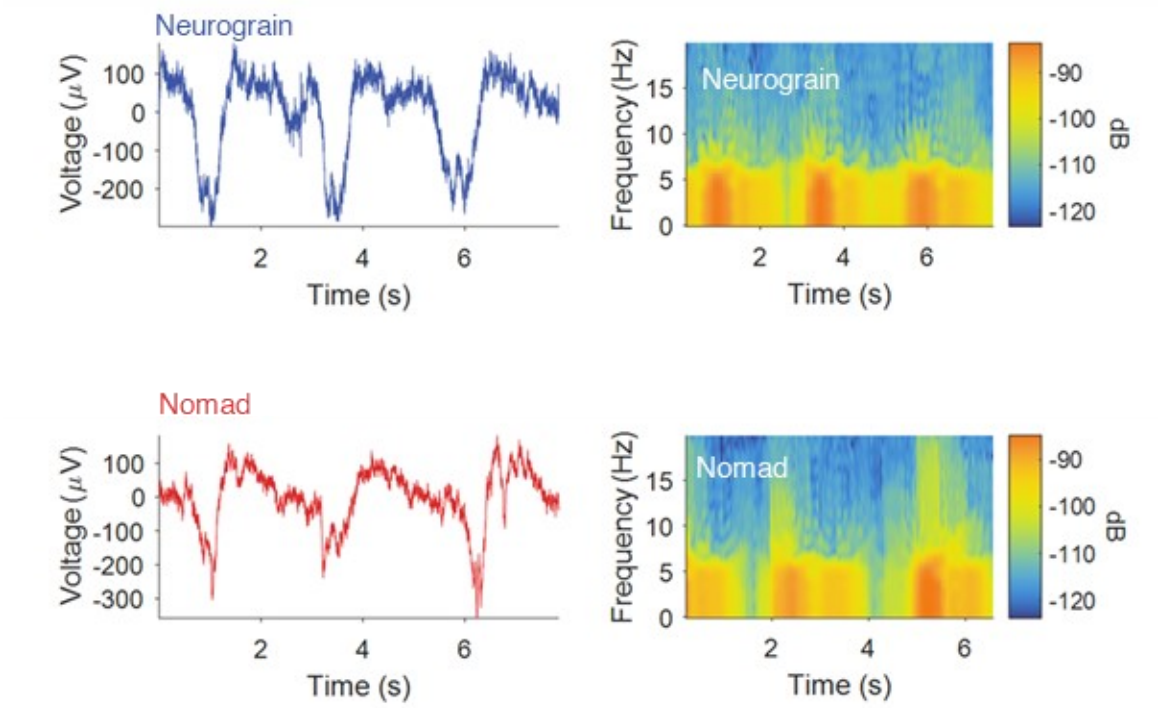

Figure 6-14 Wireless recordings by a neurograin ensemble and corresponding wired multichannel recordings by the commercial neural signal processor, of a spontaneous low-frequency oscillatory wave from anesthetized rat and their corresponding multi-taper spectrograms [96].

Figure 6-15 shows selectively recordings of spontaneous ECoG signals from the rat cortex by a subset of 12 neurograins, revealing a sub-Hz slow wave across the multiple cortical areas with a different amplitude and phase. In these experiments, while a subset of neurograins produced well-defined low-noise signals reflecting underlying spontaneous cortical activity, other neurograin sites showed higher noise levels due to possible combinations of neurograin placement, less than perfect electrode metal/tissue interfaces, and so on.

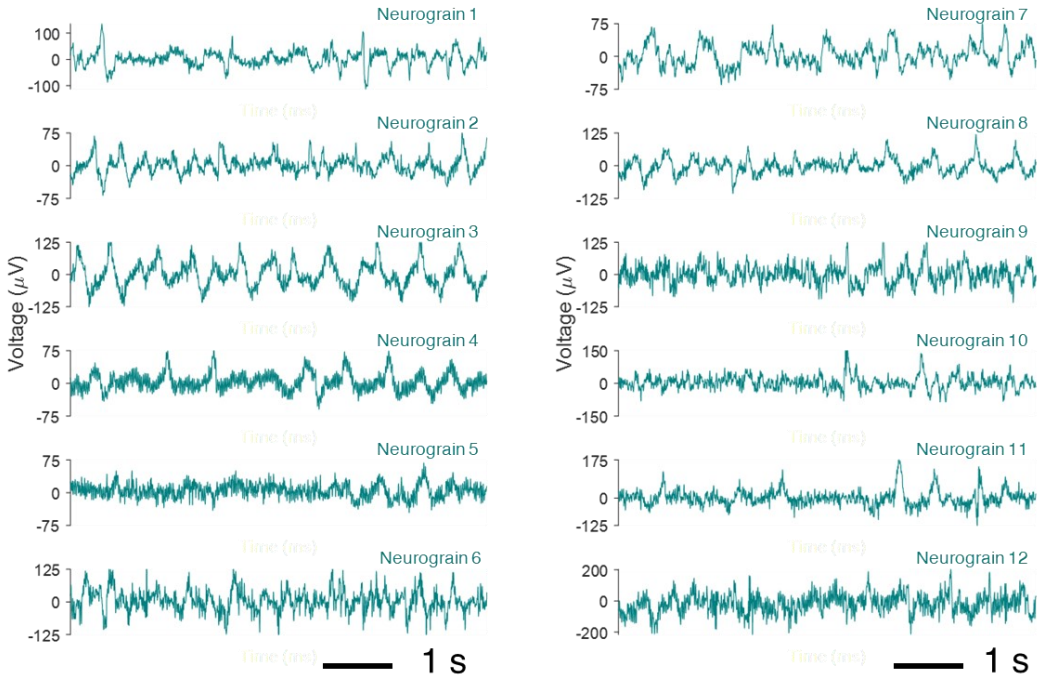

Figure 6-15. Multiple neurograin recordings of spontaneous ECoG under ketamine feature low-frequency oscillations. In these proof-of-concept experiments, we could reliably record low-noise brain activity from a fraction of the 48 channels as shown here for the 12 channels. Other channels were impaired by high noise levels due to imperfect electrode metal/tissue interface as well as practical constraints imposed by the finite size craniotomy.

After recording spontaneous cortical activities, we used the Nomad commercial system to deliver wires stimulus into the rat cortex at 5 Hz. The geometry involved a 6-channel intracortical stimulation using tungsten microwire electrodes while the ensemble of epicortical neurograins recorded wirelessly the evoked neural activity. The stimulation pulse width was set at 1 millisecond per phase, and a current up to 50 μA was delivered. An array of recording neurograins was placed on the animal's cortex, covering most motor and sensory areas in proximity to the implanted wired stimulation device. Figure 6-16 shows raw signal recording, which shows "stimulation artifacts" labeled with blue dots and

following field responses right after the stimulation. One can ask whether these field responses are real effects induced by the electrical stimulation or not. To evaluate that point, we have tested the stimulation both under ketamine and isoflurane. As shown in Figure 6-16, it is evident that the amplitude of the LFP response was much more prominent in ketamine compared to isoflurane, even if the amplitude of the "stimulation artifact" was comparable. This result suggests that the evoked LFP responses were indeed derived from the cortical tissue whereby our wireless neurograin system was able to capture the difference between responses under ketamine and isoflurane.

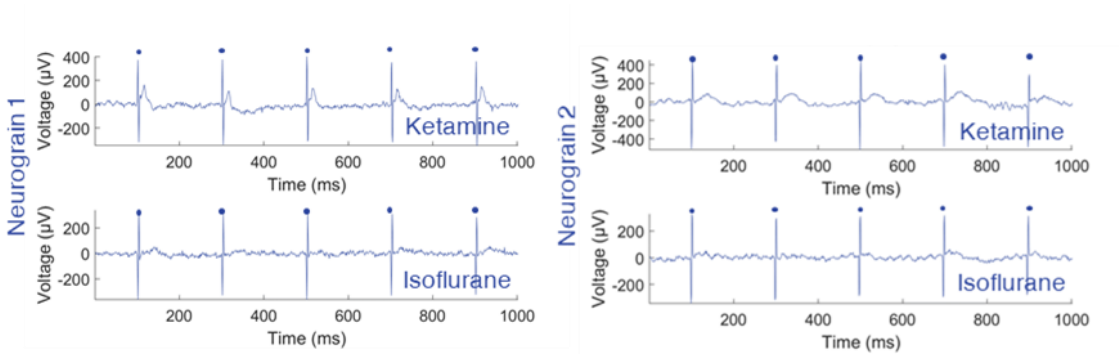

Figure 6-16. Recorded raw signals by a pair of neurograins from the wireless ensemble, in response to 5 Hz electrical stimulus (blue dots) showing evoked neural response under ketamine and 3% isoflurane. Stimulation was delivered intracortically while the neurograin ECoG-type recording was obtained epicortically.

Figure 6-17 shows a superimposed recording of stimulation artifacts and responses under ketamine and isoflurane on 4 neurograin chips. In this expanded view, we could see that post-stimulus responses under isoflurane were also noticeable but much weaker than that under ketamine. Notably, under ketamine, we were able to record distinct response patterns, such as the oscillatory response (left top), the fast depolarization (right top), and the slow hyperpolarization (bottom) while the duration of evoked responses ranged from 30 to 70 ms.

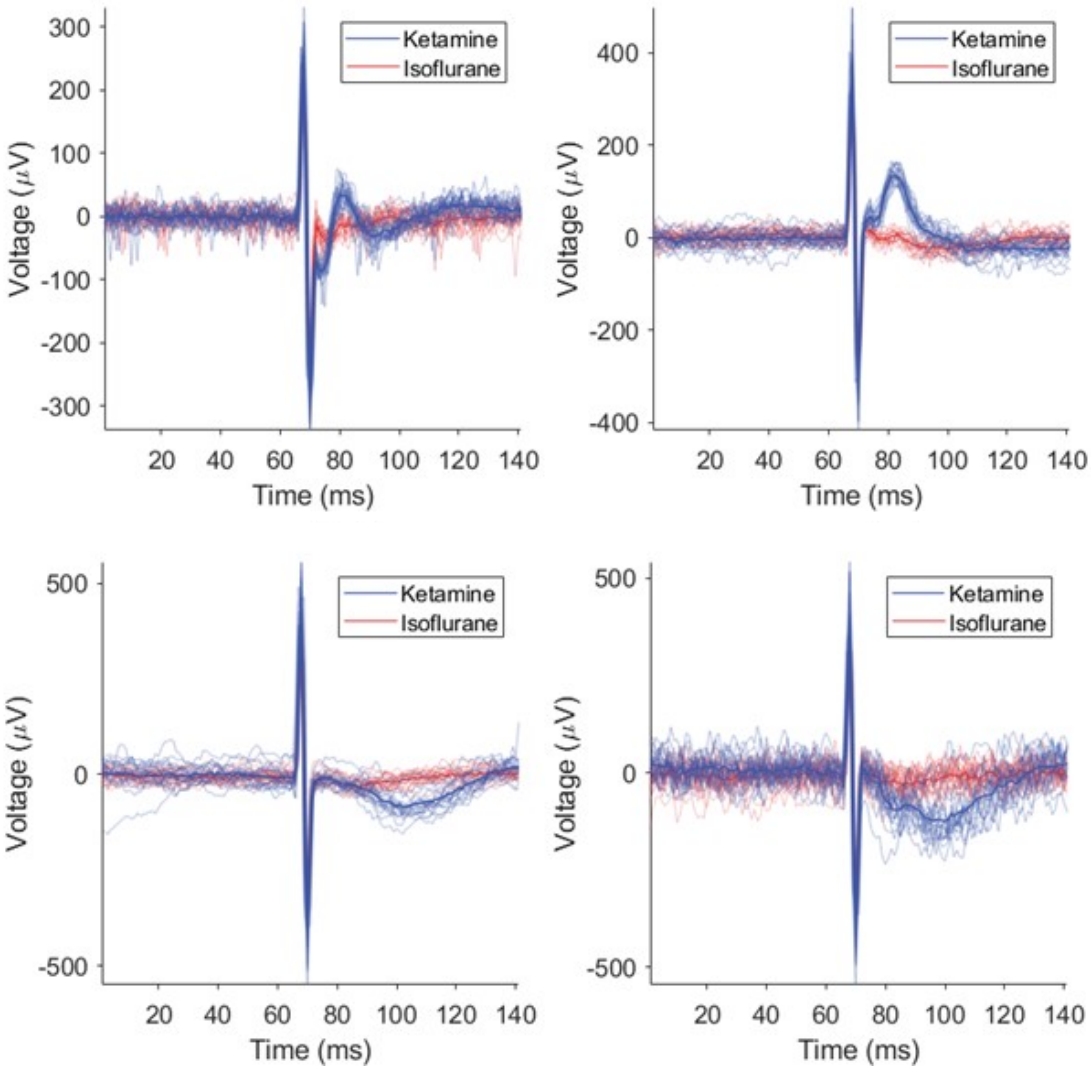

Figure 6-17 Superimposed raw ECoG signals over 20 stimulation trials showing evoked responses with 20 to 70 milliseconds duration. For this experiment, the stimulation was injected through intracortical electrodes while neurograin recorded ECoG signals in the epicortical space. The solid line shows averaged signals, and each subplot shows results from different neurograins. The electrical stimulation has an amplitude of 50 μA, and pulse width per phase is 1 ms. Evoked response after the stimulation has a higher amplitude under ketamine compared to that under isoflurane.

6.3.4. In-vivo stimulation delivered by ensembles of neurograins

In the last phase of the in vivo work, we deployed ensembles of intracortical stimulation neurograins, using the separate specially fabricated testbed assembly of Figure

6-10 above. The monolithic template assembled stimulation neurograins and a number of wired recording electrodes on the same platform connected to the Nomad commercial neural signal processor. By recording signals from several sites in proximity to the stimulating neurograins, we were able to monitor the evoked microstimulation patterns in the animal, such as in Figure 6-18. The downlink sequences, labeled by green arrows, were transmitted to initiate microstimulation and to stop it, respectively. Between the downlink commands, a given neurograin delivered 100 Hz microstimulation for 200 ms duration and the pulse width per phase of the stimulation was set to 400 μs. A subset of recording channels was able to capture stimulation artifacts and post-stimulus cortical responses. As shown in Figure 6-18, some of the channels even recorded the evoked responses while they could not capture stimulation artifacts. The results show that the wirelessly commanded neurograins actuate local neural network in vivo by electrical microstimulation, which appears to them propagate to other cortical areas. In cases where such a wider impact of stimulation is not desired, one might be able to use shorter pulse width to reduce this modulation effect on cortical circuits.

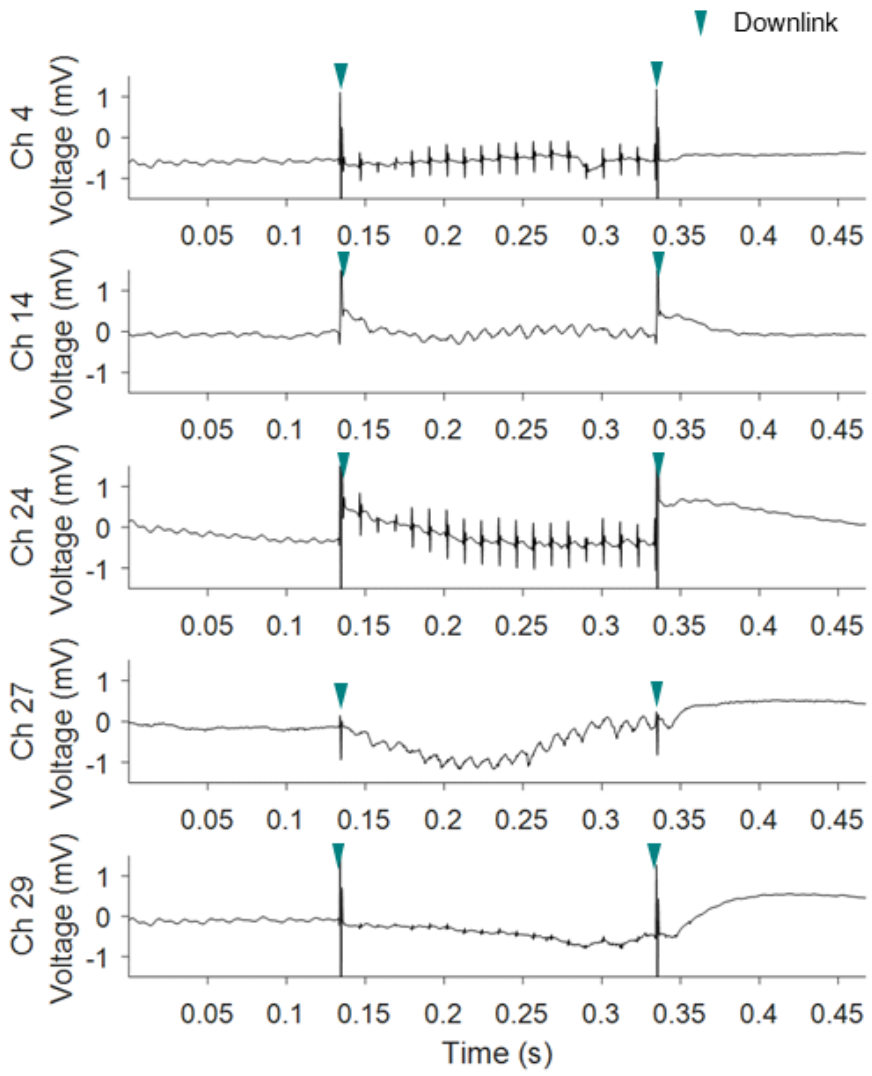

Figure 6-18 Cortical responses evoked by stimulation from a neurograin. Figure shows responses from 100 Hz neurograin wireless stimulation and evoked local field potential (LFP) responses over 5 wired recording channels.

The relationship between evoked neural responses and stimulation pulse width follows the well-known "strength-duration curve" [64, 65]. To control the modulation in the cortex, we built pulse-width programmability in stimulation neurograin and tested three pulse widths in-vivo. For that, we removed the stimulation artifacts and measured the amplitude of LFP responses for 40 trials in each pulse width. As expected, we observed

the amplitude of post-stimulus LFP responses being dependent on the stimulation pulse width, as shown in Figure 6-19.

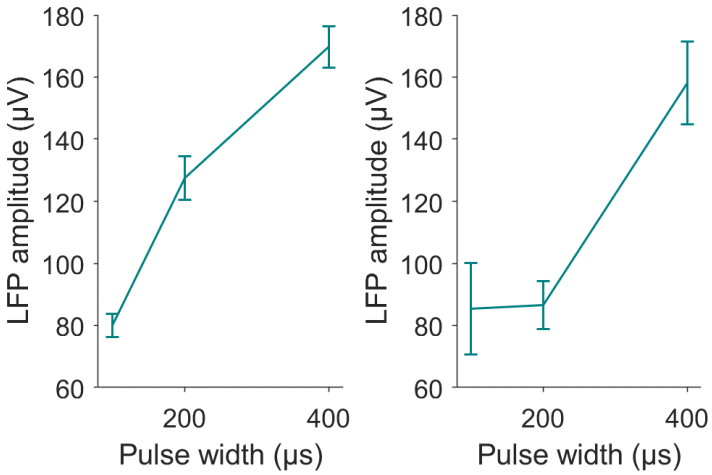

Figure 6-19. Amplitudes of post-stimulus LFP response and dependence on the stimulation pulse width (per phase) from two neurograins (left) and (right). Up to 25 μA current was delivered to the tissue from each device. The error bar represents the standard deviation (each consisting of 40 trials).

So far, we have shown results evoked by "single shot" wireless stimulation. For multiple neuromodulation purposes, the neurograin system can also deliver current in the form of high-frequency microstimulation pulse trains. To demonstrate this capability, we have delivered fast 400 Hz stimulation in 100 ms repeating blocks while recording evoked spiking activities. Under ketamine, the rat cortex typically shows spontaneous spike burst activities driven by its internal circuits as well as the low-frequency oscillation [97, 98]. We were able to induce such burst activities by delivering 400 Hz stimulation through a stimulating neurograin, as shown in Figure 6-20, regardless of the timing of previous burst activities.

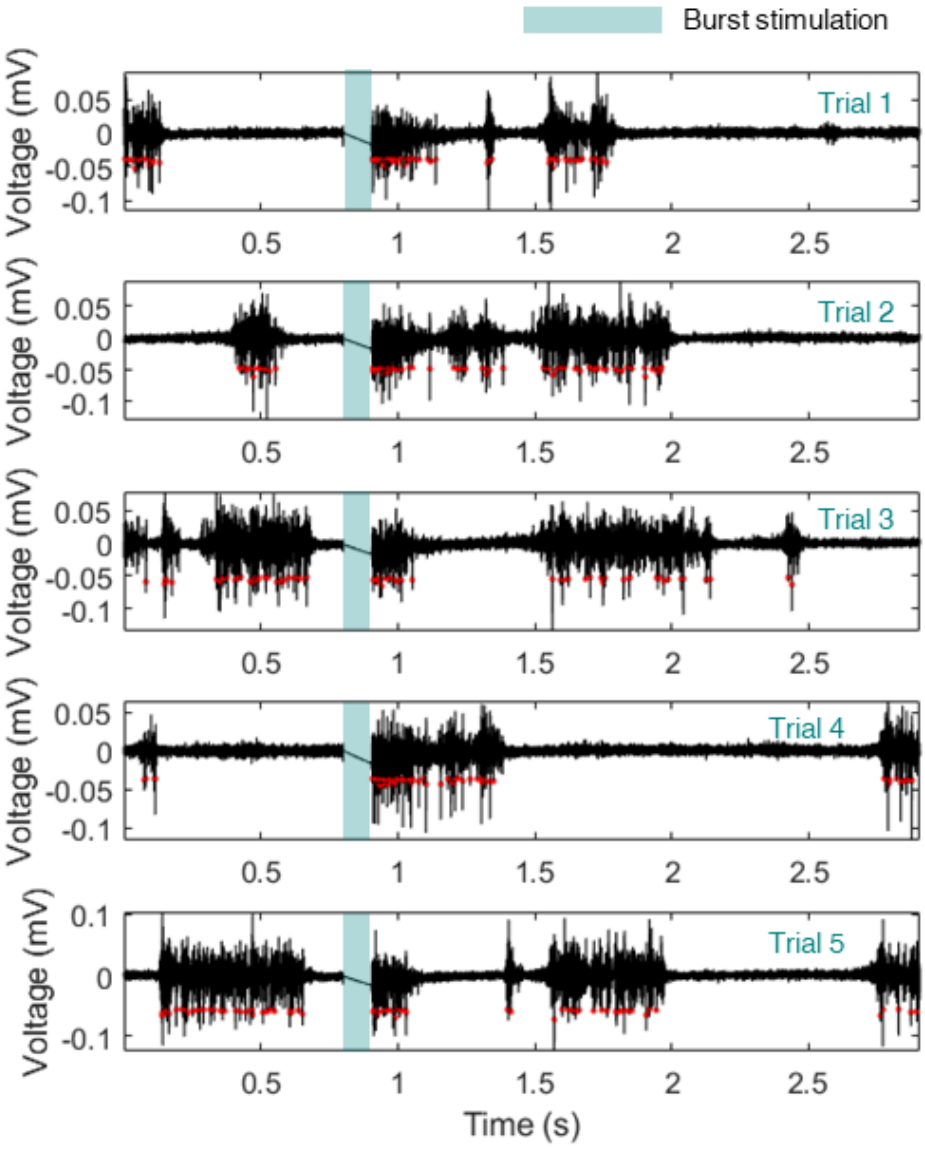

Figure 6-20. Spontaneous bursting firing activity under ketamine, artificially induced by 400 Hz stimulation delivered wirelessly by a neurograin for a 100 ms duration, recorded over 5 trials.

We evaluated the firing rate changes during such induced cortical burst activities. Figure 6-21 shows the sum of the firing rate across 3 channels (averaged over 5 trials), showing a significant firing rate increase following 400 Hz stimulation.

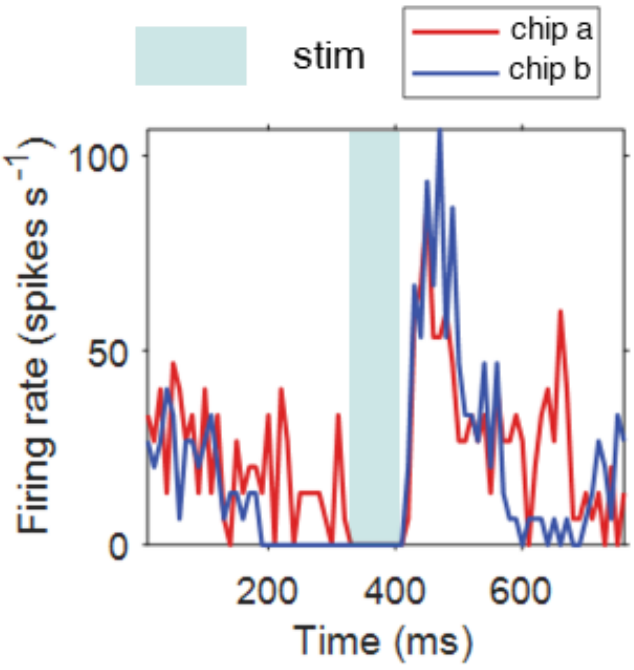

Figure 6-21. The sum of the firing rate from 3 recording channels before and after the 400 Hz burst stimulation from a neurograin showing the increased post-stimulus firing rates (averaged over 5 trials and the bin size is 10 ms).

6.4. Conclusion

In this chapter, we have demonstrated the translation of the core neurograin ideas to the benchtop and into in-vivo rat experiments. In so doing, we have validated key principles of operation of ensembles of neurograins: one-to-all networking, deployment for ECoG-type epicortical recording, and patterned intracortical microstimulation functionalities. In benchtop experiments, the neurograin ensembles successfully recorded electrical signals and delivered currents for electrical microstimulation purposes while validating the uplink and downlink operation for network communication of neurograin ensembles. Using the in-vivo rodent model, multinodal neurograins were able to record ECoG signals simultaneously or modulate neural activities delivering spatially patterned electrical microstimulations. When recording with ensembles of neurograins, we

monitored both slow-wave oscillation and post-stimulus oscillatory, depolarization and hyperpolarization responses. On the other hand, the neurograins configured for electrical microstimulation were able to evoke cortical LFP responses and induce burst activities through biphasic current excitation, by programmed remote control from the RF hub. In sum, to the best of our knowledge, this is the first demonstration of the multiple networked microsensors and microstimulator in-vivo, paving a way towards further development of spatially distributed wireless microimplants.

CHAPTER 7. CONCLUSIONS AND FUTURE DIRECTIONS

7.1. Assessing the status of wireless microscale neural implants: the case of spatially distributed neurograin sensors

We have described here what we believe are the first critical steps towards the realization of a potential new neurotechnology, namely a wireless implantable microsensor as a part of a wireless network of such devices. Each sub-mm-sized silicon chiplet (0.1 mm^3), a neurograin, combines RF energy harvesting, data communication, and recording or stimulation capability as an autonomous unit with its unique identifier. We have designed a custom time-division multiplexing approach to build a spatially distributed microsensor system that communicates with an external wireless hub. For proof-of-concept assessment of the first generation of ensembles of multifunctional chips, we performed benchtop and in-vivo experiments to demonstrate specific critical functions such as wireless power harvesting, near-field communication, and neural recording as well as patterned stimulation.

To recapitulate, we first discussed the integrated circuit principles in Chapter 2 and described the SoC ASIC design goals and specifications. One key feature is the neurograin RF backbone with its wireless power harvesting circuit specifically designed for low RF energy operation. The microimplant also leverages the high data rate uplink and downlink circuits for bidirectional communication, each device using its own on-chip device identifier. Two versions of neurograins chips were presented, for recording and stimulation, respectively.

We then described the approach to one-to-many networking protocols, adapting techniques from mobile communication to the sub-wavelength RF regime of neural implants. An autonomous TDMA protocol provides a relatively simple method to communicate with multiple chips only through the uplink but with the risk of packet collisions. The more complex call-and-response TDMA protocol requires a bidirectional communication link but is scalable to enable multinodal communication for up to 770 nodes. In early work, we measured the bit-error of these protocols to quantitatively analyze the link performance using a communication neurograin test chip with predefined bit sequences. These important test chips were also used to evaluate the wireless link efficiency as their clock frequency can be used as an indicator for on-chip power level. Then, in preparing the neurograin system for actual use, we developed scalable microfabrication post-processing techniques to build a fully functional biomedical implant from CMOS chips.

Lastly, we have demonstrated our system in-vivo rat model to record epicortical ECoG signals and intracortically stimulate neuronal microcircuits. We measured spontaneous slow-wave ECoG signals from the rat under ketamine and also recorded post-

stimulus evoked responses with recording neurograin. Using the stimulation version of the neurograin idea, we produced evoked local field responses as well as bursts of spike activities in the rat cortex. While the head size of the rodent model has limited us to 48 implanted neurograins in this work, the current networking approach has the capability to scale up to 770 nodes for future primate applications.

To provide the reader an additional perspective, Table 7-1 summarizes the performance of salient parameters of the neurograin approach and compares those with other state-of-the-art wireless microimplants in terms of wireless powering, data communication, recording and stimulation. It is important to recognize that, overall, the field in still at an early, largely exploratory stage. The neurograin system of this thesis is shown to have a considerably higher data rate to provide sufficient bandwidth for multi-node communication. Note that with the SoC design that integrates all circuit components and a coil on-chip, the present volume of Neurograin is approximately 0.1 mm^3, considerably smaller than other microdevices to date. We have elsewhere demonstrated that this volume can be reduced further to 0.01 mm^3 after mechanical thinning and encapsulation with conformal hermetic packaging based on atomic-layer deposition (ALD) [94]. In addition to their unique capability for wireless networking, the miniaturized chips were developed for dual-functionality offering recording and stimulation capability. Critically, because of the designs for ultralow-power operation, a neurograin's power budget is less than 30 μW.

Looking ahead, the electronics of the neurograin SoC circuits can be ported to further deep sub-micron CMOS process nodes at least in principle (such as the 22 nm CMOS node), reducing the chip volume further. This can provide a pathway to minimally

obtrusive implantation of large populations of devices while using e.g., recently developed intracortical insertion techniques [99]. Further, the bidirectional wireless communication approach described in this thesis lends itself to real-time 'adaptive sensing', i.e., focusing at any given time on those sub-populations of neurograins most directly engaged in information exchange with the underlying biological circuits.

	[30]	[41]	[42]	[53]	[32]	Current work
Number of node(s)	1	1	1	1	1	48 (up to 770)
Wireless Powering	1.85 MHz Ultrasonic	131 MHz RF 3-coil	1180 MHz RF 2-coil	433 MHz RF 3-coil	1.85 MHz Ultrasonic	915 MHz RF 3-coil
External Tx power (mW)	0.12	-	4000	-	19.4	Up to 500
Operation depth (mm)	8.8 (in-vivo)	-	4.6 (beef)	-	55 (ex-vivo)	5-8 (phantom)
Telemetry	Backscattering	IR-UWB	-	Backscattering LSK	Backscattering	Backscattering BPSK (TDMA)
Uplink data rate (Mbps)	0.5	0.8	-	0.205	-	10
Type	Recording	Recording	Stimulation	Recording	Stimulation	Recording/ Stimulation
Microimplants						
Technology	Discrete	350 nm CMOS+ discrete	130 nm CMOS	350 nm CMOS	65 nm CMOS	65 nm CMOS
IC Area [mm^2]	0.032	1.1	0.04	12.25	1	0.42
Total Volume (mm^3)	2.4	1	0.009	12	2.2	0.01-0.1 [1]
Energy harvester	Piezoelectric	Discrete coil	On-chip coil	Discrete coil	Piezoelectric	On-chip Coil
Power supply (V)	-	1.8	1.2	1.5	3	0.6-1
Power Consumption (uW)	<1	<300	8-38	92	65	<30
Recording						
Recorded signal	EMG	AP/LFP	-	LFP	-	LFP
Noise floor (μV_{rms})	180	3.78	-	1.8	-	2.2
Resolution (bits)	8	10	-	11	-	8
Bandwidth (kHz)	>30	10	-	0.825	-	0.5
Stimulation						
Max. current (μA)	-	-	46	-	400	10-25

Table 7-1. Comparison between neurograin and the state-of-art microimplants

7.2. Toward large scale high-resolution neural interface

A prime future research goal for the proof-of-concept work of this thesis is the development of a new brain-machine interface capability with electronic access to thousands of points across multiple cortical areas. Such an electronic interface in the brain, whether designed for neural recording or stimulation will require further miniaturization of devices towards 100-micrometer chiplet dimensions for minimizing cortical tissue displacement and associated potentially detrimental effects [48, 56]. While CMOS process nodes at 22nm and below are becoming available for researchers outside large commercial chip consumer companies, scaling neurograin circuits down in size by another order of magnitude brings on major challenges. As power and circuit area are already tightly constrained, further creative optimization of chip functionality is necessarily upon further miniaturization. Nonetheless, we envision that much smaller neurograins, fully implanted in the cortex, can be developed e.g.for neural spike (action potential) detection to realize a wireless network of thousands of microscale nodes, such as depicted in Figure 7-1.

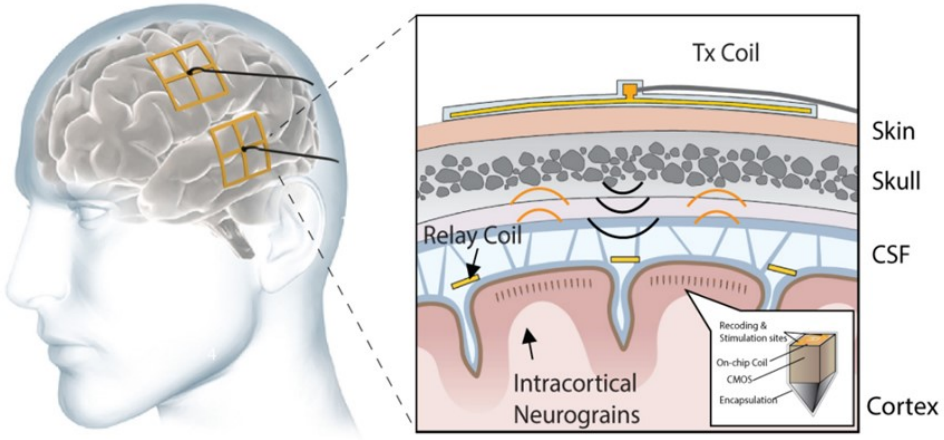

Figure 7-1. The concept of a future intracortical brain-machine interface configured as an ensemble of wireless neurograin sensors using transcutaneous RF power delivery and telecommunication data link.

CHAPTER A. NEURAL POPULATION RECORDING AND DECODING FROM THE AUDITORY CORTEX OF NON-HUMAN PRIMATES

A.1. Introduction

Electrophysiological mapping by single intracortical electrodes has provided much insight in revealing the functional neuroanatomical areas of the primate auditory cortex. Directly relevant to this work is investigating the role of the macaque secondary auditory cortex in processing complex sounds through intracortical microelectrode array (MEA) population recordings. Whereas the core of the secondary auditory cortex lies in the lateral sulcus, portions of the adjacent belt and parabelt lie on the superior temporal gyrus (STG) [100]. This area is accessible to chronic implantation of MEAs for large channel broadband recording. The prior seminal research into the STG has provided an understanding of the hierarchical processing of auditory objects by producing detailed maps of the cellular level characteristics through microwire recordings [100-105]. Results suggest that the secondary

auditory cortex contains distributed spectrotemporal representations in accordance with findings from microelectrode mapping of the macaque STG [103].

We also note relevant human research, which mainly deployed intracranial electrocorticographic (ECoG) surface electrode arrays in the STG [106-114]. While surface electrodes are thought to report neural activity over large populations of cells as field potentials, relatively accurate reconstructions of human and artificial sounds have been achieved in short-term recordings of patients during clinical epilepsy assessment. The recordings generally focus on multichannel low-frequency (0-300 Hz) local field potential (LFP) activity, such as the high-gamma band (70-150 Hz), using linear and nonlinear regression models. LFP decoding has been used to reconstruct intelligible audio directly from brain activity during single-trial sound presentations [107, 111]. When combined with deep learning techniques, electrophysiology data recorded by ECoG grids has been decoded to reconstruct English language words and sentences [115]. These studies provided insight into how the STG encodes more complex auditory aspects such as envelope, pitch, articulatory kinematics, spectrotemporal modulation, and phonetic content in addition to basic spectral and temporal content. One study augmented findings from ECoG studies by recording from an MEA implanted on the anterior STG of the patient where the single or multiunit activity showed selectivity to phonemes and words. This work provided evidence that the STG is involved in high-order auditory processing [116].

In this chapter, we investigated whether accurate low-order spectrotemporal features can be reconstructed from high spatial and temporal resolution MEA-based signals recorded in the higher-order auditory cortex (some of the results are presented in [8]). Specifically, we explored how different decoding algorithms affect reconstruction quality

when availed to large channel count neural population recordings of spikes. We implanted two 96-channel intracortical MEAs in the parabelt areas of the secondary auditory cortex in the rhesus macaque model and demonstrated the successful decoding of multiunit spiking activity to reconstruct intelligible English words and macaque call audio (Figure A-1).

Further, using a novel neural signal processing toolkit (NPT) [117, 118], we demonstrated the effects of decoding algorithm, neural preprocessing, audio representation, channel count, and array location on decoding performance by evaluating thousands of unique neural decoding models. Through this methodology, we achieved high fidelity audio reconstructions of English words and macaque calls through the successful decoding of multiunit spiking activity.

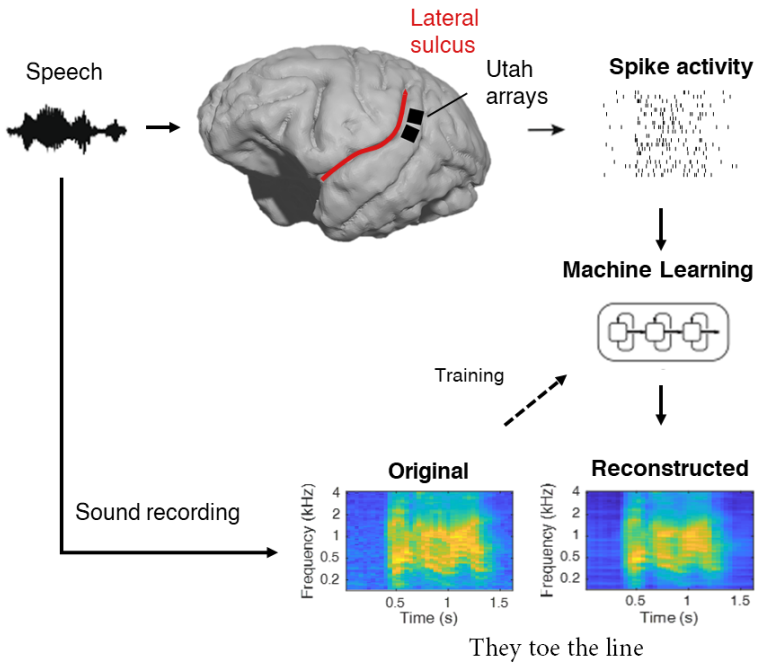

Figure A-1. The schematic for neural decoding with machine learning model aiming spectrogram reconstruction. An example of spectrogram reconstruction is shown as well.

A.2. Implantation of electrode arrays on auditory parabelt

A.2.1. MRI and CT imaging analysis

The rostral and caudal parabelt regions of STG have been shown to play a role in auditory perception [102, 119, 120]. Those two parabelt areas are closely connected to the anterior lateral belt and medial-lateral belt, which show selectivity for the meaning of sound ("what") and the location of the sound source ("where"), respectively [103]. To find this target area, we have used a 3D model built by combining images from magnetic resonance imaging (MRI) and Computed tomography (CT), as shown in Figure A-2.

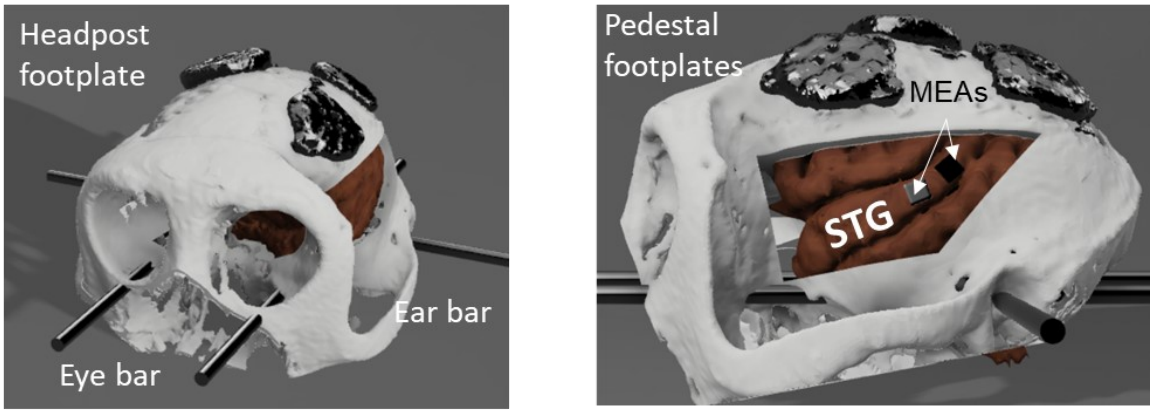

Figure A-2. 3D models for the skull, the brain, and footplates obtained by combining MRI and CT scan data.

Firstly, to get this 3D model, MRI for the animal's head has been taken before all surgeries, which provides 3D models of the skull and the brain. After that, we implanted pedestal footplates and titanium mesh (described in detail in the next section). Since we could not take MRI with these metallic structures, we took a CT scan after the surgery and obtained 3D models for metal and the skull structures. By combining MRI and CT scan and aligning them based on the skull model, we obtained 3D combined model shown

in Figure A-2. With this model, we used titanium mesh, pedestal footplates, lateral sulcus superior temporal sulcus, and central sulcus as reference locations to guide the MEA insertion on the target parabelt area during the surgery (Figure A-3).

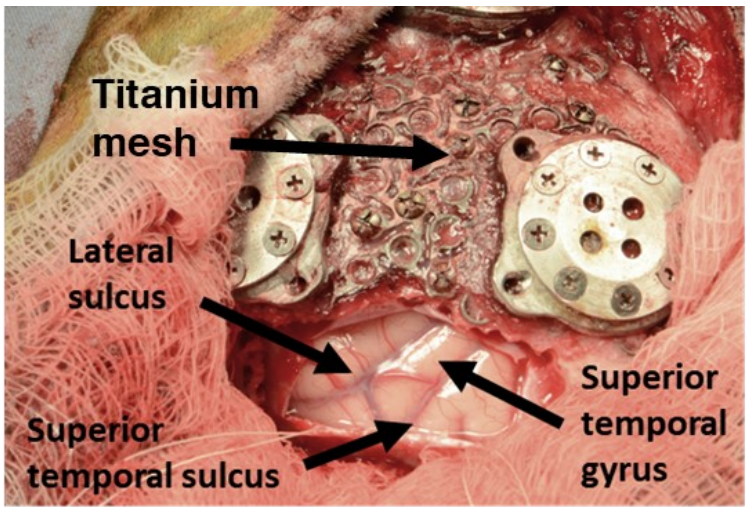

Figure A-3. Reference structures used to guide the MEA implantation on parabelt.

A.2.2. Implantation surgery

This work included two male adult rhesus macaques. The animals had two penetrating MEAs (Blackrock Microsystems, LLC, Salt Lake City, UT) implanted in STG, each providing 96 channels of broadband neural recordings. All research protocols were approved and monitored by Brown University Institutional Animal Care and Use Committee, and all research was performed in accordance with relevant NIH guidelines and regulations. Our institutional experience with implanting MEAs in NHPs has suggested that there is a non-trivial failure rate for MEA titanium pedestals [121]. To enhance the longevity of the recording environment, we staged our surgical procedure in two steps. First, we created a custom planar titanium mesh designed to fit the curvature of the skull. This mesh was designed using a 3D-printed skull model of the target area

(acquired by MRI and CT imaging) and was coated with Hydroxyapatite to accelerate osseointegration. This mesh was initially implanted and affixed with multiple screws providing a greater surface area for osseointegration on the NHP's skull.

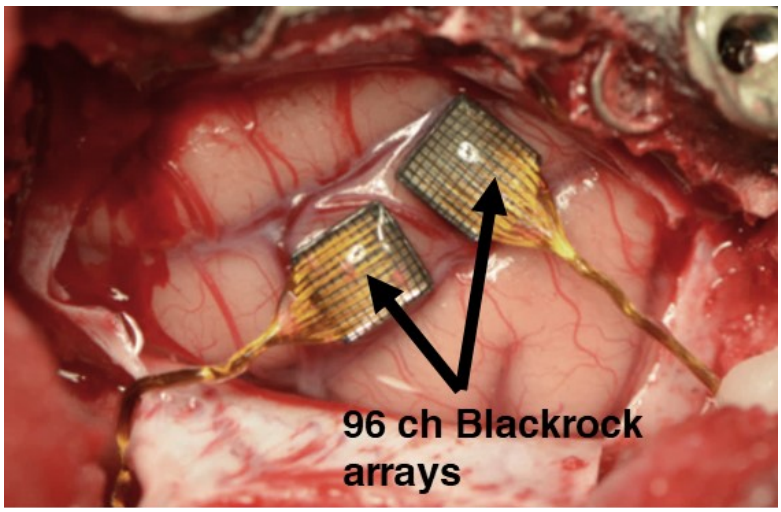

Figure A-4. Inserted two MEAs on the auditory parabelt.

Several weeks after the first mesh implantation procedure, we devised a surgical technique to access the parabelt region. A bicoronal incision of the skin was performed. The incision was carried down to the level of the zygomatic arch on the left and the superior temporal line on the right. One of the main challenges in this surgery is that the amount of temporal muscle in the rhesus macaques. Thick temporal muscle, which varied between animals, prevented an inferior temporal craniotomy. To get lower access to the skull base, we have to cut the temporal muscle in the plane perpendicular to our incision (i.e., creating two partial thickness muscle layers). After incision, we retracted the muscle anteriorly and posteriorly. This allowed us to have sufficient inferior bony exposure to plan a craniotomy over the middle temporal gyrus and the Sylvian fissure. MEA arrays were implanted with a pneumatic inserter [121], as shown in Figure A-4.

A.3. Data collection and preliminary assessment

A.3.1. NHP training and passive listening

We collected broadband neural data to characterize mesoscale auditory processing of STG during a passive listening task. While NHP subjects were restrained in a custom chair located in an echo suppression chamber using AlphaEnviro acoustic panels, complex sounds (English words sentences and a macaque call) were presented through a focal loudspeaker. The subjects were trained to remain stable during the training session to minimize audio artifacts. The passive listening task was controlled with a PC running MATLAB (Mathworks Inc., Natick, MA, Version 2018a). Animals were rewarded after every session using a treat reward. Within one session, around 30 stimuli representations were played at approximately 1 second intervals. Data from 5 to 6 sessions in total were collected in one day. Word sounds were chosen and synthesized using MATLAB's Microsoft Windows Speech API. Sentence sounds were obtained from The Texas Instruments/Massachusetts Institute of Technology (TIMIT) speech database. Trials that showed audio artifacts caused by NHP movement or environmental noise were manually rejected before processing.

A.3.2. Intracortical multichannel neural recordings

We used MEAs from Blackrock Microsystems with an iridium oxide electrode tip coating that provided a mean electrode impedance of around 50 kOhm. Iridium oxide was chosen with future intracortical microstimulation in mind. The MEA electrode length was 1.5 mm and 0.4 mm pitch for high-density grid recording. Intracortical signals were

streamed wirelessly at 30 kS/s per channel (12-bit precision, DC) using a CerePlex W wireless recording device [14, 122]. Primary data acquisition was performed with a Digital Hub and Neural Signal Processor (NSP) (Blackrock Microsystems, Salt Lake City, UT). The audio was recorded at 30 kS/s synchronously with the neural data using a microphone connected to an NSP auxiliary analog input. The NSP broadcasted the data to a desktop computer over a local area network for long-term data storage using the Blackrock Central Software Suite. Importantly, the synchronous recording of neural and audio data by a single machine aligned the neural and audio data for offline neural decoding analysis.

A.3.3. Neural preprocessing

We used multiunit spike counts or spike sorting as neural features for the neural decoders. We leveraged the Combinato [123] library or WavClus package [124] to perform spike extraction or spike sorting via a threshold crossing technique. Raw neural data was first scaled to units of microvolts and filtered using a 2nd-order bandpass elliptic filter with grid-searched cutoff frequencies. A noise level (i.e., standard deviation) of each channel was estimated over the training set using the median absolute deviation to minimize the interference of spikes [123, 124]. We set thresholds for each individual channel by multiplying the noise levels by a threshold factor. We detected negative threshold crossings on the full-broadband 30 kS/s data and then binned them into counts.

A.3.4. Audio preprocessing

We labeled raw audio to designate the begin and end time indices for single sound trials. The audio data contained English words, sentences, and macaque calls. We

subselected sounds to create target audio data sets for the neural decoders with varying complexity. Two methods were used for spectrogram calculation, 1) Mel-spectrogram and 2) auditory spectrogram based on cochlear filters. For Mel-spectrogram, the librosa audio analysis library [125] was used to calculate the short-time Fourier transform spectrogram with an FFT window of 2048. This spectrogram was then compressed to its Mel-scaled spectrogram to reduce its dimensionality to the number of Mel-bands. These Mel-bands served as the target data for the evaluated neural decoders. To solve the cochlear spectrogram, we used NSL tools [126].

Audio files were recovered from the Mel-bands by inverting the Mel-spectrogram and using the Griffin-Lim algorithm [127] to recover phase information or from cochlear spectrogram using NSL tools. All targets were standardized to zero-mean using scikit-learn transformers fit on the training data.

A.3.5. Initial assessment on parabelt spikes

We performed an initial assessment on parabelt signals before neural decoding experiments. We presented various single tone sounds to NHPs and checked the frequency selectivity of neurons, and example data from a neuron is shown in Figure A-5. The distribution of inter-spike-interval (ISI) of this neuron showed a common neural spiking pattern (Figure A-5 left). When we analyze the firing rate, we could find this neuron responded selectively to 6 kHz singe tone sounds. This result shows that neurons in the parabelt region can show frequency selectivity, which complies with previous studies [103].

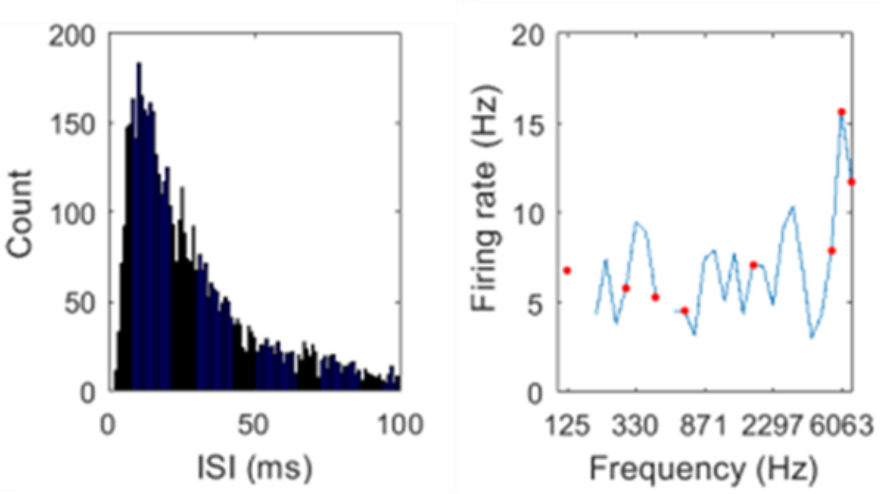

Figure A-5. (left) inter-spike intervals of the neuron in parebelt. (right) Firing rate of the neuron depending on the sound frequency of a single tone.

On the other hand, our passive listening presented English words and sentences. To check whether these sounds can activate the cortical circuit in NHP, we first solve the firing rate, presenting those sounds over 20 trials each. As shown in the example in Figure A-6, the firing rate on the neuron shows a significant increase while presenting English word sounds, which was comparable to the response observed during macaque call presentation. Particularly, this neuron was more sensitive to the onset of the sound. The results of this experiment showed that English sounds could evoke a noticeable response on the parabelt, which is a prerequisite for our decoding work described below.

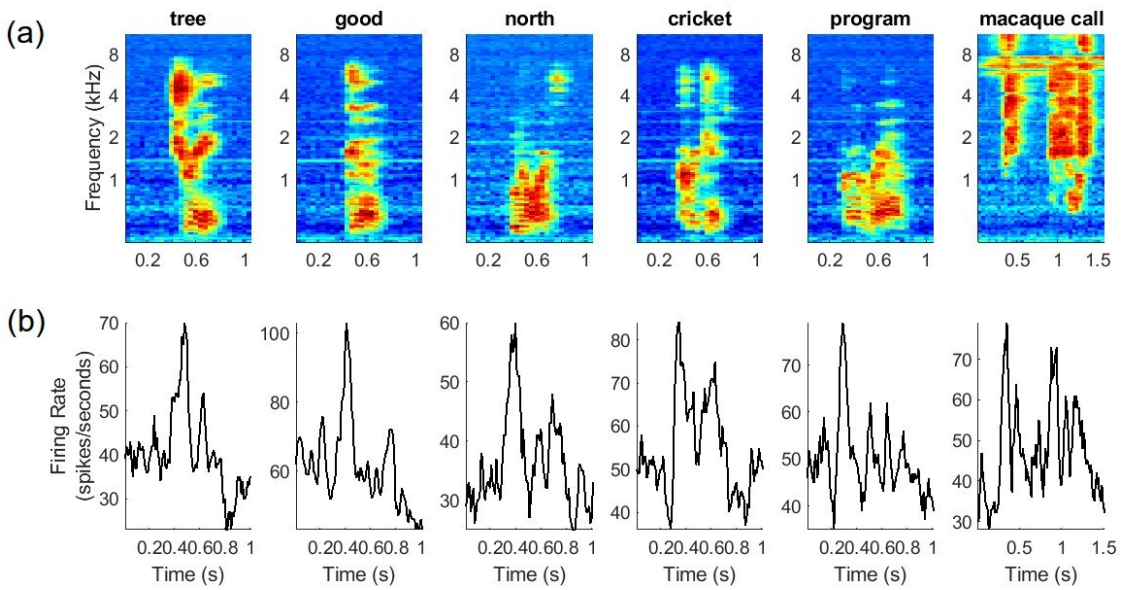

Figure A-6. Neural data from the RPB array. (a) Mel-spectrograms (128 bands) of five English word sounds and one macaque call. (b) Histogram for multiunit spiking activity on a given recording channel. In all plots, the spectrogram hop size is 40 ms and the window size for firing rates is 10 ms. Note the different scales of the horizontal axes due to different sound lengths.

A.4. Machine learning decoding

A.4.1. Decoding model comparison metrics

We utilized 6 different performance metrics for assessing and comparing the performance of neural decoding models. We selected metrics to investigate the reconstruction of different audio aspects and to quantify reconstruction speech intelligibility. For each Mel-spectrogram frequency band, Pearson's correlation coefficient (CC) was calculated between the target and predicted values. Fisher's z-transform was applied to impose additivity on the CCs before taking the mean of the transformed correlations across frequency bands. The inverse z-transform was applied to the resulting mean value to calculate the reported correlation values. This process followed previous

auditory cortex decoding work [107] and provided a spectral accuracy metric domain for audio reconstructions.

Temporal envelope

The temporal envelope of human speech includes information in the 2 to 50 Hz range and provides segmental cues to the manner of articulation, voicing, vowel identification as well as prosodic cues [128]. We found the temporal envelope of the target and reconstructed audio by calculating the magnitude of the Hilbert transform and low-pass filtering the results at 50Hz [129]. We then calculated envelope correlation by finding the Pearson correlation between the target and reconstructed envelopes.

Gross Pitch Error (GPE)

Gross Pitch Error (GPE) represents the proportion of frames where the relative pitch error is higher than 20% [130]. The Pitch of the target (F_0) and reconstructed audio ($F\hat{}_0$) was found with the Normalized Correlation Function method [131] through the MATLAB pitch function. This resulted in an estimation of the momentary fundamental frequency for the target and reconstructed audio. GPE was then calculated using the following formula:

$$GPE = \frac{N_{error}}{N} * 100\%$$

Equation A-1

where N is the total number of frames, and N_{error} is the number of frames for which $|F_0 - F\hat{}_0| > F_0 * p$ with $p = 0.2$.

Momentary loudness

Momentary loudness for the target and reconstructed audio was calculated in accordance with the EBU R 128 and ITU-R BS.1770-4 standards using the MATLAB loudnessMeter object. This resulted in loudness values with units of loudness units relative to full scale (LUFS) where 1 LUFS = 1 dB. For the purpose of comparing overall loudness reconstruction accuracy across de- coding algorithms, we calculated the absolute value of the momentary decibel difference between the target loudness (l) and reconstructed loudness (l^). By utilizing the absolute value of differences between the target and reconstructed loudness, this process equally penalized reconstructions that were too soft or too loud (analogous to Mean Absolute Percent Error). We then calculated a momentary loudness factor (LF) assuming a +10 dB change in loudness corresponds to a doubling of perceived loudness [132] using the following equation:

$$LF = 2^{\frac{|l-\hat{l}|}{10}}$$

Equation A-2

We reported the mean of the momentary loudness factor over time as the overall loudness accuracy metric where values closer to 1 represent better loudness reconstruction.

Extended Short-time Objective Intelligibility (ESTOI)

The ESTOI algorithm estimates the average intelligibility of noisy audio samples across a group of normal-hearing listeners [133]. ESTOI calculates intermediate intelligibility scores over 384 ms time windows before averaging the intermediate scores over time to calculate a final intelligibility score. Intermediate intelligibility scores are found by passing the target (i.e., clean) and reconstructed (i.e., noisy) audio signals through

a one third octave filter bank to model signal transduction in the cochlear inner hair cells. The subband temporal envelopes for both signals are then calculated before performing normalization across the rows and columns of the spectrogram matrices. The intermediate intelligibility is defined as the signed length of the orthogonal projection of the noisy normalized vector onto the clean normalized vector.

Spectro-Temporal Modulation Index (STMI)

STMI is a speech intelligibility metric that quantifies the joint degradation of spectral and temporal modulations resulting from noise [134]. STMI utilizes a model of the mammalian auditory system by first generating a neural representation of the audio (auditory spectrogram) and then processing that neural representation with a bank of modulation selective filters to estimate the spectral and temporal modulation content. Here, we utilized the STMI specific to speech samples ($STMI^T$) which averages the spectro-temporal modulation content across time and calculates the index using the following formula:

$$STMI^T = 1 - \frac{\|\mathbf{T}-\mathbf{N}\|^2}{\|\mathbf{T}\|^2}$$

Equation A-2

where T is the true audio (i.e., target) and N is the noisy audio (i.e. reconstructed). This approach captures the joint effect of spectral and temporal noise, which other intelligibility metrics (e.g., Speech Transmission Index) fail to capture.

A.4.2. Statistics and reproducibility

Statistics and Reproducibility For testing the significance of differences in spectrogram mean Pearson correlation, we followed a procedure described by Paul (1988) [135]. We first applied Fisher's z-transform to the correlation values to normalize the underlying distributions of the CCs and to stabilize the variances of these distributions. We then tested a multisample null hypothesis using an unbiased one-tailed test by calculating a chi-square value using the following equation:

$$\chi_P^2 = \sum_{i=1}^{k} \frac{n_i * (r_i - r_w)^2}{(1 - r_i r_w)^2}$$

Equation A-3

with k−1 degrees of freedom where n_i is the population sample size, r_i is the population CC, and r_w is the common CC. The resulting values rejected the null hypothesis, and we then applied a post-hoc Tukey-type test to perform multiple comparisons across decoding algorithms and hyperparameters.

For all other performance metrics, we first split the reconstructed validation audio into non- overlapping blocks of 2 seconds in length. A given performance metric was then calculated for each block, and a Friedman test was performed to test the multisample null hypothesis. For all evaluated metrics, this process rejected the null hypothesis, and we then applied Conover post-hoc multiple comparisons tests to calculate the significance of differences between decoding algorithms. We utilized the STAC [136] Python Library to perform the Friedman tests and the scikit-posthocs [137] Python library to perform the Conover tests.

A.4.3. Decoder comparison

Neural decoding models regressed audio targets on neural features. Here, we used the Mel-spectrogram [138] representation (128 bands) of the audio as target variables (one target per band) and multiunit spike counts as neural features (one feature per MEA channel). Both the Mel-spectrogram bands and multiunit spike counts were binned into 40ms time bins, and all audios were reconstructed on a bin-by-bin basis (i.e., no averaging across bins or trials).

Data were collected from a total of 3 arrays implanted in 2 NHPs (RPB and CPB arrays in NHP-0 and an RPB array in NHP-1). Data for each array were analyzed independently (i.e., no pooling across arrays or NHPs). Unless explicitly stated, all models were trained on 5 English words using all 96 neural channels from the NHP-0 RPB array. Each array data set contained ~40 repetitions of each word (words randomly interleaved) collected over multiple sessions.

To mitigate decoding model overfitting, data collected from a given array was sequentially concatenated across recording sessions, and the resulting array data sets were sequentially split into training (80%), validation (10%), and testing (10%) sets. Multiple presentations of all 5 words were present in the training, validation, and test sets.

We evaluated 7 different neural decoding algorithms, including the Kalman filter, Wiener filter, Wiener cascade, dense neural network (NN), simple recurrent NN (RNN), gated recurrent unit (GRU) RNN, and long short-term memory (LSTM) RNN. Each neural network consisted of a single hidden layer and an output layer.

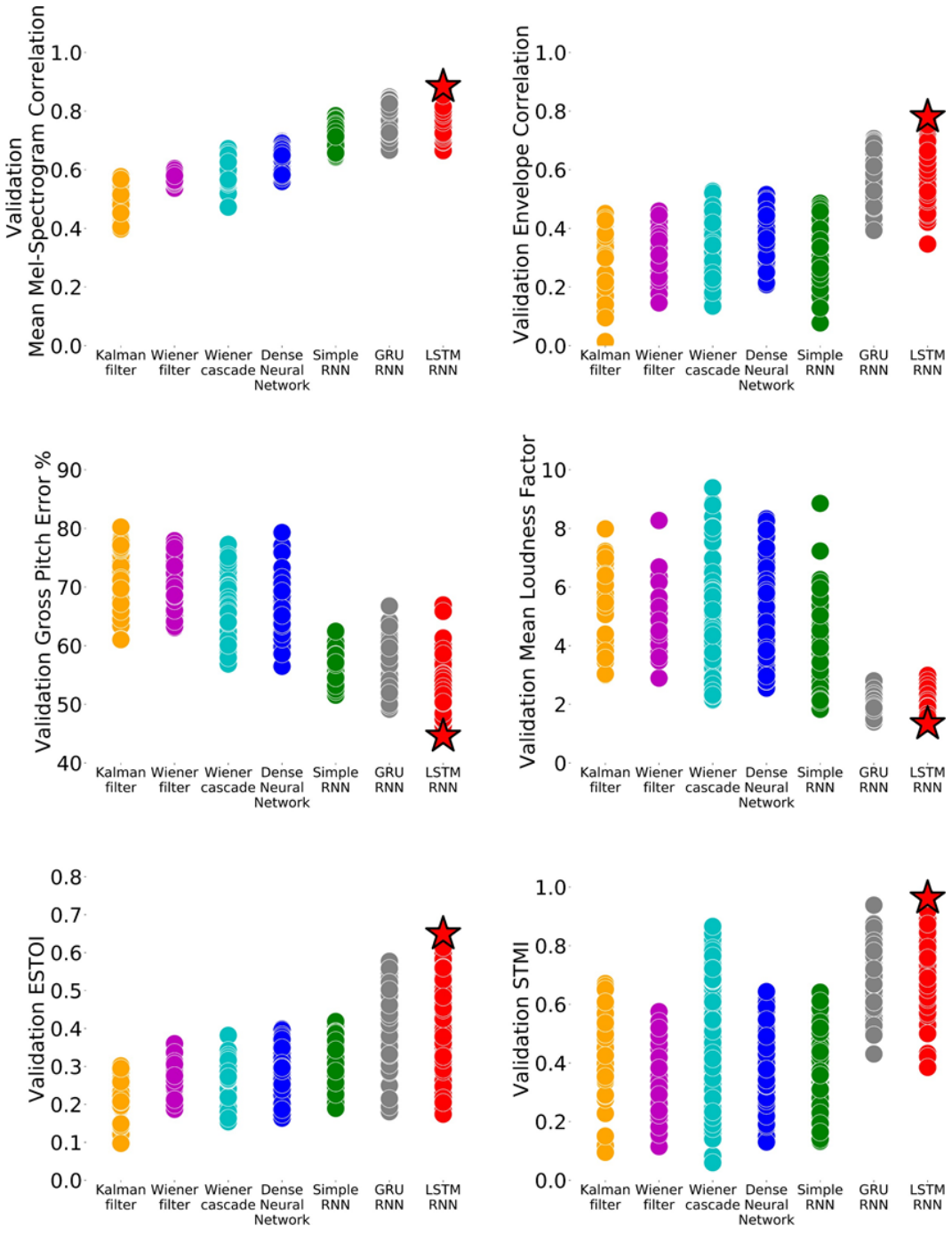

Figure A-7. We compared the effectiveness of decoding algorithms at reconstructing various audio aspects and generating intelligible audio using 6 different performance metrics. The top-performing algorithm across all metrics was the LSTM RNN (red stars).

All models were trained on Google Cloud Platform n1-highmem-96 machines with Intel Skylake processors. We calculated a mean Pearson correlation between the target and predicted Mel-spectrogram by calculating the CC for each spectrogram band and averaging across bands [107]. We applied Fisher's z-transform to CCs before averaging across spectrogram bands and before conducting statistical tests to impose additivity and to approximate a normal sampling distribution. Results are shown in Figure A-7.

We observed the Kalman filter (trained in 0.42 seconds) provided the lowest overall performance with the top model achieving a 0.57 mean validation correlation (MVC) and 0.68 mean training correlation (MTC). The top Wiener filter (trained in 2 seconds) performed similarly to the Kalman filter on the validation set (0.60 MVC) but showed an increased ability to fit the training set (0.78 MTC). A Wiener cascade (trained in 3 minutes and 33 seconds) of degree 4 beat the top Wiener filter with a MVC and MTC of 0.67 and 0.82, respectively.

A basic densely connected neural network (trained in 55 seconds) performed similarly to the top Wiener cascade decoder with a slightly improved MVC (0.69) and MTC (0.94). While the top simple RNN (trained in 2 minutes and 22 seconds) (0.94 MTC) decoder fit the training data as well as the dense neural network, it did generalize better to unseen data (0.78 MVC). Lastly, top GRU RNN (trained in 3 hours and 46 minutes) (0.85 MVC, 0.97 MTC) and LSTM RNN (trained in 3 hours and 37 minutes) (0.88 MVC, 0.98, MTC) decoders achieved similar performance with the LSTM RNN providing the best overall performance of all evaluated decoders on the validation set. The LSTM RNN also showed the highest robustness to overfitting with the top model performing only 8% worse on the validation set compared to the training set.

To examine the statistical significance of these results, we performed an unbiased multiple comparisons statistical test followed by a post-hoc Tukey-type test [135] using MVC values for the top-performing models. We found the LSTM RNN significantly outperformed all other decoding algorithms on the validation set (p-value < 0.001) except for the GRU RNN (p-value < 0.175). The GRU RNN significantly outperformed the simple RNN (p-value < 0.017), and the simple RNN significantly outperformed the dense NN (p-value < 0.017). Conversely, the dense NN did not significantly outperform the Wiener cascade (p-value < 0.900) or the Wiener filter (p-value < 0.086). These results indicate that recurrent neural networks (particularly, LSTM RNNs) provide a significant decoding improvement over traditional neural networks and other decoding methods on the performed audio reconstruction task.

In addition to examining the mean correlation performance of neural decoding algorithms on the Mel-spectrogram bands, we investigated how decoding algorithms recovered specific audio aspects including envelope, pitch, and loudness. We also quantified audio reconstruction intelligibility using the Extended Short-Time Objective Intelligibility [115, 133] (ESTOI) and Spectro-Temporal Modulation Index [134] (STMI) metrics. We found LSTM RNN decoders provided the best performance across all evaluated metrics, as shown in Figure A-7. The LSTM RNN achieved the highest envelope correlation on the validation set compared to all algorithms (p-value < 0.005) except for the GRU RNN (p-value < 0.20). For gross pitch error, the LSTM performed similarly to the GRU RNN (p-value < 0.29) and simple RNN (p-value < 0.20) but outperformed all other algorithms (p-value < 0.001). We observed a similar trend when evaluating the mean loudness factor as the LSTM RNN did not significantly outperform the GRU RNN (p-

value < 0.76) or simple RNN (p-value < 0.11), but it did out-perform all other algorithms (p-value < 0.05). For intelligibility, the LSTM RNN performed similarly to the other neural network decoding algorithms on ESTOI including the dense NN (p-value < 0.23), simple RNN (p-value < 0.09), and GRU RNN (p-value < 0.89) while performing significantly better than all other algorithms (p-value < 0.008). Lastly, the LSTM RNN achieved a significantly higher STMI score than all other algorithms (p-value < 0.001) except for the Wiener cascade (p-value < 0.08) and GRU RNN (p-value < 0.20).

We also analyzed 7 decoders aiming the reconstruction of the cochlear spectrogram. For the results in Figure A-8 left, among total 23 English words, we randomly selected 20 sets of 1, 5, 10 words. Figure A-8 right shows the result from 1 and 3 sentences selected from the total 5 sentences and the number of sets were 5 and 10, respectively. Words and sentences were randomly interleaved. The plot shows quartile distribution of CC with vertical bars showing the range of 25 % to 75 % of CC and horizontal lines in the middle of the bar showing the median CC while dots show outliers. RNN generally performs better than DNN and other decoders, as in the results shown in the previous section. Particularly, when the training model with more words and sentences, LSTM performs the best achieving the highest median CC. Generally, all decoders showed a much narrower distribution of CC with more English words and sentences, possibly due to the larger data set used during the model training. Interestingly, the Kalman filter showed higher median CC compared to that of Wiener and Wiener cascade for a single word and sentence decoding but achieved lower median CC in other cases. This may show that Kalman filter performs better at decoding a repeating feature from the neural signals.

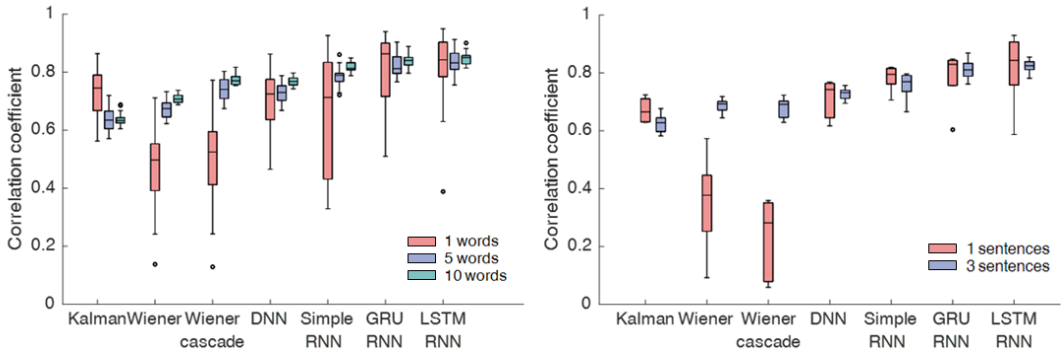

Figure A-8. Correlation coefficients between the original the reconstructed spectrogram obtained by 7 decoding algorithms. Vertical bars show the range of 25 % to 75 % of CC and horizontal lines in the middle of the bar present the median CC. For training of model 1, 5, 10 words (left) or 1, 3 sentences (right) are used.

A.4.4. Neural feature extraction comparison

To prepare neural features for decoding models, we bandpass filtered the raw neural data and calculated unsorted multiunit spike counts across all channels. We first bandpass filtered the raw 30 kS/s neural data using a 2nd-order elliptic filter in preparation for threshold-based multiunit spike extraction. To explore the effect of the filter's low and high cutoff frequencies, we performed a grid-search that evaluated 99 different bandpass filters as shown in Figure A-9. For each filter, 4 different LSTM RNN neural decoding models were evaluated. We found that using a low cutoff of 500-600 Hz and a high cutoff of 2000-3000 Hz (shown with yellow dotted lines in Figure A-9) provided a marginal improvement in decoding performance.

We used multiunit (i.e., unsorted) spike counts as neural features for decoding models. After filtering, we calculated a noise level for each neural channel over the training set using median absolute deviation [124]. These noise levels were multiplied by a scalar "threshold factor" to set an independent spike threshold for every channel (the same

threshold factor used for all channels). We extracted negative threshold crossings from the filtered neural data (i.e. multiunit spikes) and binned them into spike counts over non-overlapping windows corresponding to the audio target sampling.

While we observed optimal values for the threshold factor, we found no clear pattern across decoding algorithms. The Wiener cascade was the most sensitive to threshold factor with a 9.1% difference between the most and least optimal values. GRU RNN decoders were the least sensitive to the threshold factor with a 1.1% difference between the best and worst performing values.

Except for the Kalman filter which processed a single input at a time, we used a window of sequential values for each feature as inputs to the decoding models. Each window was centered on the current prediction with its size controlled by a "feature window span" hyperparameter (length of the window not including the current value). We evaluated 3 different feature window spans (4, 8, and 16). Both the GRU RNN and LSTM RNN showed higher performance with a larger feature window span (16); however, the other decoding algorithms achieved top performance with a smaller value (8). The LSTM RNN showed the most sensitivity to this hyperparameter with a 13.4% difference between a feature window span of 16 and 4. The Wiener filter was the least sensitive with a 3.1% difference between the best and worst performing values. These results demonstrated the importance of determining the optimal feature window span when leveraging LSTM RNN neural decoding models.

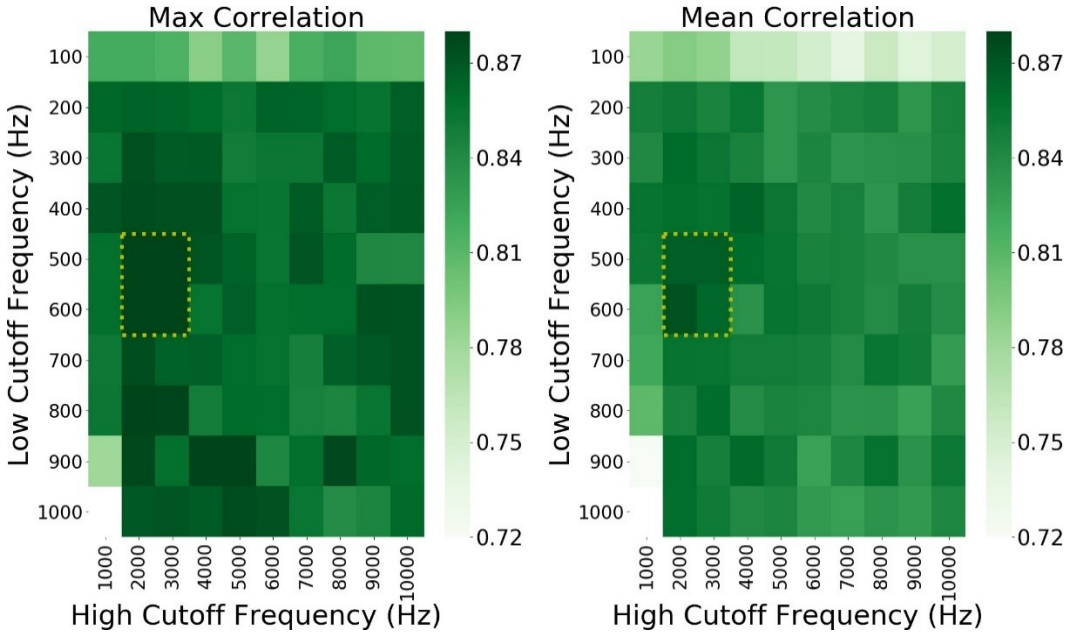

Figure A-9. A performance heat-map of the grid-searched bandpass filter cutoff frequencies. We evaluated 4 LSTM RNN neural decoding models for 99 different bandpass filters. The left and right plots show max and mean performance, respectively, for the filters. We observed a marginal improvement in decoding performance when using a low cutoff of 500-600 Hz and a high cutoff of 2000-3000 Hz.

A.4.5. Array location analysis

We implanted two NHPs (NHP-0 and NHP-1) with 96-channel MEAs in the rostral parabelt (RPB) and caudal parabelt (CPB) of the STG. We repeated the experiment described in Channel Count for 3 different arrays including the NHP-0 RPB and CPB arrays and the NHP-1 RPB array (see Figure A-10). We successfully reconstructed intelligible audio from all 3 arrays with the RPB array in NHP-0 outperforming the other 2 arrays on the validation set (p-value < 0.001). However, the successful reconstruction of intelligible audio from all 3 arrays suggests the neural representation of complex sounds is spatially distributed in the STG network. Future work will explore the benefit of

synchronously recording from rostral and caudal arrays to enable decoding of more complex audio data sets.

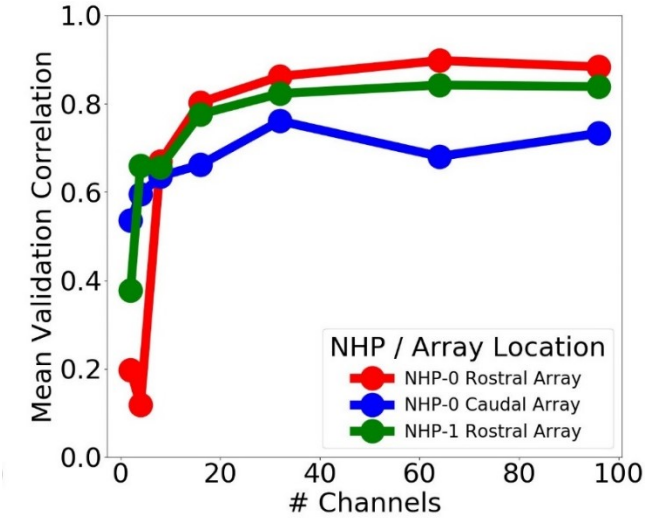

Figure A-10. We generated intelligible audio reconstructions using 3 different arrays in 2 different NHPs in the RPB and CPB regions of the STG with RPB models achieving the highest performance on the validation set.

A.4.6. Macaque call reconstruction

In addition to the English words and sentences, we investigated neural decoding models that successfully reconstructed macaque call audio from neural activity. One audio clip of a macaque call was randomly mixed in with the English audio during the passive listening task. The addition of the macaque call to the audio data set improved the achieved mean validation correlation from a 0.88 to a 0.95 despite increasing target audio complexity. This was due to the macaque call containing higher frequency spectrogram components compared to audio of the 5 English words. By successfully learning to predict these higher frequency spectrogram bands, neural decoding models achieved a higher average correlation across the full spectrogram than when those same bands contained more noise.

While adding the macaque call improved the average correlation scores, it also decreased the top ESTOI score from a 0.59 to 0.56 on the validation set.

A.4.7. Sound complexity analysis

Gao, Peiran, et al. [139] have suggested that a higher number of channels are needed to achieve better decoding performance, particularly in high neural task complexity. To evaluate this point, we have tested the LSTM model under various task complexities by selecting different types of sounds (words vs sentences) and varying the number of sounds. For this experiment, we performed spike sorting using WavClus to isolate units from channels and also randomly subsampled units for model training while testing the various number of units. As shown in Figure A-11, it was clear that a higher number of units were needed to achieve a better decoding performance and we note that more units were needed in more complex training sounds to achieve the comparable MVC, which complies with the results in the previous study [139].

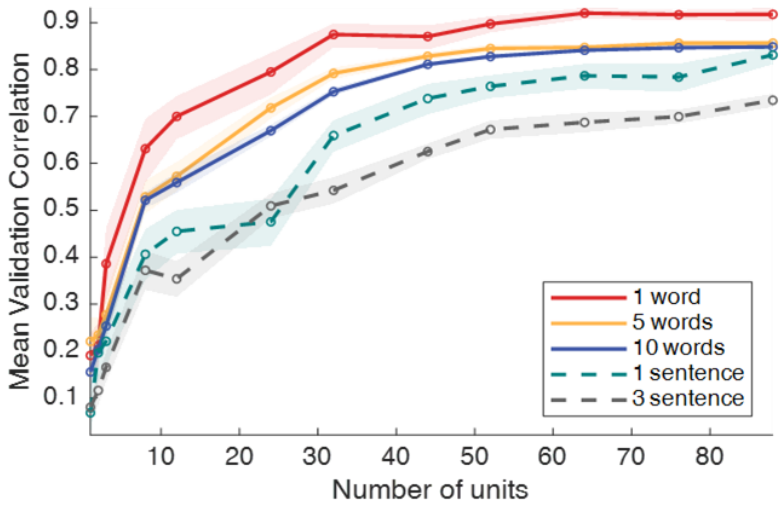

Figure A-11. The mean validation correlation coefficients depending on the number of units used for LSTM traning. English words and sentences were used for training. The shaded area shows standard error.

A.4.8. Unit selection analysis

Previous work has shown that increased neuronal firing rates correlate strongly with feature importance when using LSTM RNN models for neural decoding [140]. Therefore, to investigate the effect of channel count on decoding performance, we ordered neural channels according to the highest neural activity (i.e., highest counts of threshold crossings) or randomly selected units over the training data set. We built LSTM RNN decoding models using only the subselected units. As shown in Figure A-12, we generally observed improvements in performance as the unit count was increased in random selection. However, when we choose active units, we could achieve higher MTC even with a lower number of units and we found that adding a less active unit can negatively affect the decoding. This experiment shows that the active units carry more information than others, and it is important to choose units to improve decoding performance.

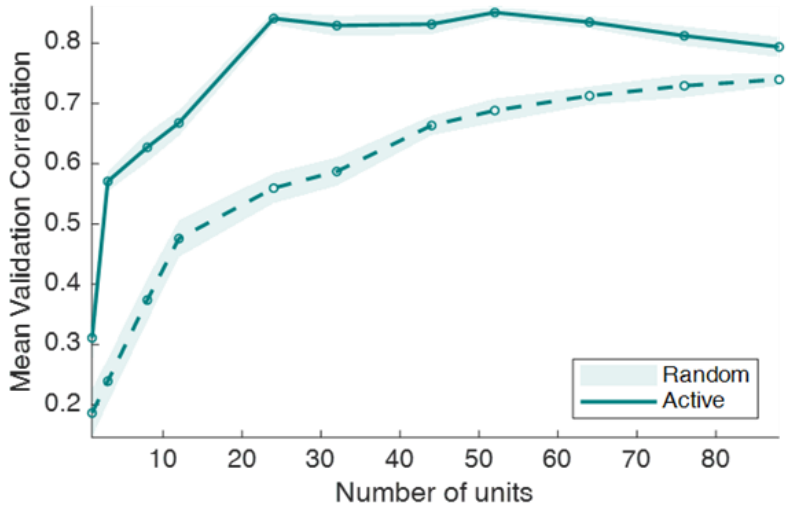

Figure A-12. Comparison of unit selection methods for LSTM training (active vs. random). The term "Active" means that the unit has a high average firing rate. And, for the active unit selection, units were sorted by their firing rate and we chose the most active units first. 10 sets of 3 sentences were used for training. The shaded area shows standard error.

A.5. Conclusion

In this chapter, we describe neural decoding studies to gaining an understanding of how sensory audio inputs such as words in the English language are processed in the most salient area of the brain: the auditory cortex. We approached the problem by training monkeys to listen to such sounds and recorded their neural population activity from the superior temporal gyrus (STG) using implanted intracortical microelectrode arrays. After training two monkeys (macaques), we used two 96 channel wired electrodes array to record neural signals from the auditory parabelt of the STG. We build a platform (called "Dockex") that handles machine learning experiments for high-dimensionality neural signals and we analyzed various factors to improve the decoding performance. We trained 7 decoding models and reconstructed 128-frequency band spectrograms from spike activities while comparing the performance of decoders. We tested parameters such as spike detection threshold, array location, sound complexity and number of units. We could achieve high decoding performance and resynthesize sound waveforms highly intelligible from reconstructed spectrograms (sound wave files are avaliable in [8]).

REFERENCES

[1] Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehs, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

[2] Afsheen Afshar, Gopal Santhanam, M Yu Byron, Stephen I Ryu, Maneesh Sahani, and Krishna V Shenoy. Single-trial neural correlates of arm movement preparation. *Neuron*, 71(3):555–564, 2011.

[3] Wilson Truccolo, Gerhard M Friehs, John P Donoghue, and Leigh R Hochberg. Primary motor cortex tuning to intended movement kinematics in humans with tetraplegia. *Journal of Neuroscience*, 28(5):1163–1178, 2008.

[4] Arto V Nurmikko, John P Donoghue, Leigh R Hochberg, William R Patterson, Yoon-Kyu Song, Christopher W Bull, David A Borton, Farah Laiwalla, Sunmee Park, Yin Ming, et al. Listening to brain microcircuits for interfacing with external

world—progress in wireless implantable microelectronic neuroengineering devices. *Proceedings of the IEEE*, 98(3):375– 388, 2010.

[5] R Bhandari, S Negi, L Rieth, RA Normann, and F Solzbacher. A novel masking method for high aspect ratio penetrating microelectrode arrays. *Journal of Micromechanics and Microengineering*, 19(3):035004, 2009.

[6] John P Donoghue. Bridging the brain to the world: a perspective on neural interface systems. *Neuron*, 60(3):511–521, 2008.

[7] Ryan C Kelly, Matthew A Smith, Jason M Samonds, Adam Kohn, AB Bonds, J Anthony Movshon, and Tai Sing Lee. Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. *Journal of Neuroscience*, 27(2):261–264, 2007.

[8] Christopher Heelan, Jihun Lee, Ronan O'Shea, Laurie Lynch, David M Brandman, Wilson Truccolo, and Arto V Nurmikko. Decoding speech from spike-based neural population recordings in secondary auditory cortex of non-human primates. *Communications biology*, 2(1):1–12, 2019.

[9] Patrick J Rousche and Richard A Normann. Chronic recording capability of the utah intracortical electrode array in cat sensory cortex. *Journal of neuroscience methods*, 82(1):1–15, 1998.

[10] Alexander T Rajan, Jessica L Boback, John F Dammann, Francesco V Tenore, Brock A Wester, Kevin J Otto, Robert A Gaunt, and Sliman J Bensmaia. The effects of chronic intracortical microstimulation on neural tissue and fine motor behavior. *Journal of neural engineering*, 12(6):066018, 2015.

[11] Guy H Wilson, Sergey D Stavisky, Francis R Willett, Donald T Avansino, Jessica N Kelemen, Leigh R Hochberg, Jaimie M Henderson, Shaul Druckmann, and Krishna V Shenoy. Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus. *Journal of Neural Engineering*, 17(6):066007, 2020.

[12] Henrique Miranda, Vikash Gilja, Cindy A Chestek, Krishna V Shenoy, and Teresa H Meng. Hermesd: A high-rate long-range wireless transmission system for simultaneous multichannel neural recording applications. *IEEE Transactions on Biomedical Circuits and Systems*, 4(3):181–191, 2010.

[13] David A Borton, Ming Yin, Juan Aceros, and Arto Nurmikko. An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. *Journal of neural engineering*, 10(2):026010, 2013.

[14] Ming Yin, David A Borton, Juan Aceros, William R Patterson, and Arto V Nurmikko. A 100-channel hermetically sealed implantable device for chronic wireless neurosensing applications. *IEEE transactions on biomedical circuits and systems*, 7(2):115–128, 2013.

[15] John D Simeral, Thomas Hosman, Jad Saab, Sharlene N Flesher, Marco Vilela, Brian Franco, Jessica Kelemen, David M Brandman, John G Ciancibello, Paymon G Rezaii, et al. Home use of a wireless intracortical brain-computer interface by individuals with tetraplegia. *medRxiv*, 2019.

[16] Y-K Song, David A Borton, Sunmee Park, William R Patterson, Christopher W Bull, Farah Laiwalla, John Mislow, John D Simeral, John P Donoghue, and Arto V Nurmikko. Active microelectronic neurosensor arrays for implantable brain

communication interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(4):339–345, 2009.

[17] Stavros Zanos, Andrew G Richardson, Larry Shupe, Frank P Miles, and Eberhard E Fetz. The neurochip-2: an autonomous head-fixed computer for recording and stimulating in freely behaving monkeys. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(4):427–435, 2011.

[18] Xilin Liu, Milin Zhang, Basheer Subei, Andrew G Richardson, Timothy H Lucas, and Jan Van der Spiegel. The pennbmbi: Design of a general purpose wireless brain-machine-brain interface system. *IEEE transactions on biomedical circuits and systems*, 9(2):248–258, 2015.

[19] Andy Zhou, Samantha R Santacruz, Benjamin C Johnson, George Alexandrov, Ali Moin, Fred L Burghardt, Jan M Rabaey, Jose M Carmena, and Rikky Muller. A wireless and artefact-free 128-channel neuromodulation device for closed-loop stimulation and recording in non-human primates. *Nature biomedical engineering*, 3(1):15–26, 2019.

[20] David A Schwarz, Mikhail A Lebedev, Timothy L Hanson, Dragan F Dimitrov, Gary Lehew, Jim Meloy, Sankaranarayani Rajangam, Vivek Subramanian, Peter J Ifft, Zheng Li, et al. Chronic, wireless recordings of large-scale brain activity in freely moving rhesus monkeys. *Nature methods*, 11(6):670–676, 2014.

[21] James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of

neural activity. *Nature*, 551(7679):232–236, 2017.

[22] Stewart Shipp. Structure and function of the cerebral cortex. *Current Biology*, 17(12):R443– R449, 2007.

[23] Jorge F Mejias, John D Murray, Henry Kennedy, and Xiao-Jing Wang. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. *Science advances*, 2(11):e1601335,2016.

[24] Roberta M Kelly and Peter L Strick. Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *Journal of neuroscience*, 23(23):8432–8444, 2003.

[25] Charles D Gilbert. Horizontal integration and cortical dynamics. *Neuron*, 9(1):1–13, 1992.

[26] Roger Lemon. The output map of the primate motor cortex. *Trends in neurosciences*, 11(11):501–506, 1988.

[27] Eric L Schwartz. Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological cybernetics*, 25(4):181–194, 1977.

[28] Brian A Wandell, Serge O Dumoulin, and Alyssa A Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383,2007.

[29] Elon Musk et al. An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194, 2019.

[30] Dongjin Seo, Jose M Carmena, Jan M Rabaey, Elad Alon, and Michel M Maharbiz.

Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces. *arXiv preprint arXiv:1307.2196*, 2013.

[31] Dongjin Seo, Ryan M Neely, Konlin Shen, Utkarsh Singhal, Elad Alon, Jan M Rabaey, Jose M Carmena, and Michel M Maharbiz. Wireless recording in the peripheral nervous system with ultrasonic neural dust. *Neuron*, 91(3):529–539, 2016.

[32] David K Piech, Benjamin C Johnson, Konlin Shen, M Meraj Ghanbari, Ka Yiu Li, Ryan M Neely, Joshua E Kay, Jose M Carmena, Michel M Maharbiz, and Rikky Muller. A wireless millimetre-scale implantable neural stimulator with ultrasonically powered bidirectional communication. *Nature Biomedical Engineering*, 4(2):207–222, 2020.

[33] Alexey Denisov and Eric Yeatman. Ultrasonic vs. inductive power delivery for miniature biomedical implants. In *2010 International Conference on Body Sensor Networks*, pages 84–89. IEEE, 2010.

[34] Shaul Ozeri and Doron Shmilovitz. Ultrasonic transcutaneous energy transfer for powering implanted devices. *Ultrasonics*, 50(6):556–566,2010.

[35] CM Langton, AV Ali, CM Riggs, GP Evans, and W Bonfield. A contact method for the assessment of ultrasonic velocity and broadband attenuation in cortical and cancellous bone. *Clinical physics and physiological measurement*, 11(3):243, 1990.

[36] Yaoyao Jia, S Abdollah Mirbozorgi, Byunghun Lee, Wasif Khan, Fatma Madi, Omer T Inan, Arthur Weber, Wen Li, and Maysam Ghovanloo. A mm-sized free-floating

wirelessly powered implantable optical stimulation device. *IEEE transactions on biomedical circuits and systems*, 13(4):608–618, 2019.

[37] Yaoyao Jia, S Abdollah Mirbozorgi, Pengcheng Zhang, Omer T Inan, Wen Li, and Maysam Ghovanloo. A dual-band wireless power transmission system for evaluating mm-sized implants. *IEEE transactions on biomedical circuits and systems*, 13(4):595–607, 2019.

[38] Philipp Gutruf, Vaishnavi Krishnamurthi, Abraham Va´zquez-Guardado, Zhaoqian Xie, Anthony Banks, Chun-Ju Su, Yeshou Xu, Chad R Haney, Emily A Waters, Irawati Kandela, et al. Fully implantable optoelectronic systems for battery-free, multimodal operation in neuroscience research. *Nature Electronics*, 1(12):652–660, 2018.

[39] John S Ho, Alexander J Yeh, Evgenios Neofytou, Sanghoek Kim, Yuji Tanabe, Bhagat Patlolla, Ramin E Beygui, and Ada SY Poon. Wireless power transfer to deep-tissue microimplants. *Proceedings of the National Academy of Sciences*, 111(22):7974–7979, 2014.

[40] Dukju Ahn and Maysam Ghovanloo. Optimal design of wireless power transmission links for millimeter-sized biomedical implants. *IEEE transactions on biomedical circuits and systems*, 10(1):125–137, 2015.

[41] Pyungwoo Yeon, Muhannad S Bakir, and Maysam Ghovanloo. Towards a 1.1 mm ^{2} free-floating wireless implantable neural recording soc. In *2018 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4. IEEE, 2018.

[42] Adam Khalifa, Yuxin Liu, Yasha Karimi, Qihong Wang, Adebayo Eisape, Milutin

Stanac´evic´, Nitish Thakor, Zhenan Bao, and Ralph Etienne-Cummings. The microbead: A 0.009^3 mm implantable wireless neural stimulator. *IEEE transactions on biomedical circuits and systems*, 13(5):971–985, 2019.

[43] Adam Khalifa, Yasha Karimi, Qihong Wang, Sahithi Garikapati, Webert Montlouis, Milutin Stanaćević, Nitish Thakor, and Ralph Etienne-Cummings. The microbead: A highly miniaturized wirelessly powered implantable neural stimulating system. *IEEE transactions on biomedical circuits and systems*, 12(3):521–531, 2018.

[44] Maysam Ghovanloo and Khalil Najafi. A wireless implantable multichannel microstimulating system-on-a-chip with modular architecture. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(3):449–457, 2007.

[45] Carolina Mora Lopez, Alexandru Andrei, Srinjoy Mitra, Marleen Welkenhuysen, Wolfgang Eberle, Carmen Bartic, Robert Puers, Refet Firat Yazicioglu, and Georges GE Gielen. An implantable 455-active-electrode 52-channel CMOS neural probe. *IEEE Journal of Solid-State Circuits*, 49(1):248–261, 2013.

[46] Takashi Tokuda, Takaaki Ishizu, Wuthibenjaphonchai Nattakarn, Makito Haruta, Toshihiko Noda, Kiyotaka Sasagawa, Mohamad Sawan, and Jun Ohta. 1 mm3-sized optical neural stimulator based on CMOS integrated photovoltaic power receiver. *AIP Advances*, 8(4):045018, 2018.

[47] Jongyup Lim, Eunseong Moon, Michael Barrow, Samuel R Nason, Paras R Patel, Parag G Patil, Sechang Oh, Inhee Lee, Hun-Seok Kim, Dennis Sylvester, et al. 26.9 a $0.19\times$ 0.17 mm^2 wireless neural recording ic for motor prediction with near-infrared-based power and data telemetry. In *2020 IEEE International Solid-State*

Circuits Conference-(ISSCC), pages 416–418. IEEE, 2020.

[48] Jihun Lee, Ethan Mok, Jiannan Huang, Lingxiao Cui, Ah-Hyoung Lee, Vincent Leung, Patrick Mercier, Steven Shellhammer, Lawrence Larson, Peter Asbeck, et al. An implantable wireless network of distributed microscale sensors for neural applications. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 871–874. IEEE, 2019.

[49] Vincent W Leung, Jihun Lee, Siwei Li, Siyuan Yu, Chester Kilfovle, Lawrence Larson, Arto Nurmikko, and Farah Laiwalla. A CMOS distributed sensor system for high-density wireless neural implants for brain-machine interfaces. In *ESSCIRC 2018-IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pages 230–233. IEEE, 2018.

[50] Vincent W Leung, Lingxiao Cui, Sravya Alluri, Jihun Lee, Jiannan Huang, Ethan Mok, Steven Shellhammer, Ramesh Rao, Peter Asbeck, Patrick P Mercier, et al. Distributed microscale brain implants with wireless power transfer and Mbps bi-directional networked communications. In *2019 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4. IEEE, 2019.

[51] Farah Laiwalla, Jihun Lee, Ah-Hyoung Lee, Ethan Mok, Vincent Leung, Steven Shellhammer, Yoon-Kyu Song, Lawrence Larson, and Arto Nurmikko. A distributed wireless network of implantable sub-mm cortical microstimulators for brain-computer interfaces. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6876–6879. IEEE, 2019.

[52] Sohmyung Ha, Abraham Akinin, Jiwoong Park, Chul Kim, Hui Wang, Christoph Maier, Patrick P Mercier, and Gert Cauwenberghs. Silicon-integrated high-density electrocortical interfaces. *Proceedings of the IEEE*, 105(1):11–33, 2016.

[53] Nur Ahmadi, Matthew L Cavuto, Peilong Feng, Lieuwe B Leene, Michal Maslik, Federico Mazza, Oscar Savolainen, Katarzyna M Szostak, Christos-Savvas Bouganis, Jinendra Ekanayake, et al. Towards a distributed, chronically-implantable neural interface. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 719–724. IEEE, 2019.

[54] James J Spilker Jr. Digital communications by satellite. *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1977. 685 p.*, 1977.

[55] Peilong Feng, Michal Maslik, and Timothy G Constandinou. Em-lens enhanced power transfer and multi-node data transmission for implantable medical devices. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2019.

[56] Jihun Lee, Vincent Leung, Ah-Hyoung Lee, Jiannan Huang, Peter Asbeck, Patrick P Mercier, Stephen Shellhammer, Lawrence Larson, Farah Laiwalla, and Arto Nurmikko. Wireless ensembles of sub-mm microimplants communicating as a network near 1 GHz in a neural application. *bioRxiv*, 2020.

[57] Ali Ersen, Stella Elkabes, David S Freedman, and Mesut Sahin. Chronic tissue response to untethered microelectrode implants in the rat brain and spinal cord. *Journal of neural engineering*, 12(1):016019, 2015.

[58] Jonas Thelin, Henrik Jörntell, Elia Psouni, Martin Garwicz, Jens Schouenborg, Nils

Danielsen, and Cecilia Eriksson Linsmeier. Implant size and fixation mode strongly influence tissue reactions in the cns. *PloS one*, 6(1):e16267, 2011.

[59] Paul T Theilmann, Calogero D Presti, Dylan J Kelly, and Peter M Asbeck. A complementary bridge rectifier with near zero turn-on voltage in sos CMOS for wireless power supplies. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 59(9):2111–2124, 2012.

[60] Ah-Hyoung Lee, Jihun Lee, Farah Laiwalla, Vincent Leung, Jiannan Huang, Arto Nurmikko, and Yoon-Kyu Song. A scalable and low stress post-CMOS processing technique for implantable microsensors. *Micromachines*, 11(10):925, 2020.

[61] Kaiyuan Yang, Qing Dong, David Blaauw, and Dennis Sylvester. 8.3 a 553f 2 2-transistor amplifier-based physically unclonable function (puf) with 1.67% native instability. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 146–147. IEEE, 2017.

[62] G Edward Suh and Srinivas Devadas. Physical unclonable functions for device authentication and secret key generation. In *2007 44th ACM/IEEE Design Automation Conference*, pages 9–14. IEEE, 2007.

[63] Jiannan Huang, Farah Laiwalla, Jihun Lee, Lingxiao Cui, Vincent Leung, Arto Nurmikko, and Patrick P Mercier. A 0.01-mm^2 mostly digital capacitor-less afe for distributed autonomous neural sensor nodes. *IEEE Solid-State Circuits Letters*, 1(7):162–165, 2018.

[64] Giles S Brindley and Walpole S Lewin. The sensations produced by electrical stimulation of the visual cortex. *The Journal of physiology*, 196(2):479–493, 1968.

[65] Sungshin Kim, Thierri Callier, Gregg A Tabot, Robert A Gaunt, Francesco V Tenore, and Sliman J Bensmaia. Behavioral assessment of sensitivity to intracortical microstimula tion of primate somatosensory cortex. *Proceedings of the National Academy of Sciences*, 112(49):15202–15207, 2015.

[66] D Yeager, William Biederman, Nathan Narevsky, Elad Alon, and J Rabaey. A fully-integrated 10.5 μw miniaturized (0.125 mm^2) wireless neural sensor. In *2012 Symposium on VLSI Circuits (VLSIC)*, pages 72–73. IEEE, 2012.

[67] Soon H Oh and Kwok H Li. Ber performance of bpsk receivers over two-wave with diffuse power fading channels. *IEEE transactions on wireless communications*, 4(4):1448–1454, 2005.

[68] Haili Cai, Mustafa Lokhandwala, Joey Zhu, Chester Kilfoyle, Jihun Lee, Lawrence Larson, Arto Nurmikko, Farah Laiwalla, and Vincent W Leung. A software-defined radio for wireless brain implants network. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 852–854, 2018.

[69] Ying Yi, Ulrich Buttner, Yiqiang Fan, and Ian G Foulds. Design and optimization of a 3-coil resonance-based wireless power transfer system for biomedical implants. *International Journal of Circuit Theory and Applications*, 43(10):1379–1390, 2015.

[70] Pyungwoo Yeon, S Abdollah Mirbozorgi, and Maysam Ghovanloo. Optimal design of a 3-coil inductive link for millimeter-sized biomedical implants. In *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 396–399. IEEE, 2016.

[71] Manjunath Machnoor, Erik Saturnino Gámez Rodríguez, Pragya Kosta, John Stang, and Gianluca Lazzi. Analysis and design of a 3-coil wireless power transmission system for biomedical applications. *IEEE Transactions on Antennas and Propagation*, 67(8):5012–5024, 2018.

[72] Mehdi Kiani, Uei-Ming Jow, and Maysam Ghovanloo. Design and optimization of a 3-coil inductive link for efficient wireless power transmission. *IEEE transactions on biomedical circuits and systems*, 5(6):579–591, 2011.

[73] Klaus Finkenzeller. *RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication*. John wiley & sons, 2010.

[74] Sanghoek Kim, John S Ho, Lisa Y Chen, and Ada SY Poon. Wireless power transfer to a cardiac implant. *Applied Physics Letters*, 101(7):073701, 2012.

[75] Yuhua Cheng, Gaofeng Wang, and Maysam Ghovanloo. Analytical modeling and optimization of small solenoid coils for millimeter-sized biomedical implants. *IEEE Transactions on Microwave Theory and Techniques*, 65(3):1024–1035, 2016.

[76] Determining the peak spatial-average specific absorption rate (sar) in the human body from wireless communications devices, 30 mhz to 6 ghz - part 3: Specific requirements for using the finite difference time domain (fdtd) method for sar calculations of mobile phones. *IEC/IEEE 62704-3:2017*, pages 1–76, 2017.

[77] IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic fields, 3 khz to 300 ghz. *IEEE Std C95.1-2005 (Revision of IEEE Std C95.1-1991)*, pages 1–238, 2006.

[78] Vikas S Shilimkar, Steven G Gaskill, and Andreas Weisshaar. Experimental characterization of metal fill placement and size impact on spiral inductors. In *2009 IEEE 18th Conference on Electrical Performance of Electronic Packaging and Systems*, pages 101–104. IEEE, 2009.

[79] Maryam Mirzaei and Mohamad Sawan. Microelectronics-based biosensors dedicated to the detection of neurotransmitters: a review. *Sensors*, 14(10):17981–18008, 2014.

[80] Alberto Bonanno, Alessandro Sanginario, Simone L Marasso, Beatrice Miccoli, Katarzyna Bejtka, Simone Benetto, and Danilo Demarchi. A multipurpose CMOS platform for nanosensing. *Sensors*, 16(12):2034, 2016.

[81] Yue Huang and Andrew J Mason. Lab-on- CMOS integration of microfluidics and electrochemical sensors. *Lab on a Chip*, 13(19):3929–3934, 2013.

[82] Francisco López-Huerta, Agustín L Herrera-May, Johan J Estrada-López, Carlos Zuñiga-Islas, Blanca Cervantes-Sanchez, Enrique Soto, and Blanca S Soto-Cruz. Alternative post-processing on a CMOS chip to fabricate a planar microelectrode array. *Sensors*, 11(11):10940– 10957, 2011.

[83] Junseok Chae, Haluk Kulah, and Khalil Najafi. A CMOS-compatible high aspect ratio silicon-on-glass in-plane micro-accelerometer. *Journal of Micromechanics and Microengineering*, 15(2):336, 2004.

[84] Jun-Chau Chien, Ali Ameri, Erh-Chia Yeh, Alison N Killilea, Mekhail Anwar, and Ali M Niknejad. A high-throughput flow cytometry-on-a-CMOS platform for single-cell dielectric spectroscopy at microwave frequencies. *Lab on a Chip*,

18(14):2065–2076, 2018.

[85] McKay Lindsay, Kevin Bishop, Shaan Sengupta, Megan Co, Michael Cumbie, Chien-Hua Chen, and Matthew L Johnston. Heterogeneous integration of CMOS sensors and fluidic networks using wafer-level molding. *IEEE transactions on biomedical circuits and systems*, 12(5):1046–1055, 2018.

[86] Michele Sessolo, Dion Khodagholy, Jonathan Rivnay, Fabien Maddalena, Melanie Gleyzes, Esther Steidl, Bruno Buisson, and George G Malliaras. Easy-to-fabricate conducting polymer microelectrode arrays. *Advanced Materials*, 25(15):2135–2139, 2013.

[87] Benoˆıt Charlot, Gilbert Sassine, Alexandra Garraud, Brice Sorli, Alain Giani, and Philippe Combette. Micropatterning pedot: Pss layers. *Microsystem technologies*, 19(6):895–903, 2013.

[88] Shihong Ouyang, Yingtao Xie, Dongping Wang, Dalong Zhu, Xin Xu, Te Tan, and Hon Hang Fong. Surface patterning of pedot: Pss by photolithography for organic electronic devices. *Journal of Nanomaterials*, 2015, 2015.

[89] Stuart F Cogan. Neural stimulation and recording electrodes. *Annu. Rev. Biomed. Eng.*, 10:275–309, 2008.

[90] Steven M Wellman, James R Eles, Kip A Ludwig, John P Seymour, Nicholas J Michelson, William E McFadden, Alberto L Vazquez, and Takashi DY Kozai. A materials roadmap to functional neural interface design. *Advanced functional materials*, 28(12):1701269, 2018.

[91] Mehran Ganji, Erik Kaestner, John Hermiz, Nick Rogers, Atsunori Tanaka, Daniel Cleary, Sang Heon Lee, Jospeh Snider, Milan Halgren, Garth Rees Cosgrove, et al. Development and translation of pedot: Pss microelectrodes for intraoperative monitoring. *Advanced Functional Materials*, 28(12):1700232, 2018.

[92] Mehran Ganji, Ahmed T Elthakeb, Atsunori Tanaka, Vikash Gilja, Eric Halgren, and Shadi A Dayeh. Scaling effects on the electrochemical performance of poly (3, 4-ethylenedioxythiophene (pedot), au, and pt for electrocorticography recording. *Advanced Functional Materials*, 27(42):1703018, 2017.

[93] So Hyun Bae, Jeong-Hwan Che, Jong-Mo Seo, Joonsoo Jeong, Eui Tae Kim, Seung Woo Lee, Kyo-in Koo, Gregg J Suaning, Nigel H Lovell, Sung June Kim, et al. In vitro biocompatibility of various polymer-based microelectrode arrays for retinal prosthesis. *Investigative ophthalmology & visual science*, 53(6):2653–2657, 2012.

[94] Joonsoo Jeong, Farah Laiwalla, Jihun Lee, Riina Ritasalo, Marko Pudas, Lawrence Larson, Vincent Leung, and Arto Nurmikko. Conformal hermetic sealing of wireless microelectronic implantable chiplets by multilayered atomic layer deposition (ald). *Advanced Functional Materials*, 29(5):1806440, 2019.

[95] Joonsoo Jeong, So Hyun Bae, Kyou Sik Min, Jong-Mo Seo, Hum Chung, and Sung June Kim. A miniaturized, eye-conformable, and long-term reliable retinal prosthesis using monolithic fabrication of liquid crystal polymer (lcp). *IEEE Transactions on Biomedical Engineering*, 62(3):982–989, 2014.

[96] Hemant Bokil, Peter Andrews, Jayant E Kulkarni, Samar Mehta, and Partha P Mitra. Chronux: a platform for analyzing neural signals. *Journal of neuroscience methods*,

192(1):146–151, 2010.

[97] Mircea Steriade, Diego Contreras, Florin Amzica, and Igor Timofeev. Synchronization of fast (30-40 hz) spontaneous oscillations in intrathalamic and thalamocortical networks. *Journal of Neuroscience*, 16(8):2788–2808, 1996.

[98] Richárd Fiáth, Patrícia Beregszászi, Domonkos Horváth, Lucia Wittner, Arno AA Aarts, Patrick Ruther, Hercules P Neves, Hajnalka Bokor, László Acsády, and István Ulbert. Large-scale recording of thalamocortical circuits: in vivo electrophysiology with the two-dimensional electronic depth control silicon probe. *Journal of neurophysiology*, 116(5):2312–2330, 2016.

[99] Stefan A Sigurdsson, Zeyang Yu, Joonhee Lee, and Arto Nurmikko. A method for large-scale implantation of 3D microdevice ensembles into brain and soft tissue. *Microsystems & Nanoengineering*, 6(1):1–13, 2020.

[100] TA Hackett, I Stepniewska, and JH Kaas. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology*, 394(4):475–495, 1998.

[101] Jon H Kaas and Troy A Hackett. 'what'and 'where' processing in auditory cortex. *Nature neuroscience*, 2(12):1045–1047, 1999.

[102] Josef P Rauschecker, Biao Tian, and Marc Hauser. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207):111–114, 1995.

[103] Biao Tian, David Reser, Amy Durham, Alexander Kustov, and Josef P Rauschecker. Functional specialization in rhesus monkey auditory cortex. *Science*,

292(5515):290–293, 2001.

[104] Biao Tian and Josef P Rauschecker. Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *Journal of Neurophysiology*, 92(5):2993–3013, 2004.

[105] Lizabeth M Romanski and Bruno B Averbeck. The primate cortical auditory system and neural representation of conspecific vocalizations. *Annual review of neuroscience*, 32:315– 346, 2009.

[106] Edward F Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428, 2010.

[107] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS Biol*, 10(1): e1001251,2012.

[108] Andrew R Dykstra, Eric Halgren, Thomas Thesen, Chad E Carlson, Werner Doyle, Joseph R Madsen, Emad N Eskandar, and Sydney S Cash. Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial eeg. *Frontiers in human neuroscience*, 5:74, 2011.

[109] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[110] Christopher R Holdgraf, Wendy De Heer, Brian Pasley, Jochem Rieger, Nathan

Crone, Jack J Lin, Robert T Knight, and Frédéric E Theunissen. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature communications*, 7(1):1–15, 2016.

[111] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[112] David A Moses, Matthew K Leonard, and Edward F Chang. Real-time classification of auditory sentences using evoked cortical activity in humans. *Journal of neural engineering*, 15(3):036005, 2018.

[113] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3D convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

[114] Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9:217, 2015.

[115] Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1):1–12, 2019.

[116] Alexander M Chan, Andrew R Dykstra, Vinay Jayaram, Matthew K Leonard, Katherine E Travis, Brian Gygi, Janet M Baker, Emad Eskandar, Leigh R Hochberg, Eric Halgren, et al. Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex*, 24(10):2679–2693, 2014.

[117] Chris Heelan. Nurmikkolab-brown/mikko: Initial release (2019). url https://doi.org/10.5281/zenodo.3525273.

[118] Chris Heelan. Connexonsystems/dockex: Initial release (2019). url https://doi.org/10.5281/zenodo.3527651.

[119] Yoshinao Kajikawa, Stephen Frey, Deborah Ross, Arnaud Falchier, Troy A Hackett, and Charles E Schroeder. Auditory properties in the parabelt regions of the superior tem poral gyrus in the awake macaque monkey: an initial survey. *Journal of Neuroscience*, 35(10):4140–4150, 2015.

[120] Daniel Bendor and Xiaoqin Wang. The neuronal representation of pitch in primate auditory cortex. *Nature*, 436(7054):1161–1165, 2005.

[121] James C Barrese, Naveen Rao, Kaivon Paroo, Corey Triebwasser, Carlos Vargas-Irwin, Lachlan Franquemont, and John P Donoghue. Failure mode analysis of silicon-based intracortical microelectrode arrays in non-human primates. *Journal of neural engineering*, 10(6):066014, 2013.

[122] Ming Yin, David A Borton, Jacob Komar, Naubahar Agha, Yao Lu, Hao Li, Jean Laurens, Yiran Lang, Qin Li, Christopher Bull, et al. Wireless neurosensor for full-spectrum electrophysiology recordings during free behavior. *Neuron*, 84(6):1170–1182, 2014.

[123] Johannes Niediek, Jan Boström, Christian E Elger, and Florian Mormann. Reliable analysis of single-unit recordings from the human brain under noisy conditions: tracking neurons over hours. *PloS one*, 11(12):e0166598, 2016.

[124] R Quian Quiroga, Zoltan Nadasdy, and Yoram Ben-Shaul. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural computation*, 16(8):1661– 1687, 2004.

[125] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015.

[126] P Ru. Multiscale multirate spectro-temporal auditory model. *University of Maryland, College Park, MD*, 2001.

[127] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[128] Stuart Rosen. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1278):367–373, 1992.

[129] Kirill V Nourski, Richard A Reale, Hiroyuki Oya, Hiroto Kawasaki, Christopher K Kovach, Haiming Chen, Matthew A Howard, and John F Brugge. Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience*, 29(49):15564–15574, 2009.

[130] Sofia Strömbergsson. Today's most frequently used f0 estimation methods, and their accuracy in estimating male and female pitch in clean speech. In *INTERSPEECH*, pages 525–529. Dresden, 2016.

[131] Bishnu S Atal. Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, 52(6B):1687–1697, 1972.

[132] Stanley S Stevens. The measurement of loudness. *The Journal of the Acoustical Society of America*, 27(5):815–829, 1955.

[133] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[134] Mounya Elhilali, Taishih Chi, and Shihab A Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41(2-3):331–348, 2003.

[135] Jerrold H Zar. *Biostatistical analysis*. Pearson Education India, 1999.

[136] Ismael Rodríguez-Fdez, Adrián Canosa, Manuel Mucientes, and Alberto Bugarín. Stac: a web platform for the comparison of algorithms using statistical tests. In *2015 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2015.

[137] Maksim A Terpilowski. scikit-posthocs: pairwise multiple comparison tests in python. *Journal of Open Source Software*, 4(36):1169, 2019.

[138] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[139] Peiran Gao, Eric Trautmann, M Yu Byron, Gopal Santhanam, Stephen Ryu, Krishna

Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, page 214262, 2017.

[140] Ardi Tampuu, Tambet Matiisen, H Freyja Ólafsdóttir, Caswell Barry, and Raul Vicente. Efficient neural decoding of self-location with a deep recurrent network. *PLoS computational biology*, 15(2):e1006822, 2019.

[141] Jihun Lee, Farah Laiwalla, Joonsoo Jeong, Chester Kilfoyle, Lawrence Larson, Arto Nur-mikko, Siwei Li, Siyuan Yu, and Vincent W Leung. Wireless power and data link for ensem-bles of sub-mm scale implantable sensors near 1GHz. *In 2018 IEEE Biomedical Circuits andSystems Conference (BioCAS),* pages 1–4. IEEE, 2018.