Sensor signal conditioning circuit design for multi-electrode intra-cortical recording

Yi, Chen

2014

Yi, C. (2014). Sensor signal conditioning circuit design for multi-electrode intra-cortical recording. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62570>

<https://doi.org/10.32657/10356/62570>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**Sensor Signal Conditioning Circuit Design for
Multi-Electrode Intra-Cortical Recording**

by

Yi Chen

A Thesis

Submitted to the Faculty
of the

NANYANG TECHNOLOGICAL UNIVERSITY

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Electrical and Electronic Engineering

2014

ACKNOWLEDGEMENTS

The thesis not only indicates the end of four years of doctoral study in NTU, but also commence a new exciting journey to a destiny still unknown to me. It might be pointless to look back and evaluate gain and loss along the way, since life is just a peaceful and never-ending stream that will eventually carry you away no matter how hard or struggling you seek for meaning or reason. However, I feel grateful towards those who guide me, help me or just share the experience with me along the way.

I would like first to thank my advisor, Assistant Professor Arindam Basu for teaching, guiding and advising me in the four years. He is a very nice guy with agreeable personality, as known to many students and staffs in the lab. He is smart and considerate, always capable of finding the best way to help throughout the process. I would never forget in those countless afternoons, he sit with me in the lab, discussing about problems and confusion I was facing either in research or in life.

I would also like to express my gratitude to my supervisors and colleagues: Dr. Minkyu Je, Dr. Kevin Chai, Dr. Xiaodan Zou, Dr. Lei Yao and Dr. Yuan Gao in Institute of Microelectronics (IME), Agency for Science, Technology and Research, for meaningful guidance, discussion and advices.

I would also like to thank my friends and colleagues for all the joy and suffer, sympathy and support we shared together. This experience will never be forgotten. Thanks also to the staffs in VIRTUS IC Design Centre for their help with the technical or administrative support. Special gratitude is given to VIRTUS footballers for all the fun we had together in every Saturday afternoon and night.

I would like to thank my parents and my dear Lucy for their support throughout the past few years. Their kind understanding is what makes all these possible.

Finally, I want to acknowledge the financial support from NTU Research Scholarship, and the UMC tapeout sponsored by MediaTek Inc, and GF tapeout support from IME.

ABSTRACT

Brain-Machine Interfaces (BMIs) have been developed in the past few decades to establish a bridge between brain and external electronics and computing devices. Driven by an increasing demand in health care industry and the rapid development of complementary metal oxide semiconductor (CMOS) and micro-electro-mechanical systems (MEMS) technology, significant progress has been achieved in the BMIs that record and process the electrical signals from the brain for the purpose of studying the brain functions or controlling an actuator. The trend is to develop cortex-implantable miniature devices integrating entire system that are bio-compatible and are able to operate chronically and autonomously. This raises significant challenge to various aspects of systems and circuits design of the BMIs due to limited resource and budget.

One promising vision of BMIs is sensing neural extracellular action potentials with an array integrating 100 or more miniature probes implanted into the cortex. To guarantee the quality of the neural signal recorded by these high impedance electrodes, customized integrated circuits (IC) are developed with multiple amplifiers on the same die, achieving on-site signal recording. Since power consumption is one of the key limiting factors of the implanted system, significant effort has been dedicated to reduce the power consumption of the front-end analog and mixed-signal circuits. However, high channel count also yields high data transmission rate, leading to considerable power consumption in wireless transmission circuits. Methods of on-chip data compression are therefore highly desirable in the implanted system. In this thesis, circuits are presented for sensing and processing the neural signals robustly

while dissipating minimal power and reducing data rate for enhancing scalability of the designs.

The first contribution is the design of a novel signal folding and reconstruction scheme for neural recording applications that exploits the $1/f^n$ characteristics of the neural signals is proposed to reduce the dynamic range in the front-end circuits. The amplified output is 'folded' into a predefined range of voltage by using comparison and reset circuits along with the core amplifier. After this output signal is digitized and transmitted, a reconstruction algorithm is applied in the digital domain to recover the amplified signal from the folded waveform. This scheme enables the use of an analog-to-digital convertor with less number of bits for the same effective dynamic range at final output. It, therefore, reduces the transmission data rate of the recording chip. Both of these features allow power and area saving at the system level. Other advantages of the proposed topology are increased reliability due to the removal of pseudo-resistors, less harmonic distortion and low-voltage operation. An analysis of the reconstruction error introduced by this scheme is presented along with a MATLAB-based behaviour model for signal folding amplifier introduced to provide estimates for reconstruction error. Measurement results from two designs in two different CMOS processes are presented to prove the generality of the proposed scheme in neural recording applications. In-vivo testing results on anaesthetized rat are also conducted to show the capability of the proposed neural amplifier to simultaneously record local field potential and spike.

Another emerging trend of BMIs in recent years is to implement neural stimulator with neural recording circuits to build a brain-machine-brain closed loop in the implanted system. These closed-loop BMIs find vast applications in therapies of neurological disorders, neural prosthesis and neuro-scientific research, where certain action potential in the brain tissue are observed as a response to the neural stimulation delivered. One challenge of building the closed-loop BMIs is to record as much of

the action potential waveform as possible in the presence of large artifacts introduced by the neural stimulation. These artifacts, much larger than the action potential, usually cause saturation of the high-gain neural recording amplifier and create a long blind period of tens of milliseconds during and right after the stimulation, in which action potential cannot be properly recorded. Applying the proposed signal folding and reconstruction scheme mentioned above, a proof-of-concept IC is presented in this thesis, targeting simultaneous neural stimulation and recording. The testing results shows that the signal folding scheme creates a feedback loop that prevents the amplifier from saturation by resetting the amplifier even with artifacts at the input, leading to a reduced recovery time from the interference of the stimulation artifacts. This is the second contribution in this thesis.

Signal processing circuits can also be integrated with microelectrode array (MEA) to enable real-time on-chip processing and reduce the data transmission rate as well as the power consumption. In the BMIs used in neural prosthesis, one key component is the motor intention decoder that extracts the subjects' intention of moving from the neural signals recorded in the motor cortex of the brain. A hardware-friendly motor intention decoding algorithm based on the Extreme Learning Machine (ELM) and its mixed-signal CMOS implementation is presented in this thesis as the final contribution. ELM is a single hidden-layer neural network, in which input weights are randomly assigned and remain unchanged in training. Only output weights are trained in a one-time manner, leading to a very fast training progress and reduced computing efforts. This neural network is realized in a mixed-signal architecture, in which random weights are implemented by exploiting transistor random mismatch in the CMOS process, while output weights are implemented in digital domain for robustness and programmability. The results of the decoded neural signal data (recorded in previously conducted animal experiments) with the proposed algorithm are shown and analyzed.

In conclusion, BMIs with multiple electrodes implanted in the cortex have experienced fast development in recent years. Next generation BMI will possibly integrate up to one thousand micro-electrodes with circuit components including neural recording, signal processing as well as sensory feedback, leading to a better performance and possibly more applications. In the BMI for neural prosthesis, for instance, better control accuracy and robustness can be obtained with more intra-cortical recording and feedback. Types of prosthesis can also be extended from, for instance, upper limb to bipedal. In this great interdisciplinary endeavor, power consumption and robustness of the electronics system will always be a limiting factor to the practical use of the devices. In this thesis, some problems of the front-end signal amplifying circuit of intra-cortical BMIs have been addressed and low-power solutions have been proposed. Architectural solutions to enhance scalability and reducing transmission data rate by integrating a movement classifier on-chip is also shown for the first time.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	v
LIST OF FIGURES	xii
LIST OF TABLES	xv
LIST OF ABBREVIATIONS	xvi
CHAPTER	
1. Introduction and Motivation	1
1.1 Background	1
1.2 Brain-Machine Interfaces	3
1.3 Motivation	8
1.4 Thesis Contribution	9
1.5 Thesis Organization	10
2. Literature Review	13
2.1 Neural Signal	13
2.2 System architecture of BMI	19
2.3 Microelectrode Array	20
2.4 Neural Recording Amplifiers	22
2.5 Analog-to-digital Convertor	35
2.6 Spike Detection and Spike Sorting	36
2.7 Neural Signal Decoding	38
2.7.1 Fundamental	38
2.7.2 State-of-the-art decoding algorithms	38
2.7.3 Filtering techniques	39
2.7.4 Issues	40
2.8 Summary	42

3. A Digital Assisted Signal Folding Amplifier for Intra-Cortical Neural Recording Applications	45
3.1 Signal Folding	46
3.1.1 Concept	46
3.1.2 Version-One Implementation	48
3.1.3 Version-Two Implementation	51
3.1.4 Transient Simulation Results	51
3.2 Reconstruction	53
3.3 Reconstruction Error Analysis	55
3.4 Measurement Results	61
3.4.1 Small signal measurements	61
3.4.2 Results of Reconstruction Algorithm	63
3.4.3 Distortion and Interference	64
3.4.4 Noise and Reconstruction Error	70
3.4.5 <i>In vivo</i> Measurements	75
3.5 System Level Savings	75
3.6 Summary	78
 4. A Simultaneous Neural Recording and Stimulation System Using Signal Folding in Recording Circuits	 83
4.1 Overview	84
4.2 Proposed Design	86
4.3 Experiment Setup	87
4.4 Measurement Results	89
4.5 Conclusion	94
 5. A Neuromorphic IC for Neural Signal Decoding	 95
5.1 Experiment	96
5.2 Extreme Learning Machine	99
5.3 A neuromorphic IC for neural signal decoding based on Extreme Learning Machine	102
5.3.1 Architecture	102
5.3.2 Sub-block circuits	107
5.4 System Simulation and Parameters Setting	111
5.4.1 Parameters Setting	112
5.4.2 System simulation results	114
5.5 Verification of neural signal decoding IC	115
5.5.1 Circuit Blocks Verification	116
5.5.2 Verification of Neural Signal Decoding	119
5.5.3 Power Estimation	123

5.6	Conclusion	126
6.	Conclusion and Future Work	127
6.1	Conclusion	127
6.2	Future Work	130
APPENDICES		132
A.	Time-Domain Behavior Model of Signal Folding Amplifier	133
B.	Author's Publication	135
B.1	Journal Papers	135
B.2	Conference Papers	136
BIBLIOGRAPHY		136

LIST OF FIGURES

Figure

1.1	Various types of brain-machine interfaces with different neural signal acquisition methods	3
2.1	The diagram of the neuron	14
2.2	The formation of the action potential	15
2.3	The model of neural signal propagating through extracellular medium	17
2.4	A waveform segment of neural recording from cat motor cortex . .	18
2.5	A simplified diagram of BMI for the neural prosthesis	19
2.6	A generic system diagram of intra-cortical recording system . . .	21
2.7	A typical topology proposed for neural amplifier	23
2.8	Frequency response and power spectrum of the AC-coupled capacitive feedback topology	25
2.9	A modified folded-cascode amplifier with denegerated current sources	26
2.10	A fully differential amplifier channel	27
2.11	A self-biased fully differential OTA	28
2.12	The schematic of neural amplifier proposed by R. Harrison	29
2.13	Two ways of tuning the resistance of pseudo-resistor	29
2.14	Diagram of a DC-coupled neural recording channel employing mix-signal feedback loop	30
2.15	Diagram of a capacitive array based SAR-ADC	35
2.16	The signal processing in the spike detection and spike sorting . . .	37
2.17	A computing architecture for neural signal decoding proposed in [1]	41
2.18	Diagram of the signal chain in the intra-cortical BMI for neural prosthesis	44
3.1	Signal Folding Concept	47
3.2	Version-One Design of the Proposed Amplifier	49
3.3	Version-Two Design of the Proposed Amplifier	50
3.4	Transient Simulation Results	52
3.5	Reconstruction Algorithm	53
3.6	Output spectrum and reconstruction error	56
3.7	Behavioral Model	60
3.8	Spectrum of reconstructed signal in simulation based on behavioral model	60

3.9	Frequency response and output noise spectrum	63
3.10	Measured output waveforms and spectrum for sinusoidal input . .	65
3.11	Measured output waveforms and spectrum for a two-tone input signal	66
3.12	Neural recording in the presence of motion artifact	68
3.13	Harmonic distortion in the proposed and conventional amplifiers .	69
3.14	Output spectrum and reconstruction error	71
3.15	Total input-referred noise and SNDR	72
3.16	Time domain characterization of reconstruction error	73
3.17	Experimental setup for <i>in vivo</i> measurement	74
3.18	Neural Activity Recorded <i>in vivo</i>	76
4.1	Die photo and circuits diagram of the simultaneous neural recording and stimulation IC	86
4.2	The <i>in vitro</i> testing setup of simultaneous neural stimulation and recording	88
4.3	Measured waveforms of the simultaneous neural stimulation and recording <i>in vitro</i> testing	89
4.4	Aligned waveforms of the simultaneous neural stimulation and recording <i>in vitro</i> testing	90
4.5	Comparison of the conventional amplifier and signal folding amplifier in the <i>in vitro</i> testing	90
4.6	Comparison of the conventional amplifier and signal folding amplifier with increased sampling rate	92
4.7	Aligned waveforms of the simultaneous neural stimulation and recording <i>in vitro</i> testing with increased sampling rate	93
4.8	Comparison of the conventional amplifier and signal folding amplifier with increased stimulation duration	93
5.1	Setup of the neural signal recording and decoding in monkey finger movement experiment	97
5.2	Timing diagram of a successful trial	98
5.3	Architecture of the single hidden-layer feedforward neural network used in the Extreme Learning Machine	100
5.4	Architecture of the proposed neuromorphic IC for the neural signal decoding	102
5.5	Timing diagram of neural decoding IC	106
5.6	Schematic of the input processing circuit	108
5.7	Schematic of the current DAC used in the decoding IC	109
5.8	Schematic and waveforms of the silicon neuron used in the decoding IC	111
5.9	Flow chart of parameters setting using system simulation in MATLAB	113
5.10	Decoding accuracy of proposed decoding IC obtained by system simulation in MATLAB	115
5.11	Die photo of the proposed neural signal decoding IC	116

5.12	The output waveform of the window counter processing an input spike sequence	117
5.13	The operation of the silicon neuron	117
5.14	The transfer curves of the silicon neurons	118
5.15	2-dimension mismatch graph	119
5.16	Histogram of mismatch in the input weights	119
5.17	Change of the decoding accuracy of the proposed IC with number of hidden-layer neurons used	120
5.18	Input and output of the proposed neural signal decoding	121
5.19	Decoding accuracy of the neural signal data with and without the delayed samples	122
5.20	Decoding accuracy of neural signal data from Monkey C, G and K	123

LIST OF TABLES

Table

1.1	Summary of signal acquisition methods for BMI	5
2.1	Summary of Performance Comparison among Different Works . .	33
3.1	Noise performance and power breakdown	64
3.2	Comparion of signal-folding and conventional topology in system level	78
3.3	Performance summary and comparison of neural amplifiers	79
5.1	Measured Current consumption of DVDD with varying number of input channel and hidden-layer neurons	124
5.2	Comparison among different works on neural decoding IC	125

LIST OF ABBREVIATIONS

CMOS	Complementary metal oxide semiconductor
MEMS	Micro-electro-mechanical systems
BMI	Brain-machine interface
fMRI	functional magnetic resonance imaging
NIRS	Near infrared spectroscopy
EEG	Electroencephalogram
ECoG	Electrocorticography
MEA	Micro-electrodes array
ADC	Analog-to-Digital Converter
SNDR	Signal-to-Noise and Distortion Ratio
DAC	Digital-to-Analog Converter
SAR	Successive Approximation Register
MSB	Most Significant Bit
ANN	Artificial Neural Network
LFP	Local Field Potential
ELM	Extreme Learning Machine
NOB	number-of-bits
PBS	phosphate-buffered saline
ASIC	application-specific integrated circuit
SoC	system-on-a-chip
NEF	noise efficiency factor

Chapter 1

Introduction and Motivation

1.1 Background

The human brain consists of up to 200 billion neurons and 125 trillion connections between these huge amount of neurons [2], making it the most powerful and intelligent ‘machine’ in the world. Its magic power in cognition, learning and decision making are far beyond the reach of any computer or machine ever made by human beings. It is generally believed that the key of the power of the human brain lies in the enormous number of connections or ‘synapses’ between neurons. With decades of great efforts by neuroscientists, people have attained a reasonably clear understanding of a single neuron. The model of the single neuron has been taught in the courses of neurobiology, computational neuroscience, etc. in universities and colleges all around the world. Unfortunately, unlike the mechanism that guides the beautiful starry sky above us has been largely explained by the theories of relativity and quantum mechanics, the mechanism of another fabulous universe in the head of every human being termed as ‘biological neural network’ is yet to be understood and explored. But now it may be close to the dawn. Eighty years ago, the discoveries of relativity theory and quantum mechanics had greatly expanded the boundary of human knowledge and gave birth to the grand technology revolution that brought us

to where we are today. In the near future, the discovery of the theory underneath the operation of the biological neural network will undoubtedly become another grand breakthrough of sciences and technology that will again revolutionize the human society, our lives and ourselves.

The discovery of new knowledge is often boosted by the advancement of technology. The artificial neural network—a computational model inspired by biological neural network—was first proposed in 1940s [3]. But its research went into stagnation in 1970s due to insufficient computer capability of that time [4]. It regained attention later partially because of the advancement of computer power and it has now been vastly applied by scientists and engineers in machine learning and artificial intelligence. The same story happens again and again in the history of science and technology. Our era has seen the phenomenal advancement of semiconductor manufacture capacity, summarized by then well-known Moore’s law [5], giving rise to an exponential growth of computing power and other tools that enhance our ability of knowledge discovery. With the rapid development of Complementary metal oxide semiconductor (CMOS), Micro-electro-mechanical systems (MEMS) technology and other biomedical imaging technology, we may have reached the observatory on which a deeper and full picture of the ‘starry sky’ of the brain can be captured.

Human brain project [6] was announced in 2013 as one of the flagship research programmes in Europe in the next 10 years. It was largely funded by the European Union and involves many European research facilities. It is dedicated to achieve a deeper understanding of how human brain functions by large-scale computer emulation of the entire brain networks, and to boost technology advancements in brain disease treatment and neuromorphic computing. In the United States, the White House also launched the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative with the goal of mapping every neuron in the brain using advanced technology [7]. Often referred to as being analogous to the successful Hu-

Figure 1.1: Various types of brain-machine interfaces with different neural signal acquisition methods

man Genome Project, it plans to invest 300 million dollars per year in the next 10 years, expecting to even generate a greater impact on various aspects of human society. To quote Paul Alivisatos [8], the director of the Lawrence Berkeley National Laboratory, “Understanding how the brain works is arguably one of the greatest scientific challenges of our time.”

1.2 Brain-Machine Interfaces

Brain-machine interfaces (BMIs) have been playing a rather active and important role in the grand challenge of brain research. BMI, as implied by its name, is essentially an interfacing agent between the brain and the external electronics or mechanical devices. Rather than the indirect ways of communication through sensory, motor, expressions, speech, etc., the brain-machine interface builds a direct communication between the brain and the external devices, opening up the possibility of directly observing and studying brain activity rather than studying them indirectly

through the behavior. Besides its scientific significance, the brain-machine interfaces also provide a new technology of analyzing, diagnosing and curing brain diseases. Prototype of a brain-machine interface dedicated to neural disease detection and prevention has been shown in [9] [10] [11]. In these systems, electrical signal of the neurons is recorded by the recording circuits, enabling real-time diagnosis and applying on-site stimulation to prevent the on-set of the epilepsy. Furthermore, brain-machine interfaces, served as the key component in neural prosthesis, can be used to help the disabled regain body functions by reconnecting damaged nerve or establish directly control of a prosthesis by the brain. Some proof of concept experiments has been conducted on primates, successfully achieving 'mind-control' of prosthesis by subject animals in task such as reaching and grasping [12] [13]. Neural prosthesis controlled by human being is also proved to be feasible in [14] [15].

One emerging trend of neural prosthesis is to integrate artificial sensory feedback to establish a closed-loop system [16], leading to a better control accuracy, robustness and possibly more applications. Types of prosthesis can also be extended from, for instance, upper limb to bipedal. Based on this, the concept of whole body neural prosthesis has been proposed with the goal of building a clinical whole-body exoskeleton controlled by brain through BMI [17]. This concept describes an ambitious vision of grand engineering that requires inter-disciplinary efforts from life science, material science, mechanical engineering and electronics engineering. It can help not only disabled people who lose part of body function, but also those who are severely paralyzed or 'locked in' by motor neuron disorder [18], such as amyotrophic lateral sclerosis (ALS), or spinal cord injury. With this picture in mind, one may also imagine that this technology can be developed furthermore to extend the limitation of the human body and mind that nature gives us.

BMI usually consists of three function blocks: signal acquisition that collect signal generated by brain activity, signal processing unit that extract features from the

Table 1.1: Summary of Signal Acquisition Methods for BMI [19]

methods	Neural Activity	Direct /Indirect	Invasiveness	Portability	Temporal resolution	Spatial resolution
fMRI [20]	Metabolic	Indirect	Non-invasive	No	~ 1 s	~ 1 mm
NIRS [21]	Metabolic	Indirect	Non-invasive	Yes	~ 1 s	~ 5 mm
EEG [22]	Electrical	Direct	Non-invasive	Yes	~ 0.05 s	~ 10 mm
ECoG [23]	Electrical	Direct	Invasive	Yes	~ 0.003 s	~ 1 mm
Intracortical recording [24]	Electrical	Direct	Invasive	Yes	~ 0.003 s	~ 0.05 mm

recorded signal as well as decode the motor intention from the extracted features, and finally, the actuator interface that controls the actuator according to the decoded results from the signal processing unit. According to the modality of the recorded signal, the mainstream BMIs can be divided into different categories, as shown in Table 1.1 and Fig. 1.1. These methods essentially form a spectrum of tools for neural signal acquisition or neural imaging, in perspective of a variation of temporal and spatial resolution.

On one end of the spectrum, with large range of imaging and low temporal and spatial resolution, there are two metabolic based imaging methods: functional magnetic resonance imaging (fMRI) and Near infrared spectroscopy (NIRS). fMRI gives a mapping of the brain activities by detecting the associated fluctuation in local blood flow in the tissue. It relies on the fact that blood flow increases with increasing neuron activity of certain cortical regions, leading to an indirect acquisition [25]. It has gained large popularity in cognitive studies to localize active regions inside the brain [26], because it is generally safe, surgery-free and capable of generating a full brain mapping. The disadvantage of the fMRI is its low temporal and spatial resolution [20]. It can only localize the neuron activities within a voxel of a few millimeters and a time window of 1 s, missing details of how a small group or a single neuron behaves and responses to an event. It is, therefore, not suitable for appli-

cations that requires high temporal and spatial resolution such as neural prosthesis. Furthermore, the huge and clumsy magnet required by magnetic resonance make it impossible for free-moving subject or long-term use. NIRS is another indirect way of imaging by referring neuron activity in the tissue to associate blood flow changes by monitoring infrared absorption rate of the tissue. It has similar temporal and spatial resolution to fMRI due to similar signal modality [21] [27]. Comparing to fMRI, NIRS is portable with wireless device available now, thus, capable of application to free-moving tasks [28].

In the center of the spectrum, Electroencephalogram (EEG) collects electrical signal of the neuron activity from a set of electrodes populated on the scalp. The electrical signals measured by EEG are the voltage fluctuation caused by ion currents flowing in and out of the neuron membrane [29]. Since it is a direct way of measuring neuron activity, EEG can provide a higher temporal resolution of around 0.02 s [30]. EEG has been applied vastly in clinical diagnosis, such as diagnosis of epilepsy, sleep disorder, coma [31] [32], as well as neural prosthesis. The EEG-based BMI is a relatively mature technology with many prototypes and commercial products available [33] [34] [35] [36]. The limitation of EEG lies on its low temporal and spatial resolution [22]. Since the electrodes are placed on the scalp, the electrical potential generated by neurons are coupled into the electrode through tissue, scalp, skin and hair. This process makes EEG signal essentially an average of simultaneous electrical activity of a cluster of neurons. The averaging or filtering effects is imposed spatially and temporally, making it impossible to separate the activity of a single neuron and a single action potential event of the neuron. Furthermore, signal quality is also degraded with multiple sources of interference, noise and artifacts being coupled into the signal path [37]. These disadvantages make EEG-based BMI not suitable for neural prosthesis with multiple degrees of freedom, such as upper limb or bipedal prosthesis. To accomplish this type of task, we need to ‘watch’ neurons

more closely [38].

Until now, all the signal modality introduced above use non-invasive acquisition. Non-invasive acquisition rules out the risk of the infection and the complexity introduced by an implant surgery. However, none of present non-invasive acquisition can provide enough information for a motor neural prosthesis to accomplish complicated work in daily-life. Electrocorticography (ECoG) is basically a invasive version of EEG, implanting multi-electrodes on the surface of the cortex right beneath the scalp. It therefore has higher temporal and spatial resolution [23] and has less vulnerability to interferences and artifacts than EEG [39]. ECoG-based brain-machine interface has been shown to successfully control a two-dimension cursor [40]. Recent study reveals its possibility of getting higher classification accuracy and generalization of movement types [41]. It is a compromise between non-invasive ECG and more invasive intra-cortical neural recording that requires electrodes penetrating the brain tissue.

Intra-cortical neural recording has gained vast popularity in research in recent years, thanks to the rapid development of CMOS and MEMS technology. Smaller electrodes can be manufactured enabling higher acquisition resolution, higher integration and less invasiveness to the tissue. Modern Micro-electrodes array (MEA) usually integrates up to 100 electrodes, each of which has a size of tens of micrometers [24]. CMOS frond-end circuits are integrated with the MEA to improve signal acquisition quality and enable wireless data transmission. Very high selectivity of the electrode makes it possible to realize single-unit acquisition so that the action potential of even a single neuron can be recorded and identified, while a large number of electrodes are able to simultaneously monitor a large group of neurons in a certain cortical region [42]. Intra-cortical brain-machine interfaces has been developed to perform various intriguing tasks including dexterous finger movements [43], reaching and grasping [44] [12], bimanual operation [45], etc.

The drawbacks of the intra-cortical recording lies mainly on the fact that it is highly invasive. Considerable amount of effort has been devoted to electrode design to manufacture more biocompatible electrodes that produce less tissue damage and provide stable signal quality for a long time, and on the circuits side, to build low power front-end circuits and wireless data link for long-term operation. Another drawback is that it is largely a localized signal acquisition based on current technology, hence unable to give a full image of the brain activity. More electrodes are therefore implanted to simultaneously monitor more neurons in the cortex so that neural signals from more than one cortical region can be involved during modeling and decoding [16].

1.3 Motivation

Despite the drawbacks, intra-cortical BMI is the most promising technology available for future neural prosthesis and a very powerful tool as well in neuroscientific research to study the connectivity between neurons. For the circuit designers, significant challenge is raised from the system requirement of the intra-cortical recording. Power consumption of the implanted circuits are highly restricted to prevent tissue damage caused by the heat dissipation of the circuits [46]. Furthermore, implanted devices are predominantly supplied by small battery or wireless power link, making the power budget even more restricted, assuming a long-term operation of the devices. As the number of electrodes increases, higher channel count makes the intra-cortical recording a more challenging task, calling for the optimization of each function block as well as system architecture. Significant effort has been dedicated into the front-end circuit design to reduce power consumption [47] [48]. Beside the front-end circuits, the wireless data link, usually required to reduce the infection risk and to facilitate free-moving application, is another function block that

consumes significant amount of power because of the increasing data transmission workload. Methods of data compression are thus highly desirable [49] [50] [51].

1.4 Thesis Contribution

In this thesis, several problems of the front-end signal conditioning and signal processing circuits of the intra-cortical BMIs are addressed and low-power solutions are proposed. A number of contributions that have been achieved are listed below, mainly in two areas: (1) the sensor signal conditioning, and (2) the neural signal decoding.

The contribution on the sensor signal conditioning side includes:

- Proposed a signal folding scheme as a general architecture that can be applied to various types of amplifiers with benefits of dynamic range reduction, system power saving and interference resilience in the neural recording application.
- The design of a low power front-end amplifier in Global Foundries (GF) 0.18 μm CMOS process for weak neural signal boosting with the signal folding scheme to reduce dynamic range requirement and to save system power up to 20%.
- The design of a non-Nyquist reconstruction algorithm to recover the amplified signal for further processing and building a behavioral model for estimating reconstruction error and Signal-to-Noise and Distortion Ratio (SNDR) of the signal folding amplifier.
- Implemented a monolithic IC in GF-0.18 μm CMOS process that integrates neural stimulation and neural amplifier.

- Applied signal folding scheme in simultaneous neural stimulation and recording, hence demonstrated the proposed signal folding recording circuits' capability in fast recovery from large stimulation artifacts.

The contributions on the signal decoding side includes:

- The design of a neural network algorithm based on Extreme Learning Machine to decode motor intentions embedded in the recorded neural signal and improving the classification accuracy by increasing feature dimension by including delayed spike input.
- Implemented a mixed-signal machine learning IC in AMS-0.35 μ m CMOS for real-time on-chip decoding, featuring neuromorphic realization of random input weights and hidden-layer-neurons in analog domain and tunable output weights in digital domain, being the first to report a motor intention decoder ASIC in the world.
- Verified the neural signal decoding IC with monkey cortex data of finger movement tasks, resulting in real-time motor intention classification accuracy of 97.6% and energy consumption of 307 nJ per classification.

1.5 Thesis Organization

This thesis is organized as follows. Chapter 1 introduces the background of this work and gives an overview on various brain-machine interface signal acquisition methods, leading to the motivation of this work. The contributions of this work are also listed.

Chapter 2 reviews the state-of-the-art neural amplifiers for the intra-cortical recording. Thereafter, various algorithms and systems of the motor intention decoding used in the intra-cortical BMIs are described.

Chapter 3 presents the proposed signal folding scheme and its application to intra-cortical neural recording, demonstrating its benefits including dynamic range reduction, system power saving, resilience to interference, less harmonic distortion and higher robustness.

Chapter 4 presents a simultaneous neural stimulation and recording prototype chip with the proposed signal folding scheme applied to the recording circuits.

Chapter 5 proposes a algorithm based on the Extreme Learning Machine for motor intention decoding. A mixed-signal machine learning IC is implemented using the proposed algorithm for real-time on-chip decoding. The verification of this chip is also presented in this chapter.

Chapter 6 concludes the whole thesis. Some future works on the design of low-power intra-cortical neural recording and signal processing are also discussed.

Chapter 2

Literature Review

As discussed in the previous chapter, multi-channel intra-cortical BMI has become a promising candidate as a key component in the future neural prosthesis. The power consumption and robustness of the electronics system is and will always be a limiting factor in the practical use of the BMI devices. In this chapter, a literature review is provided to introduce the various aspects of the intra-cortical BMI. The formation and characteristics of the neural signal collected by the microelectrodes is described first. After a brief review of the architecture of the existing intra-cortical BMI, different types of neural recording amplifiers are discussed. In the last section, the state-of-the-art of the neural signal decoding for neural prosthesis is presented and discussed.

2.1 Neural Signal

The source of neural signal is a series of physiological actions that the neuron takes to receive, process and transmit information. A typical neuron is comprised of a soma, dendrites and an axon, as shown in Fig. 2.1. The soma is the main body of the neuron, from which dendrites and axon arise. The dendrites usually split into many branches, extending up to hundreds of micrometers and forming a

Figure 2.1: The diagram of the neuron

structure called 'dendritic tree'. Several locations on the dendritic tree are connected chemically to synapses from other neurons, where the information transmitted from other neurons are received. The axon is a special cell structure extending from a site called 'axon hillock', to some distance away from the soma that can be as long as 1 m in human. In contrast to the fact that the neuron can have multiple dendrites, only one axon arises from the neuron. The axon is in charge of propagating action potential—an electrical signal usually generated on axon hillock, to its terminal, where the axon is divided into multiple branches ending at synapses. Excited by the action potential, the synapses release the chemical neurotransmitters to the dendrites of the other neurons, completing the process of information transmission.

The neuron cells, just as other cells in animal body, are electrically polarized across the cell membrane due to different ion concentrations inside and outside the cell membrane. The cell membrane of the neuron is an electrically insulating lipid bi-layer with various types of ion channels embedded. The ion channels can be triggered chemically or electrically, allowing certain type of ions to flow in or out of

Figure 2.2: The formation of the action potential: (a) Synapse signal processing (b) The diagram of the action potential.

neuron. The voltage across the cell membrane, usually called membrane potential, rests at around -70 mV when there is no turbulence or stimulation from outside.

As shown in Fig. 2.2 (b), the action potential is a pulse-shaped event on membrane potential that propagates through axon, carrying information of neuron signal processing to other neurons. The formation of the action potential is described below, with the illustration of Fig. 2.2 (a). Though there can be a wide variety of ionic channels responsible for the neuronal excitability, here we describe one of the most common and basic mechanisms involving a transient sodium and a persistent potassium channel. The synapses, triggered by an action potential traveling through the axon of pre-synaptic neuron, releases chemical neurotransmitters into the gap between the synapse and post-synaptic receptor on the dendrite. The neurotransmitters, which can either be excitatory or inhibitory, accepted by the receptors, cause the depolarization or hyperpolarization of the membrane potential respectively, by turning on different types of ion channels accordingly. The term depolarization refers to an increase in the membrane potential, while hyperpolarization is a decrease in

the membrane potential. The increase of the membrane turns on or activates the fast voltage-dependent sodium ion channels which allow the influx of the sodium ions, followed by the opening of the slower voltage-dependent potassium ion channels which allow the efflux of the potassium ions. The influx of the sodium ions will cause depolarization of the membrane, while efflux of the potassium ions will cause hyperpolarization. In a certain time window, if the membrane potential crosses a certain threshold, usually around -55 mV, the opening of the sodium channels will be dominant, increasing the membrane potential which in turns opens more sodium channels. The positive feedback will make the membrane potential increase sharply in a very short time up to its peak at around 100 mV, as shown in Fig. 2.2 (b). On a slower time scale, the sodium channels start inactivating or closing and the potassium channels start activating. This leads to a decrease of the membrane potential to a level usually even lower than the resting potential. The membrane potential then recovers back to the resting potential, ready for the next firing of the action potential. The action potential usually arises at the axon hillock, which has a high population of sodium channels, propagating through axon to its multiple terminals where synapses are triggered and release neurotransmitters to other neurons.

The shape of action potential events are usually approximately identical for a given neuron. Typically, the duration of an action potential event is around 1 ms in most neurons with a event rate of up to 10-100 per second. It is widely believed that the thinking process of the brain is encoded by the timing of the action potentials in this massive network of the neurons. It is, therefore, of crucial importance to record and study the action potentials of the neurons either in neuro-scientific research or other applications such as neural prosthesis.

Action potentials can be recorded intracellularly or extracellularly. In intracellular recording, an electrode is inserted into the neuron cell for better recording quality. The advantage of intracellular recording is greatly compromised, however, by its dis-

Figure 2.3: The model of neural signal propagating through extracellular medium

advantages of difficulty in experiment setting, destruction of the cell membrane and incapability of long-term recording. The extracellular recording, in which the electrodes are placed close to the neuron, keeps the neuron intact and therefore is more suitable for the action potential recording in BMI. The action potential waveform recorded extracellularly will be called a spike in this thesis to specifically differentiate it from intracellular recordings. Compared to the action potential recorded by intracellular recording, the amplitude of the spike is 2-3 order smaller, because of the filtering effect introduced by neuron membrane. The neuron membrane can be modelled as a serial capacitor, making extracellular spike a high-pass filtered version of the intracellular signal [52].

Except for the spikes generated by the neurons that are close to the electrodes, neural signals from distant sources can also be coupled into the electrode through extracellular medium. As depicted in Fig. 2.3, the extracellular medium can be modelled as a continuous resistance. Signals from the distant sources are coupled resistively but with higher frequency components that have largely been attenuated by the capacitive neuron membranes on the signal path. Therefore, the electrode

also collects lower frequency signal from a larger range, involving the activity from a cluster of neurons. This lower frequency component, usually called local field potential (LFP), represents a spatial average of electrical activity generated by an ensemble of neurons around the recording site. Generally, the spike recorded by MEA has an amplitude ranging from $20\ \mu\text{V}$ to $200\ \mu\text{V}$ and a frequency range of 200 Hz to 5 kHz. Spikes recorded by an extracellular electrode could belong to a few different neurons in close proximity of the recording site. Since each neuron generates spikes with nearly identical features (e.g. duration and amplitude), spikes can be sorted based on those features so that timing of spike generated by a single neuron can be extracted. This timing information, or metrics derived from it like 'firing rate', bears desired information for BMIs. The LFP recorded by the MEA contains low frequency components ranges from 1-200 Hz and has an amplitude up to a few mV.

Figure 2.4: A Waveform Segment of Neural Recording from Cat Motor Cortex [53]

Fig. 2.4 shows a segment of neural signal obtained by extracellular recording from an MEA [53]. Pulse-like event here are spikes while the LFP appears as an

Figure 2.5: A simplified diagram of BMI for neural prosthesis

oscillating baseline for the spikes. Generally speaking, the LFP signal is ‘slow but large’ while the spikes are ‘fast but small’. From the perspective of power spectrum, the power density of neural signals collected by MEA rolls off towards higher frequency with a characteristics of $1/f^n$ approximately, where n varies from 2 to 4. A filter can be applied to select desired signal component.

In addition to these information-bearing neural signals, an offset voltage is generated by the electrochemical processes at the electrode-tissue interface. This offset voltage is not biologically relevant. Since its amplitude can rise up to tens of mV, it can probably saturate the analog front-end or consume large dynamic range if it not rejected.

2.2 System architecture of BMI

Fig. 2.5 is a simplified diagram of a typical BMI for neural prosthesis based on the intra-cortical neural recording. The neural recording system consisting of an MEA and neural signal conditioning and processing electronics are implanted on

the cortex through surgery. The recorded data including LFP and spikes are wirelessly transmitted to external computing devices where recorded neural signals are decoded into control signals of prosthesis. The control signal can be continuous 3-D trajectory [54] or discrete decisions [55] of the prosthesis according to different tasks assigned. With the electronic link built between brain and prosthesis, the disable or paralyzed people can control the prosthesis just by thinking. The learning process happens in electronics also and in the brain. Typically, the neural signal decoding algorithm used here involves some kind of learning process that generalize mapping between input signal patterns and the desired movements. On the other hand, the brain is by nature a learning machine with high flexibility that can adapt to collaborate with the BMI through the help of feedback [54] [56] [57] [58], so that very high accuracy of decoding may not be needed. Besides the visual feedback, sensor can be installed on the prosthesis to provide tactile sensory feedback for the brain. For instance, the pressure data collected by the sensor in the process of grasping will be helpful for the subject to apply a suitable amount of strength to the object. The feedback information should be encoded in a neurological way that can be understood by the brain and delivered by neural stimulation circuits. Since this work focuses on signal conditioning and signal decoding in the intra-cortical recording, the following sections of the literature review will discuss mainly on these two components.

2.3 Microelectrode Array

The microelectrode array (MEA) is the very foundation of the capability which BMI possesses today. Empowered by fast advancement of MEMS technology, recording sites with a size of tens of micrometer can be fabricated so that only a few neurons are selected in the recording. As illustrated in Fig. 2.6, several types of microelectrodes are used in the multichannel intra-cortical neural recording. Microwires, as

Figure 2.6: A generic system diagram of intra-cortical recording system: (a) Microwire; (b) Michigan Probe; (c) Utah microelectrode array.

shown in Fig. 2.6 (a), are comprised of a core of stainless steel or tungsten connection and an insulating coat of Teflon or polyimide [59]. The diameter of microwires are around $30\text{-}50\text{ }\mu\text{m}$. They are assembled into an array with a spacing of $200\text{-}330\text{ }\mu\text{m}$ to enable multi-channel recording. A planary silicon microelectrode developed by University of Michigan [60] is shown in Fig. 2.6 (b). The recording sites are fabricated by photolithography technology, distributed along a silicon shank of $15\text{ }\mu\text{m}$ thick, $50\text{-}100\text{ }\mu\text{m}$ wide and a few mm long. The neural recording electronics can be implemented on the planar base from which the shank extends. Several shanks can be assembled in parallel to form an array with recording sites number up to 256 [61]. The arrangement facilitates recording from different depth beneath the cortex surface. Fig. 2.6 (c) is the Utah electrodes array (UEA) –a three-dimensional silicon array fabricated by MEMS technology [62]. It is a $10\text{-by-}10$ array of silicon electrodes resting on a platform of $4\times 4\times 1\text{ mm}^3$, with $25\text{-}50\text{ }\mu\text{m}$ long platinum tip at

the end of each 1.5 mm long needle. UEA has advantage of having high density of recording sites but it has a fixed recording depth after fabrication. UEAs with different recording depths must therefore be used to meet the requirement of recording from various depths [43].

Based on the MEAs mentioned above, researchers have been trying to develop fully implanted neural recording and stimulation system that enables chronic real-time operation in recent years [48] [63] [64]. These systems take advantages of the continuously scaling modern CMOS processes to integrate signal conditioning and data process circuit with MEA. The signal conditioning circuits usually contains amplifiers to boost weak neural signal, band-pass filter to reject noise and offset voltage and Analog-to-Digital Converter (ADC) to digitize amplified signal for data transmission and signal processing. In order to reduce chances of infection, information and power of the implanted device is transmitted by wireless telemetry link, as illustrated in the generic system diagram of the neural recording system in Fig. 2.6.

2.4 Neural Recording Amplifiers

The High channel count of MEA has placed severe constraints on the design of the signal conditioning circuits. One of these constraints is the limited circuit area. Synchronous recording of the neural signal on multiple sites requires implementing a front-end amplifier for each electrode channel. The pixel pitch of these devices is only a few hundred μm , which limits the area available for amplifier and signal conditioning circuits for each channel. Another challenge is to suppress the noise of the amplifier with limited power budget. Power consumption of the amplifier should be minimized to prevent the damage of the brain issue caused by overheating of the implanted device. On the other hand, input-referred noise of the amplifier should be kept low enough in order not to degrade the input signal. Since there is a design trade-

Figure 2.7: A typical topology proposed by [53] for front-end amplifier in neural recording applications integrated with MEA.

off in low power and low noise level, noise efficiency factor (NEF) is introduced to evaluate the performance of amplifiers and facilitate the comparison among different designs. NEF is defined by

$$NEF \equiv V_{ni,rms} \sqrt{\frac{2I_{tot}}{\pi \cdot U_T \cdot 4kT \cdot BW}}, \quad (2.1)$$

where $V_{ni,rms}$, I_{tot} and BW denote input-referred noise, total bias current and bandwidth of amplifier respectively. Finally, high channel count creates heavy data traffic in wireless transcutaneous transmission, consuming large amount of power. To sum up, a front-end amplifier is integrated close to the probe to boost the neural signal in every channel with various requirements in the following aspects according to the systematic configuration:

- **Noise**

Circuit noise from amplifier should be low enough to guarantee adequate SNR

for data processing to extract neural information.

- **Input Impedance**

The amplifier should have input impedance larger than impedance of signal source in order not to degrade weak neural signal. Furthermore, the input DC current should be negligible to guarantee the safety of brain tissue.

- **Dynamic Range**

The dynamic range of neural amplifier should be large enough to cope with both the large LFP signal with maximum amplitude of a few mV and the small spike signal with minimum amplitude of $20 \mu V$.

- **Input DC Offset Immunity**

Large DC offset voltage generated by electrode-tissue interface will cause saturation if it passes through the amplifier. Hence, this offset voltage should be blocked by the amplifier.

- **Band-Pass Filtering**

The amplifier is to amplify the desired signal component (around 200-5 kHz for the spike signal and 1-200 Hz for the LFP) as well as to suppress noise and signal outside interested band.

- **Minimize Resource Consumed**

The amplifier has limited power and silicon area budget due to high-level integration and the target of chronic intra-cortical operation. Moreover, the off-chip components should be avoided to minimize system size.

- **High Common-Mode Rejection Ratio (CMRR) and Power-Supply Rejection Ratio (PSRR)**

High CMRR and PSRR are expected to suppress 50/60 Hz interference and power supply noise.

Figure 2.8: (a) Frequency response of the AC-coupled capacitive feedback topology (b) Noise power spectrum of the AC-coupled capacitive feedback topology [53]

To meet these challenges, an AC-coupled capacitive feedback topology was proposed in [65]. Input coupling and feedback capacitors are implemented around an operational transconductance amplifier (OTA) to amplify the neural signals. The schematic of this topology is depicted in Fig. 2.7. Frequency response of this amplifier is shown in Fig. 2.8 (a). The gain of this amplifier is determined by the ratio of input and feedback capacitors. A very large resistive element in the order of $T\Omega$ is needed in parallel with the feedback capacitor to bias the amplifier and create a large time constant high-pass corner to eliminate DC offset voltage and low frequency noise. The low-pass corner of this amplifier is determined by the transconductance of the OTA, load capacitance and closed-loop gain. Thermal noise of amplifier consists of noise from OTA and resistive elements that provide stable DC biasing, as shown in Fig. 2.8 (b). Thermal noise from the OTA is shaped by the frequency response of the amplifier while thermal noise due to resistive elements demonstrates a feature of first-order rolling off of which the corner is determined by high-pass corner of the amplifier. Flicker noise, or so called ‘1/f noise’ is not depicted in this figure but is dominant in low frequency domain. This topology has potential of achieving low noise floor with small power consumption due to simple structure and capability

Figure 2.9: A modified folded-cascode amplifier with denegerated current sources proposed by [66]

of blocking the offset voltage with input capacitors. Therefore it is widely accepted as a basic structure for the front-end amplifier in neural recording application.

Generally, a pre-amplifier adopting this topology with a gain of 33 dB or 40 dB serves as low-noise first stage in the recording channel, followed by a band-pass filter selecting desired signals. This band-pass filter may have a mid-band gain of 10 to facilitate digitization. A buffer might be needed in some cases to provide enough driving capability for analog multiplexers or ADCs that followed.

Since then, considerable effort has been made to bring innovation in the OTA design to achieve lower NEF. A current mirror OTA is used in [65]. Large $pFET$ s are used as input differential pair of OTA to reduce $1/f$ noise. Proper sizing is chosen so that input differential pair and current mirror are biased in sub-threshold and above-threshold region respectively to enhance noise performance. It achieves a NEF of 4.0 with supply voltage of $\pm 2.5V$. A modified fold-cascode structure is introduced in the OTA design of [66], as shown in Figure. 2.9. The proposed folded-cascode struc-

Figure 2.10: A fully differential amplifier channel proposed in [64]

ture utilizes a large scaling of 16:1 in current distribution between input branch and output branch. It results in large g_m in input differential pair and small g_m of current source in output branch, achieving reduced circuits noise. Matching issue, however, is introduced by the high current scaling ratio between two branches. Source degeneration technique is proposed, using on-chip resistors as degenerating elements for current mirror, converting matching of transistors into that of resistors so that better matching can be achieved. This work achieves an NEF of 2.67, while the theoretical limit of this type of amplifier is 2.02. Power consumption of the amplifier in this work is $7.56 \mu\text{W}$.

Fully differential structure can increase common-mode rejection ratio (CMRR) and power supply rejection ratio (PSRR) by suppressing common-mode signal. A fully differential version of AC-coupled capacitive feedback amplifier is used in [64], as depicted in Fig. 2.10. The fully differential structure requires common-mode feedback (CMFB) circuits to stabilize the DC biasing of OTA, leading to increased power and area consumption. The NEF of the amplifier channel in this work is around 5.0. To confront the issue introduced by CMFB while benefiting from fully differential structure, a self-biased fully differential OTA is proposed by [67], as depicted in

Figure 2.11: A self-biased fully differential OTA proposed in [67]: The removal of CMFB due to self-biasing technique reduces area and power consumption while using nFET as well as pFET in differential pair boost g_m generated by input bias current

Fig. 2.11. A class-AB type modified differential input pair is implemented in this OTA, exploiting *nFET* as well as *pFET* to increase the effective g_m/I in input pair. CMFB circuit is not required in this OTA due to a simple self-biasing scheme provided by negative feedback from differential output to upper and lower current source directly. The area and power consumption is therefore reduced. Multiple VDDs are used in [68] to achieve low power consumption. An on-chip DC-DC convertor generates a $2 \cdot VDD$ supply for VGA to meet output swing requirement, leading to a VDD as low as 0.45 V. The pre-amplifiers and ADC work under VDD with a lower signal swing. The output of the VGA will be subtracted by VDD if it is larger than VDD, thus fitting it to the smaller input range of the ADC. This design reports a NEF of 1.57, which is the smallest up to date.

Large resistive elements used for DC biasing are usually realized on chip by MOS-bipolar pseudo-resistor or its modified version, as shown by transistors M_a –

Figure 2.12: The schematic of neural amplifier proposed in [65]: Transistors $M_a - M_d$ are connected as MOS-bipolar pseudo-resistors to provide DC path for biasing capacitive amplifier

Figure 2.13: Two ways of tuning the resistance of pseudo-resistor [69]

M_d in Fig. 2.12. Each transistor acts as a diode-connected pMOS transistor when V_{GS} is negative. On the other hand, a positive V_{GS} activates the parasitic lateral p-n-p bipolar junction transistor (BJT), making the transistor a diode-connected BJT. Generally speaking, with its special connections, the transistor serves as reverse-biased diode, yielding extremely high small signal resistance of the order of $T\Omega$. There are also amplifier designs in bio-medical application using transistors biased in subthreshold region as large resistive elements [70] [71]. This technique requires additional biasing circuits causing additional power and area consumption. Tunabil-

Figure 2.14: Diagram of a DC-coupled neural recording channel employing mix-signal feedback loop to suppress offset voltage and separate LFP and spike signal [74]

ity of high-pass corner is desired to give flexibility of recording the spike and the LFP signal with same amplifier. This can be achieved by tuning the resistance of the pseudo-resistors by adjusting gate potential in a structure depicted in Fig 2.13 (a), as proposed in [66] [67] [72]. This method, however, yields asymmetric and nonlinear behavior leading to distortion in large signal. A balanced tuneable pseudo-resistors structure is proposed by [69], as depicted in Fig. 2.13 (b), where control voltage $v_{ctrl(b)1}$ and $v_{ctrl(b)2}$ are generated by additional biasing circuits. This structure yields tunable resistance of pseudo-resistors with favorable symmetric feature. Although on-chip large resistive elements are widely used in neural recording applications, no resistive element larger than $10\text{ G}\Omega$ has passed reliability tests, prohibiting its use in critical signal paths of commercial implanted devices [73].

It is desirable to record the LFP and spike signals simultaneously, since both signals contains relevant information in both experimental neuroscience and clinical applications [10] [75] [76]. This, however, leads to a large dynamic range *and* bandwidth requirement for the analog front-end. In the worst case, we assume that spikes

with an amplitude of tens of μV is added on LFPs with amplitudes of 2 mV appears at the input of the recording channel. If a noise floor of 2 μV is needed to meet minimum signal-to-noise ratio requirement of spike signal, the dynamic range of the channel is around 60 dB, resulting in a 10-bit analog-to-digital conversion. Also, this sampling has to be done fast enough to capture the information in spikes, 20 kHz being a typical choice. In general, the area and worst-case power consumption of successive approximation register (SAR) ADC, which is a typical choice in the neural recording system, increase exponentially with increasing number of bits [77] rendering the joint acquisition of spike and LFP challenging. Moreover, high channel count of the implanted neural recording system creates heavy data traffic in wireless transcutaneous transmission and consumes a large amount of power. Obviously, the bit rate (BR) of data transmission is also proportional to the number of bits (NOB) of the ADC. Therefore, it is desirable that the NOB of ADC could be reduced. Recent works show increased interest in joint recording. [78] employs switch-capacitor filtering and a 10-bit SAR ADCs to achieve $2.2 \mu\text{V}_{\text{rms}}$ input-referred noise and a SNDR of 60.3 dB suitable for joint recording. An unique mixed-signal architecture is proposed in [74] to build a DC-coupled recording channel in 65 nm CMOS. Exploiting advantages of digital circuits in deep sub-micrometer CMOS process, this architecture introduces two mixed-signal feedback loops to cancel offset voltage and separate the LFP and the spikes, reducing NOB of ADC and area consumption since large input AC-coupled capacitors are removed. As shown in Fig. 2.14, outer feedback loop embedded a 7-bit offset DAC to oppress offset voltage under 1 Hz or less, while inner loop performs fine residual offset correction and LFP filtering of which high-pass corner is defined by an IIR filter. The LFP and the spike are thus separated and are available at the input of the inner DAC and the output of ADC respectively. This design consumes an area of 0.013 mm^2 and a power consumption of $5 \mu\text{W}$ for a full recording channel under a supply voltage of 0.5 V.

As can be concluded from the above review section, people has been competing heavily on NEF in the neural amplifier design, pushing the power efficiency to the limit. However, the wireless data transmission, as required by low infection risk and long term operation, is also a rather significant contributor to the system power consumption. Due to very large number of recording channels in the system, large amount of data throughput puts a heavy burden for the wireless data transmission. High temporal resolution in the intra-cortical recording requires higher sampling rate of ADC, according to Nyquist theory of sampling. In a neural recording system of 100 recording channels, for instance, the ADC typically samples each amplifier output at a rate higher than 20 Ksa/s to avoid aliasing, leading to a transmission rate larger than 20 Mb/s assuming an ADC resolution of 10 bits. Therefore, besides the efforts made to develop low power transmitter, data compression is also required to reduce the power consumption of the wireless transmission.

The sparseness of the spike waveform provides ample opportunity for data compression. Inspired by compressive sensing first applied in the field of image sensor, a compressive sensing method using random bases is proposed in [50]. This method is essentially to project the neural data segments at ADC output into a set of random bases and to transmit only significant components of the projection. At the receiver side, the signal is reconstructed by solving a convex optimization problem with respect to a set of signal base in which recorded signals have a sparse representation. This method achieves a compression rate of 10-40, but requires a complicated reconstruction algorithm at the receiver side. The signal quality also degrades a lot when the compression rate is high.

Since the information carried by spike waveform is largely contained in the timing of the spike firing and the waveform of the spike, spike detection and feature extraction is a natural way of data compression. Spike detection circuits, either analog or digital, detects the spike event in the recorded waveform so that only the waveform

around a spike event and the timing of the spike event are sent by transmitter. The waveform between spikes, essentially noise, is therefore discarded for power saving. The process of data compression can happen in digital domain [79] or in analog domain [51]. Implementing in digital domain is more flexible and robust. The analog implementation requires an analog spike detection circuit, which increases the complexity, but also reduces ADC power consumption, since the ADC is only turned on when the spike event is detected.

Further data compression can be achieved by on-chip spike feature extraction. As mentioned earlier, the spike signals from a few neurons can be coupled into one electrode. In this case, spike sorting is often used to cluster spikes into different groups so that timing of spike firing from a single can be found. Feature extraction is a pre-processing of spike sorting in which the features used to differentiate between spikes are extracted. With on-chip feature extraction, large data compression can be achieved by transmitting only features extracted for the spike sorting. This can also be realized either in digital domain [67] or analog domain [80] [81].

Table 2.1: Summary of Performance Comparison among Different Works

Work	CMOS Process	Fully Diff.	IRNoise (μV_{rms})	NEF	CMRR/ PSRR(dB)	BW (kHz)	Gain (dB)	VDD/ Power(V/ μW)	Area (mm ²)
[65]	1.5 μm	No	2.2	4.0	83/85	7.2	39.5	$\pm 2.5/80$	0.16
[66]	0.5 μm	No	3.06	2.67	66/75	5.3	40.8	2.8/7.56	0.16
[64]	0.35 μm	Yes	6.08	5.55	60/NA	5	33	3/8.4	0.02
[67]	0.35 μm	Yes	4.9		80/90	20	40	$\pm 1.65/6.7$	0.184
[74]*	65nm	No	4.9	5.99	75/64	10	NA	0.5/5.04	0.013
[68]**	0.18 μm	Yes	3.2	1.57	73/NA	10	52	0.45/0.73	0.202

* and **: Data of [74] and [68] are corresponding to full recording channel, including pre-amplifier, filter and ADC.

To conclude this literature review and facilitate comparison, a performance table for different designs concerning neural amplifier is presented in Table 2.1. As mentioned before, considerable effort has been made to reduce power consumption

of neural amplifier while keeping the noise floor low. A competition of achieving lower NEF has been going on for recent years. However, aiming at an implanted device capable of chronic recording, more systematic consideration in amplifier design is required, as power consumption of ADC and wireless communication is also a limiting factor. Several works address this consideration [74] [82]. [82] adopts the conventional system architecture and tries to re-examine and define the requirements of both the front-end amplifier and ADC from a perspective of system optimization. [74] starts from considering signal acquisition channel as a whole, taking advantages of better digital performance in deep sub-micrometer CMOS to assist the analog front-end. It leads to a very compact design of signal recording channel in which large input capacitors and pseudo-resistors are removed; NOB of ADC is reduced while joint recording of the LFP and spike signals is achieved. Various data compression methods are also proposed to reduce the power consumption of the wireless transmission circuits.

In this work, we also try to approach neural amplifier design in a more systematic approach, leveraging the power of digital back-end processing to ease the burden of the front-end implanted device, focusing more on resistance against interference, reliability issue of pseudo-resistors and NOB reduction in the joint recording of the LFP and spike signals, rather than competing NEF. In the following chapters, a novel neural amplifier featuring signal folding is proposed with a reconstruction algorithm in the digital domain to recover the amplified signal. The concept is illustrated and the circuits and algorithm implementation is described in the next chapter. The measurement results are also presented. The signal folding and reconstruction scheme introduces error additional to circuits noise. Therefore, we conduct measurement and analysis to get a better understanding of this error, which is vital in estimating performance of the proposed scheme.

Figure 2.15: Diagram of a capacitive array based SAR-ADC [74]: (a) Block diagram of SAR-ADC, (b) A binary-weighted capacitor array DAC

2.5 Analog-to-digital Convertor

Generally, ADCs are needed in a neural recording system to digitize the amplified neural signal, since it is easier to store, transfer and process digital signal rather than analog signal. The vision of analog signal processing [83] suggests that on-chip low power analog processing circuits can greatly reduce wireless data transmission rate and enable long-term operation of the device. However, an ADC is still needed in data verification and calibration at least.

Due to the restricted power and area budget, low resolution (8-10 bits) and low speed requirement (>20 Ksa/s for each channel), ADCs used in the neural recording systems are predominantly Successive Approximation Register (SAR). Some designs with sigma-delta ADC are proposed as an alternative, but they suffer from very large power consumption due to static basing of the operational amplifier required by sigma-delta modulation [84] [85].

The structure of a capacitive array based SAR-ADC [86] is shown in Fig. 2.15. The analog input is sampled, initiating the conversion process consisting of several approximation operations. In every approximation operation, the Digital-to-Analog Converter (DAC) output V_{DAC} is updated by the determined bits in the previous ap-

proximation operation. Then, the next Most Significant Bit (MSB) is determined by comparing the sampled analog value V_{in} to V_{DAC} . The capacitive array is widely used in the DAC so that no static current flows through the DAC. The only operation that consumes power in the DAC is charging and discharging capacitors. This feature make the capacitive array based SAR-ADC very power efficient and thus suitable for the intra-cortical multi-channel neural signal recording.

2.6 Spike Detection and Spike Sorting

As mentioned in the previous section, information carried in the spike signal is the timing of the spiking firing of individual neuron. In order to extract this information, the first step is to detect the spike events from the noisy background. The second step is to cluster spikes into different groups according to differences in the shape of the spikes. Each group, in which spikes have similar shape, is believed to be generated by one neuron. The reasoning behind spike sorting is that the shape of spikes generated by neurons and recorded by a electrode is stereotypical, determined by geometry feature, ion channel distribution of the neuron and the transmission pathway to the electrode. It is therefore believed that shape of spikes from different neurons are distinct from each other and does not change over time, or in a long enough time period at least. The signal processing involved in the spike detection and spike sorting is depicted in Fig. 2.16.

The spike detection and spike sorting were largely done in software in the past. With advancement of CMOS technology and requirements of power saving and on-the-fly operation, many spike detection and spike sorting circuits are now appearing in literature. The simplest and yet reasonably effective way of spike detection is applying a threshold to the spike waveform. Crossing of the threshold triggers a spike event. The performance of this method degrades when signal SNR is low or noise

Figure 2.16: The signal processing in the spike detection and spike sorting

background is time-variant. Adaptive threshold method is proposed to automatically change threshold according to the noise level [87]. Double threshold is also proposed by add a negative threshold so that negative phase of the spike is taken into consideration [51] [79] [88]. All threshold methods mentioned above have issue of degrading performance at low signal SNR. In the case of low SNR, methods of pre-processing has bee proposed to pre-emphasize the spike before applying the threshold. Non-linear energy operator [81] is one such method that has gained a lot of popularity. It emphasizes the localized energy change at high frequency that makes the spikes pro-founder in the noisy background, leading to a better detection. It can be implemented both in analog [81] and in digital circuits [89].

The spike sorting can be accomplished in two steps—extracting features from the spike shape and then clustering the spikes according to the extracted features. Feature extraction is addressed in many works with circuit implementation in either analog or digital domain to extract various sets of features [81] [88] [89] [90]. On the other hand, only few works have achieved on-chip clustering of the spikes [91] [92]. These spike sorting processors consume significant amount of chip area and demands computational effort. Clustering also faces issues such as lack of ground truth, non-stationarities, non-Gaussian noise, etc [93]. Furthermore, a debate is still going on over whether spike sorting is necessary at all for neural signal decoding in the neural prosthesis. Experiments has been conducted to prove that without spike sorting,

multi-unit activity recorded by the electrodes can yield comparable or even better performance of decoding than those achieved using single-unit activity with spike sorting [94] [95].

2.7 Neural Signal Decoding

The recorded neural signal—sequences of spikes from different neurons around the electrodes—carries the information of motor intention of the subject. Neural signal decoding is thus a process of extracting the motor intention embedded in the recorded neural signal, the output of which is used as a command to control the prosthesis. Through this process, the subject can move the prosthesis simply through mind control.

2.7.1 Fundamental

The fundamental of current decoding algorithms can be referred back to the work done by Georgopoulos and his colleagues [96] [97]. It is revealed in the experiment that the activity intensity of some neurons in the motor cortex is tuned to be a sinusoidal function of the movement direction of the arm with respect to a preferred direction where the activity reaches its maximum. They therefore proposed to represent each neuron by a vector indicating its preferred direction. The population vectors can be obtained by linear combination of all preferred vectors in the group weighted by the firing rate in the short time period of tens of millisecond, leading to a prediction on the velocity of upcoming arm movement [98].

2.7.2 State-of-the-art decoding algorithms

State-of-the-art decoding algorithms for mapping population activity into motor intention can be categorized into two broad subgroup: inferential decoders and

classifiers. The inferential decoders rely largely on behavioral models of neural activity which give an empirical understanding of the relationship between activity and movement. Optimal linear estimation [99], Bayesian inference [100], Gaussian mixture [101] and particle filtering [102] all fall into this category. The classifiers, on the other hand, rely on a consistent mapping pattern between neural activity and movement rather than a behavioral model. Various Artificial Neural Network (ANN) algorithms [13] [103] [104] and maximum likelihood methods [105] [106] are main choices in this category.

2.7.3 Filtering techniques

The performance of the decoding algorithm can be improved by filtering techniques that combine historical states with current output of the core algorithm to predict new state. Essentially, the state space-the ensemble of all possible output states obtained by the filter-is limited by the previous states. A simple intuition of this principle is that the movement of the arm tends to be smooth. A predicted state implying a very large change in direction, for instance, would be less probable. Learning is another way of improving the performance of neural decoding algorithm. A closed-loop is formed when brain receives feedback from the decoder output. The learning happens when the brain adapts to the BMI and reduces the error between direct controlled movement and target movement. Studies show that learning in such a closed-loop BMI improves the decoding performance [107] [108] [109]. With the help of the learning, the performance of a simple decoding algorithm may be found to be close to, if not better than, that of a sophisticated algorithm. The feedback in the closed-loop system is predominantly visual feedback nowadays. It is suggested that sensory feedback from the prosthesis should be included into the closed-loop learning of BMI for better performance [110] [111]. This feedback path involves physical sensors and processing circuits for collecting the states of the prosthesis,

neural encoder for converting collected data into neural signal patterns, and neural stimulator for delivering stimulator to the brain according to the signal patterns. A similar method has already gained significant clinical success in cochlear implant, which recovers auditory sensation for the hearing impaired [112].

A set of interesting experiments of neural prosthesis has been conducted on the primate animals to perform reaching and grasping [44], dexterous finger movements [43], bimanual operation [45], etc. Human testing is also conducted in [14] and [15]. In the amazing experiment described by [15], a paralyzed woman is able to grasp a bottle on a table through a direct controlled robotic arm and drink coffee inside the bottle.

2.7.4 Issues

As mentioned above, various elaborate models and methods are developed in past decades, in the pursuit of better decoding performance. These highly sophisticated decoding algorithms work reasonably well in the experiments but requires significant amount of computation efforts. Therefore, the state-of-the art neural signal decoding are mainly conducted on software, consuming a considerable amount of power, making it impractical for the long-term and daily use of the neural prosthesis. As discussed above, the next generation neural prosthesis calls for a miniaturized and less power hungry neural signal decoding that achieves real-time decoding. With an on-chip real-time neural signal decoding, the size and the power consumption of the computing device can be reduced effectively. Furthermore, up to a thousand micro-electrodes will be implanted in the next generation neural prosthesis. Integrating the neural decoding algorithm with neural recording device is also desired to reduce the wireless data transmission rate.

Until now, very few works give a solution for the on-chip neural decoding. A low-power neural decoding architecture using analog computing is proposed [83],

featuring optimization of mapping parameters in the training mode by continuous feeding of recorded neural signal and using the optimized mapping to generate the output in the operational mode. The architecture is largely an active filtering method involving massive parallel computing through low power analog filters and memories. Complicated learning algorithm of a modified gradient-descent approach is adopted on chip to minimize the error in a least-squares sense, adding to the complexity of the design. To achieve low power operation, sub-threshold design is used for lower biasing current, magnifying the mismatch and robustness issue in the analog circuits. Furthermore, except some SPICE simulation results, no measurement results are published to support the silicon viability of the architecture.

[113] proposes a neural decoding algorithm using spiking neuron network Kalman filter (SNN-KF). The proposed Kalman filter is realized by mapping it into a network of 20,000 spiking neurons. The decoding algorithm is used to decode 2-D cursor trajectory from neural data recorded by a 96-electrode array. The simulation using *Nengo*, a spiking neuron software package is conducted showing that the proposed algorithm deviates the results of the standard decoder within 0.03%. Future plan of this work is to implement the proposed algorithm on-chip with low-power neuromorphic circuits. The area of the chip, however, might be rather big if 20,000 spiking neurons are implemented in silicon.

A recent work proposes a universal computing architecture for neural signal decoding [1]. As depicted in Fig. 2.17, the architecture consists of internal part integrating with implanted neural recording device and external part. The internal part pre-processes the neural signal at each time step by doing binary classification for a series of possible states according a set of pre-defined rules. Only classification decision vector is transmitted to external device, reducing data rate by a factor of 10000. The transmitted data is further processed by external device in a non-causal manner, selecting a most probable state. The computation power is distributed unbalanced

Figure 2.17: A computing architecture for neural signal decoding proposed in [1]. Unit I applies a set of pre-defined rule to raw neural data, with thresholds learned externally. Only features generated by internal computation are transmitted and further processed by external computation unit E to generate decoding output.

between internal and external part, where internal part performs only light-weight logic but reduce data rate effectively and external part with less power constraint finishes more complicated computation required by the algorithm. The architecture is claimed to be universal, capable of implementing various decoding algorithm. An example using pattern matching algorithm is shown and implemented in field programmable gate array (FPGA) is shown with verification in rodent animal experiment. The power consumption of the FPGA for implementing this example is 537 μW .

In this work, a hardware-friendly neural signal decoding algorithm is proposed based on the Extreme Learning Machine. The algorithm is essentially a single hidden-layer feed-forward neural network. A neuromorphic IC is implemented in CMOS process to realize the proposed algorithm, being the first motor intention decoding ASIC in the world. The detail of the algorithm and circuits implementation is described in Chapter 5.

2.8 Summary

To sum up this chapter, a diagram of signal chain in the intra-cortical BMI for neural prosthesis is shown in Fig. 2.18. The dashed lines indicate the alternative paths. To be specific, the ADC can be shut down for the sake of power saving, if analog processing circuits are available. As pointed out earlier in 2.7.4, the spike sorting usually involves a rather complicated algorithm and considerable computing effort, though its necessity is still open to debate. Presence of the Local Field Potential (LFP) in the neural signal makes the spike detection more complicated. Therefore, the LFP is usually removed or separated from amplified neural signals before the spike detection. Studies have shown that the LFP can also be used in motor intention decoding [75] [76]. Therefore, a signal processing and decoding path can also be extended from the LFP side to build a neural signal decoding system that uses the LFP as input signal. This work focuses on motor intention decoding using the spikes, due to higher temporal and spatial resolution of the spike signals. Hence, the details of motor intention decoding using LFP is not presented here and it is not in the scope of this work.

Figure 2.18: Diagram of the signal chain in the intra-cortical BMI for neural prosthesis

Chapter 3

A Digital Assisted Signal Folding Amplifier for Intra-Cortical Neural Recording Applications

As mentioned in the literature review, significant amount of research work had focused on optimizing the neural amplifier design and also competing on the NEF. In contrast, this work places emphasis on other problems such as robustness, resistance to interference while power savings are obtained at the system level without hampering NEF. In this chapter, a neural amplifier featuring signal folding is proposed with a reconstruction algorithm in the digital domain to recover the amplified signal. This scheme provides the benefits of reducing the resolution requirement of the ADC and the transmission bit rate as well as eliminating the need for pseudo-resistors in the design. Two versions of CMOS implementation of the proposed signal folding neural amplifier are presented in this chapter, followed by presenting the measurement results, which includes *in vivo* data from the pre-frontal cortex of rat. Furthermore, a detailed analysis of the reconstruction error is presented. A behavioral model is also built for rapid prediction on the signal to noise and distortion ratio (SNDR) of the system for arbitrary inputs. Lastly, an estimation on the power saving of the en-

tire front-end circuits with multiple recording channels is given, demonstrating that the proposed signal folding scheme can reduce the power per channel to 80% of the power consumed by the conventional design [114].

3.1 Signal Folding

3.1.1 Concept

The proposed signal folding concept is illustrated in Fig. 3.1. Here a neural recording system with a noise floor of $2\ \mu\text{V}$ is considered. Assuming the maximum input voltage range for LFP to be 2 mV, the dynamic range is 60 dB, leading to a 10-bit ADC resolution. With a gain of 500 provided by the amplifier, the input range of the ADC should be 1 V in order to cover the whole range of neural signal, between the two solid lines as shown in Fig. 3.1 (a). Reducing the NOB by 2, the ADC input range narrows to 250 mV. As indicated by the horizontal dashed lines in Fig. 3.1 (a), applying this 8-bit ADC directly to the neural signal, the information beyond the ADC input range is lost. However, in the proposed amplifier, the output signal is reset to a common DC level (vertical dashed lines) whenever it goes higher than an upper comparison threshold or goes below a lower threshold. Setting the difference between the upper and lower thresholds to be 250 mV, as shown in the lower part of Fig. 3.1(b), the output signal is folded into a smaller range that fits the input range of an 8-bit ADC. All the information can thus be digitized by the ADC with the reduced NOB. A reconstruction process is needed to recreate an amplified version of the original input signal after the data is transmitted out of the body.

As shown in Fig. 3.3(b), comparison and reset circuits are implemented together with a fully differential amplifier to realize the signal folding scheme. The core amplifier may employ the typical AC-coupled capacitive feedback topology with the closed-loop gain G fixed to C_1/C_2 and the low-pass corner of the amplifier set to

Figure 3.1: Signal folding concept: (a) A 10-bit ADC is needed to cover the whole range of the large output from the amplifier. (b) Applying the proposed scheme, the large signal is folded into a smaller range that fits the input range of a 8-bit ADC.

$G_m/2\pi GC_L$, where G_m is the transconductance of the OTA and C_L is the load capacitance. The outputs are compared with a threshold voltage $V_{th} = V_{cm} - \Delta V_{th}$ (separate thresholds V_{hi} and V_{lo} may also be used as shown in Fig. 3.2 [115] [116]). Whenever any one of the differential outputs (V_{o+} and V_{o-}) is lower than V_{th} , a narrow pulse will be generated by the ‘reset pulse generator’ block, resetting the input and output nodes of the amplifier to V_{incm} and V_{cm} respectively. The amplifier continues its amplifying function after the reset. The output signal of the amplifier is then folded into a voltage range defined by $2\Delta V_{th}$. Since the DC level of the capacitive amplifier is defined by the reset operation, pseudo-resistors used for DC biasing in [65] can be removed. The input nodes of the OTA become floating after the reset with the DC bias being provided by the charge stored on these nodes. Due to leakage current on the input node, the output common-mode voltage will shift away from V_{cm} until the reset circuit is triggered again. This scheme allows us to maintain the input common-mode level within a workable range in the absence of pseudo-resistors. The low-frequency components added to the output signal due to the slow common-mode change can be easily eliminated by a high pass filter after the reconstruction in the digital domain. The proposed neural recording system can benefit from the signal

folding concept in the following ways:

- **Reduced ADC resolution and transmission bit rate**

As shown earlier, the proposed scheme can reduce the required input dynamic range of the ADC by folding a large output signal of the neural amplifier into a smaller range, leading to area and power savings. This also proportionally reduces the number of bits to be transmitted.

- **Lowered high pass corner for LFP acquisition**

The leakage current at the input nodes of the OTA, which can be modeled as a parasitic resistor R_{leak} , forms a high pass filter with the effective capacitance at the input node. The corner frequency of this filter, $1/2\pi R_{leak}(C_1 + A_o C_2)$, is much lower than $1/2\pi R_{pseudo}C_2$ which is formed by pseudo-resistors and feedback capacitor C_2 . A_o denotes open-loop gain of OTA. This guarantees a high-pass corner frequency low enough to capture LFP signals.

- **Reduced harmonic distortion and operation with lower supply voltages**

The amplifier output signal is confined in a small voltage range so that the harmonic distortion caused by the amplifier can be reduced. Moreover, since a large signal is folded in to a smaller range, reduced power supplies can still provide the large ‘effective’ dynamic range.

This reset and reconstruction scheme is general and can work with any core amplifier topology in any CMOS process. Two different versions of the proposed design are implemented in silicon to verify this scheme. The detailed design of circuits used in the two versions are described in the following sub-sections.

3.1.2 Version-One Implementation

A version-one implementation fabricated using UMC 65-nm process has been described in detail in [115] [116]—here, it is briefly described for the sake of com-

Figure 3.2: Version-one design of the proposed aAmplifier - die photo and schematics: (a) The rectangle indicates the die area where the proposed amplifier has been fabricated using UMC 65-nm process. (b) Comparison and reset circuits are used to fold the output signal of the amplifier. Unlike the conventional AC-coupled capacitive feedback amplifier, pseudo-resistors are not used. (c) OTA using a fully differential folded-cascode amplifier topology. (d) Comparator used for upper threshold voltage comparison

pleteness. The schematic of version-one implementation is shown in Fig. 3.2. The OTA used in this design is a single stage fully differential folded-cascode amplifier. A large current scaling (8:1) between the input and output branches is employed to reduce the power consumption [66]. Since this is a fully differential circuit unlike the circuits in [65] and [66], a common mode feedback (CMFB) circuit is used to define the output DC level to be V_{cm} . Thick-oxide devices are used for the input transistors in the differential pair to minimize the gate leakage current at the input node. C_1 and C_2 are set to be 20 pF and 200 fF respectively, aiming at a gain of 40 dB. Two different comparators are used in the comparison and reset circuits. Hysteresis is induced in both comparators to prevent the amplifier from being reset frequently. In the comparator used for the upper-threshold comparison, NMOS transistors are

3.1.3 Version-Two Implementation

A version-two implementation is fabricated using GlobalFoundries 0.18- μm process with a gain of 500 to enhance the capability of recording small spike signals. The die photo and schematics of the version-two design are presented in Fig. 3.3. A fully differential two-stage topology is employed in the OTA to achieve sufficient open-loop gain and output driving capability. As depicted in Fig. 3.3(c), two NMOS transistors (M_1 and M_2) along with two PMOS transistors (M_3 and M_4) are biased at the weak inversion region as input transistors to boost g_m/I_d and obtain lower thermal noise [117]. Large input gate areas are chosen to reduce the flicker noise. V_{o1+} and V_{o1-} are simply connected to the gate of M_5 and M_6 respectively to form a CMFB path for the input stage. At the output stage, a resistive-sensing CMFB utilizing pseudo-resistors is used to define the output DC level. The CMFB benefits from the large resistance of the pseudo-resistors in achieving negligible power consumption and loading effect. To achieve a gain of 500, C_1 and C_2 are set to be 20 pF and 40 fF respectively. C_2 is implemented by five 200 fF capacitors connected in series to reduce the gain inaccuracy caused by parasitic capacitances.

The version-two design has simpler comparison circuits, as illustrated in Fig. 3.3(b). It takes advantage of the characteristics of differential outputs so that only one threshold voltage V_{th} and one type of comparator are used with proper connections. Only two inputs are needed for the comparator input stage, as shown in Fig. 3.3 (d).

3.1.4 Transient Simulation Results

The signal-folding operation is demonstrated by SPICE simulation results of the version-one design depicted in Fig. 3.4. The input signal is a 300 Hz 6 mV_{pp} sinusoid. V_{Io} and V_{hi} are set to be 0.5 V and 0.7 V respectively with a VDD of 1.2 V and a V_{cm} of 0.6 V. The folded output waveform is shown in Fig. 3.4 (b). The resetting

(a)

(b)

(c)

Figure 3.4: Transient simulation results: (a) Input signal of the amplifier—a 300 Hz 6 mV_{pp} sinusoid, (b) Output waveform of the version one amplifier in simulation, (c) Resetting pulses generated by the comparison and reset circuits.

pulses generated by the comparison and reset circuits are shown in Fig. 3.4 (c).

3.2 Reconstruction

In order to recover the amplified signal, a non-Nyquist reconstruction process is applied to the signal digitized by the ADC. This is the crux of our ability to reduce the required dynamic range and relax the ADC design without losing any information in large signals. The algorithm is described as follows and with the folded output values, the reconstructed output values and the additive correction terms denoted as $y(n)$, $u(n)$ and $d(n)$ respectively where n is the sample index.

Figure 3.5: Reconstruction algorithm: (a) Folded output waveform y and the operation of setting the reset flag. (b) Derivative of the folded output waveform and the operation of interpolation and prediction. (c) Reconstructed output waveform u and the operation of DC level correction.

As illustrated in Fig. 3.5, the proposed reconstruction algorithm consists of three steps:

1) Setting the reset flags

To find the sample point at which a reset event occurred, the derivative of the output signal y is computed and compared with a threshold. If it is larger than the threshold, a reset event is declared. This is possible as the bandwidth of neural signals is limited and is much narrower than the bandwidth of fast reset events.

2) Interpolation and prediction

Due to the bandwidth and slew rate limitation of the OTA, the neural amplifier needs a short period of settling time to recover from the reset mode back to the amplification mode. Also, there is an added offset right after reset due to charge injection from the switches. So a certain number of data points n_{stl} in this short period are discarded. Fig. 3.5 (b) shows derivative of the recorded data before and after a reset. In order to predict and recover the value of discarded data, a polynomial interpolation is performed to link the *derivative* values of the waveform before and after the reset. Suppose n_{rst} is the last sample before the reset occurred and n_0 and n_1 points respectively are chosen before and after the reset for curve fitting. Then $y(n_{rst} - n_0, n_{rst})$ and $y(n_{rst} + n_{stl}, n_{rst} + n_{stl} + n_1)$ denote the two waveform segments before and after reset. Finally, by integrating the derivative values obtained by the polynomial interpolation process, the value of the discarded data points as well as the first data after settling ($u(n_{rst} + n_{stl} + 1)$) can be predicted.

3) Correction

Every reset introduces a DC level shift between the data points before and after it. To recover the original waveform, a correction should be made to all the data after every reset. The amount of correction that needs to be added is given by $d(n_{rst} + n_{stl} + n) = u(n_{rst} + n_{stl} + 1) - y(n_{rst} + n_{stl} + 1)$. After the correction, the waveform segments before and after the reset will join and form a smooth curve.

This algorithm has been implemented in MATLAB and can be moved to hardware in the future.

3.3 Reconstruction Error Analysis

The reconstruction process does not create a perfect replica of the original neural signal; instead, some error is introduced in the prediction due to the finite polynomial order used in the curve fitting process as well as thermal noise and limited bandwidth of the amplifier. This reconstruction error, added to the source and circuit noise, can be a limiting factor in the achievable signal-to-noise ratio (SNR) of the neural recording system. As an example, consider the measured spectrum of the reconstructed signal for an amplifier (the version-two design) for different input sinusoids as shown in Fig. 3.6. The details of the measurement setup will be provided in the next section; but it is interesting to note the variation in the spectrum depending on input signal parameters. First, the spectrum of the reconstructed output for the same input signal frequency (100 Hz) but with different amplitudes are shown in Fig. 3.6(a). With increasing amplitude, the magnitude of the reconstruction error, mostly limited in the low frequency band, increases. The second trend observed in Fig. 3.6(b) is when the input signal frequency increases and amplitude stays constant at 0.8 mV_{pp}, the error increases again. It can also be observed that these errors do not affect the noise in the spike band due to its low-pass characteristic in the spectrum. Hence, analysis and modeling of these non-trivial characteristics of the reconstruction error are necessary to predict how the error affects the performance of the neural recording amplifier.

As mentioned in the previous section, the reconstruction algorithm is basically the addition of a predicted correction term to each data point in a recording sequence, which is represented as

$$u(n) = y(n) + d(n), \quad (3.1)$$

where $u(n)$, $y(n)$ and $d(n)$ are as described earlier. After the reconstruction, the

(a)

(b)

Figure 3.6: Output spectrum and reconstruction error: Plots of the output spectrum for input signals having (a) the same frequency (100 Hz) but different amplitudes or (b) same amplitude (0.8 mV_{pp}) but different frequencies.

signal passes through a digital high-pass filter. The final output signal of the whole recording system is given by

$$v(n) = hpf(n) * u(n) = hpf(n) * y(n) + hpf(n) * d(n), \quad (3.2)$$

where $hpf(n)$ denotes the impulse response of the high pass filter and ‘ $*$ ’ denotes the convolution operation. Applying the z-transform to both sides of the above equation

gives:

$$V(z) = HPF(z)Y(z) + HPF(z)D(z), \quad (3.3)$$

where $HPF(z)$, $Y(z)$ and $D(z)$ denote the z-transforms of $hpf(n)$, $y(n)$ and $d(n)$ respectively.

According to the reconstruction algorithm, $d(n)$ for a particular data point is calculated by accumulating the correction terms induced by all the preceding reset operations, which can be described by:

$$d(n) = d(n-1) + c(n) + \Delta c(n), \quad (3.4)$$

where $c(n)$ and $\Delta c(n)$ denote the ideal correction term and the correction error introduced by the reconstruction respectively. Intuitively, at every reset, $c(n)$ is the ideal value of correction to be applied to merge two segments of the waveform before and after the particular reset operation. Hence, to obtain the actual desired waveform, all such correction terms at the resets that have occurred till the current time are needed to be accumulated. However, due to the imprecision in estimating $c(n)$, an error $\Delta c(n)$ is introduced here. Both $c(n)$ and $\Delta c(n)$ are pulse sequences in which all the data points except those at reset time instants are zero. Both the time and amplitude of the pulses at the reset time instants are random. Applying the discrete Fourier transform (DFT) to both sides of (3.4) gives:

$$D(z) = \frac{C(z)}{1-z^{-1}} + \frac{\Delta C(z)}{1-z^{-1}}, \quad (3.5)$$

where the second term is the error term. Substituting $D(z)$ in (3.3) by the expression

in (3.5), the final output signal can be expressed as

$$V(z) = HPF(z) \cdot Y(z) + HPF(z) \cdot \frac{C(z)}{1 - z^{-1}} + HPF(z) \cdot \frac{\Delta C(z)}{1 - z^{-1}}. \quad (3.6)$$

The third term is the reconstruction error that appears in the final output signal. In this work, the high-pass filtering is realized by subtracting the moving average from the original signal. So the transfer function of the high-pass filter is given by:

$$HPF(z) = 1 - \frac{1 + z^{-1} + z^{-2} + \dots + z^{-(k-1)}}{k}. \quad (3.7)$$

The parameter k determines the high-pass corner given by the expression $f_{hi} = 1/kT_s$ where T_s is the sampling time in sec. When k is equal to the number of samples per second, the high-pass corner is 1 Hz.

If the resets happen at random instants introducing random errors in amplitude, it can be assumed that the spectrum of $\Delta c(n)$ is white, i.e. $\Delta C(f) = \bar{c}df$. Thus the spectrum of reconstruction error is a white noise spectrum shaped by a high-pass filter and integrator $\frac{1}{1-z^{-1}}$, which results in a low-pass-shaped spectrum where the major power components stay at low frequencies. It can be concluded that the total error power is proportional to the value of $\bar{c}$ for a fixed transfer function $HPF(z)$. However, the value of $\bar{c}$ is dependent on the input signal. This can be understood intuitively by noting that the magnitude of error will depend on the frequency of resets (more resets imply more error), which in turn depends on the amplitude and rate of change of the input. Hence, $\bar{c}$ will be denoted as $\bar{c}(V_{in})$ in the following sections to explicitly denote this dependence. Finally, an expression can be derived

for the spectrum of the total input referred noise and error ($V_{n,e}^{in}$) for this amplifier as:

$$\begin{aligned}
V_{n,e}^{in}(f) &= V_n^{in}(f) + V_e^{in}(f) \\
&= V_n^{in}(f) + \frac{1}{G} \cdot HPF(f) \cdot \frac{\bar{c}(V_{in})}{1 - e^{-j2\pi f T_s}} df \\
&\approx V_n^{in}(f) + \frac{1}{G} \cdot \frac{\bar{c}(V_{in})}{2\pi f T_s} (f \gg \frac{1}{kT_s})
\end{aligned} \tag{3.8}$$

where V_n^{in} and V_e^{in} denote the input referred noise and reconstruction error, respectively; f denotes the frequency variable and other symbols have the same meaning as defined earlier. In the following sections, we shall also use $\overline{V_{n,e}^{in}}$ to denote the rms quantity. V_n^{in} contains both thermal and flicker noise while V_e^{in} contains a signal dependent part c_2 (described earlier) and a signal independent (but process dependent) part c_1 due to the leakage current induced resets. As discussed in Section 3.1.1, we use reset scheme instead of pseudo-resistors to define the DC bias of the neural amplifier. When reset switches are off, due to leakage current on the input node, the output common-mode voltage will shift away from V_{cm} until the reset circuit is triggered again. These reset are denoted as leakage current induced resets in the following sections, different from the resets caused by the large amplitude input signals. The resets caused by the large amplitude input signals are denoted as signal induced resets. However, the signal independent part is very small (shown in section 3.4) and can be neglected for most cases. In other words, we can write:

$$\bar{c}(V_{in}) = c_1 + c_2(V_{in}) \approx c_2(V_{in}) \tag{3.9}$$

A corner frequency f_{cor} can be defined here, similar to the flicker noise corner, where the spectral density of V_e^{in} is same as that of the thermal noise component in V_n^{in} .

For pure sinusoids, estimates of the reconstruction error can be provided; however, due to the non-linear nature of the reconstruction, it is non-trivial to use those

Figure 3.7: Behavioral model: Circuits parameters such as bandwidth, input-referred noise, reset pulse width and gain are provided by Cadence simulation.

Figure 3.8: Spectrum of reconstructed signal in simulation based on behavioral model: The reconstructed signal spectrum and best fit theoretical curve with 10 Hz 2.2 mV_{pp} input signal.

estimates to predict error for a non-sinusoidal signal (e.g. LFP which has a broad spectrum). In order to evaluate the reconstruction error in design phase, a MATLAB-based time-domain behavioral model of the proposed amplifier, as depicted in Fig. 3.7 is introduced to generate the output waveforms according to given input waveforms. The behavioral model uses 4-th order Runge-Kutta method to emulate transient behavior of a one-pole amplifier with the output being reset upon threshold crossing events (details are presented in the appendix). Parameters needed in this behavioral model, e.g. low pass filter corner, gain, input referred noise level and

reset pulse width are derived from SPICE simulations. The output waveform of the behavioral model is then processed by the algorithm mentioned in section III. The resulting spectrum of the reconstructed waveform can be used to evaluate $\bar{c}$ and the reconstruction error for a given input waveform. An example of the output spectrum of behavioral model and the corresponding $\bar{c}$ curve fitting are shown in Fig. 3.8. Measurement results are shown later in Section 3.4 matching very well with the results given by the behavioral model. This behavioral model thus gives a quick estimate of $V_{n,e}^{in}$ in the design phase without consuming significant amount of time needed by transient noise simulations in SPICE.

3.4 Measurement Results

As explained in Section 3.1, two different versions of design employing different core amplifiers and reset circuits have been fabricated. The measurement results shown in this section are obtained from version-two fabricated in GF 0.18- μm process. It has a gain high enough for direct neural recording. Similar results are also obtained from the version-one design in UMC 65-nm. Details and performance of version-one design are presented in [115]. Also, pseudo-resistors (P-res) are implemented on chip and can be connected into the circuits as feedback elements by switches to facilitate the comparison between the proposed and conventional topologies. In the following text, conventional amplifier measurement is referred when the P-res are enabled, while they are disabled for the measurements of the proposed amplifier.

3.4.1 Small signal measurements

The frequency response and output noise spectrum of the conventional and proposed neural amplifier with a bias current of $2.37\ \mu\text{A}$ in the OTA are shown in Fig.

3.9. The frequency response and noise spectrum of the proposed amplifier are measured after reconstruction. The mid-band gain of the proposed amplifier is 54.2 dB, higher than the mid-band gain of 53.4 dB in the conventional topology. This is due to reduced parasitic capacitor in feedback path by disconnecting P-res. With P-res removed, the gain of the proposed amplifier does not decrease significantly even in frequency below 30 mHz, indicating a much lower high-pass corner than that of 1.7 Hz in the conventional amplifier. This high-pass corner, determined by leakage current at the input node of the OTA, is much lower than the lower end of desired neural signal frequency—hence variations in this corner frequency do not affect the final high-pass corner that can be set by a digital filter. The low-pass corner in both configurations is 5.7 kHz.

The output noise spectrum of both configurations are shown in Fig. 3.9 (b). Both spectra are virtually overlapping each other thus validating our claim in the earlier section that $c_1 \ll c_2$, i.e. the reconstruction noise due to leakage current induced reset is negligible compared to V_n^{in} , the thermal and flicker noise of the amplifier. The total input-referred noise in the spike band is obtained by integrating the output noise spectrum from 1 Hz to 100 kHz and dividing by the mid-band gain after applying a digital high-pass filter with 200 Hz cut-off. The input-referred noise of LFP band is obtained in the same way but after applying a digital low-pass filter with 200 Hz cut-off. The power dissipation of the proposed amplifier is measured with a 2 mV_{pp} input signal to include dynamic power of comparators. With a total power consumption of 2.52 μ W, NEF of the proposed neural amplifier in spike band is 3.03, still competitive among other recent designs. The details of noise performance as well as power breakdown is summarized in Table 3.1.

Figure 3.9: Frequency response and output noise spectrum: (a) Frequency response curves and (b) output noise spectrum of the proposed and conventional amplifier topologies.

3.4.2 Results of Reconstruction Algorithm

To verify the viability of the proposed signal folding and reconstruction scheme, sinusoids with different amplitudes and frequencies are first chosen as test inputs. Aiming to reduce ADC NOB by 2 bits, the output range of the proposed amplifier is limited by proper threshold voltage to one fourth of VDD (full scale voltage range of 10-bit ADC is VDD). The output waveform of the proposed amplifier is digitized by an Agilent measurement module with a sampling rate of 20 ksps and is then processed by the MATLAB-based reconstruction algorithm on a PC to recover the

Table 3.1: Noise performance and power breakdown

Parameter	Conventional amplifier	Signal-folding amplifier
VDD	1 V	1 V
Input-referred noise (LFP)	$1.35 \mu\text{V}$	$1.22 \mu\text{V}$
Input-referred noise (spike)	$3.76 \mu\text{V}$	$3.74 \mu\text{V}$
Total power	$2.37 \mu\text{W}$	$2.52 \mu\text{W}$
OTA power	$2.37 \mu\text{W}$	$2.37 \mu\text{W}$
Comparator power	NA	$0.15 \mu\text{W}$
NEF (LFP)	5.65	5.27
NEF (spike)	2.95	3.03

Power dissipations of the signal-folding amplifier and its comparator are measured with an input signal amplitude of 2.2 mV_{pp} .

amplified signal. Fig. 3.10 (a) and (b) depict the digitized output signal and the recovered signal respectively when a 10 Hz 1 mV_{pp} sinusoid is applied to the input, while Fig. 3.10 (c) and (d) show the same in frequency domain. Though the tone at 10 Hz is visible in Fig. 3.10(c), a lot of high frequency noise can be observed due to the reset phenomenon which is removed after the reconstruction in Fig. 3.10(d). This underlines the non-Nyquist nature of the reconstruction algorithm. Also, note that $f_{cor} \approx 150$ Hz implying large and slow LFP signals will not add extra noise to the spike band data.

3.4.3 Distortion and Interference

Two types of measurements for distortion are considered. First, in Fig. 3.11, the input signal is sum of a 1 kHz 100 μV_{pp} sinusoid and a 50 Hz 4 mV_{pp} sinusoid. This is similar to the actual neural signal in the sense that it is the sum of a small

Figure 3.10: Measured output waveforms and spectrum for 10 Hz 1 mV_{pp} sinusoidal input: Digitized differential output signal in (a) time domain and (c) frequency domain, (b) Reconstructed output signal in (b) time domain and (d) frequency domain.

and fast signal and a large and slow signal. It can also represent the situation of a large 50/60-Hz interference coupled with neural signal. Note that this is a differential input signal and hence represents an extreme case of 50/60-Hz interference where the common mode signal is converted into an effective differential signal due to electrode mismatch. For a conventional amplifier operating linearly, the differential output swing in this case would be 2 V_{pp} which is twice the VDD used here (and hence equal to the maximum theoretical swing for a differential amplifier). Therefore, for a realistic amplifier, clipped outputs are expected, in the conventional case as seen in Fig. 3.11(b). However, the reconstructed output from the proposed amplifier shown in Fig. 3.11(a) does not exhibit clipping since its dynamic range can be extended beyond VDD by the reconstruction algorithm. An even more interesting observation can be made by high-pass filtering the output to display the small signals

Figure 3.11: Measured output waveforms and spectrum for a two-tone input signal (1 kHz $100 \mu\text{V}_{\text{pp}}$ sinusoid added on a 50 Hz 4 mV_{pp} sinusoid): (a) Reconstructed output of the proposed amplifier. (b) Digitized output of the conventional amplifier. Spike band signal at the output of the (c) proposed and (d) the conventional amplifiers. Output spectrum for the (e) proposed and (f) conventional amplifiers.

in the spike band. Fig. 3.11(c) and (d) plot the spike band outputs for the proposed and conventional amplifiers. It can be clearly seen that for the conventional amplifier, the small 1 kHz signal is severely distorted whenever the amplifier is saturated by the large 50 Hz signal, while this effect is significantly less for the proposed amplifier. Fig. 3.11(e) and 3.11(f) plot the spectrum of the output in both cases. The large amplitude 50 Hz signal can be clearly separated from the small amplitude 1 kHz one

after reconstruction with remaining intermodulation products at least 36 dB smaller than the 1 kHz tone and around 60 dB smaller than the interference. In case the interference is smaller, this intermodulation product will also be smaller and get buried in the noise floor. The 50/60-Hz interference can be removed later by a digital notch filter without degrading the desired signal. However, for the conventional amplifier, the intermodulation product is only 7 dB lower than the 1 kHz tone indicating the severity of distortion. Also, there are many more intermodulation products within the spike band for the conventional amplifier. Therefore, this example demonstrates a big advantage of the proposed amplifier in spike recording applications.

Another example of a commonly encountered interference signal in the recording of bio-potentials is motion artifact [118] [119]. For example, a sudden movement of the biological subject can create a large and fast pulse-like common mode motion artifact (MA) signal. However, the repetition rate of such signals are small (<50 Hz) due to mechanical speed limitations. Hence, in analogy with the previous example, the proposed amplifier should be able to record neural spikes even in presence of such large infrequent interferers. To validate this, an artificial signal is created comprising 6.5 mV_{pp} MA pulses superposed on pre-recorded $100 \text{ } \mu\text{V}_{pp}$ neural spikes and 2 mV_{pp} LFP. Again, the presumption made here is that mismatch has resulted in common-mode to differential conversion of the MA signal. This signal is provided as an input to both the conventional and proposed amplifiers using an NI PXI-5422 arbitrary waveform generator. Fig. 3.12(a) plots the result of this experiment and it can be seen that as expected the conventional amplifier is saturated while the proposed one is not. Fig. 3.12(b) shows that the proposed amplifier can preserve the information in the LFP signal before and after the MA event—intuitively, extra reset events are generated by MA only in a limited time and hence do not corrupt the signal before and after it. Finally, Fig. 3.12(c) depicts that the proposed amplifier can record spikes even during a large MA while a conventional amplifier gets saturated

Figure 3.12: Neural recording in the presence of motion artifact: A synthesized waveform containing LFP, spike and motion artifact signals is the input to both the proposed amplifier and conventional amplifier. LFP band signals of the output of both amplifiers as well as LFP band signal of the gained input are compared in (a). A zoomed in view in (b) shows that the proposed amplifier can preserve the information in LFP band very well except in the period when motion artifact happens. Spike band signals of the output of both amplifiers as well as spike band signal of the gained input are compared in (c).

Figure 3.13: Harmonic distortion in the proposed and conventional amplifiers: Output spectrum of the (a) conventional and (b) proposed amplifier with a 3 Hz 2 mV_{pp} sinusoid input. (c) THD at 3 Hz and 1 kHz.

and cannot record small spike signals at that time.

The total harmonic distortion (THD) is also measured for our proposed amplifier and compared it with the conventional one. Fig. 3.13 compares the output spectrum from the proposed and conventional amplifiers when the input signal is a 3 Hz 2 mV_{pp} sinusoid. The output spectrum of the conventional amplifier shown in Fig. 3.13(a) has obvious harmonic components at 6 Hz, 9 Hz and 12 Hz, etc., indicating significant distortion from the circuit when the input amplitude is large. However, the output spectrum of the proposed amplifier, as shown in Fig. 3.13(b) is free from these harmonic components. THD of the proposed amplifier ($< 1\%$) is considerably

smaller than the conventional one (Fig. 3.13(c)) for the large signals in the LFP band. For higher frequency inputs in the spike band, the proposed amplifier has THD of ($< 1\%$) for small inputs and is comparable to the conventional one. But its performance becomes worse ($\text{THD} \approx 3\%$) than the conventional for amplitudes larger than $0.5 \text{ mV}_{\text{pp}}$ due to signal induced resets. But this is not a problem in the neural recording applications since spike band signals are typically less than $0.2 \text{ mV}_{\text{pp}}$ [53]. In fact, this increased distortion provides a guideline for setting ΔV_{th} —it has to be large enough to prevent signal induced reset for the largest spike signal. However, it can be noted that though the proposed amplifier does not cause significant harmonic distortion, it does introduce reconstruction errors at low frequencies which will be analyzed next.

3.4.4 Noise and Reconstruction Error

As mentioned in the earlier section, the reconstruction error V_e^{in} varies with the input signal and can be parameterized by $\bar{c}(V_{in})$. To justify this assertion, the spectrum of the reconstructed signal obtained by varying amplitude and frequency parameters of the input sinusoid is shown in Fig. 3.14(a) and (b). Clearly, the error changes depending on input V_{in} . It can be observed that the shape of the reconstructed signal spectrum of measurement and behavioral model match quite well. To further validate this, $\bar{c}$ derived from measurement and behavioral model is compared in Fig. 3.14(c). It can be seen that the behavioral model predicts $\bar{c}$ quite accurately for large error magnitudes and hence can be used to get a quick estimate of the error in design phase. The discrepancy for small errors can be attributed to inaccurate modeling of the leakage induced resets contributing to the term c_1 in eq. 3.9 and flicker noise which would contribute to V_n^{in} in eq. 3.8.

Next, the rms input referred noise $\overline{V_{n,e}^{in}}$ is measured for both spike and LFP bands with different sinusoidal inputs. The bands are defined by using a digital band-pass

Figure 3.14: Output spectrum and reconstruction error: plots of the output spectrum for input signals having (a) the same frequency (100 Hz) but different amplitudes or (b) same amplitude (0.8 mV_{pp}) but different frequencies. (c) Comparison of $\bar{C}$ derived from measurement results with the behavioral model.

filter from 1 Hz to 200 Hz for the LFP band, and a high-pass filter with corner of 200 Hz for the spike band. For input signals located in spike band, only the situation that the amplitude is small and it does not trigger a reset by itself is considered. This is consistent with the fact that spike signals are much smaller than LFP. The total input-referred noise varies with different input amplitude and frequencies as shown in Fig. 3.15(a) and (b). Observing the curves in these figures, it can be found that $\overline{V_{n,e}^{in}}$ does not change much with different input signal amplitudes and frequencies

Figure 3.15: Total input-referred noise and SNDR; total rms Input-referred noise $\overline{V_{n,e}^{in}}$ in: (a) LFP band and (b) spike band, (c) variation of SNDR with input amplitude.

when the output signal amplitude is smaller than the threshold with an average of 1.3 μV_{rms} in the LFP band and 3.83 μV_{rms} in the spike band. Hence, it is obvious that for spike signals, $V_{n,e}^{in} \approx V_n^{in}$ as could be expected from earlier noise measurements presented in Fig. 3.9(b).

When the output signal amplitude is larger than the threshold as shown in Fig. 3.15(a), $\overline{V_{n,e}^{in}}$ increases with the input signal amplitude and frequency as expected. The SNDR for every input is shown in Fig. 3.15(c). With input amplitude smaller than threshold voltage (around 0.5 mV referred back to input), SNDR is signal-

Figure 3.16: Time domain characterization of reconstruction error in LFP band shows strong correlation with frequency of reset.

limited, increasing with larger input amplitude. At larger amplitude end beyond threshold voltage, increasing number of resets introduces more reconstruction error. Hence, SNDR becomes error-limited and tends to saturate or even reduces. For 1 Hz inputs, SNDR saturates at ≈ 8 mV_{pp} inputs, to a value of 46 dB. Combining this maximum input amplitude with the noise floor in LFP band of $1.3 \mu\text{V}_{\text{rms}}$, the dynamic range in LFP band for this amplifier is evaluated to be around 66 dB. This difference in maximum value of SNDR and dynamic range is typical of systems where the noise floor increases with increasing signal level [120] [121] [122].

To demonstrate the performance for realistic LFP signals, the amplifier is also tested with a segment of pre-recorded LFP waveform as input. The maximal amplitude of this waveform was increased from 2-8 mV (smaller amplitudes cannot be set due to equipment limitations). The resulting SNDR plotted in Fig. 3.15(c) increases with amplitude reaching approximately 38 dB for 8 mV_{pp} inputs. The prediction from the behavioral model, also shown in the figure, is close to the measured one

Figure 3.17: Experimental setup for *in vivo* measurement.

with a maximum SNDR value of 33 dB. This difference is probably due to inaccurate modeling of thermal and flicker noise in the behavioral model.

Finally, the reconstruction error for LFP band inputs in time domain is also characterized. Intuitively, input signals generating a higher number of resets should have larger V_e^{in} and hence larger $\overline{V_{n,e}^{in}}$. So the earlier data for $\overline{V_{n,e}^{in}}$ in LFP band with sinusoidal inputs are plotted against the frequency of resets generated by the signal as shown in Fig. 3.16. It can be seen that for inputs producing reset frequency $< \approx 1$ Hz, $\overline{V_{n,e}^{in}}$ is almost constant at $1.3 \mu\text{V}_{\text{rms}}$ implying again, that leakage current induced resets cause minimal distortion. However, when frequency of resets increase beyond this, it produces a roughly linearly increasing error profile clearly displaying the strong correlation suggested in the earlier section.

3.4.5 *In vivo* Measurements

The proposed neural amplifier is further verified by an *in vivo* recording. The experimental setup is depicted in Fig. 3.17. Sprague-Dawley rats were anaesthetized with chloral hydrate (400 mg/kg) and fixed in a stereotaxic frame. A burr hole was drilled at a coordinate of +3.3 mm anterior to the bregma and 0.8 mm lateral to the midline. A Starbore glass electrode (Radnoti, USA) of 5-10 M Ω impedance, filled with 2M sodium chloride was gradually lowered into the brain using a microdrive (IVM Scientifica, UK). At 3-4 mm beneath the skull, pyramidal neurons of the pre-frontal cortex were encountered and identified according to their electrophysiological characteristics as described previously [123–125]. The local field potentials were recorded simultaneously from the same brain region. The neural signal waveform recorded by the proposed amplifier is shown in Fig. 3.18. Fig. 3.18(a) plots the raw digitized data and reconstructed data showing both spikes and LFP, while separated spike and LFP band data are plotted in Fig. 3.18(b)-(d). The digital filters used to separate the two bands are two band-pass filters passing frequencies from 1-200 Hz for the LFP signal and 200 Hz- 5 kHz for the spike signal. This experiment proves the viability of the proposed amplifier for neural recording applications.

3.5 System Level Savings

As mentioned earlier, the proposed neural amplifier reduces the NOB of ADC by 2 bits. To evaluate the savings due to this over a conventional amplifier, system level power dissipation for both cases is estimated, based on the details of a 100-channel neural recording system described in our earlier work [126].

To complete the recording system, the analog outputs from neural amplifiers presented in this work are buffered and multiplexed for analog-to-digital conversion. In the case of conventional amplifier, buffer and 10-bit SAR-ADC will be the same as

Figure 3.18: Neural activity recorded *in vivo*: The neural signal recorded by the proposed amplifier (referred to input) is shown. (a) The raw recorded signal and reconstructed output with both LFP and spikes (waveforms are shifted for ease of viewing). It is filtered into (a) spike and (c) LFP bands, which are plotted after passing the recorded signal through digital band-pass filters. (b) and (d) are zoomed versions of (a) and (c) respectively.

those in [126]. With one ADC shared by 10 analog channels, total power and area consumption of one neural recording channel are thus $5.70 \mu\text{W}$ and 0.124 mm^2 respectively. In the case of signal-folding amplifier, NOB of ADC can be reduced by 2 bits, resulting in an area reduction of 75% in charge-redistribution capacitor array used in SAR-ADC. Since the capacitor array occupies $\approx 80\%$ area in SAR-ADC implemented in [126], the ADC area can be reduced to 40% of original by using an 8-bit ADC. Two 8-bit ADCs can now approximately fit into same area previously taken by one 10-bit ADC. Therefore, number of ADCs used in the system is doubled while the sampling rate of a single ADC is halved. This leads to a 50% reduction in power consumption of single ADC but does not change the ADC power for a single channel. However, the power required by the driving buffer becomes 50% due to increased settling time. 2-bit reduction can furthermore cut power consumption of ADC by 75% in capacitor switching activity due to reduced capacitor sizes and 20% in digital (SAR logic and time-domain comparator) switching activity due to reduced conversion cycles. With a power breakdown of 1 : 3 between capacitor and digital switching, the power saving is thus 66.9% for a single ADC while power per channel is reduced to 80% of the original.

As seen in the preceding calculation, signal-folding scheme enables a power saving of up to $1.17 \mu\text{W}$ per channel compared with the conventional topology with only a negligible increase in area due to additional comparison and reset circuits. This is a significant amount in an implanted neural recording system with large number of channels. The savings are even more if the bit-rate reduction in the transmitter are taken into account. The calculation details used in this comparison are listed in Table 3.2. It should be noted that the proposed signal-folding scheme is a general strategy to reduce the dynamic range of the neural recording channel and can be applied to other neural recording systems such as [47] [48] [64], with a similar system level topology to [126]. The specific figure of power and area saving differs from case to

Table 3.2: Comparison of signal-folding and conventional topology in system level

Parameter	Conventional topology	Signal-folding
Process	0.18- μm	0.18- μm
VDD	1 V	1 V
Sampling rate	20 kSa/s	20 kSa/s
Neural amp. area	0.102 mm ²	0.109 mm ²
Neural amp. power	2.37 μW	2.52 μW
Buffer area	0.01 mm ²	0.01 mm ²
ADC resolution	10-bit	8-bit
No. of ch.s per ADC	10	5
ADC power (single)	21.6 μW (200 kSa/s)	7.16 μW (100 kSa/s)
ADC power per ch.	2.16 μW	1.43 μW
Total power per ch. (amp.+buf.+ADC)	5.70 μW	4.53 μW
Total area per ch. (amp.+buf.+ADC)	0.124 mm ²	0.129 mm ²
Data rate per 100 channels	20 Mbps	16 Mbps
DR	49.4 dB	66 dB
SNDR _{max}	49.4 dB	43 dB

case due to different system parameters and specifications of building blocks.

3.6 Summary

A signal folding and reconstruction scheme is presented here, reducing the dynamic range requirement of the ADC in an implantable neural recording system by exploiting the $1/f^n$ characteristics of neural signals. This is realized by folding the large output waveform into a smaller range and recovering it back by a non-Nyquist reconstruction algorithm in the digital domain after the data transmission.

Table 3.3: Performance summary and comparison of neural amplifiers

Ref	Process	VDD (V)	Power (μW)	Area (mm^2)	BW (kHz)	Gain (dB)	CMRR (dB)	PSRR (dB)	Noise (μV)	NEF	NEF ² · VDD	DR (dB)	THD
This work	0.18 μm	1	2.52	0.109	5.7	54.2	65	67	3.83 (200Hz-10kHz)	3.09	9.72	66	0.2% @ 1.54·VDD ³
[65]	1.5 μm	+/-2.5	80	0.16	7.2	40	83	85	2.2 (0.5Hz-50kHz)	4.0	80	60	1% @ 0.25·VDD
[66]	0.5 μm	2.8	7.56	0.16	7.2	40.8	66	75	3.06 (10Hz-98kHz)	2.67	19.96	63.7	1% @ 0.26·VDD
[74]	65nm	0.5	4.79	0.0082	10	36 ²	75	64	4.9 (1Hz-10kHz)	5.99 ¹	17.96	NA	NA
[78]	0.13 μm	1.2	35	0.035	10	40/56	NA	NA	2.2 (100Hz-10kHz)	4.5	24.3	60.3	NA
[68]	0.18 μm	0.45	0.73	0.2025	6.2	52/58	73	NA	3.2 (1Hz-10kHz)	1.57	1.12	47.9	1% @ 0.9·MAXo/p

¹ NEF and NEF²·VDD in [74] is based on measurement of whole recording channel.

² Figure listed here is the gain of first-stage amplifier in [74].

³ THD is quoted for 3 Hz input signal. For 100 Hz input, THD is 1.5% @ 1.54·VDD

The signal-folding amplifier described in this thesis achieves simultaneous recording of the LFP and spike signal with reduced signal dynamic range and bit rate. Therefore, it provides an alternative way of neural signal acquisition as well as other bio-potential recording where interference is significant. Measurement results are presented to prove the viability of the proposed signal folding and reconstruction scheme. Other benefits of this scheme include and lower high-pass corner due to the elimination of P-res as well as reduced harmonic distortion and lower supply operation due to smaller output signal swing.

Table 3.3 summarizes and compares performance of the neural amplifier presented in this thesis and other state of the art designs. A DC-coupled front-end circuit is proposed in [127] to reduce the area consumed by large input capacitors in AC-coupled solutions as in [65]. This work exploits advantage of deep sub-micron process in digital domain to achieve a compact realization of simultaneous digitization of the LFP and spike signals with high power efficiency. Another work described in [78] is dedicated to achieve instrument grade signal acquisition. The LFP and spike signals are digitized simultaneously by a per-pixel 10-bit ADC. Switched-capacitor filtering technique is used to provide a well-controlled frequency response in the acquisition front-end. Both [127] and [78] achieve simultaneous recording of the LFP and spike signals but do not reduce signal dynamic range as well as wireless transmission rate. By detecting and folding output signal with $2 \cdot V_{DD}$ swing from VGA, [68] enables recording channel operation under a supply voltage as low as 0.45 V, leading to a power reduction of 50% without degrading dynamic range performance. This recording system, however, suffers from large area consumption of analog front-end and increased circuit complexity due to charge-pump and LDO required to generate $2 \cdot V_{DD}$ power supply for VGA. The neural amplifier presented in this thesis achieves good trade-off between noise performance, area and power consumption, providing a competitive solution for neural recording front-end. Fur-

thermore, the signal-folding scheme enables simultaneous recording of the LFP and spike signals while reducing power per recording channel and bit-rate to 80% of the conventional one.

Chapter 4

A Simultaneous Neural Recording and Stimulation System Using Signal Folding in Recording Circuits

As mentioned in the previous chapters, one of the trends in intra-cortical BMI is to integrate neural stimulation circuits with neural recording channels so that a feedback loop from external environment or computing devices to the brain can be built and coordinated closely with the feed-forward path. In neural prosthesis, for instance, the sensory feedback from sensors is applied to the brain cortex by the neural stimulation integrated with the MEA, as shown in Fig. 2.18. In other applications [11] [128], the implanted BMI is implemented to apply stimulation current to the brain as a preventive or therapeutic treatment for various neurological disorders including epilepsy, Parkinson's disease, etc. The integration of neural recording and stimulation as well as signal processing circuits can provide a closed loop for long term monitoring of patient status and automatically applying treatment according to results of real-time recording and signal processing. Finally, the integration of neural stimulation and recording is also essentially for neuro-scientific experiment where not only intrinsic behavior of neurons but also how these neurons response to a stimulation is

observed [129].

4.1 Overview

Several works have been done with an effort to realize integration of both the neural recording and the neural stimulation [11] [64] [128] [130]. A big challenge here is stimulation artifacts produced by neural stimulation. In order to pick up weak signals attenuated by extracellular volume, a high gain amplifier is used in the recording channel. Moreover, stimulation with high magnitude of current or voltage is used in neural stimulation to overcome path attenuation so that a response can be triggered. These two factors naturally make the neural recording circuits very vulnerable to stimulation artifact. The artifact produced by the stimulation would cause saturation of the sensitive recording channel in the surrounding area during and after the stimulation for a long time period of time (up to hundreds of milliseconds), blocking the recording circuits from monitoring neural activity.

Effort has been made to eliminate the interference on the recording circuits during stimulation. The simplest way is just physically separating the recording site and stimulation site. This method, however, seriously limits the functionality of the neural recording and stimulation circuits, avoiding the problem of signal acquisition at the stimulation site. Effort is also made in the circuit design to cope with the problem. One method is to restore voltage on the electrode back to its pre-stimulation level after stimulation either by storing pre-stimulation in a sample-and-hold capacitor [131] or discharging the electrode by an auxiliary amplifier in the feedback loop [130]. An alternative method commonly used is to blank the recording amplifier at or around stimulation sites during stimulation and to keep the amplifier disabled for a few milliseconds right after stimulation [132]. Other methods involve post-processing in software to remove stimulation artifacts using spectral or spatial

correlation [133] [134]. The problem here is that the data very close to the stimulation is discarded in all existing method mentioned above. A big issue is that the duration of blanking needed varies across the different channels and also over time with formation of scar tissue and changes of electrode impedance. Hence, the typical method of setting one global blanking time results in loss of data across the array either due to amplifier saturation or due to it being blanked even when it is not necessary. The discarded data can be significant if response to the stimulation happens within the period right after stimulation. Therefore, the time period of the lost data should be as short as possible to provide a complete picture of evoked neural response.

In this work, an integration of neural stimulation and recording on the same chip is implemented, aiming at simultaneous neural stimulation and recording using the same electrode. Front-end circuit technique and post-processing are combined to reduce the recovery time period from stimulation artifact. The signal folding scheme proposed in the last chapter is used to prevent the neural recording amplifier from saturation, while reconstruction described in the previous chapter and a simple filtering are adopted to retrieve spike arrived in post-stimulation period. Even in this case some data is lost right after stimulation since the recovery algorithm is unable to handle very frequent reset event. However, due to the local feedback loop in every channel, each channel only loses a small amount of data independent of settings or conditions of other channels. Also, if electrode impedances change over time, the effect is again handled naturally due to the feedback without requiring external intervention. *In vitro* experiment is conducted to verify effectiveness of the proposed concept. The circuits and experimental setup are described in following sections. The measurement results of the *in vitro* testing are also presented, showing the effectiveness of the proposed method.

Figure 4.1: Die photo and circuits diagram of the simultaneous neural recording and stimulation IC: (a) die photo; (b) neural recording amplifier using signal folding; (c) circuit diagram and transistors sizes of neural stimulation circuit.

4.2 Proposed Design

The proposed design is implemented in Global Foundries 0.18- μm CMOS process. Die photo and circuit diagram of the proposed design are presented in Fig. 4.1. The neural recording amplifier design is similar to the version two design described in Chapter 3. OTA design presented in the version two design is reused here. Two comparators are adopted to compare differential output of the amplifier to two threshold voltages, v_{hi} and v_{lo} . Whenever the differential outputs exceed voltage range defined by the threshold voltages, reset pulse will be sent to switches, resetting the outputs back to v_{cm} . Therefore, the signal folding scheme provides a local

feedback that prevents the amplifier from saturation caused by stimulation artifacts.

The neural stimulation circuit adopted in this design is first proposed in [135]. As demonstrated in Fig. 4.1 (c), the neural stimulation circuit consists of three stages: a current DAC stage, a DAC range setting stage and a high voltage output stage. The current DAC stage is a set of binary weighted current mirrors controlled by 5-bit complementary digital signal, $D\langle 4:0\rangle$ and $D_n\langle 4:0\rangle$. The DAC range setting stage is essentially another 5-bit current DAC with cascode transistor to reduce mismatch of the current mirrors. $R\langle 4:0\rangle$ and $R_n\langle 4:0\rangle$ define output range of the stimulation current, and are usually fixed in one stimulation. Arbitrary output current waveform can be achieved by changing $D\langle 4:0\rangle$ and $D_n\langle 4:0\rangle$ according to a pre-defined sequence. The LSB of the DAC is 5-bit adjustable from 100 nA to 3.1 μ A, if I_{ref} is 100 nA. The current DAC stage and the DAC range setting stage are supplied by a lower voltage supply V_{DDL} , which is set to 1.8 V in this design. The output current of the DAC (I_{DAC}) is further amplified by a high voltage push-pull output stage operated under a high supply voltage V_{DDH} for larger output range. Therefore, thick gate transistors are used to prevent gate break-down under high voltage. Phase of the output current is controlled by digital signals POS_n and NEG_n .

4.3 Experiment Setup

Diagram and photo of the *in vitro* the experiment setup is illustrated in Fig 4.2. Bi-phasic current stimulation, commonly used in neural stimulating experiment is adopted here, as shown in Fig. 4.2. Current amplitude (I_{STM}) are usually kept constant during two phases, while durations of two phases (t_n and t_p) are set to be the same, in order to achieve charge balance to some extent. t_n , t_p , and t_m are all adjustable through timing control by FPGA. One of the inputs of the differential amplifier is chosen as the reference of recording and is connected to the circuits ground.

Figure 4.2: The *in vitro* testing setup of simultaneous neural stimulation and recording: (a) diagram and (b) photo.

The other input of amplifier is connected to the output of stimulation circuits. A DC reference is also required to provide a DC level for stimulation. For this proof-of-concept experiment, all three nodes described above are connected to phosphate-buffered saline (PBS) through thin metal wires that are commonly used in electronic experiment. In the future, *in vitro* testing with more realistic electrode model and

Figure 4.3: Measured waveforms of the simultaneous neural stimulation and recording *in vitro* testing, measured with $I_{STM} = 100 \mu\text{A}$, stimulation duration 0.5 ms and sampling rate 20 ksa/s.

even *in vivo* testing will be conducted to further verify the proposed method in simultaneous neural stimulation and recording. An extra connection is made from function generator to the PBS, providing signal source that mimics a neuron in tissue. The amplifier output is digitized by an ADC in an Agilent measurement module (U2352A) and transferred to a PC in which the digital output is post-processed by reconstruction and filtering. On-chip stimulation circuits is controlled by a FPGA, which in turn communicates with PC.

4.4 Measurement Results

I_{STM} is first set to $100 \mu\text{A}$. The duration of the stimulation is set to be around 0.5 ms, with t_n , t_p and t_m being 0.2 ms, 0.2 ms and 0.1 ms, respectively. A spike-like waveform is generated by the function generator with pulse width of around 1 ms to mimic the firing of biological neurons. The output of the signal folding amplifier is digitized with a sampling rate of 20 ksa/s. Measurement results with this setting are presented in Fig. 4.3. The red waveform on the top shows the stimulation waveform

Figure 4.4: Aligned waveforms of the simultaneous neural stimulation and recording *in vitro* testing, measured with $I_{STM} = 100 \mu\text{A}$, stimulation duration 0.5 ms and sampling rate 20 ksa/s.

Figure 4.5: Comparison of the conventional amplifier and signal folding amplifier in the *in vitro* testing, measured with $I_{STM} = 100 \mu\text{A}$, stimulation duration 0.5 ms and sampling rate 20 ksa/s.

presented at the output of stimulation circuit as well as the input of the recording amplifier. Both conventional amplifier and signal folding amplifier are affected by the artifact. Saturation happens in the conventional amplifier, distorting the spike

waveform, losing the ability of recording if a spike arrives at the recovery period. The signal folding amplifier does not saturate due to the reset scheme, but produces many resets during and right after the stimulation. The frequent resets caused by stimulation artifact, produce a lot of frequency component, beyond 10 kHz (Nyquist frequency in this case), resulting in information lost in stimulation period under 20 ksa/s sampling rate. Since reconstruction based on this under-sampled waveform is unable to produce meaningful results, the data within 1.8 ms after start of stimulation is discarded. The data is replaced by simply connecting the remaining data using a straight line. The reconstructed waveform of the signal folding amplifier is shown at the bottom of Fig. 4.3.

In order to facilitate the comparison between conventional and signal folding amplifier, waveforms proximate to stimulations are aligned and overlapped in Fig. 4.4 to clearly show what would happen to the spike that arrives right after the stimulation. In Fig. 4.4, the beginning of stimulation is aligned to 0 s. Spikes generated by function generator have exactly the same shape: same magnitude and same pulse width. Therefore, comparing spikes near stimulation with spikes away from the stimulation, we get an estimate of how the artifact is affecting the spikes. From the output of the conventional amplifier shown in the middle part of Fig. 4.4, the spike that arrives at within 4 ms after the beginning of the stimulation experiences distortion. In the recovered output of the signal folding amplifier shown at the bottom of Fig. 4.4, spike arriving at 2 ms can be recorded by the amplifier with significant fidelity, but still affected by the base-line fluctuation caused by stimulation artifact. High-pass filtered waveforms are presented in Fig. 4.5 to further illustrate the point. The red vertical dash line denotes the beginning of the stimulation. The original data between red dash line and red solid line is discarded. It can be clearly seen that a spike arriving right after the data discarding period can be captured with little distortion.

Since a sampling of 20 ksa/s commonly used in the neural recording is insuffi-

Figure 4.6: Comparison of the conventional amplifier and signal folding amplifier with increased sampling rate, measured with $I_{STM} = 100 \mu A$, stimulation duration 0.5 ms and sampling rate 200 ksa/s.

cient for capturing adequate information proximate to the stimulation, the sampling rate can be temporarily increased proximate to the stimulation to involve more data in the reconstruction so that the performance of simultaneous recording can be improved. Fig. 4.6 shows the comparison using 200 ksa/s sampling rate under the same setting. Only data in the stimulation period (0.5 ms) is discarded in this case. As seen from Fig. 4.6, by increasing sampling rate, the boundary of valid recording can be extended very close to the stimulation. This leads to a possible future design in which multiple sampling rate ADC is implemented and controlled in different scenarios, providing sufficient information for simultaneous recording with stimulation.

Finally, t_n and t_p are increased to 0.8 ms to check performance of the proposed design under different stimulation duration. In this case, I_{STM} is $50 \mu A$ and sampling rate of ADC is 200 ksa/s. It can be observed from Fig. 4.7 that more charge imbalance prolong the recovery time, for both amplifiers. From the comparison of the conventional output and reconstructed output of signal folding amplifier in Fig. 4.8, the proposed amplifier is less affected. The proposed amplifier regain recording

Figure 4.7: Aligned waveforms of the simultaneous neural stimulation and recording *in vitro* testing with increased sampling rate, measured with $I_{STM} = 50 \mu A$, stimulation duration 1.7 ms and sampling rate 200 ksa/s.

Figure 4.8: Comparison of the conventional amplifier and signal folding amplifier with increased stimulation duration, measured with $I_{STM} = 50 \mu A$, stimulation duration 1.7 ms and sampling rate 200 ksa/s.

function at around 4 ms after stimulation, while in conventional one, the spikes are affected by the resettling behavior until 10 ms after stimulation.

4.5 Conclusion

In this chapter, the possibility of simultaneous neural stimulation and recording through the same electrode using signal folding scheme is investigated. An *in vitro* testing is conducted on an integrated neural stimulation and recording chip. The measurement results shows that the proposed signal folding scheme can reduce the recovery time of amplifier after stimulation happens. In order to fully verify the feasibility of simultaneous recording and stimulation, a more realistic model of electrode is needed in future *in vitro* testing. *In vivo* testing on animal will also be conducted to give a more practical proof of the concept.

Chapter 5

A Neuromorphic IC for Neural Signal Decoding

After amplifying and filtering by front-end circuits of the intra-cortical recording system, the recorded neural signal is further processed by a spike detection circuit, as mentioned in Chapter 2. Several works have been done in this part with either an analog or digital solution [51] [79] [81] [89]. The output of spike detection is essentially the time at which spike happens. It can be presented as a digital waveform with sparse positive pulses indicating spike firings. Spike sorting is also required if further processing or analysis is based on single-unit recording. It typically consists of feature extraction and clustering. A detailed review of recent works on spike sorting and its hardware implementation is given in [93].

Rather than spike detection and spike sorting, this work focuses on neural signal decoding—extracting motor intentions from spike sequences derived from spike detection and sorting. It is also possible that the output of the spike detector can be directly fed into the decoder bypassing the spike sorter [94] [95]. In this work, a neural signal decoding algorithm is proposed based on Extreme Learning Machine (ELM). A neuromorphic mixed-signal IC is implemented to realize the proposed algorithm. The neural decoding algorithm and chip is verified using recorded neural signals

from monkey cortex in an individuated finger movements experiment. The reason for using this data set is that we can benchmark our results with published ones using software analysis on the same data set [104]. The experiment, design and verification are described in detail in the following sections.

5.1 Experiment

The experiment is described in detail in a previous study [104] [136]. As illustrated in Fig. 5.1, male rhesus monkeys are trained to perform finger and wrist movement with the guidance of a visual cue. The monkey puts its right hand into a pistol-grip manipulandum. The manipulandum has different slots. Individual fingers of the monkey can be placed into each slot. The monkey is trained to follow the visual cue given by performing flexion or extension of the individual finger and wrist. A flexing and an extending switch would be closed respectively by the flexing and extending movement of each finger and wrist. If correct movement is performed by the monkey in a trial, reward is given to enhance the training.

The different movements of fingers and wrist are labeled by using a number to denote which finger is instructed to move and by first letter of which direction is required. For instance, e1 denotes extending of thumb while f4 and ew stands for flexing of ring finger and extending of wrist respectively. Totally 12 types of individuated movements are instructed and trained during the experiment, including e1, e2..., e5, ew, f1, f2..., f5, fw. The data set from Monkey K also includes 6 combined finger movements involving moving 2 fingers simultaneously: f1+f2, f2+f3, f4+f5, e1+e2, e2+e3 and e4+e5.

A trial begins when all fingers and wrist of the monkey are in neutral position where all switches are open. As depicted in Fig. 5.2, an yellow LED in each column is turned on indicating that the hand is on ready status. After a ready period lasting

Figure 5.1: Setup of the neural signal recording and decoding in monkey finger movement experiment: Monkey moves finger according to visual cues given, while spike sequences are recorded from M1 region; Motor intention decoder proposed in this work is used to identify finger movement type from the spike sequences.

for 500 to 750 ms, a visual cue is given to the monkey to flex or extend corresponding finger or wrist by lighting one red LED in one of the columns. If the monkey closes the correct switch indicated by the cue in a period of 700 ms from lighting of the red LED and holds it for 500 ms, the trial is regarded as successful and the monkey is rewarded. The successful trial leads to start of a new trial with a different cue. If a wrong switch is closed or the monkey is unable to respond within 700 ms of the cue given, the trial is failed and the same cue will be given in the next trial. The timing of a successful trial is given in Fig. 5.2.

A single-unit recording device is implanted into the brain of the monkey by surgery, enabling real-time monitoring of single neuron activity in M1 region during the experiment. Three monkeys, labeled C, G and K, are engaged in the experiment. Multiple neurons in a constrained volume are accessed in different trials by

Figure 5.2: Timing diagram of a successful trial: Neural data is saved as a spike sequence with a label of actual movement type, which is used to verify movement type predicted by the decoder.

the recording device in a way that mimics simultaneous multi-channel recording of an MEA [43]. Totally 312 units in M1 of monkey C, 125 units in monkey G and 115 units in monkey K are obtained in the recording. After amplifying and spike detection, the recorded neural signal is presented as a spike sequence for further data analysis and used as input samples for motor decoding. The spike sequence of a trial is saved only when this trial is successful.

Each trial contains spike sequence in a duration of approximately 2 s and a label indicating which finger movement is conducted when the trial is recorded. Therefore, it is possible to implement a motor intention decoding algorithm with supervised learning to classify the movement type, given a particular trial. As illustrated in Fig. 5.1, A motor intention decoder is proposed and implemented in this work to predict the movement type given a testing trial. Comparing the predicted movement type with actual movement type observed in the experiment, classification performance of the decoder can be estimated. In a word, task of the motor intention decoder is to identify which finger movement the monkey performs according to the spike sequence of a given trial, and is essentially a multi-class classification problem.

Data analysis in the previous study [104] reveals the complexity of the input

space of the neural signal decoding. Some neurons respond universally to all types of movements, while some other neurons only respond to a particular movement type. Furthermore, an analysis on time evolution of spike activity of neurons shows changes in the intensity of the activity in different time period. An increase in the spike activity can be observed in period right preceding the closure of the switch. The non-trivial input space calls for an appropriate algorithm for the neural signal decoding. Many works are done before to map complex input space into movement intention, including inferential decoders and classifiers [99] [100] [101] and classifiers [13] [103] [105] [106]. In this work, an ANN classifier that facilitate hardware implementation is used for neural signal decoding with detailed description in following sections.

5.2 Extreme Learning Machine

In this work, the neural signal decoding algorithm is based on the ELM, proposed first in [137]. The ELM is essentially a single hidden-layer feed-forward neural network, as shown the Fig. 5.3. The output of the neural network can be expressed as follows,

$$o = \sum_i^L \beta_i g(\mathbf{w}_i^T \mathbf{x} + b_i), \mathbf{w}_i, \mathbf{x} \in R^d, \beta_i, b_i \in R, \quad (5.1)$$

where $\mathbf{x}$ denotes the input feature vector, b_i is the bias for each hidden-layer neuron, $\mathbf{w}_i$ and β_i are input and output weights respectively. $g()$ is the activation function of the hidden-layer neuron. A non-linear activation function is needed for non-linear classification. Multiple output nodes with same formation can be implemented to accomodate a multi-class classification task.

The salient feature of the ELM is that it uses random input weight and bias for hidden-layer projection. A classification boundary is drawn in the new feature space

Figure 5.3: Architecture of the single hidden-layer feedforward neural network used in the Extreme Learning Machine. It consists of input processing for feature extraction, current mirror array for random project and silicon neurons driving output counters for hidden layer operation.

determined by the random hidden-layer nodes during the training phase. Different from conventional neural network using back propagation in the training, the input weight remains unchanged after their initialization. Therefore, the hidden-layer output matrix $\mathbf{H}$ is actually unchanged after initialization of the input weights, reducing the training of this single-layer feed-forward neural network into a linear optimization problem of finding a least-square solution of β for $\mathbf{H}\beta = \mathbf{T}$, essentially:

$$\hat{\beta} = \min_{\beta} \|\mathbf{H}\beta - \mathbf{T}\| \quad (5.2)$$

where β is output weights and $\mathbf{T}$ is the target of the training. The β can be found analytically and in one-time manner by

$$\beta = \mathbf{H}^{\dagger} \mathbf{T} \quad (5.3)$$

where, $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of the matrix $\mathbf{H}$. The simple training algorithm brings advantages such as fast training speed and better general-

ization. It achieves significantly faster learning speed and a better generalization than the gradient-based learning in most cases [138].

The output of the ELM can be trained to approximate a continuous target function, resulting in a regression task. Only one output is needed in this case. In case of multi-class classification task, a commonly used form for the ELM is making number of output nodes equal to number of classes. For instance, in a m -class classification task, output of the ELM is a m -dimension vector. If label for a input sample is p , meaning this input sample belongs to p th class, only p th element of the expected output vector is set to one, while the rest of the elements are zero. In testing phase, testing sample is categorized to the class which has maximal response at the corresponding output node, among all output nodes.

Specific advantages of the ELM for this application are:

(1) the fixed random input weights can be realized by a current mirror array conducting multiply-accumulate (MAC) operation on input feature current, exploiting fabrication mismatch of the CMOS process;

(2) one-step training that is necessary for quick weight update to address change in input statistics, a commonly faced situation in the neural recording [104].

The hidden-layer neurons can be realized in an efficient way using neuromorphic circuits. The output weights can be implemented in digital circuits to facilitate tuning and better accuracy. The neuromorphic chip design for the decoding is described in the next subsection to elaborate these points.

Figure 5.4: Architecture of the proposed neuromorphic IC for the neural signal decoding

5.3 A neuromorphic IC for neural signal decoding based on Extreme Learning Machine

5.3.1 Architecture

Architecture of the neuromorphic IC developed for the neural signal decoding are presented in Fig. 5.4. The input and hidden layers of the ELM are mapped into the neural decoding IC fabricated by AMS-0.35 μ m CMOS process, where high computation efficiency is achieved by exploiting fabrication mismatch and massively parallel analog computing, while output layer is currently implemented in digital domain on a PC to achieve high resolution output weights. The neural decoding IC consists mainly of four parts: input processing circuits, current mirror array for input weights, integrate-and-firing silicon neurons and output counters. Up to 128 input channels and 128 hidden layer nodes are supported by the neural decoding IC, with each input channel embedding an input processing circuit that extracts input

feature from the incoming spike sequences. In this regard, this work follows the method in [104] to extract a moving average of the spike count as the input feature of interest. If needed, the feature extraction block may be changed to implement other functions in different applications.

Whenever there is a spike in the spike sequence of the recorded trial, a $100\ \mu\text{s}$ pulse via *SPK* and 7-bit channel address ($A\langle 6:0 \rangle$) to the DEMUX in the neural decoding IC for row-decoding. $100\ \mu\text{s}$ is found to be a small enough width for this case, but if needed, this pulse width can easily be reduced to $1\ \mu\text{s}$ to quickly handle collisions among incoming spikes on different channels. Each row of the neural decoding IC has a 6-bit window counter (*WinCNT*) to count the total number of input spikes in a moving window with length of $5 \times t_s$ and a moving step of t_s . The length of t_s , normally set to 20 ms for a maximum input spike rate of 640 Hz, is determined by the period of *CLK_{in}*, which serves as a synchronization signal of the input processing circuits driving all internal registers of these circuits. The counter value in j -th row is converted into input feature current I_{DACj} for the ELM, corresponding to the input x_j in Fig. 5.3.

The system simulation in MATLAB shows that increasing input feature dimension by introducing delayed spike sequence of the existing channel can improve the accuracy of movement classification in the neural signal decoding. This is to be expected from results in time series analysis where some of the past samples are used to better predict the current output. In the proposed IC, this feature, as will be described later, can be realized by exploiting various delayed counter value in the register chain. The ‘effective’ number of input channels can thus be made larger than the actual number of neural recording channels producing an increased input dimension. A 1-bit control signal ($\text{Sext}\langle j \rangle$) stored in each row determines whether the j -th row’s input to the moving window circuit is an external spike count or a delayed spike count from the previous channel. The delay length can be selected from among

5 delay steps ranging from 20 ms to 100 ms, based on $SDL\langle 2 : 0 \rangle$. This feature of time-delay based dimension increase (TDBDI) is useful to extract more information from functional electrodes when signal quality degrades in others.

Input feature current from each row is further mirrored into all hidden-layer nodes by a current mirror array, as shown in Fig. 5.4. In each row, the gates of the transistors are all connected to the diode-connected nFET that sinks the output current from DAC of the input processing circuit. And in each column, the drains of the transistors are connected to the input of the hidden-layer node at the top of the column, so that the input current of each row is mirrored into the hidden-layer node. The summation of weighted input features is automatically done due to the current-mode operation. Hence, ratios of the current mirrors are essentially the input weights. Since the size of all transistor in the matrix are the same, in theory the input weights are all the same. However, due to inherent statistic variation of device size, doping and oxide thickness and other parameters in the fabrication of the CMOS circuits, even the devices designed to be identical have random mismatch after fabrication [139]. The device random mismatch has long been a serious issue and limiting factor in the traditional analog and mixed signal design and in neuromorphic IC design as well [140]. Furthermore, in sub-threshold design, in which the transistors are biased with very small current to reduce power consumption, the exponential relation between bias current and threshold voltage exacerbates this effect, comparing to conventional above-threshold design with quadratic dependence between current and the threshold. Huge amount of effort has been dedicated in the analog and mixed signal circuits design to calibrate or compensate this effect. It, however, becomes a benefit in the ELM decoding circuits design, since the random input weights are required.

To quantify the device random variation, the threshold voltage of the transistors is usually modeled as a statistic term following Gaussian distribution to represent all random effects. So considering transistors in the current mirror matrix, the transistor

threshold voltage can be expressed in the equation below:

$$V_{th,ij} = V_{th,m} + \delta V_{th,ij}, \quad (5.4)$$

where i and j denote the position of transistor in the matrix, $V_{th,m}$ is the expectation of the distribution and $\delta V_{th,ij}$ is a Gaussian distribution centered at zero. Thus, the input weights are determined by the exponential relation as depicted in:

$$w_{th,ij} = ne^{\frac{\delta V_{th,ij}}{U_T}}, \quad (5.5)$$

where n is the designed ratio of the current, which equals to 1 in this design. U_T is the thermal voltage. The input weight is thus a random variable with log-normal distribution and is fixed after fabrication. In order to fully exploit the mismatch, all transistors in the array are of minimal size of the CMOS process used to implement the circuits. Therefore, the random input weights are realized in a very low ‘cost’ way that requires only one transistor per weight. It is the fixed random input weights of the ELM that makes this unique design possible.

Each hidden layer node is implemented by a integrate-and-firing silicon neuron driving a 14-bit counter with a 3-bit programmable stop value f_{max} to implement a saturating nonlinearity in the activation function $g()$. The advantage of choosing this nonlinearity is that it can be digitally set and also some neurons can be configured to be linear as well to achieve good performance in linearly separable problems [141]. The computation of hidden layer nodes is activated by setting NEU high. The output of the neuron is a pulse frequency modulated signal with the frequency proportional to total input current. Therefore, the 14-bit output counter counts number of the pulses generated by the neuron when NEU is high. The counter outputs are latched and serially read using the CLK_{out} signal when NEU is low with the neuron dis-

Figure 5.5: Timing diagram of neural decoding IC

abled to save power.

Timing diagram of the neural decoding IC is shown in Fig. 5.5. All registers and counters in the input processing circuits are reset to be 0, when *RN_in* is low. Setting *RN_in* to be 1 initiates a classification operation. *SPK* contains spikes from all channels and are distributed to different input channels by *A < 6 : 0 >*. The input processing circuits are driven and synchronized by input clock, *CLK_in*. After a short settling time for current DAC, *NEU* is pulled high enabling silicon neurons and output counters. The width of positive pulse of *NEU* therefore determines operation time of the hidden-layer neuron in each cycle and is adjustable according to different classification task. After pulling *NEU* down, the output counter value is latched. *CLK_out* then drives the output scanner, putting output counter values on 13-bit output bus one by one at each rising edge for read-out. After read-out, *RN_cnt* reset all counter value to zero, to get ready for operation in next cycle. The pulling down of *RN_in* is the closure of a complete classification operation.

Currently, the output weights and output layer are not implemented on-chip. The output data of the hidden-layer nodes are transferred to a PC, where the computation of ELM output is done in MATLAB. It can be implemented easily by a digital cir-

cuit with output weights stored in on-chip memory, facilitating tuning of the output weights. The operation of the proposed neural signal decoding is characterized by two operation phases. In the training phase, training samples are fed to the neural decoding IC while the outputs of the hidden-layer neurons are captured accordingly. The outputs of all samples form the matrix H as mentioned before. The target is derived from correct movement class of each sample. The output weights are calculated in MATLAB by multiplying Moore-Penrose generalized inverse of the matrix H with the target, as mentioned earlier. In the decoding phase, the operation of the proposed neuromorphic chip is actually the same. After the counter data read-out, the results of decoding can be calculated with the output weights derived in the training phase.

The neural decoding IC utilizes multiple VDDs in its design with DVDD supplying the input processing circuits, the current mirror matrix, silicon neuron and output counter and AVDD supplying the current DACs and current reference for the DACs. Since DVDD supplies mainly digital or digital-like circuit blocks, it can be at a lower level than AVDD from which analog circuits blocks get their supply. Furthermore, power consumption of digital circuits depends on its VDD in a quadratic way, if power consumption from switching activity is dominant. It is therefore high desirable to scale down DVDD.

5.3.2 Sub-block circuits

Fig. 5.6 shows two adjacent input processing circuits with $WinCNT_m$ configured to receive an external spike train by setting $Sext\langle m \rangle = 0$ and $WinCNT_{m+1}$ configured as time delay based channel by setting $Sext\langle m+1 \rangle = 1$. The corresponding signal flows are also depicted in the figure by red dash lines. The function of the input processing circuit, is to count the number of spikes in a moving window and convert the counter value by a DAC to a current that can be processed by the decoder. The moving window counter is realized by (1) counting spike in a sub-window in a length

Figure 5.6: Schematic of the input processing circuit. The red dash line denotes the signal flow line in which one channel (channel m) uses input spike sequence from spike detection and sorting block and the other channel (channel $m+1$) uses delayed spike count from the previous channel.

of t_s ; (2) storing sub-window counter value in a delay chain made of shift registers; and (3) adding and subtracting previous 6-b output value with corresponding sub-window counter values in the delay chain to get new 6-b output value of $WinCNT$. This calculation can be represented as:

$$Q_n \langle 5:0 \rangle = Q_{n-1} \langle 5:0 \rangle + D_n \langle 3:0 \rangle - D_{n-5} \langle 3:0 \rangle. \quad (5.6)$$

where $Q_n \langle 5:0 \rangle$ and $D_n \langle 3:0 \rangle$ are 6-b output value and 4-b sub-window counter value at time instance n respectively. All registers in the input processing circuits toggle at the rising edge of CLK_{in} . The advantage of this structure is that the delay chain for sub-window counter value is reused in the proposed TDBDI feature, lead-

Figure 5.7: Schematic of the current DAC used in the decoding IC

ing to a compact design. As depicted in Fig. 5.6, $\text{Sext}\langle m+1 \rangle = 1$ make WinCNT_{m+1} receive the delayed sample from WinCNT_m , rather than count spikes in SPK_{m+1} . The delay length is determined by $\text{SDL}\langle 2:0 \rangle$ through MUX5to1.

Schematic of the current DAC used in the input processing circuit is shown in Fig. 5.7. To achieve a low power consumption of the decoding chip, as well as maximal mismatch in the current mirrors, the reference current of the DAC should be smaller than 100 nA. Furthermore, a compact design is required here due to the requirement of multiple channels. Therefore, the current division technique using MOS transistors proposed in [142] is used in this work. The transistors of the same size forms a R-2R like network to distribute reference current in a binary way. Comparing to binary weighted array design in which layout area increases exponentially with number-of-bits (NOB), this current division method offers a more compact solution with array size increasing linearly with NOB. A 6-bit DAC is used in this work to split the reference current I_{ref} (also 6-bit programmable in range of 1 nA to 63 nA) according to the WinCNT output value to generate the input feature current I_{DAC} to the current mirrors.

The silicon neuron, firstly proposed by Mead [143], is essentially a current-controlled oscillator, with output spike firing rate roughly proportional to input current. Diagram of the silicon neuron is depicted in Fig. 5.8 (a), while waveforms which describes dynamics of the silicon is demonstrated in Fig. 5.8 (b). The capacitance of $C_{int} = 400$ fF sets oscillation frequency based on summed input current,

while $C_f = 100$ fF provides hysteresis through positive feedback. When NEU is pulled high, pFET M_2 is turned off. If M_1 is off, I_{in} from the current mirrors starts to discharge v_{mem} until it crosses the threshold voltage of the INV_1 , leading to transition of all invertors. Then, v_{mem} is pulled down very quickly through a positive feedback loop formed by C_f . At the same time, M_3 turns on, charging v_{mem} towards $DVDD$ until it crosses the threshold voltage of INV_1 from low to high. INV_2 toggles again, pushing up v_{mem} and turning off M_3 , starting a new cycle of discharging and charging. The INV_3 serves as a buffer for driving the output counter. Neglecting higher order effects, the time for each cycle of the neuron operation is determined by the sum of the charging and discharging time constant of v_{mem} , and can be expressed by:

$$T_{CCO} = \frac{k_1}{I_{in}C_{int}} + \frac{k_2}{(I_{rst} - I_{in})C_{int}}, \quad (5.7)$$

where k_1 and k_2 are constant coefficients determined by the electrical characteristic of the discharging and charging path respectively, and I_{rst} is the charging current when M_3 is on. Normally I_{rst} is much larger than I_{in} , making the first term in the equation dominant. Hence, the frequency of the CCO output can be written as:

$$f_{CCO} = \frac{1}{T_{CCO}} \approx \frac{1}{k_1} I_{in} C_{int}. \quad (5.8)$$

The output counter counts the number of firing of the neuron in a certain time window. The counter has a maximum 14-bit resolution with stop value $(f_{max}) 2^m - 1$ with integer m programmable from 7 to 14. When the counter reaches its stop value, it will stop counting and hold the value until it is read and reset. This saturating feature gives non-linearity to the neuron transfer function and reduces power consumption when only a low resolution is required. Another way of adding non-linearity is

Figure 5.8: Schematic and waveforms of the silicon neuron used in the decoding IC: (a) The schematic of the silicon neuron, (b) The waveform showing operation of the silicon neuron.

through driving the gate of the pFET M_1 to impose a leakage current at the input, making minimal firing current of the neuron larger than 0.

5.4 System Simulation and Parameters Setting

In order to evaluate decoding performance of the algorithm before chip fabrication, system level simulation of the decoding IC is required. Since SPICE simulation becomes rather slow for the system level simulation and is unable to have a good modeling of random mismatch of the current mirror matrix, a behavioral model for the decoding IC is built in MATLAB for emulation and evaluation.

In the simulation and later silicon verification as well, period of CLK_{in} is set to be 20 ms so that a moving counting window with window length of 100 ms and moving step of 20 ms is adopted. Consistently, period of NEU is also 20 ms, leading to a classification rate of 50 classifications per second. For movements classification, neural activity right preceding the closure of switch is stronger than other time instance, implying that it has strong correlation with on-going movement of arm.

Therefore, only spike count within 100 ms before the closure of switch is chosen as input feature for movement type decoding. The hidden-layer neurons is modeled by an current-to-counter-value transfer function, with the output values being integers with a maximum saturating value. This transfer function can be written as follows:

$$C = \begin{cases} \mathbf{floor}(k_{i2f}t_{cnt}I_{in}), & \text{if } k_{i2f}t_{cnt}I_{in} < f_{max} \\ f_{max}, & \text{if } k_{i2f}t_{cnt}I_{in} \geq f_{max} \end{cases}, \quad (5.9)$$

where k_{i2f} is the linear coefficient of the current-to-frequency transfer function of the silicon neuron, t_{cnt} is the output counting window and $\mathbf{floor}()$ takes integer part of its input value. Here, k_{i2f} is derived from SPICE simulations of the silicon neuron. In the simulation, only random mismatch of the current mirror array is addressed by distributing current mirror ratio according to Eq. 5.5. Mismatch from other circuit blocks is neglected for simplicity.

5.4.1 Parameters Setting

System level simulation of the decoding IC in MATLAB can also be used to assist the setting of the parameters. The parameters that need to be set for proper operation of the neural signal decoding IC include reference current of DAC (I_{ref}), NOB of the output counter and output counting time (t_{cnt}). NOB of the output counter defines dynamic range of hidden-layer output, thus affecting the decoding accuracy directly. Meanwhile, f_{max} in the proposed design is defined by $2^{\text{NOB}}-1$. I_{ref} determines how the input features are mapped into the input current of the hidden-layer neurons. According to Eq. 5.9, these three parameters defines how much non-linearity can be derived from the hidden-layer neurons, thus determining the classifying capacity of the hidden-layer neurons when facing a non-linear classification problem.

The performance of classification, on the other hand, also depends on charac-

Figure 5.9: Flow chart of parameters setting using system simulation in MATLAB.

teristic of data set. Some data set could be intrinsically easier to be separated than others. Taking NOB of the output counter as an example, a low resolution may be enough for a particular data set. Although higher resolution always gives no worse result than those of lower resolution, it may increase circuit power consumption for no obvious improvement in performance. Therefore, in this work, system simulation with the recorded data for the experiment is heavily involved in parameter setting.

Procedure of finding a set of suitable parameters for the neural decoding IC using system simulation is illustrated in Fig. 5.9. The NOB is first empirically set to be 10 in order to guarantee enough dynamic range in the hidden-layer neuron output. f_{max} is therefore 1023. Maximum possible value of I_{in} , denoted as I_{max} also needs to be large enough to generate a 10-bit dynamic range at output counter. Since the proposed decoding IC is targeted for low power BMI, I_{ref} is first set to be a very low level at 1 nA. In the algorithm, maximum number of input dimension (D) is 120, rendering a I_{max} of 120 nA. I_{max} is fixed so that roughly same range of mapping from current to counter value is obtained regardless to number of input dimension. For instance, if D is scaled down to 40, I_{ref} is scaled up accordingly to 3 nA. From SPICE simulation, k_{2f} can be obtained, which is round 1000 Hz/nA. Hence, from

Eq. 5.9, a counting window of 8.5 ms is required to map 120 nA to 1023 at counter output. Since non-linearity may be required in the transfer function, the window length is round up to 10 ms.

After primary setting is done, a set of simulations is run with I_{max} adjusted locally around 120 nA to optimize the classification accuracy. If satisfactory accuracy is still not achieved. NOB then can be increased and the entire setting procedure starts again.

5.4.2 System simulation results

The number of recording channels used in neural decoding is an important factor that directly affects decoding accuracy. Generally, increasing number of recording channels will increase the information volume available for the decoding algorithm but consumes more resource in the decoding circuits and in the recording system as well. Hence, it is desirable to improve the decoding accuracy when the number of recording channels is small. As discussed in Sub-section 5.3.1, delayed information of the recorded data is used to increase the input feature dimension with an aim of improving the decoding accuracy when the number of recording channels used is small.

Fig. 5.10 show the system level simulation of the decoding IC in MATLAB with parameter setting obtained by the paradigm discussed before. Number of hidden-layer neurons used in the ELM is 60. 12 types of movements are selected (flexing and extending of all fingers and wrist) with 10 trials in each type. Half of the trials is used as training set. The other half is used as testing set. Both lines show that decoding accuracy increases with increasing number of recorded neurons. The blue does not use delayed samples to increase the input dimension while the red line adds two sequential delayed samples with a delay of 40 ms between them. As clearly shown in Fig. 5.10, increasing input dimension does not improve decoding accuracy for

Figure 5.10: Decoding accuracy of proposed decoding IC obtained by system simulation in MATLAB. Results of decoding with or without TDBDI are compared.

number of recorded neurons larger than 30 since large enough information is already embedded in 30 neurons, but improves the performance with number of recorded neurons smaller than 30.

This is especially promising for long term implants where some of the initially available recording channels may be unusable later due to electrode degeneration (due to formation of scar tissue etc).

5.5 Verification of neural signal decoding IC

The proposed neural signal decoding IC is implemented using AMS-0.35 μ m CMOS process, with a die area of 4.95×4.95 mm², including pad frame. The die photo of the chip is shown in Fig. 5.11. In the following subsections, the verification of separate circuits blocks are presented first, followed by neural signal decoding using cortical data of monkey. Finally, the measured power consumption is also presented.

Figure 5.11: Die photo of the proposed neural signal decoding IC.

5.5.1 Circuit Blocks Verification

Fig. 5.12 verifies operation of the input processing by showing output of the window counter. The yellow waveform is the waveform of CLK_{in} with a cycle period of 20 ms. The white dash line denotes start of the operation. In Fig. 5.12 (a) where $S_{ext} = 0$, the input processing circuits in the tested channel processes an input spike sequence with a spike rate of 630 Hz, starting to 6-bit spike count after a delay of 40 ms and increasing in every 20 ms until 140 ms. In Fig. 5.12 (b), on the other hand, the input processing circuit receives delayed spike counting value from the channel in (a). Here, $SDL < 2 : 0 >$ is set to be 001, giving an additional delay of 40 ms in the waveform compared to that in (a).

Fig. 5.13 shows operation of the silicon neuron. The waveform on the top is the input spike sequence. The membrane voltage v_{mem} and output $v_{o,neu}$ are presented in

Figure 5.12: The output waveform of the window counter processing an input spike sequence: (a) $S_{ext} = 0$, (b) $S_{ext} = 1$.

Figure 5.13: The operation of the silicon neuron: Waveform on top is spike signal on SPK ; Waveform of v_{mem} is presented on middle part of the figure; Bottom part shows output spike of the silicon neuron.

Figure 5.14: The transfer curves of the silicon neurons.

the middle and bottom part respectively. The rising and falling ramp in the waveform of v_{mem} is the result of charging and discharging of capacitor. In every cycle of charging and discharging, a spike is fired at the output of neuron.

Fig. 5.14 shows transfer curves of the silicon neurons in terms of input spike frequency to output spike frequency. Only one input channel is selected with a input spike rate changing from 10 Hz to 630 Hz. A reference current of 32 nA is given to the current DAC. As presented in Fig. 5.14, the transfer curves of different neurons shows a linear relation between input and output frequency, with different slopes due to mismatch in the current mirror array as well as silicon neurons.

Furthermore, 128 input channels are swept with a spike sequence with a firing rate of 320 Hz. The reference current for each channel is set to be 32 nA. The output frequency of each neuron with respect to input spike sequence at each input channel is measured, resulting in a 2-dimension mismatch graph in Fig. 5.15. Normalizing this mismatch matrix by the average value, the distribution of the mismatch is shown in the histogram of Fig. 5.16. The shape of the histogram implies a log-normal distribution as suggested by Equ. 5.5.

Figure 5.15: 2-dimension mismatch graph.

Figure 5.16: Histogram of mismatch in the input weights.

5.5.2 Verification of Neural Signal Decoding

The proposed decoding IC is further verified by the recorded neural signal from monkey cortex in the experiment. As mentioned before, neural signal from three monkeys, labeled as Monkey C, K and G, is collected with trials belongs to 12 individuated movements and 4 combined movements. In the verification, equal number

Figure 5.17: Change of the decoding accuracy of the proposed IC with number of hidden-layer neurons used

of trials are selected from each movement type. Entire data set used in the verification is distributed equally into training set and testing set. The results shown in the decoding verification is derived using neural data from Monkey K if not pointed out.

An important reasoning behind the random projection of the ELM is that a linear separation can be done in the random feature space into which input features are projected as long as dimension of the random feature space is high enough. Therefore, number of hidden-layer neurons (L) is an important parameter in the ELM. Typically the classification accuracy will increase with increasing L and saturate after enough number of hidden-layer neurons are used. Fig. 5.17 presents this trend of measured testing accuracy of classification. At each point, the testing accuracy is derived by averaging over 50 iterations in which corresponding number of output counters are randomly selected. The small bar on each point indicates standard deviation of the iterations. It can be concluded from this curve that 60 hidden-layer neurons provide enough accuracy for the classification problem while further increasing beyond this figure gives no obvious improvement.

Fig. 5.18 shows input and output of the proposed neural signal decoding. Bottom

Figure 5.18: Input and output of the proposed neural signal decoding

As discussed before, one of challenges for neural signal decoding is to use less number of recording channels to get enough accuracy in the decoding, equivalently to improve accuracy of the decoding with small number of recording channels. It is proposed in this work to use delayed information of the recorded spike sequence for increasing the input dimension so that the decoding accuracy with small number

Figure 5.19: Decoding accuracy of the neural signal data with and without the delayed samples

of recording channels can be improved. This method has been verified with the system simulation in MATLAB. Similar curves are drawn from measured decoding accuracy in Fig. 5.19. In the decoding using delayed information of spike sequences, two sequential delay channels are added next to the input channel that receives spike sequence. By setting $SDL < 2 : 0 >$ to be 001, the delay between two adjacent channels is 40 ms. It can be concluded from Fig. 5.19 that the decoding accuracy improves by introducing delayed information. An improvement larger than 5% in decoding accuracy can be obtained when number of recorded neurons are equal to or less than 20. The decoding accuracy with delayed samples using 40 recording channels is 99.2% while decoding using same number of recording channels without using delayed samples gives an accuracy of 97.6%.

Finally, the decoding IC is verified by more data from Monkey C, Monkey G and data set of Monkey K that includes combined finger movement type. The decoding accuracy derived using 60 hidden-layer neurons and delayed information for each data set is presented in Fig. 5.20.

Figure 5.20: Decoding accuracy of neural signal data from Monkey C, G and K: All decoding is conducted by the chip using 60 hidden-layer neurons and 2 sequential 40-ms delayed channel for each recorded neuron.

5.5.3 Power Estimation

There are two power domain in the proposed decoding IC. DVDD mainly supplies digital and digital-resembled circuits including window counters, neurons, mirror array and output counters. AVDD supplies analog circuits including internal reference circuits and current DACs. The power of digital circuits can be significantly reduced by scaling down DVDD due to quadratic relation between power consumption and DVDD, while certain level of AVDD may still be required due to DC biasing condition of the analog circuits. Separate power domain design is used to facilitate different power scaling down strategy. Hence, on-chip level shifter circuit is needed to accommodate signal level transition between two power domains.

Average current consumption of digital and digital-assembled circuits during decoding operation can be modeled by the following equation:

$$\bar{I}_{dvdd} = \alpha_0 + \alpha_1 D \bar{f}_{in} + \alpha_2 D \bar{f}_{in} L B \frac{t_{cnt}}{T_{in}}. \quad (5.10)$$

The first term in the equation is a static current that does not change with switching

activity of the digital circuits. One possible source of this term is leakage current in the circuits. The second term is average current consumption of switching activity of the window counter. It is therefore linearly related to the number of input processing circuits used (D) in the decoding and average spike frequency of the input data set ($\bar{f}_{in}$) with a coefficient α_1 . The third term is average current consumption of neurons, current mirror array and output counter. When neurons are enabled (NEU is high), all these circuit blocks have current consumption proportional to average input current of the neurons which is in turn determined linearly by $\bar{f}_{in}D$ and B which is I_{ref} normalized to 1 nA. When neurons are off (NEU is low), all these circuit blocks are off and consume only leakage current. The ratio t_{cnt}/T_{in} is multiplied here to account for the time t_{cnt} for which these circuits are enabled compared to the entire classification period of T_{in} . Finally, unused output counter, neurons and current mirror that connected to the unused neurons can be turned off so that only chosen L hidden-layer neurons consumes power.

Three coefficients α_0 , α and α_2 are determined by doing linear regression using a set of measured current consumption with a $DVDD = 0.6$ V, $AVDD = 2.1$ V and with varying D and L . The measured current consumption and corresponding setting used is listed in Tab. 5.1.

Table 5.1: Measured Current consumption of DVDD with varying number of input channel and hidden-layer neurons

$\bar{I}_{dvdd}$ (nA)	D				
	1	21	41	61	81
1	20	52	90	121	156
11	33	410	770	1150	1527
L 21	45	669	1297	1944	2549
31	65	969	1892	2795	3729
41	71	1147	2263	3421	4497

Measured under condition: $DVDD=0.6$ V,
 $AVDD=2.1$ V, $B=8$, $\bar{f}_{in}=80$ Hz, $t_{cnt}=10$ ms,
 $T_{in}=20$ ms.

Linear regression on these data results in $\alpha_0 = 14.4$ nA, $\alpha_1 = 18.4$ pA/Hz and $\alpha_2 = 4.3$ pA/Hz, under the condition of DVDD = 0.6 V and AVDD = 2.1 V.

Current consumption of AVDD (I_{avdd}) is static biasing current in the reference circuit and current DAC, and stays constant during the decoding operation. The measured current consumption of AVDD is $6.67 \mu\text{A}$ when AVDD = 2.1 V.

Given period of the classification $T_{in} = 20$ ms, energy consumed per classification can be derived by:

$$E_{per,clas} = (\bar{I}_{dvdd} \cdot DVDD + I_{avdd} \cdot AVDD) T_{in}. \quad (5.11)$$

A typical setting for the classification with the monkey data set is as follows: $D = 40$, $L = 60$, $\bar{f}_{in} = 100$ Hz, $B = 4$, $t_{cnt} = 10$ ms and $T_{in} = 20$ ms. Under this setting, I_{dvdd} is calculated as $2.2 \mu\text{A}$, leading to an energy per classification of 307 nJ.

As can be seen energy consumption of the analog circuits takes larger part in the total energy consumption. It can be reduced in future by shutting down the current DACs in input channels when neurons are disabled and by scaling down AVDD as well.

Table 5.2: Comparison among different works on neural decoding IC

Work	Method	Hardware Realization	Task	Input Dim.	Output Rate(ops)	E per Op.(μJ)	Accuracy
This work	ELM classifier	0.35- μm CMOS	Movement types	40	50	0.307	0.976
[1]	Receptive field	FPGA	Place cells	32	0.69	773.3	0.94
[83]	Adaptive filter	0.18- μm (SPICE sim.)	Head direction	100	10	2.5	N.A.
[113]	SNN-KF	N.A.	2-D cursor trajectory	96	20	N.A.	N.A.

Here, AVDD is reduced to 1.2 V, a minimal level under which the decoding accuracy is still high enough for the application. In this decoding operation, the

dimension and parameter setting remains the same as before, except exact value of B is unknown since the on-chip current reference may be comprised under 1.2 V. $\bar{I}_{dvd}$ and I_{avdd} are measured to be $0.09 \mu\text{A}$ and $0.30 \mu\text{A}$ respectively, rendering a energy per classification of 8.3 nJ.

5.6 Conclusion

In this chapter, the proposed neural signal decoding IC based on the ELM is described and verified by pre-recorded spike data from M1 cortex region of the monkeys in the finger movement experiment. Using 40 recording channels and 60 hidden-layer neurons, The decoding IC achieves accuracy of 97.6% in classifying 12 different types of individuated finger movement with 307 nJ consumed per classification. Many algorithms have been developed for neural signal decoding for intra-cortical BMI. However, most of these algorithm run on PC or general-purpose processor, and requires significant computing efforts. On-chip integration of the neural signal decoding is highly desirable for next generation BMI where only miniaturized and low-power computing device can be implemented. The proposed decoding IC in this work is the first effort, to the knowledge of the author, to realize ASIC implementation of neural signal decoding and verify the decoding IC with real data. A comparison of this work with other state-of-the-art works on neural decoding is given in Tab. 5.2.

The future work of the neural signal decoding IC includes achieving on-chip integration of output weights and output neurons, further reducing the power consumption by shutting down unused analog circuit block during the classification and applying the proposed decoding method for movement onset detection in the pre-recorded data from monkey.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Next generation BMI calls for chronically implantable devices with ultra low power, high integration, more functionality and robustness in chronic use. In this work, effort is mainly dedicated to address issues in power consumption, robustness and data compression in the intra-cortical neural recording and neural signal processing.

In the area of intra-cortical neural recording circuits, work has been mainly focused on the front-end amplifier. Rather than competing in terms of noise efficiency factor (NEF) as many other works in literature, the signal folding scheme is proposed as a general method that can be applied to any low power amplifier. The signal folding scheme folds amplifier output into a pre-defined range by continuous monitoring of the output and resets whenever output exceeds the range. By doing this, the signal dynamic range presented at the input of ADC is reduced, leading to a reduced NOB of the ADC. The area and power consumption of ADC used in the recording system can therefore be saved by reduced NOB. Since data rates of wireless transmission is reduced accordingly, power consumption of wireless data transmission which has been a bottleneck of the future intra-cortical recording system can be also reduced.

The signal folding can be perceived as a spectrum shaping method from the perspective of frequency domain. The neural signal has a characteristic of $1/f^n$ rolling off in frequency spectrum with larger amplitude in low frequency component and smaller amplitude in high frequency component. By folding large and slow signal, the signal folding essentially reshapes the spectrum of the neural signal, pushing energy in low frequency into higher frequency domain in a predictive manner, therefore reducing dynamic range of the signal. After data transmission, the amplified signal is retrieved from the folded signal by a non-Nyquist reconstruction algorithm. The dynamic range is regained and spectrum restored. Of course, some error is introduced in this process. A behavioral model is built in MATLAB to estimate the reconstruction in simulation using circuits parameter derived from SPICE simulation. Besides power and area saving, the signal folding scheme provides other benefits including: lowered high pass corner for LFP, reduced harmonic distortion, low supply voltage operation and resilience to interference and artifact.

It has become a trend for future intra-cortical BMI that neural stimulation and recording cooperate closely to build a closed loop from monitoring neural activity to interacting with external environment and then back to changing the neural activity. Since the resets in signal folding scheme provides a way of quick recovery from big artifact and interference, it is applied to simultaneous neural stimulation and recording, in an attempt to integrate neural stimulation and recording circuits on the same die and conduct stimulation and recording using the same electrode. *In vitro* testing in PBS is conducted, showing that signal folding scheme can reduce the recovery time of recording amplifier to within 2 ms after stimulation with a current level of 100 μA , which is commonly used in neural stimulation of central nerve system.

For the signal conditioning in the intra-cortical recording, further processing steps after amplifying signal may include analog-to-digital conversion, spike detec-

tion and spike sorting. Huge amount of work has been done in literature to realize these processing steps. In this work, data compression as a way to reduce power consumption has been a major focus to enable long term operation of the devices, in which complete transmission of raw recording data would be an extreme heavy burden. As discussed in Chapter 3, the signal folding has provided a way of data compression in neural recording. By cutting down the NOB of the ADC from 10 to 8, the recording data is compressed by a factor of 0.8.

On-chip neural signal decoding has provided another highly efficient way of data compression. Furthermore, for practical use of next generation BMI devices, power efficient and miniaturized neural decoding is highly desired rather than the current decoding algorithms running on PC. Very few works have been done in attempting to integrate neural signal decoding on-chip. To fill this void, a neural decoding algorithm based on the ELM is proposed in this work with the world's first application-specific integrated circuit (ASIC) realization of neural decoding. The ELM is a single hidden-layer feed-forward neural network with fixed random input weights and with output weights tuned during training phase. The training can be done in an analytic and one-time manner by linear optimization of output of hidden-layer neurons to the targets of training, leading to a very fast training speed and better generalization. The fixed random input weights are found to be easily implemented if random mismatch of transistors intrinsic to all CMOS process is used. The multiplying of the input weights is therefore realized in current mode by current mirrors with low power and very small area. The output weights can be realized in digital circuits to facilitating tuning. Silicon neuron and output counter serves as hidden-layer neuron, performing non-linear operation in the neural network and converts analog current mode input into digital output.

In this work, the input weights and hidden-layer neurons of the ELM is implemented in AMS-0.35 μ m CMOS process. The output of hidden-layer neurons is read

out and transferred to a PC, where training is conducted and output of ELM is calculated. The decoding IC is measured and verified using pre-recorded neural data from M1 cortex region of monkey conducting finger movement tasks. The measurement shows that using 40 recording channels and 60 hidden-layer neurons, the decoding IC achieves accuracy of 97.6% in classifying 12 different types of individuated finger movement with 307 nJ consumed per classification.

One of the challenges of neural signal decoding is to achieve reasonable accuracy using limited number of recording channels. A limited number of recording channels typically means less information is provided from the recording side to the decoding algorithm and therefore less input dimensions. In this work, the delayed information of recorded neural signal is used to increase the input dimension of the decoding algorithm so that time evolution of the neural signal can be partially presented to the decoding algorithm. The increased information does lead to higher decoding accuracy, as suggested by simulation results as well as measurement results. Since the input processing circuits proposed in this work uses a delay chain of registers to realize window counting of the spikes. It can be easily reused to provide the delayed spike count to the next input channel. In this way, the proposed method of increasing input dimension does not cause any penalty in hardware implementation.

6.2 Future Work

Future work that can be done to improve the existing circuits on neural signal recording and decoding are listed as follows.

On the neural recording side, a complete neural recording channel that uses signal folding scheme can be implemented by integrating ADC and reconstruction algorithm on chip. As discussed in Section 2.5, a capacitive array based SAR-ADC can be used to digitize the amplified signal. The capacitive array based SAR-ADC

is a common choice in various works on intra-cortical neural recording, due to its high power efficiency. The reconstruction algorithm can be first implemented on a field-programmable gates array (FPGA) for verification. Then it can be integrated on the same chip with neural signal processing circuits, such as spike detection and sorting.

More experiments can be conducted in future to further verify simultaneous neural stimulation and recording with signal folding scheme. An *in vivo* testing is meaningful in verifying the concept in a more realistic scenario. Higher integration can also be realized by adding ADC and digital control circuits on the same chip. As suggested by *in vitro* testing results presented in Chapter 4, an ADC with multiple sampling rates is desirable here to acquire more information for the reconstruction, when neural stimulation is applied.

On the neural signal decoding side, other neural signal data set involving different experiment protocol will be used to further verify the proposed neural decoding IC. For instance, spike pattern of place cells in hippocampal region of rat cortex has been used to predict future position of the rat [144].

Many works can be done in future to improve circuit performance and expand system functionality. Firstly, power consumption of the neural decoding IC can be further reduced by shutting down unused static bias current on chip. Circuits realizing different types of hidden layer activation function can also be added to improve classification performance in different scenarios. Furthermore, output weights and output layer of the ELM can be integrated on chip to realize complete function of the ELM on a single device. Alternatively, the proposed decoding IC can cooperate with a conventional digital micro-processor, serving as a co-processor which conducts massive computing work of hidden layer in a more power-efficient way. The micro-processor, on the other hand, provides flexibility in implementing the output layer in which various operations and algorithms can be implemented in firmware.

Finally, as a general classification method, the decoding IC has potential of being used in a wide range of applications, including wireless sensor node, other biomedical devices, internet-of-things and data science.

It is also very interesting to integrate the proposed neural decoding circuits with front-end circuits of neural recording to accomplish a system-on-a-chip (SoC) on which a complete signal chain from neural signal recording to neural signal decoding can be realized by a single implanted device. This makes the implanted device capable of on-site neural signal decoding so that only results of neural signal decoding is transmitted out of body, leading to a very large data compression.

APPENDIX A

Time-Domain Behavior Model of Signal Folding Amplifier

In order to measure the reconstruction error, a test-bench of an ideal amplifier and an amplifier with reset behavior is needed, so that comparison of the two outputs would provide the error. SPICE was not viable option due to the long simulation times required for systems that have slow (e.g. LFP waveform) and fast (e.g. reset of the amplifier) time scales simultaneously. Moreover, to predict reconstruction error, the simulations need to include noise in transient analysis making SPICE simulation times prohibitive. Hence, we developed a behavioral description of the ideal and resetting amplifier in MATLAB using differential equations.

First, the ideal amplifier is approximated by a system with a closed loop gain of G and a single-pole low-pass transfer function.

Then, the change in its output, $\Delta y'$ can be derived from the one-pole low-pass filter transfer function $H(s)$ as shown below:

$$\begin{aligned} H(s) &= \frac{\Delta Y(s)}{\Delta X(s)} = \frac{G}{1+s/2\pi f_{LP}} \\ \Delta y'(t) &= 2\pi f_{LP}(G\Delta x(t) - \Delta y(t)) \end{aligned} \tag{A.1}$$

where f_{LP} is the low-pass corner of the proposed amplifier respectively. This equa-

tion is integrated numerically using a fourth order Runge-Kutta method with a fixed time step of $50 \mu s$.

Next, the reset behavior of the proposed amplifier can be easily modeled by a 'if structure' in MATLAB code. Here, we ignore the finite reset time of the actual implementation since we found it did not affect the reconstruction error spectrum significantly.

Finally, to get the correct reconstruction error spectrum, we added random noise and a leakage current component at the input of the amplifier. The values of these components were chosen based on measurement results so that the reconstruction error spectrum of this model matched the measurements closely.

APPENDIX B

Author's Publication

B.1 Journal Papers

1. Yi Chen, Arindam Basu, Lei Liu, Xiaodan Zou, Rajkumar Ramamoorthy, Gavin Dawe and Minkyu Je, "A Digitally Assisted, Signal Folding Neural Recording Amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 8, pp. 528-542, August 2014.
2. Yi Chen, Enyi Yao, Arindam Basu, "A 128 channel 290 GMACs/W Machine Learning based Co-processor for Intention Decoding in Brain Machine Interfaces," submitted to *IEEE Journal of Solid State Circuits*.
3. Yi Chen, Arindam Basu, Lei Yao, Minkyu Je, "An Integrated Simultaneous Neural Stimulation and Recording System with Fast Recovery from Stimulation Artifacts", submitted to *IEEE Transactions on Circuits and Systems, II*.
4. Enyi Yao, Yi Chen, Arindam Basu, "A 0.7 V, 40 nW Compact, Current-Mode Neural Spike Detector in 65 nm CMOS", submitted to *IEEE Transactions on Biomedical Circuits and Systems*.

B.2 Conference Papers

1. Yi Chen, Enyi Yao, Arindam Basu, “A 128 channel 290 GMACs/W Machine Learning based Co-processor for Intention Decoding in Brain Machine Interfaces,” submitted to *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015.
2. Yi Chen, Arindam Basu and Minkyu Je, “A Signal Folding Neural Amplifier Exploiting Neural Signal Statistics”, *IEEE Biomedical Circuits and Systems Conference (BioCAS 2012)*, pp. 224-227, 2012.
3. Yi Chen, Arindam Basu and Minkyu Je, “A Digitally Assisted Pseudo-resistor-Less Amplifier in 65nm CMOS for Neural Recording Applications”, *the 55th International Midwest Symposium on Circuits and systems (IEEE MWSCAS 2012)*, pp. 366-369, 2012.

BIBLIOGRAPHY

- [1] B. Rapoport, L. Turicchian, W. Wattanapanitch, T. Davidson, and R. Sarpeshkar, “Efficient universal computing architectures for decoding neural activity,” *PLoS ONE*, vol. 7, p. e42492, 2012.
- [2] K. Micheva, B. Busse, N. Weileer, N. O’Rourke, and S. Simith, “Single-synapse analysis of a diverse synapse population: proteomic imaging methods and markers,” *Neuron*, vol. 68, pp. 639–653, 2010.
- [3] W. McCulloch and P. Walter, “A logical calculus of ideas immanent in nervous activity,” *Bulletin of Mathematical Biophysics*, vol. 5, p. 115133, 1943.
- [4] “Artificial neural network”, Wikipedia.
- [5] G. Moore, “Cramming more components onto integrated circuits,” *Electronics Magazine*, p. 4, 1965.
- [6] Human Brain Project Official Website.
- [7] The BRAIN Initiative, NIH website.
- [8] A. Alivisatos, M. Chun, G. Church, R. Greenspan, M. Roukes, and R. Yuste, “The brain activity map project and the challenge of functional connectomics,” *Neuron*, vol. 74, pp. 970–974, 2012.
- [9] N. Verma, A. Shoeb, J. Bohorquez, J. Dawson, J. Guttag, and A. Chandrakasan, “A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system,” *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 804–816, 2010.
- [10] A. Avestruz, W. Santa, D. Carlson, R. Jensen, S. Stanslaski, A. Helfenstine, and T. Denson, “A 5uW/channel spectral analysis IC for chronic bidirectional brain-machine interfaces,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 3006–3024, 2008.
- [11] J. Lee, H. Rhew, D. Kipke, and M. Flynn, “A 64 channel programmable closed-loop neurostimulator with 8 channel neural amplifier and logarithmic ADC,” *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 1935–1945, 2010.

- [12] M. Mollazadeh, V. Aggarwal, A. Davidson, A. Law, and N. V. Thakor, "Spatiotemporal Variation of Multiple Neurophysiological Signals in the Primary Motor Cortex during Dexterous Reach-to-Grasp Movements," *The Journal of Neuroscience*, vol. 31, pp. 15331–43, Oct 2011.
- [13] J. Wessberg, C. Stambaugh, J. Kralik, P. Beck, M. Laubach, J. Chapin, J. Kim, J. Biggs, M. Srinivasan, and M. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," *Nature*, vol. 408, pp. 361–365, 2000.
- [14] L. Hochberg, M. Serruya, G. Friehs, J. Mukand, M. Saleh, A. Caplan, A. Branner, D. Chen, R. Penn, and J. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, p. 164171, 2006.
- [15] L. Hochberg, D. Bacher, B. Jarosiewicz, N. Masse, J. Simeral, J. Vogel, S. Haddadin, J. Liu, S. Cash, P. der Smagt, and J. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, pp. 372–375, 2012.
- [16] M. Lebedev and M. Nicolelis, "Brain-machine interfaces: past, present and future," *Trends in Neurosciences*, vol. 29, no. 9, pp. 536–546, 2006.
- [17] M. Lebedev and M. Nicolelis, "Toward a whole-body neuroprosthetic," *Progress in Brain Research*, vol. 194, pp. 47–60, 2011.
- [18] A. Kbler, B. Kotchoubey, J. Kaiser, J. Wolpaw, and N. Birbaumer, *Brain-Computer Communication: Unlocking the Locked in; American Psychological Association: Washington, DC, USA, 2001*. American Psychological Association, 2001.
- [19] L. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, pp. 1211–1279, 2012.
- [20] N. Weiskopf, K. Mathiak, S. Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer, "Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI)," *IEEE Transactions on Biomedical Engineering*, vol. 51, p. 966970, 2004.
- [21] S. Coyle, T. Ward, and C. Markham, "Brain-computer interface using a simplified functional near-infrared spectroscopy system," *Journal of Neural Engineering*, vol. 4, pp. 219–226, 2007.
- [22] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 9, p. 283288, 2001.

- [23] E. Asano, C. Juhasz, A. Shah, O. Muzik, D. Chugani, J. Shad, S. Sood, and H. Chugani, "Origin and propagation of epileptic spasms delineated on electrocorticography," *Epilepsia*, vol. 46, p. 10861097, 2005.
- [24] A. Hoogerwerf and K. Wise, "A three-dimensional microelectrode array for chronic neural recording," *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 1136–1146, Dec. 1994.
- [25] S. Huettel, A. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*. Massachusetts: Sinauer, 2009.
- [26] R. deCharms, K. Christoff, G. Glover, J. Pauly, S. Whitfield, and J. Gabrieli, "Learned regulation of spatially localized brain activation using real-time fMR," *Neuroimage*, vol. 21, p. 436443, 2004.
- [27] G. Taga, F. Homae, and H. Watanabe, "Effects of source-detector distance of near infrared spectroscopy on the measurement of the cortical hemodynamic response in infants," *NeuroImage*, vol. 38, p. 452460, 2007.
- [28] R. Kennan, S. Horovitz, A. Maki, Y. Yamashita, H. Koizumi, and J. Gore, "Simultaneous recording of event-related auditory oddball response using transcranial near infrared optical topography and surface eeg," *Neuroimage*, vol. 16, p. 587592, 2002.
- [29] E. Niedermeyer and F. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincot Williams & Wilkins, 2004.
- [30] M. Teplan, "Fundamentals of EEG measurement," *Meas. Sci. Rev.*, vol. 2, pp. 1–11, 2002.
- [31] B. Abou-Khalil and K. Misulis, *Atlas of EEG & Seizure Semiology*. Elsevier, 2006.
- [32] EEG, MedlinePlus.
- [33] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan., "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, p. 767791, 2002.
- [34] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kbler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, p. 297298, 1999.
- [35] B. Obermaier, G. Muller, and G. Pfurtscheller, "'Virtual keyboard' controlled by spontaneous EEG activity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, p. 422426, 2003.

- [36] IntendiX: Words first Personal BCI Speller.
- [37] A. Usakli, "Improvement of EEG signal acquisition: An electrical aspect for state of the art of front end," *Computational Intelligence and Neuroscience*, vol. 2010, p. 630649, 2010.
- [38] P. Konrad and T. Shanks, "Implantable brain computer interface: Challenges to neurotechnology translation," *Neurobiol. Dis.*, vol. 38, p. 369375, 2010.
- [39] T. Ball, M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage, "Signal quality of simultaneously recorded invasive and non-invasive EEG," *Neuroimage*, vol. 2009, p. 708716, 2009.
- [40] G. Schalk, J. Kubnek, K. Miller, N. Anderson, E. Leuthardt, J. Ojemann, D. Limbrick, D. Moran, and L. Gerhardt, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 4, pp. 264–75, 2007.
- [41] P. Shenoy, K. Miller, J. Ojemann, and R. Rao, "Generalized features for electrocorticographic BCIs," *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 273–280, 2008.
- [42] C. Nordhausen, E. Maynard, and R. Normann, "Single unit recording capabilities of a 100-microelectrode array," *Brain Research*, vol. 726, pp. 129–140, 1996.
- [43] S. Acharya, F. Tenore, V. Aggarwal, R. Etienne-Cummings, M. Schieber, and N. Thakor, "Decoding individuated finger movements using volume-constrained neuronal ensembles in the M1 hand area," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, pp. 15–23, 2008.
- [44] J. Chapin, K. Moxon, R. Markowitz, and M. Nicolelis, "Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex," *Nature Neuroscience*, vol. 2, pp. 664–670, July 1999.
- [45] P. Ifft, S. Shokur, Z. Li, M. Lebedev, and M. Nicolelis, "A brain-machine interface enables bimanual arm movements in monkeys," *Science: Translational Medicine*, vol. 5, pp. 1–13, 2013.
- [46] S. Kim, R. Normann, R. Harrison, and F. Solzbacher, "Preliminary study of the thermal impact of a microelectrode array implanted in the brain," in *Proc. IEEE Engineering in Medicine and Biology Conf.*, pp. 2986–2989, 2006.
- [47] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher, "A low-power integrated circuits for a wireless 100-electrode neural recording system," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 123–133, Jan. 2007.

- [48] R. Sarpeshkar, W. Wattanapanitch, S. K. Arfin, B. I. Rapoport, S. Mandal, M. W. Baker, M. S. Fee, S. Musallam, and R. A. Adersen, "Low-power circuits for brain-machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 2, pp. 173–183, Sept. 2008.
- [49] J. Aziz, K. Abdelhalim, R. Shulyzki, R. Genov, B. Bardakjian, M. Derchansky, D. Serletis, and P. Carlen, "256-channel neural recording and delta compression microsystem With 3D Electrodes," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 995–1005, 2009.
- [50] F. Chen, A. P. Chandrakasan, and V. Stojanovic, "A signal-agnostic compressed sensing acquisition system for wireless and implantable sensors," in *IEEE Custom Integrated Circuits Conference (CICC)*, 2010.
- [51] S. Mitra, J. Putzeys, F. Battaglia, C. Lopez, M. Welkenhuysen, C. Pennartz, C. van Hoof, and R. Yazicioglu, "24-channel dual-band wireless neural recorder with activity-dependent power consumption," in *Digest of Technical Papers of IEEE International Solid-State Circuits Conference*, 2013.
- [52] Z. Nadasdy, J. Csicsvari, M. Penttonen, J. Hetke, K. Wise, and G. Buzski, *Neuronal Ensembles: Strategies for Recording and Decoding*. New York: Wiley-Liss, 1998.
- [53] R. Harrison, "The design of integrated circuits to observe brain activity," *Proceedings of the IEEE*, vol. 96, pp. 1203–1216, Jul. 2008.
- [54] M. Velliste, S. Perel, M. Spalding, A. Whitford, and A. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, pp. 1098–1101, 2008.
- [55] G. Santhanam, S. Ryu, B. Yu, A. Afshar, and K. Shenoy, "A high-performance braincomputer interface," *Nature*, vol. 442, pp. 195–198, 2006.
- [56] D. Taylor, S. Tillery, and A. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," *Science*, vol. 296, pp. 1829–1832, 2002.
- [57] S. Masallam, B. Corneil, B. Greger, H. Scherberger, and R. Andersen, "Cognitive control signals for neural prosthetics," *Science*, vol. 305, pp. 258–262, 2004.
- [58] J. Carmena, M. Lebedev, R. Crist, E. Odoherly, D. Santucci, D. Dimitrov, P. Patil, C. henriquez, and M. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *Public Library of Science Biology*, vol. 1, pp. 1–16, 2003.
- [59] M. Nicolelis, D. Dimitrov, J. Carmena, R. Crist, G. Lehew, J. Kralik, and S. Wise, "Chronic, multisite, multielectrode recordings in macaque monkeys," *PNAS*, vol. 100, pp. 11041–11046, 2003.

- [60] K. Najafi, K. Wise, and T. Mochizuki, "A high-yield IC-compatible multi-channel recording array," *IEEE Transactions on Electron Devices*, vol. 32, pp. 1206–1211, 1985.
- [61] K. Wise, A. Sodagar, Y. Yao, M. Gulari, G. Perlin, and K. Najafi, "Microelectrodes, microelectronics, and implantable neural microsystems," *Proceedings of the IEEE*, vol. 96, pp. 1184–1202, 2008.
- [62] P. Campbell, K. Jones, R. Huber, K. Horch, and R. Normann, "A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array," *IEEE Transactions on Biomedical Engineering*, vol. 38, pp. 758–768, 2002.
- [63] R. Harrison, R. Kier, C. Chestek, V. Gilja, P. Nuyujukian, S. Ryu, B. Greger, F. Solzbacher, and K. Shenoy, "Wireless neural recording with single low-power integrated circuit," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, pp. 322–329, 2009.
- [64] F. Shahrokhi, K. Abdelhalim, D. Serletis, P. L. Carlen, and R. Genov, "The 128-channel fully differential digital integrated neural recording and stimulation interface," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 4, pp. 149–161, Jun. 2010.
- [65] R. Harrison and C. Charles, "A low-power low-noise cmos amplifier for neural recording applications," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 958–965, Jun. 2003.
- [66] W. Wattanapanitch, M. Fee, and R. Sarpeshkar, "An energy-efficient micropower neural recording amplifier," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, pp. 136–147, Jun. 2007.
- [67] M. Chae, Z. Yang, M. Yuce, L. Hoang, and W. Liu, "A 128-Channel 6 mW Wireless Neural Recording IC With Spike Feature Extraction and UWB Transmitter," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 4, pp. 312–321, 2009.
- [68] D. Han, Y. Zheng, R. Rajkumar, G. Dawe, and M. Je, "A 0.45V 100-channel neural-recording IC with sub-uW/channel consumption in 0.18um CMOS," in *IEEE International Solid-State Circuits Conference*, pp. 290–291, 2013.
- [69] X. Zou, X. Xu, L. Yao, and Y. Lian, "A 1-v 450-nw fully integrated programmable biomedical sensor interface chip," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1067–1077, 2009.
- [70] A. Chandran, K. Najafi, and K. Wise, "A new DC baseline stabilization scheme for neural recording microprobes," in *Proc. IEEE BMES/EMBS Conf.*, p. 386, 1999.

- [71] P. Mohseni and K. Najafi, "A low power fully integrated bandpass operational amplifier for biomedical neural recording applications," in *Proc. IEEE EMBS/BMES Conf.*, pp. 2111–2112, 2002.
- [72] I. R. H. Olsson, D. L. Buhl, A. M. Sirota, G. Buzsaki, and K. D. Wise, "Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 7, pp. 1303–1311, 2005.
- [73] T. Jochum, T. Denison, and P. Wolf, "Integrated circuit amplifier for multi-electrode intracortical recording," *Journal of Neural Engineering*, vol. 6, p. 012001, 2009.
- [74] R. Muller, S. Gambini, and J. Rabaey, "A 0.013mm², 5uW, DC-coupled neural signal acquisition IC with 0.5V supply," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 1–12, January 2012.
- [75] C. Mehring, J. Rickert, E. Vaadia, S. C. de Oliveira, A. Aertsen, and S. Rotter, "Inference of hand movements from local field potentials in monkey motor cortex," *Nature Neuroscience*, vol. 6, pp. 1253–1254, 2003.
- [76] H. Scherberger, M. Jarvis, and R. Andersen, "Cortical local field potential encodes movement intentions in the posterior parietal cortex," *Neuron*, vol. 46, p. 347354, April 2005.
- [77] M. Scott, B. Boser, and K. Pister, "An ultralow-energy ADC for smart dust," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 7, pp. 1123–1129, 2003.
- [78] H. Gao, R. Walker, P. Nuyujukian, K. Makinwa, K. Shenoy, B. Murmann, and T. Meng, "Hermes: a 96-channel full data rate direct neural interface in 0.13 m cmos," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 1043–1055, April 2012.
- [79] B. Gosselin, A. Ayoub, J. Roy, M. Sawan, F. Lepore, A. Chaudhury, and D. Guitton, "A mixed-signal multichip neural recording interface with bandwidth reduction," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, pp. 129–141, 2009.
- [80] A. Rodriguez-Perez, J. Ruiz-Amaya, M. Delgado-Restituto, and A. Rodriguez-Vazquez, "A low-power programmable neural spike detection channel with embedded calibration and data compression," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, pp. 87–100, 2012.
- [81] J. Holleman, A. Mishra, C. Diorio, and B. Otis, "A micro-power neural spike detector and feature extractor in .13um CMOS," in *IEEE Custom Integrated Circuits Conference (CICC)*, 2008.

- [82] M. Chae, W. Liu, and M. Sivaprakasam, "Design Optimization for Integrated Neural Recording Systems," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 1931–1939, 2008.
- [83] B. Rapoport, W. Wattanapanitch, H. Penagos, S. Musallam, R. Andersen, and R. Sarpeshkar, "A biomimetic adaptive algorithm and low-power architecture for implantable neural decoders," in *31st Annual International Conference of the IEEE EMBS*, 2009.
- [84] A. Zabihian and A. Sodagar, "A new architecture for multi-channel neural recording microsystems based on delta-sigma modulation," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2009.
- [85] M. Mollazadeh, K. Murari, G. Cauwenberghs, and N. Thakor, "Micropower CMOS integrated low-noise amplification, filtering, and digitization of multi-modal neuropotentials," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, pp. 1–10, 2009.
- [86] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*. Oxford University Press Inc., 2002.
- [87] R. Harrison, "A low-power integrated circuit for adaptive detection of action potentials in noisy signals," in *Proceeding of the 25th Annual International Conference of the IEEE EMBS*, 2003.
- [88] A. Bonfanti, T. Borghi, R. Gusmeroli, G. Zambra, A. Oliyink, L. Fadiga, A. Spinelli, and G. Baranauskas, "A low-power integrated circuit for analog spike detection and sorting in neural prosthesis systems," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2008.
- [89] M. Chae, Z. Yang, M. Yuce, L. Hong, and W. Liu, "A 128-channel 6 mW wireless neural recording IC with spike feature extraction and UWB transmitter," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, pp. 312–321, 2009.
- [90] T. Horiuchi, D. Tucker, K. Boyle, and P. Abshire, "Spike discrimination using amplitude measurements with a low-power CMOS neural amplifier," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2007.
- [91] T. Chen, W. Liu, and L. Chen, "128-channel spike sorting processor with a parallel-folding structure in 90nm process," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2009.
- [92] V. Karkare, S. Gibson, C. Yang, H. Chen, and D. Markovic, "A 75 uW, 16-channel neural spike-sorting processor with unsupervised clustering," in *IEEE Symposium on VLSI Circuits Digest of Technical Papers*, 2011.

- [93] S. Gibson, J. Judy, and D. Markovic, "Spike sorting: the first step in decoding the brain," *IEEE Signal Processing Magazine*, vol. 29, pp. 124–143, 2012.
- [94] V. Ventura, "Spike train decoding without spike sorting," *Neural Computing*, vol. 20, pp. 923–963, 2008.
- [95] E. Stark and M. Abeles, "Predicting movement from multiunit activity," *Journal of Neuroscience*, vol. 27, pp. 8387–8394, 2007.
- [96] A. Georgopoulos, J. Kalaska, R. Caminiti, and J. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motorcortex," *Journal of Neuroscience*, vol. 2, pp. 1527–1537, 1982.
- [97] A. Georgopoulos, J. Kalaska, R. Caminiti, and J. Massey, "Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations," *Experimental Brain Research*, vol. Suppl. 7, pp. 327–336, 1983.
- [98] A. Georgopoulos, A. Schwartz, and R. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, pp. 1357–1440, 1986.
- [99] E. Salinas and L. Abbott, "Vector reconstruction from firing rates," *Journal of Computational Neuroscience*, vol. 1, pp. 89–107, 1994.
- [100] W. Wu, Y. Gao, E. Bienenstock, J. Donoghue, and M. Black, "Bayesian population decoding of motor cortical activity using a Kalman filter," *Neural Computation*, vol. 18, pp. 80–118, 2006.
- [101] W. Wu, M. Black, D. Mumford, Y. Gao, E. Bienenstock, and J. Donoghue, "Modeling and decoding motor cortical activity using a switching Kalman filter," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 933–942, 2004.
- [102] A. Brockwell, A. Rojas, and R. Kass, "Recursive bayesian decoding of motor cortical signals by particle filtering," *Journal of Neurophysiology*, vol. 91, pp. 1899–1907, 2004.
- [103] S. Lin, J. Si, and A. Schwartz, "Self-organization of firing activities in monkey's motor cortex: trajectory computation from spike signals," *Neural Computation*, vol. 9, pp. 607–621, 1997.
- [104] V. Aggarwal, S. Acharya, F. Tenore, H. Shin, R. Etienne-Cummings, M. Schieber, and N. Thakor, "Asynchronous decoding of dexterous finger movements using M1 neurons," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, pp. 3–14, 2008.
- [105] A. Pouget, K. Zhang, S. Deneve, and P. Latham, "Statistically efficient estimation using population coding," *Neural Computation*, vol. 10, pp. 373–401, 1998.

- [106] C. Kemere, K. Shenoy, and T. Meng, "Model-based neural decoding of reaching movements: A maximum likelihood approach," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 925–932, 2004.
- [107] M. Lebedev, J. Carmena, J. O'Doherty, M. Zacksenhouse, C. Henriquez, J. Principe, and M. Nicolelis, "Cortical Ensemble Adaptation to Represent Velocity of an Artificial Actuator Controlled by a BrainMachine Interface," *The Journal of Neuroscience*, vol. 25, pp. 4681–493, 2005.
- [108] B. Jarosiewicz, S. Chase, G. Fraser, M. Velliste, R. Kass, and A. Schwartz, "Functional network reorganization during learning in a braincomputer interface paradigm," *Proc. Natl. Acad. Sci. USA*, vol. 105, p. 486491, 2008.
- [109] R. Heliot, K. Ganguly, J. Jimenez, and J. Carmena, "Learning in closed-loop brain-machine interfaces: Modeling and experimental validation," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 40, pp. 1387–1397, 2010.
- [110] A. Abbott, "Neuroprosthetics: in search of the sixth sense," *Nature*, vol. 442, pp. 125–127, 2006.
- [111] A. Suminski, D. Tkach, A. Fagg, and N. Hatsopoulos, "Incorporating feedback from multiple sensory modalities enhances brain-machine interface control," *The Journal of Neuroscience*, vol. 30, pp. 16777–16787, 2010.
- [112] J. Rubinstein, "How cochlear implants encode speech," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 12, pp. 444–448, 2004.
- [113] J. Dethier, V. Gilja, P. Nuyujukian, S. Elassaad, K. Shenoy, and K. Boahen, "Spiking neural network decoder for brain-machine interfaces," in *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on*, 2011.
- [114] Y. Chen, A. Basu, L. Liu, X. Zou, R. Rajkumar, G. Dawe, and M. Je, "A digitally assisted, signal folding neural recording amplifier," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 8, pp. 528–542, Aug 2014.
- [115] Y. Chen, B. Arindam, and M. Je, "A digitally assisted, pseudo-resistor-less amplifier in 65nm CMOS for neural recording applications," in *IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2012.
- [116] Y. Chen, B. Arindam, and M. Je, "A signal folding neural amplifier exploiting neural signal statistics," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2012.
- [117] L. Liu, X. Zou, W. Goh, R. Ramamoorthy, G. Dawe, and M. Je, "800 nW 43 nV/sqrt Hz neural recording amplifier with enhanced noise efficiency factor," *Electronics Letters*, vol. 48, pp. 479–480, 2012.

- [118] J. G. Webster, "Reducing Motion Artifacts and Interference in Biopotential Recording," *IEEE Transactions on Biomedical Engineering*, vol. 31, pp. 823–6, Dec 1984.
- [119] Y. M. Chi, T.-P. Jung, and G. Cauwenberghs, "Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 106–119, 2010.
- [120] A. Basu, R. Robucci, and P. Hasler, "A Low-power, Compact, Adaptive Log-TIA operating over Seven decades of current," *IEEE Transactions on Circuits and Systems-I*, vol. 54, no. 10, pp. 2167–77, 2007.
- [121] A. Basu and P. Hasler, "A Fully Integrated Architecture for Fast and Accurate Programming of Floating-gates," *IEEE Transactions on VLSI*, vol. 19, no. 6, pp. 953–62, 2011.
- [122] S. ODriscoll, K. Shenoy, and T. Meng, "Adaptive resolution ADC array for an implantable neural sensor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 2, pp. 120–130, 2011.
- [123] B. Lewis and P. O'Donnell, "Ventral tegmental area afferents to the prefrontal cortex maintain membrane potential up states in pyramidal neurons via D(1) dopamine receptors," *Cerebral Cortex*, vol. 10, pp. 1168–75, 2000.
- [124] H. Trantham, K. K. Szumlinski, K. McFarland, P. W. Kalivas, and A. Lavin, "Repeated cocaine administration alters the electrophysiological properties of prefrontal cortical neurons," *Neuroscience*, vol. 113, pp. 749–753, 2002.
- [125] M. V. Puig, F. Artigas, and P. Celada, "Modulation of the activity of pyramidal neurons in rat prefrontal cortex by raphe stimulation in vivo: involvement of serotonin and GABA," *Cerebral Cortex*, vol. 15, pp. 1–14, 2005.
- [126] X. Zou, L. Liu, J. Cheong, L. Yao, P. Li, M. Cheng, W. Goh, R. Rajkumar, G. Dawe, K. Cheng, and M. Je, "A 10-channel 1-mW implantable neural recording IC," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. PP, pp. 1–13, 2013.
- [127] R. Muller, S. Gambini, and J. Rabaey, "A 0.013mm² 5uW DC-coupled neural signal acquisition IC with 0.5V supply," in *IEEE International Solid-State Circuits Conference*, 2011.
- [128] M. Azin, D. Guggenmos, S. Barbay, R. Nudo, and P. Mohseni, "A battery-powered activity-dependent intracortical microstimulation IC for brain-machine-brain interface," *IEEE Journal of Solid-State Circuits*, vol. 46, pp. 731–745, April 2011.

- [129] J. L. Novak and B. C. Wheeler, "Multisite hippocampal slice recording and stimulation using a 32 element microelectrode array," *Journal of Neuroscience Methods*, vol. 23, no. 2, pp. 149 – 159, 1988.
- [130] R. Blum, J. Ross, E. Brown, and S. DeWeerth, "An integrated system for simultaneous, multichannel neuronal stimulation and recording," *IEEE Transactions on Circuits and Systems–I: Regular Papers*, vol. 54, pp. 2608–2618, 2007.
- [131] Y. Jimbo, N. Kasai, K. Torimitsu, T. Tateno, and H. Robinson, "A system for MEA-based multisite stimulation," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 241–248, 2003.
- [132] J. Minzly, J. Mizrahi, N. Hakim, and A. Liberson, "Stimulus artefact suppressor for EMG recording during FES by a constant-current stimulator," *Medical and Biological Engineering and Computing*, vol. 31, no. 1, pp. 72–75, 1993.
- [133] J. Gnadat, S. Echols, A. Yildirim, H. Zhang, and K. Paul, "Spectral cancellation of microstimulation artifact for simultaneous neural recording in situ," *Biomedical Engineering, IEEE Transactions on*, vol. 50, pp. 1129–1135, Oct 2003.
- [134] D. A. Wagenaar and S. M. Potter, "Real-time multi-channel stimulus artifact suppression by local curve fitting," *Journal of Neuroscience Methods*, vol. 120, no. 2, pp. 113 – 120, 2002.
- [135] L. Yao and M. Je, "Cmos based 16-channel neural/muscular stimulation system with arbitrary waveform and active charge balancing circuit," in *Integrated Circuits (ISIC), 2011 13th International Symposium on*, pp. 456–459, Dec 2011.
- [136] M. Schieber, "Individuated finger movements of rhesus monkeys: A means of quantifying the independence of the digits," *Journal of neurophysiology*, vol. 65, pp. 1381–1391, 1991.
- [137] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, p. 489501, 2006.
- [138] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, pp. 513–529, April 2012.
- [139] A. Hasting, *The Art of Analog Layout*. Prentice Hall, 2006.
- [140] E. Neftci and G. Indiveri, "A device mismatch compensation method for VLSI neural networks," in *Proceeding of the IEEE Biomedical Circuits and Systems Conference*, 2010.

- [141] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, “Op-elm: Optimally pruned extreme learning machine,” *Neural Networks, IEEE Transactions on*, vol. 21, pp. 158–162, Jan 2010.
- [142] K. Bult and G. Ceelen, “An inherently linear and compact MOST-only current division technique,” *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 1730–1735, 1992.
- [143] C. Mead, *Analog VLSI and Neural Systems*. Addison-Wesley, Reading, MA, 1989.
- [144] R. U. Muller and J. L. Kubie, “The firing of hippocampal place cell predict the future position of freely moving rats,” *The Journal of Neuroscience*, vol. 9, no. 12, pp. 4101–4110, 1989.