Metal Oxide Memristors with Internal Dynamics for Neuromorphic Applications

by

Chao Du

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2017

Doctoral Committee:
Professor Wei D. Lu, Chair
Assistant Professor Ronald Dreslinski Jr.
Associate Professor Zhengya Zhang
Associate Professor Zhaohui Zhong

ACKNOWLEDGEMENTS

Enormous gratitude is extended to my advisor, Prof. Wei D. Lu, who has given me insightful suggestions and constant support throughout my course of study. He has always been available and helpful, and has really set up a great example of blending smart mind and hard work into success. I would like to thank my committee members for their useful discussions: Prof. Dreslinski, Prof. Zhang and Prof. Zhong. I would also like to express my appreciation to my former and current group members: Dr. Yang, Dr. Gaba, Dr. Chang, Dr. Kim, Dr. Sheridan, Dr. Zhu, Dr. Lin, Dr. Zhou, Wen Ma, Jihang Lee, Fuxi Cai and Yeonjoo Jeong for their helpful discussions and timely assistances.

I would like to express my deepest gratitude to my family, especially my wife Zhaoqi Xu, for their relentless support and encouragement during my course of study.

I would also like to thank the Lurie Nanofabrication Facility (LNF) staff for their technical support: Sandrine Martin, Matthew Oonk, Vishva Ray, Gregory Allion, David Sebastian, Anthony Sebastian, Brian Armstrong, Nadine Wang, Shawn Wright, Katharine Beach, Robert Hower, and Kevin Owen.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS……………………………………...……………………………ii

LIST OF FIGURES…………………...……………………………………………………vi

LIST OF TABLES………………………………………………………………………...x

ABSTRACT……………………………….……………………………………….......xi

Chapter 1 Introduction 1

1.1 Memristors 1

1.2 Biological Synapses and Memristor Based Artificial Synapses 2

1.3 Memristor Crossbar Network for Neuromorphic Applications 4

1.4 Organization of The Dissertation 5

Chapter 2 Memristors Fabrication and Basic Characteristics 7

2.1 WO_x Memristor Fabrication Process... 7

2.2 Basic Characteristics of WO_x Memristor 9

2.3 WO_x Memristor Switching Mechanism and 1st Order Dynamics 11

2.4 Ta_2O_5-TaO_x Memristor Fabrication 12

2.5 Ta_2O_5-TaO_x Memristor Switching Behaviors and Mechanism.................................... 13

2.6 Complimentary Resistive Switching 15

2.7 Conclusion.. 16

Chapter 3 Second Order Memristor Dynamics.. 18

3.1 Basic Memristor Model.. 18

3.2 Second Order Memristor Model 19

3.3 WO_x (volatile) Memristor as 2nd Order Memristor 20
3.3.1 From Measurement Results to Modeling 20
3.3.2 Equations Describing Memristor Dynamics 23
3.3.3 SPICE Code 25
3.3.4 Simulation Results 27
3.4 Ta_2O_5-TaO_x (nonvolatile) Memristor as 2nd Order Memristor 30
3.5 Conclusion 33
Chapter 4 Bio-realistic Implementation of Synaptic Functions Using Internal Ionic Dynamics of Memristors 35
4.1 Short-term Synaptic Behaviors 36
4.1.1 Paired-pulse Facilitation 36
4.1.2 Frequency-dependent Weight Change 38
4.2 Long-term Synaptic Behaviors 39
4.2.1 Spike-timing Dependent Plasticity 40
4.2.2 Frequency-dependent Long-term Weight Change 46
4.3 Metaplasticity 48
4.3.1 Experience-dependent Plasticity in WO_x Memristor 49
4.3.2 Metaplasticity in Ta_2O_5-TaO_x Memristor 53
4.4 Conclusion 55
Chapter 5 Memristor Array for Sparse Coding 58
5.1 Sparse Coding 58
5.2 Sparse Coding Algorithm 59
5.3 Memristor Network for Sparse Coding 60
5.4 Sparse Coding Experiment Results 63
5.5 Conclusiton 64

Chapter 6 Memristor Array for Reservoir Computing 67
6.1 Reservoir Computing 67
6.2 Liquid State Machine 68
6.3 WO_x Memristor Based Synapses as the Liquid 71
6.4 System and Task Design 73
6.5 LSM for Simple Digit Recognition 73
6.6 LSM for Handwritten Digit Recognition 80
6.7 Discussion 83
6.8 Conclusion 85
Chapter 7 Memristor Performance Improvement 87
7.1 Single Device Performance Improvement 87
7.1.1 TaO_x Memristor Performance Improvement 88
7.1.2 WO_x Memristor Fabrication Process Optimization 90
7.2 Memristor Array Performance Improvement 93
7.3 Transition of WO_x Memristor from Analog to Digital Type Switching 95
7.4 Conclusion 99
Chapter 8 Future Works 101
8.1 Device Performance Improvement 101
8.1.1 TaO_x Memristor Analog Behavior Improvement 101
8.1.2 WO_x Memristor Fabrication Methods Optimization 102
8.2 Memristor Array for Reservoir Computing 102
8.2.1 Memristor Array Optimization 102
8.2.2 Other Neuromorphic Applications 102

LIST OF FIGURES

Figure 1-1: Memristor as the forth electrical element ... 1
Figure 1-2: Crossbar structure for memristor ... 2
Figure 1-3: Schematic of synapse ... 3
Figure 1-4: Memristors as electrical synapses ... 3
Figure 1-5: Memristor crossbar architecture ... 4
Figure 2-1: Schematic of a WO_x memristor ... 7
Figure 2-2: Broken TE at the crosspoint ... 8
Figure 2-3: SEM image showing spacer formation of WO_x memristor ... 9
Figure 2-4: Gradual conductance increase of WO_x memristor by DC sweep ... 9
Figure 2-5: Gradual conductance increase and decrease of WO_x memristor by pulses ... 10
Figure 2-6: Schematic illustration of the internal V_O dynamics in WO_x memristor ... 11
Figure 2-7: Schematic of a Ta_2O_5-TaO_x memristor ... 13
Figure 2-8: I-V curve of a Ta_2O_5-TaO_x memrsitor ... 13
Figure 2-9: Schematic illustration of the internal V_O dynamics in Ta_2O_5-TaO_x memristor ... 14
Figure 2-10: Complimentary resistive switching in Ta_2O_5-TaO_x memristor... 15
Figure 3-1: Read current decay of WO_x memristor after stimulation... 20
Figure 3-2: Conductance saturation by repeated stimulations ... 22
Figure 3-3: Conductance change by programming pulse pairs with different intervals ... 23
Figure 3-4: Simulation of the two-stage decay using the 2nd order WO_x memristor model ... 27
Figure 3-5: AC sweep measurement and simulation results ... 28
Figure 3-6: DC sweep measurement and simulation results ... 28
Figure 3-7: Simulation results showing the evolution of short-term and long-term state variables ... 29

Figure 3-8: Simulation results showing the effect of short-term dynamics on long-term plasticity. 29
Figure 3-9: Temporal effect of heat on weight change in Ta_2O_5-TaO_x memristor 30
Figure 3-10: Simulated transient temperature evolution during and after the application of only heating or programming pulse 31
Figure 3-11: Simulated transient temperature evolution during and after the application of a heating pulse followed by a programming pulse 31
Figure 4-1: Results of PPF 36
Figure 4-2: PPF effect obtained in WO_x memristor 37
Figure 4-3: PPF ratio for different pulse intervals obtained from WO_x memristor. 38
Figure 4-4: Frequency-dependent weight change in WO_x memristor 39
Figure 4-5: Waveforms with overlap for STDP 40
Figure 4-6: Stimulation protocol for STDP on WO_x memristor 41
Figure 4-7: STDP implemented in WO_x memristor 42
Figure 4-8: Illustration of weight change under different pulse pair timing by simulation based on WO_x memristor model 43
Figure 4-9: Illustration of state variable dynamics for different Δt by simulation based on WO_x memristor model 44
Figure 4-10: Stimulation protocol for STDP on Ta_2O_5-TaO_x memristor 45
Figure 4-11: STDP implemented in Ta_2O_5-TaO_x memristor 46
Figure 4-12: Frequency-dependent long-term weight change of Ta_2O_5-TaO_x memristor 47
Figure 4-13: The standard paradigm of metaplasticity 48
Figure 4-14: Experience-dependent synaptic weight change of biological synapse 50
Figure 4-15: Experience-dependent weigh change of WO_x memristor 51
Figure 4-16: Experience-dependent weight change by different stimulation strength on WO_x memristor 52
Figure 4-17: Experience-dependent plasticity as shown by hold voltage of WO_x memristor 53
Figure 4-18: CRS of Ta_2O_5-TaO_x memristor 54
Figure 4-19: Metaplasticity observed in Ta_2O_5-TaO_x memristor 55
Figure 5-1: Schematic of the sparse coding concept 59
Figure 5-2: Schematic of memristor crossbar based computing 61

Figure 5-3: Memristor crossbar network for sparse coding 62
Figure 5-4: Sparse coding results by WO_x memristor crossbar array 63
Figure 6-1: Reservoir computing. 67
Figure 6-2: Schematic of a Liquid State Machine 70
Figure 6-3: Schematic of a Liquid State Machine with synapses as the liquid. 71
Figure 6-4: Expected synapse's response to a spike train 72
Figure 6-5: Memristor's temporal response to a pulse train 72
Figure 6-6: Simple digit images 73
Figure 6-7: LSM for simple digit recognition 74
Figure 6-8: Experimental setup for LSM 75
Figure 6-9: Memristors' variation and response to six pulse trains 76
Figure 6-10: Memristor's response to ten pulse trains 76
Figure 6-11: Liquid's internal states after subjected to the ten digit inputs 77
Figure 6-12: The effect of decay in separating different digit images 79
Figure 6-13: Recognition of noisy digits 79
Figure 6-14: Samples from the MNIST database. 80
Figure 6-15: LSM for handwritten digit recognition 80
Figure 6-16: Liquid states for three MNIST digit images 82
Figure 6-17: Device degradation during digit recognition task 84
Figure 7-1: Better linearity in a TaO_x single layer memristor 89
Figure 7-2: Heat decay measured in Ta_2O_5-TaO_x memristor with Pt TE. 90
Figure 7-3: Issue of residues along the sidewalls of WO_x memristors 91
Figure 7-4: Broken TE of WO_x memristor without spacer. 91
Figure 7-5: Schematic of spacer formation 92
Figure 7-6: SEM image of spacer formed 92
Figure 7-7: 32x32 WO_x memristor array with 500 nm line width and 4 μm pitch 94
Figure 7-8: 32x32 WO_x memristor array with 200 nm line width and 500 nm pitch 94
Figure 7-9: Transition of WO_x memristor from analog to digital type switching by multiple pulses 96
Figure 7-10: Full range DC sweep after transition 97
Figure 7-11: DC sweep before transition 98

Figure 7-12: Abrupt resistance state change by pulses after transition 98
Figure 7-13: Retention test after transition 99

LIST OF TABLES

Table 3-1: The values of parameters used in the simulation 25
Table 6-1: Experimental and simulation results of handwritten digit recognition by memristor-based LSM 83

ABSTRACT

Inspired by the advantages of biological systems, especially high efficiency, parallel processing ability and the combination of computing with memory, solid-state neural network systems have attracted much attention for neuromorphic applications. Memristors, with several unique properties, are exceptional candidates for emulating artificial synapses, a critical component for neural networks. This thesis work explores the device characteristics and dynamics for synaptic functions implementation and networks for neuromorphic applications, with device performance improvement and fabrication process optimization.

Device fabrication, electrical studies for both WO_x and Ta_2O_5-TaO_x memristors are presented. Bipolar resistive switching is observed due to oxygen vacancy redistribution within the switching layer upon an applied electric field. In WO_x memristor, oxygen vacancies drift by electric filed and spontaneous diffusion result in a gradual resistance change with decay (volatile), while in Ta_2O_5-TaO_x memristor, filament formation/rupture lead to a distinct, abrupt resistance change (nonvolatile).

The traditional framework for memristor swtiching mechanism, which has only one state variable, is fundamentally a 1st order dynamic model. Based on more experimental results showing more detailed switching behaviors, 2nd order models are proposed for both devices by introducing a second state variable, that is, the oxygen vacancy mobility for WO_x memristor and temperature for Ta_2O_5-TaO_x memristor, to quantitively capture the ionic dynamics and describe device responses over a large range of programming conditions.

Based on the understanding of the switching dynamics of these two memristors, important synaptic functions are demonstrated with natural, bio-realistic implementations in these electric devices. Both short-term synaptic plasticity, including paired-pulse facilitation, rate-dependent weigh change, and long-term plasticity, including spike-timing dependent plasticity, experience-dependent plasticity and metaplasticity are implemented.

Apart from achieving synaptic functions in single device, memristor crossbar network has also been utilized for neuromorphic applications.

The crossbar network, by forming memristor cell at each crosspoint, enables an easy way of conducting matrix operation, specifically the dot product, and can store trained weights by the adjustable resistances of memristors. Sparse coding algorithm has been implemented by the crossbar network, both through simulation based on WO_x memristor model and experiment on fabricated 32-by-32 WO_x array.

Further, the dynamics of WO_x memristor, especially the decay of state caused by spontaneous oxygen vacancy diffusion, is utilized to do temporal information processing. The handwritten digit recognition task is achieved by converting the spatial information of digit image to temporal information and fed into a network, composed of memristors in the configuration inspired by an improve reservoir computing structure, i.e., the liquid state machine, to generate the desired recognition results through the training of a simple linear readout function.

Finally, despite the successful demonstration of some synaptic functions and neuromorphic applications, the device performance could be further improved through fabrication process optimization and regulation of switching region. Memristor array fabrication process also requires optimization for better performance and higher yield with large scale for neuromorphic applications. Some progress has been achieved, which has shed light upon future research.

Chapter 1

Introduction

1.1 Memristors

Memristive devices, or so called memristors, are two-terminal electrical devices whose states, normally represented by resistance or conductance, can be regulated by the history of applied stimulations. The device states are described by one or a few internal state variable(s) and are governed by dynamic ionic processes. The concept of memristor was initially proposed in the 1970s[1,2] as shown in Figure 1-1 and has been intensively investigated in the last a few years.

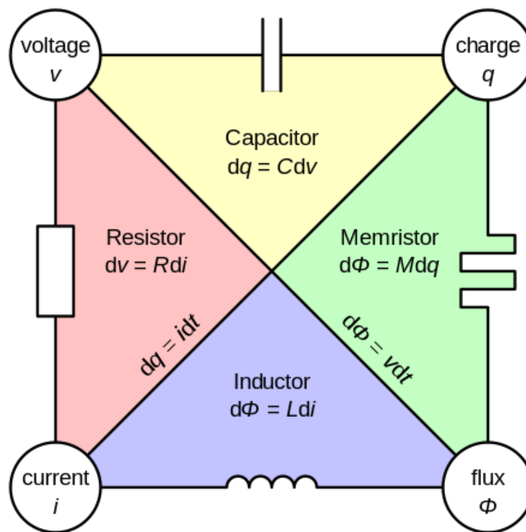

Figure 1-1: Memristor as the forth electrical element, besides resistor, inductor and capacitor.

A conventional memristor has a sandwiched metal-insulator-metal (MIM) structure, in which the switching happens in the insulator layer, or so-called switching layer. Due to the simplicity of MIM structure, memristors can be easily fabricated by inserting the switching

material between two crossing metal lines, thus forming the cell at the crosspoint, as illustrated in Figure 1-2. This is the so-called "crossbar" structure and commonly used in research.

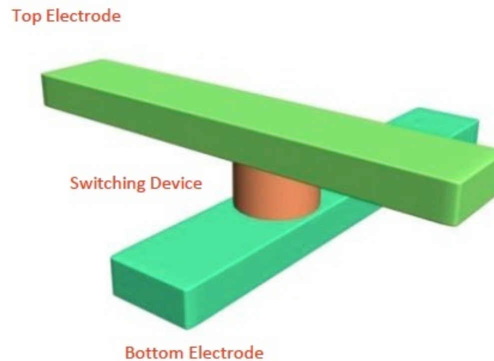

Figure 1-2: Crossbar structure for memristor. The cell is formed at the crosspoint of two metal lines by inserting the switching material.

The key advantages of memristors include the simple structure, fast speed, low power, complementary metal oxide semiconductor (CMOS) compatibility and the ability for hybrid circuit and 3D integration, making them attractive for a broad range of applications including memory, analog and reconfigurable circuits, as well as neuromorphic computing.

1.2 Biological Synapses and Memristor Based Artificial Synapses

Neurons and synapses together make up neural networks, which are the building blocks that empower humans to learn, think and remember. Synapses are connections between neurons, and can transfer and regulate signals between them as illustrated in Figure 1-3. There are $\sim 10^{11}$ neurons and $\sim 10^{14}$ synapses in a human brain[3]. A key attribute of the brain's computing power is that the synapses are "plastic" – that is, the synaptic weight, or the connection strength between neurons, can be modulated and new weight can be retained. Since the synaptic weight regulates the transmission of signals between neurons, synaptic plasticity along with the very large synaptic connectivity empowers the efficient brain-based parallel computing paradigm and lays the foundation of neuromorphic computing.

So far, most previous attempts to implement neural networks by CMOS technology suffered from the difficulty of building large networks with massive plastic connections and low

power consumption. Therefore, the prospect of building biologically inspired neuromorphic computing systems with memristor-based synapses[4] has generated significant interest since memristors can phenomenologically and bio-realistically emulate synaptic plasticity, and offer the desired large connectivity and low power budget, as illustrated in Figure 1-4[4].

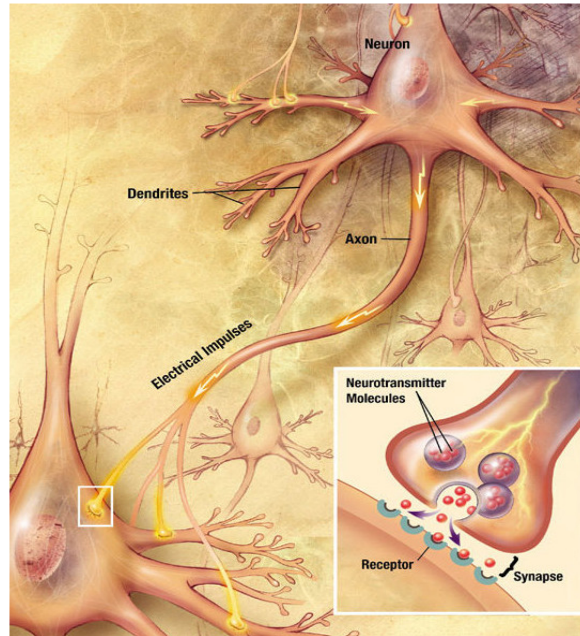

Figure 1-3: Schematic of synapse. Synapse is the connection between two neurons and can transmit signal through it from axon of one neuron to dendrites of another neuron.

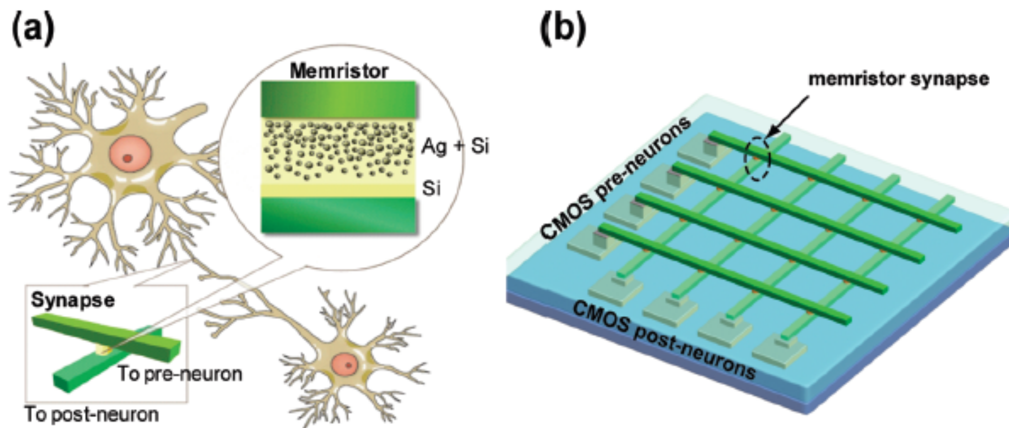

Figure 1-4: Memristors as electrical synapses. a) Schematic illustration of the concept of using memristors as synapses between neurons. The insets show the schematics of the two-terminal device geometry and the layered structure of the memristor. b) Schematic of a neuromorphic network with CMOS neurons and memristor-based synapses in a crossbar configuration.

1.3 Memristor Crossbar Network for Neuromorphic Applications

Besides being utilized as synapses in neural networks, the network of memristors itself has also been intensively investigated for neuromorphic applications recently because the crossbar architect can implement certain matrix operation in a natural and elegant fashion. As illustrated in Figure 1-5, if a vector ***X*** is input with each element x_i applied on the top electrode (the row metal lines on the left) while keeping the bottom electrode (the column metal lines at the bottom) ground, the current flowing through each memristor at the crosspoint (i, j) will be:

$$i_{i,j} = x_i w_{i,j} \quad (1\text{-}1)$$

where x_i represents the vector element which for example could be a pulse with a fixed amplitude but width modulated according to the input, and w_j represents the state of memristor, i.e., the conductance (but more often called weight in neuromorphic research). Then the current is measured at the bottom electrode and since all memritors on one column share the same bottom electrode, the current collected is the sum of all the currents flowing through all the memristors on this column:

$$I_j = \sum_{i=1}^{n} x_i w_{i,j} = \vec{X} \cdot \overrightarrow{W_j} \quad (1\text{-}2)$$

Therefore, the dot product can be easily implement in a memristor crossbar network.

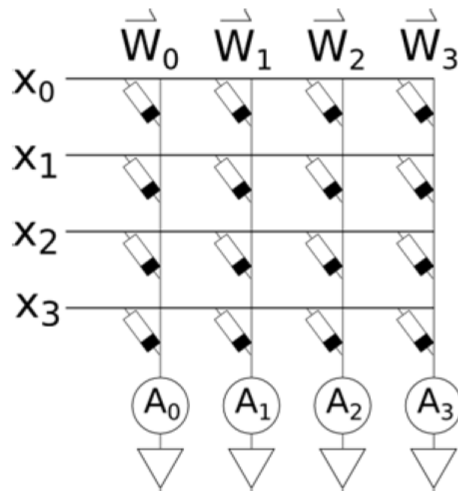

Figure 1-5: Memristor crossbar architecture. Inputs are indicated on the rows as x_i, while the charge is collated on the columns, schematically shown as A_j. Memristors are formed at the crosspoints with the weight $w_{i,j}$.

Moreover, because the resistances of memristors can be easily modulated, the information of a neuromorphic system can be stored in the network by programming the resistances of the memristors on one column corresponding to one weight vector (or called receptive field of a neuron) for further computing.

These two merits make memristor crossbar network very suitable for certain neuromorphic applications, which will be discussed in later chapters.

1.4 Organization of The Dissertation

In this chapter the fundamental background of memristors and the advantages of memristor-based approach for neuromorphic computing are introduced.

Chapter 2 discusses the fabrication, basic characteristics and switching mechanisms of WO_x memristor and Ta_2O_5-TaO_x memristor.

Chapter 3 discusses the second order effect observed in WO_x memristor and Ta_2O_5-TaO_x memristor, and the improved dynamic models that incorporate critical second order state variables for both memristors.

Chapter 4 discusses different synaptic functions we have implemented with memristors using their internal second order dynamics.

Chapter 5 and 6 discuss utilizing the switching characteristics of memristor array, especially WO_x memristor array, for some neuromorphic applications.

Chapter 5 discusses a sparse coding algorithm that has been implemented by WO_x memristor crossbar network, through both simulation based on dynamic model and experiment with fabricated devices.

Chaoter 6 discusses the handwritten digit recognition by memristor-array-based reservoir computing, emphasizing the temporal information processing ability of WO_x memristor through its internal dynamics.

Chapter 7 discusses the improvement of memristor performance for neuromorphic applications through the optimization of fabrication processes, both for single cell and large scale array.

Chapter 8 discusses the remaining issues and further research.

Reference

1. Chua, L. O. Memristor-the missing circuit element. *Circuit Theory, IEEE Trans.* **18,** 507–519 (1971).
2. Chua, L. O. & Kang, S. M. Memristive devices and systems. *Proc. IEEE* **64,** 209–223 (1976).
3. Azevedo, F. A. C. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* **513,** 532–541 (2009).
4. Jo, S. H. *et al.* Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Lett.* **10,** 1297–1301 (2010).

Chapter 2

Memristors Fabrication and Basic Characteristics

In this research project, two types of metal oxide memristors, using different metal oxides as the switching layer, are investigated: 1) WO_x memristor and 2) Ta_2O_5-TaO_x memristor. Here we introduce the fabrication process and basic characteristics of these devices.

2.1 WO_x Memristor Fabrication Process

WO_x memristors are based on a metal-insulator-metal (MIM) structure similar to those reported earlier[1,2]. The device has a palladium (Pd) top electrode (TE), a tungsten oxide (WO_x) switching layer, and a tungsten (W) bottom electrode (BE), schematically shown in Figure 2-1.

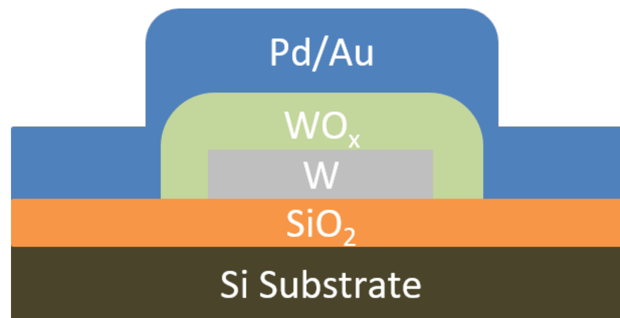

Figure 2-1: Schematic of a WO_x memristor. The device has a MIM structure, with W as the bottom electrode, WO_x as the switching layer and Pd as the top electrode.

First, a 60 nm W film is deposited on a Si/SiO_2 substrate by RF sputtering at room temperature. Then the bottom electrodes are patterned by ebeam lithography, Ni deposition by evaporation and lift-off, followed by reactive ion etching (RIE) using Ni as a hard mask to etch uncovered W. After removing Ni by wet etching, rapid thermal annealing (RTA) in pure oxygen at temperatures ranging from 375 ^oC to 450 ^oC, with annealing times ranging from 30 s to 90 s, is performed to partially oxidize the W film and form the nonstoichiometric tungsten oxide layer as

the switching layer. The thickness of the WO_x layer ranges from 40 nm to 90 nm depending on the oxidation condition, which in turn leads to different switching behaviors and allows tuning of the device performance for different applications. Finally, the Pd/Au top electrodes, where Au acts as a cover layer for better ohmic contact and a protective layer for probe station test and wire-bonding, are patterned by ebeam lithography and evaporation of metal materials. After lift-off, the tungsten oxide regions outside the crosspoints formed between the TEs and the BEs are etched away by RIE, using the TEs as a hard mask. Another photography and metal deposition process, usually by evaporating NiCr (5 nm) and Au (140 nm), may be performed to form the bonding pads for both BEs and TEs for wire-bonding the chip to a chip carrier for measurement using customized testing board of our group.

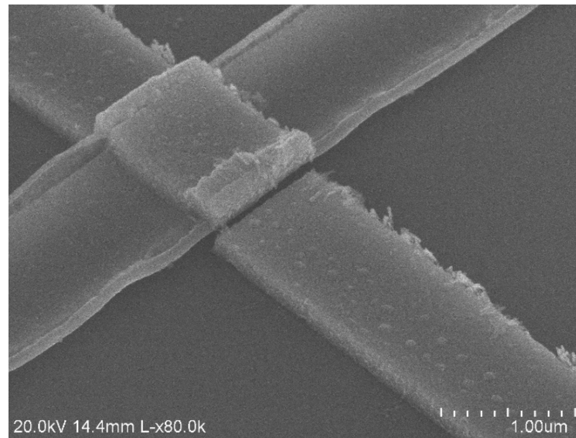

Figure 2-2: Broken TE at the crosspoint. The TE breaks due to its bad step coverage and the height of BE.

It should be noted that after oxidation of the W BE, the total thickness of the remaining BE and the switching layer could be as large as 100 nm, which may cause the TE to break at the crosspoint due to its bad step coverage when using evaporation for metal deposition, as shown in Figure 2-2. This will cause TE discontinuity and more severely, no switching behavior at all. Therefore, for a better yield and reliably repeatable device fabrication, we introduced a spacer structure along the sidewall of BE, which will be discussed in more detail in later chapter. Briefly speaking, after BE formation a conformal SiO_2 layer (~250 nm - 400 nm) is deposited by plasma-enhanced chemical vapor deposition (PECVD) and then directionally etched by precisely measuring the etch rate and controlling the etching time using RIE. The isotropic deposition by PECVD leads to uniform coverage of the SiO_2 in all directions, while the following anisotropic RIE etching removes SiO_2 in the vertical direction. These processes leave residual SiO_2 only

along the sidewalls of the BEs, creating a ramp like, self-aligned structure termed the spacer, as shown in Figure 2-3.

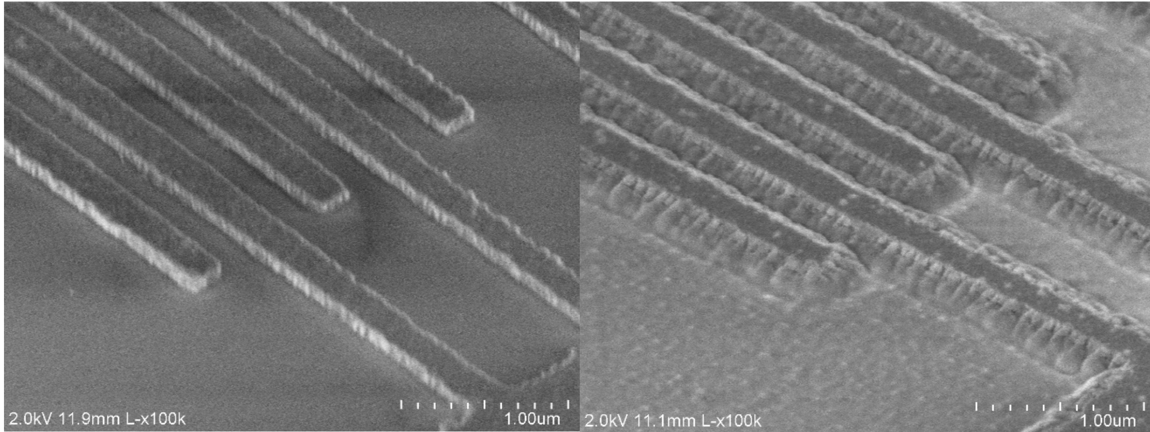

Figure 2-3: SEM image showing spacer formation of WO_x memristor. Left: W bottom electrode with thickness of ~60-70 nm and width of 200 nm. Right: Spacer formed along the side walls of W bottom electrodes after isotropic deposition of 400 nm SiO_2 and directional etch that fully removes SiO_2 on top of the BEs but leaves residual SiO_2.

2.2 Basic Characteristics of WO_x Memristor

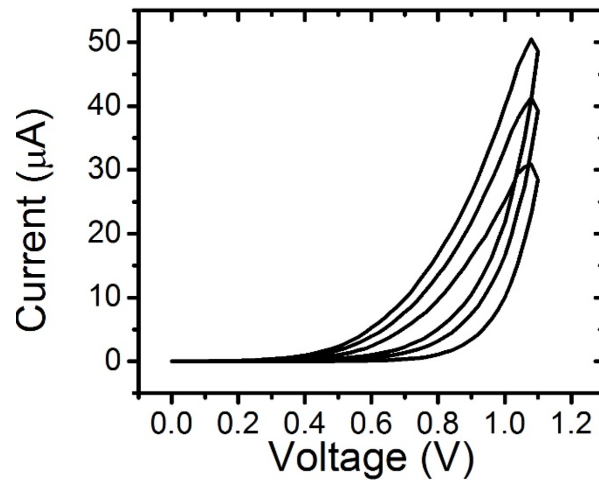

Figure 2-4: Gradual conductance increase of WO_x memristor by DC sweep. Three consecutive positive DC sweeps, from 0 V to 1.1 V were applied to the device. Gradual conductance change and overlaps between each sweep were observed.

The WO_x memristor was characterized by DC and AC electrical measurements. A typical DC measurement involves sweeping the voltage on the TE from 0 V to 1.1 V at a ramp rate of 1.5 V/s and then back to 0 V, with the BE grounded (all signals are applied to the TE with the BE grounded, if not otherwise specifically mentioned, in the following context). The current

through the device was monitored and recorded during the voltage sweep. A typical result is shown in Figure 2-4, highlighting two characteristics: 1) gradual conductance (or current) increase and 2) overlaps between the loops. The gradual conductance increase indicates an analog type of resistance modulation (switching), which is an intrinsic characteristic of WO_x memristor.

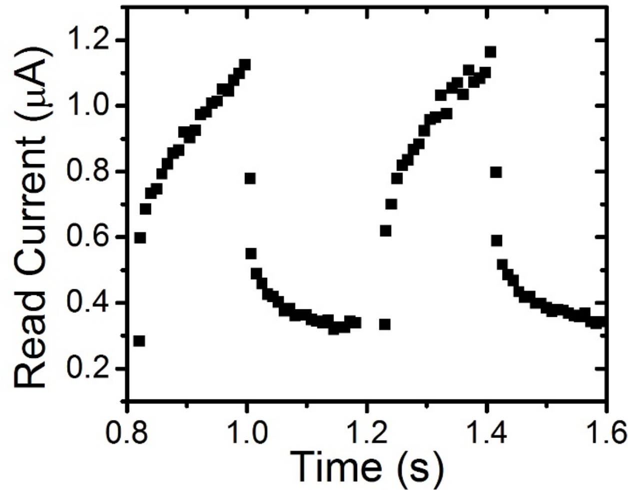

Figure 2-5: Gradual conductance increase and decrease of WO_x memristor by pulses. 20 positive pulses (+1.2 V, 100 µs) followed by 20 negative pulses (-1.2 V, 100 µs) were applied to the device and the conductance was obtained by a small read pulse (0. 5 V ,1 ms) after each operation pulse. Gradual conductance increase(decrease), controlled by positive(negative) pulses, was observed.

The analog switching behavior was more clearly demonstrated through pulse measurement. Figure 2-5 shows results obtained by applying 20 positive pulses (+1.2 V, 100 µs) followed by 20 negative pulses (-1.2 V, 100 µs) as programming and resetting pulses, respectively. The device conductance was obtained through the read current by applying a read pulse (0.5 V, 1 ms) after each programming/resetting pulse. The read current, representing the state of device, showed a gradual increase, which we call a SET process with positive pulses and decrease, referred as RESET process with negative pulses. As shown in Figure 2-5, multiple states between the low read current and high read current can be achieved by controlling the number of applied programming/resetting pulses (or equivalently, the duration of the pulses).

2.3 WO_x Memristor Switching Mechanism and 1st Order Dynamics

The above behaviors can be explained by the redistribution of ions[2,3], here in the form of oxygen vacancies (V_Os), in the switching layer as schematically illustrated in Figure 2-6.

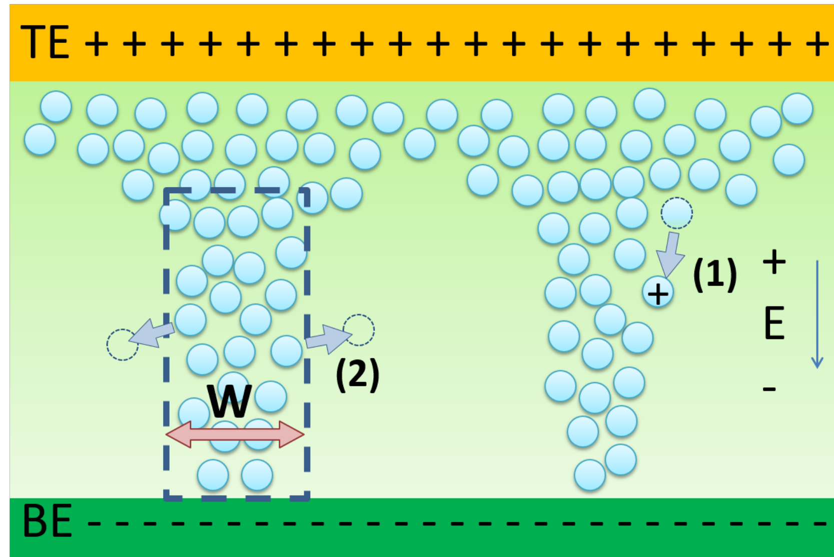

Figure 2-6: Schematic illustration of the internal V_O dynamics in WO_x memristor. Two competing effects are shown: (1) electric field driven V_O drift and (2) spontaneous diffusion of V_Os.

Upon the application of a positive voltage, positively charged oxygen vacancies will be driven toward the BE under a high electric field at a speed exponentially dependent on the electric field[4] to form regions rich with oxygen vacancies as indicated by the dash square in the figure. These regions form high conductance channels since the V_Os act as dopants that increase the local conductivity, resulting in a tunneling-like conduction mechanism through the V_O defects[5,6]. The rest of the regions remain at low conductance and form a Schottky contact with the W electrode. Therefore, the device can be modeled as having two conduction paths in parallel, with a state variable w representing the relative area of the more conductive region. When positive voltage is applied continuously, either by consecutive positive DC sweeps as in Figure 2-4 or by multiple positive pulses in Figure 2-5, the conductive region will keep expanding, resulting in the increasing read current. When negative pulses are applied, as in Figure 2-5, the electric field, now with opposite direction, drives the oxygen vacancies away from the BE thus reducing the conductive region, resulting in the decreasing read current.

Meanwhile the non-uniform distribution of the oxygen vacancies will lead to a spontaneous diffusion. Specifically, the more resistive state is the thermodynamically ground state[4] and the diffusion of oxygen vacancies will lead to a natural decay toward that state, causing the decrease of conductance therefore the overlapping of the neighboring hysteresis loops in Figure 2-4.

The memristor dynamics can now be described by the following equations:

$$I = (1 - w)\alpha[1 - exp(-\beta V)] + w\gamma \sinh(\delta V) \quad (2\text{-}1)$$

$$\frac{dw}{dt} = \lambda sinh(\eta V) - \frac{w}{\tau} \quad (2\text{-}2)$$

where Equation (2-1) is the I-V equation which includes the Schottky term (1st term) and the tunneling term (2nd term). The two conduction channels are in parallel and their relative weight is determined by the internal state variable w, which is the normalized area index of the conductive region, i.e., $w = 0$ indicates fully Schottky-dominated conduction (no conductive channels) while $w = 1$ indicates fully tunneling-dominated conduction. Equation (2-2) is the dynamics equation which describes the rate of change of the state variable w with respect to the applied voltage, including the drift under electric filed (1st term) and the spontaneous diffusion (2nd term). $\alpha, \beta, \gamma, \delta, \lambda, \eta$ are all positive-valued parameters determined by material properties. τ is the diffusion time constant characterizing the state decay.

2.4 Ta_2O_5-TaO_x Memristor Fabrication

The TaO_x-based memristor devices consist of a Au/Pt/Ta_2O_5/TaO_x/Pt crossbar structure as shown in Figure 2-7 with a line width from 200 nm up to 2 µm. Pt bottom electrodes with 50 nm thickness were firstly fabricated on the SiO_2/Si substrate by e-beam or photo lithography and e-beam evaporation of the Pt metal, followed by the lift-off process. A 35 nm TaO_x layer was deposited by DC reactive sputtering of a Ta metal target in an Ar/O_2 environment at room temperature, followed by the deposition of a 5 nm-thick Ta_2O_5 switching layer through sputtering a Ta_2O_5 ceramic target in the same chamber but without O_2. Subsequently, top electrodes, with 30 nm Pt and 25nm Au, were fabricated following the BE fabrication processes. Finally, a reactive ion etching process using SF_6/Ar was performed to expose the contact area of the bottom electrodes.

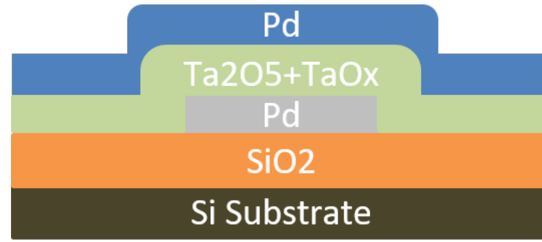

Figure 2-7: Schematic of a Ta_2O_5-TaO_x memristor. The device has a MIM structure, with Pd as the bottom electrode, Ta_2O_5-TaO_x as the switching layer and Pd as the top electrode.

2.5 Ta_2O_5-TaO_x Memristor Switching Behaviors and Mechanism

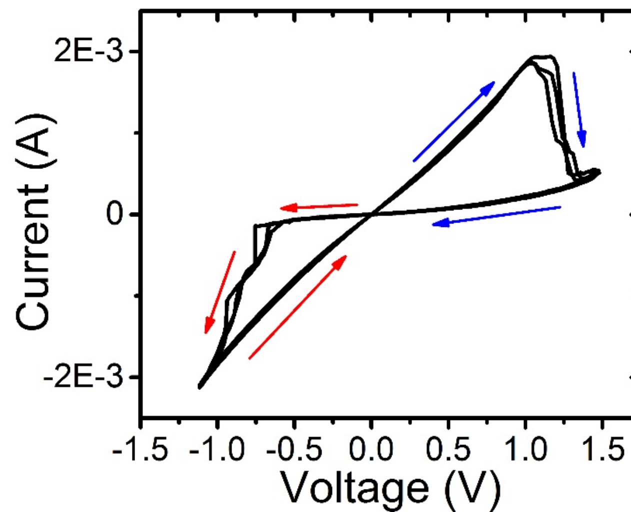

Figure 2-8: I-V curve of a Ta_2O_5-TaO_x memrsitor. Typical bipolar resistive switching, with negative voltage SET the device (conductance increase) and positive voltage RESET the device (conductance decrease).

The Ta_2O_5-TaO_x memristor was characterized by DC measurement. The voltage on the TE was first swept in a negative loop (0 V to -1.3 V to 0 V) then a positive loop (0 V to 1.5 V to 0 V), with the BE grounded. The current through the device was monitored and recorded during the voltage sweep. A typical hysteresis loop is shown in Figure 2-8. The device, initially in a high resistance state (HRS) can be SET by negative voltage to a low resistance state (LRS) as indicated by red arrows. Then positive voltage can RESET the device back to HRS as indicated by blue arrows.

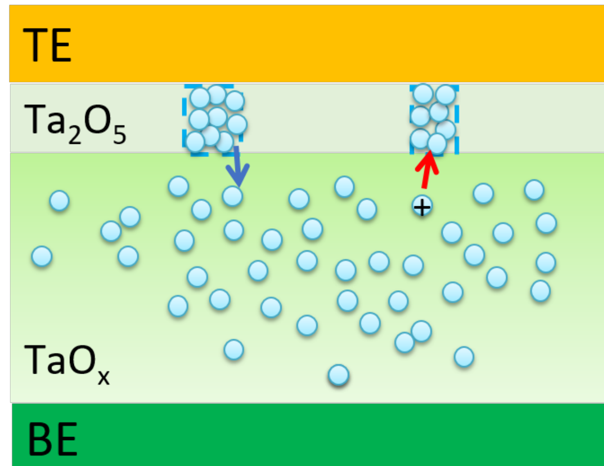

Figure 2-9: Schematic illustration of the internal V_O dynamics in Ta_2O_5-TaO_x memristor. The oxygen vacancies (cyan spheres) in TaO_x layer can be driven into or away from Ta_2O_{5-y} layer, resulting in the formation or rupture of conduction filaments (blue rectangles).

Similar to the WO_x memristor, the resistive switching in Ta_2O_5-TaO_x memristor can be understood based on the V_O redistribution, specifically, V_O exchange between the Ta_2O_{5-y} (the subscript 5-y indicates the as-fabricated stoichiometric Ta_2O_5 layer will lose some oxygen ions during the switching process) and TaO_x layers[7,8]. As illustrated in Figure 2-9, because the two tantalum oxide layers are in series, the device resistance is mainly determined by the more resistive, oxygen-rich Ta_2O_{5-y} layer. A negative voltage attracts V_Os from the oxygen-deficient TaO_x layer into the Ta_2O_{5-y} layer, forming conduction filaments connecting the TE and the conductive TaO_x layer and in turn switches the device to LRS (SET process). When a positive voltage is applied, it repels the V_Os from the Ta_2O_{5-y} layer and breaks those filaments, thus switching the device to HRS (RESET process).

A major difference compared with the switching dynamics of WO_x memristors is that in the Ta_2O_5-TaO_x case V_Os of very high concentration are typically accumulated in the conductive regions, essentially making the conductive region metallic where the high concentration of V_Os form an extended state for electrons to move through. This conducting region is analogous to a "filament". Its formation/rupture leads to dramatic resistance changes thus the "digital" type resistive switching behavior. While in WO_x memristors the effect of V_O is similar to a doping near the oxide-metal interface, therefore the resistance change is gradual when the amount of V_Os are modulated during the SET and RESET processes, leading to the incremental "analog" type resistive switching behavior.

2.6 Complimentary Resistive Switching

Besides conventional resistive switching behaviors shown above, another effect was observed in Ta_2O_5-TaO_x memristor at certain conditions. Different from typical batches which used photolithography for both TE and BE patterning, one batch of device was fabricated by using ebeam lithography for TE and BE patterning to reduce the line width to 200 nm. As a result, the switching is restricted to a much smaller area and the heat generated during switching is also concentrated, leading to more elevated temperature and more dramatic filament modulation.

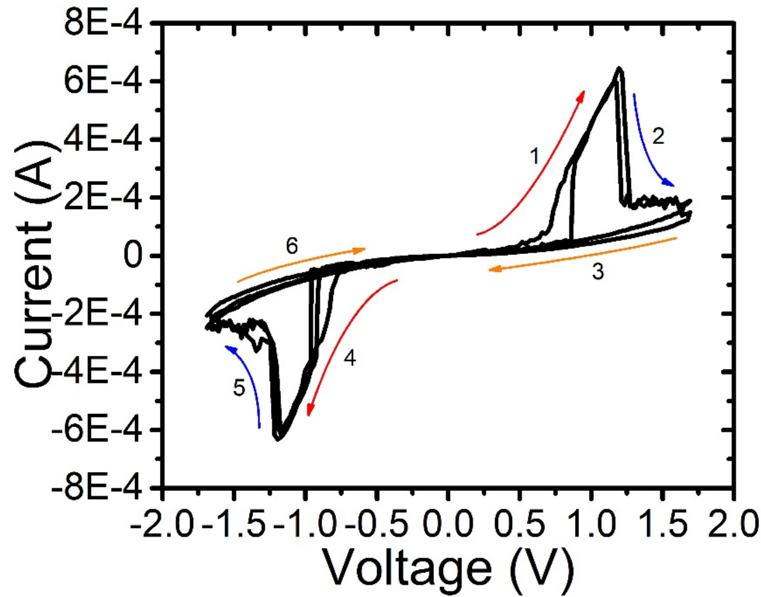

Figure 2-10: Complimentary resistive switching in Ta_2O_5-TaO_x memristor. The device was made by ebeam lithography to achieve narrower line width thus confining the switching location and heat generation, leading to a more dramatic filament modulation thus different switching behavior by consecutive DC sweeps.

After a forming and subsequence RESET process, several DC voltage sweeps were applied on the device and its current-voltage (I-V) characteristics were shown in Figure 2-10. In the positive voltage sweep, the device, initially in a high resistance state (HRS1), was first switched into low resistance state (LRS) as shown in step 1 when the voltage was increased from 0.5 V to 1.0 V. Then it was switched back to another high resistance state (HRS2) when the voltage was further increased to 1.8 V (step 2). After sweeping back to 0 V (step 3), a similar resistance change was observed during the sweep at the negative cycle: the device was first switched from HRS2 into LRS by an intermediate negative voltage sweep (step 4) then fell back

to HRS1 as the voltage was further increased (step 5) and remained in HRS1 (step 6), consistent with the complimentary resistive switching (CRS) effect observed in memristive devices[9,10]. The underlying mechanism of the observed CRS can be understood based on the movement and redistribution of a limited amount of oxygen vacancies under electric field, resulting in the formation of filaments and gaps at the two opposite interfaces between the oxide layer and either TE or BE[10]. This characteristic provides an insight into tunable resistance switching for synaptic function implementation, as will be discussed later in Chapter 4.

2.7 Conclusion

The fabrication processes for WO_x memristor and Ta_2O_5-TaO_x memristor are described above. The switching behaviors of both types of memristors are shown. The bipolar resistive switching behaviors are attributed to the movement of oxygen vacancies inside the switching layer, though with different dynamics. In WO_x memristor, the drift of oxygen vacancies under electric filed and the spontaneous diffusion, equivalent to doping/undoping near the metal-insulator interface, result in the gradual resistance change (analog type switching) upon an applied stimulation and the decay behavior after the simulation is removed. In Ta_2O_5-TaO_x memristor, the electric field driven oxygen vacancy movement can form/rupture the filaments inside the insulating layer, resulting in a more abrupt resistance change (digital type switching). By restricting the filament modulation location, a complementary switching behavior can also be observed.

Reference

1. Chang, T., Jo, S.-H. & Lu, W. Short-Term Memory to Long-Term Memory Transition in a Nanoscale Memristor. *ACS Nano* **5,** 7669–7676 (2011).
2. Chang, T. *et al.* Synaptic behaviors and modeling of a metal oxide memristive device. *Appl. Phys. A Mater. Sci. Process.* **102,** 857–863 (2011).
3. Waser, R. & Aono, M. Nanoionics-based resistive switching memories. *Nat. Mater.* **6,** 833–840 (2007).

4. Strukov, D. B. & Williams, R. S. Exponential ionic drift: fast switching and low volatility of thin-film memristors. *Appl. Phys. A* **94,** 515–519 (2008).

5. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453,** 80–83 (2008).

6. Yang, J. J. *et al.* Memristive switching mechanism for metal/oxide/metal nanodevices. Supp. *Nat. Nanotechnol.* **3,** 429–433 (2008).

7. Lee, M.-J. *et al.* A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures. *Nat. Mater.* **10,** 625–630 (2011).

8. Kim, S., Choi, S. & Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. *ACS Nano* **8,** 2369–2376 (2014).

9. Linn, E., Rosezin, R., Kügeler, C. & Waser, R. Complementary resistive switches for passive nanocrossbar memories. *Nat. Mater.* **9,** 403–406 (2010).

10. Yang, Y., Sheridan, P. & Lu, W. Complementary resistive switching in tantalum oxide-based resistive memory devices. *Appl. Phys. Lett.* **100,** 1–5 (2012).

Chapter 3

Second Order Memristor Dynamics

3.1 Basic Memristor Model

Following the memristor theoretical framework[1], a typical memristive device can be described by the following equations:

$$I(t, V) = V * G(V, w) \tag{3-1}$$

$$\frac{dw(t, V)}{dt} = F(V, w) \tag{3-2}$$

Here Equation (3-1) is the current-voltage equation, and the state of the device (conductance G) is determined by the internal state variable *w*; while Equation (3-2) is the dynamics equation of the state variable *w*, and highlights the fact that the present input cannot deterministically control *w* – rather it can only control the change rate of *w*. As a result, *w*, hence the device state, is controlled by the cumulative effects of the inputs, leading to the history-dependence of the device.

In models proposed previously, for example the WO_x memristor model in Chapter 2, the conductance/resistance is always expressed by an equation containing one state variable *w*. This variable normally represents a physical element that will be changed during operation and determines the device state, such as the effective conductive region area, or tunneling gap length and conduction filament width. This state variable directly determines the resistance, while the dynamics equation is usually a first order differential equation. This kind of memristors can be categorized as 1st order memristors. The dynamics equation, or the change rate of the state variable *w*, can be re-written as:

$$\frac{dw(t, V)}{dt} = F(w(t), V) \tag{3-3}$$

In these 1st order memristors, the state variable, *w*, directly modulating the device conductance (through Equation (3-1)) is in turn directly modulated by the inputs (through Equation (3-3)).

3.2 Second Order Memristor Model

In biology and other nonlinear systems, the state variable that determines the output is typically different from the state variable that directly responds to the input, leading to much richer behaviors. In the memristor theoretical framework, these devices can be described as 2nd order memristors[2]. In a 2nd order memristor, there are more than one state variables: the 1st order state variable directly determines the conductance and is usually a physical element related to the conduction structure in the switching layer, while the 2nd order state variable may not directly control the device conductance but still responds to stimulation and more importantly it will regulate the dynamic process of the 1st order state variable during and after stimulations.

The dynamic equations for a 2nd order memristor can be written as:

$$I(t, V) = V * G(V, w) \quad (3\text{-}4)$$

$$\frac{dw(t,V)}{dt} = F(w, s, V) \quad (3\text{-}5)$$

$$\frac{ds(t,V)}{dt} = Y(s, V) \quad (3\text{-}6)$$

Here *s* is the 2nd order state variable and it only indirectly determines the device conductance. As will be shown in more detail below, the dynamics of the 2nd order stable variable (Equation (3-6)) can strongly modulate the dynamics of the 1st order state variable (and hence the device conductance modulation) through Equation (3-5), allowing the bio-realistic implementation of synaptic functions naturally, which can be hardly achieved in simple 1st order memristors. Understanding and utilizing the different dynamics in 2nd order memristors thus provide an elegant path towards bio-inspired neuromorphic hardware.

3.3 WO_x (volatile) Memristor as 2nd Order Memristor

3.3.1 From Measurement Results to Modeling

The volatile behavior in WO_x memristors were systematically studied. In particular, a two-stage decay was observed, leading to the development of a 2nd order memristor model with the oxygen vacancy mobility as the 2nd order state variable[3].

3.3.1.1 Two-Stage Decay

To study the dynamics of the WO_x memristor, we first analyzed the temporal decay characteristics of the device. In one experiment, ten positive write pulses (1.2 V, 1 ms) at 5 ms interval were applied first, followed by small read pulses (0.4 V, 1 ms) to track the memristor conductance change as shown below.

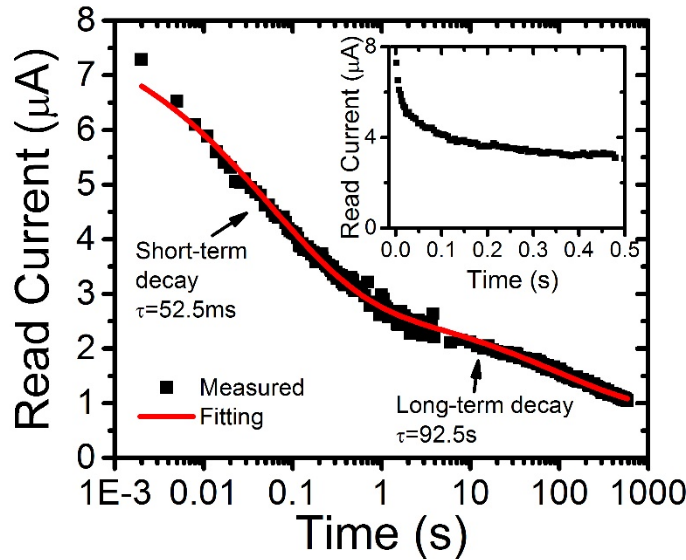

Figure 3-1: Read current decay of WO_x memristor after stimulation.

The stimulation drives the memristor conductance higher (as represented by the higher read current). However after stimulation is stopped the memristor conductance decays. A clear read current decay could be observed, shown in Figure 3-1. This decay is likely due to the spontaneous diffusion of V_Os[4–8]. Significantly, careful analysis of the data shows that the decay appears to occur at two very different time scales: right after stimulation (within ~200 ms after

stimulation is removed), the read current shows a very fast decay and the decay becomes much slower after a few hundreds of milliseconds (up to 1000 s), as shown in Figure 3-1 in both log and linear time scales.

Specifically, the data can be well fitted with stretched exponential functions having two time constants: a short-term time constant of ≈52.5 ms and a long-term time constant of ≈92.5 s

$$I = A_1 * I_{short} + A_2 * I_{long} = A_1 I_{0s} exp\left[-\left(\frac{t}{\tau_s}\right)^{\beta_s}\right] + A_2 I_{0l} exp\left[-\left(\frac{t}{\tau_l}\right)^{\beta_l}\right] \qquad (3\text{-}7)$$

Here, stretched exponential functions that describe the relaxation in a disordered system are used to model both the short-term and the long-term decays, where τ_s (τ_l), I_{0s} (I_{0l}), β_s (β_l) are the characteristic relaxation time, prefactor, and the stretch index for the short-term (long-term term) process, respectively. Experimentally, this behavior can be explained by the fact that the state variable directly governing memristor conductance (w_c) is affected by how mobile the oxygen vacancies are. It has been found that the mobility of oxygen vacancies (ions) increases when they are driven out of equilibrium right after a stimulation pulse, possibly due to the local lattice distortion and strain, followed by slow relaxation after certain period[8]. The temporary higher mobility, represented by another state variable w_m in our model, may explain the initial fast decay of the memristor conductance and also affect how the conductance determining state variable w_c responds to stimulations.

Physically, the migration of oxygen vacancies is driven by electrochemical gradients[9], including the field-driven drift process by an electrical potential gradient and the diffusion process driven by an internal chemical potential gradient. Additional factors (e.g., protons provided by moisture, local morphology change, etc.) can in turn affect the dynamics of such processes. The formation of electrochemical potential gradients in both electrochemical metallization memory (ECM) and valence change memory (VCM, to which the WO_x memristor belongs), and the relaxation that lead to the experimentally observed nanobattery effect[9] have been extensively discussed previously[9].

Borrowing terms used in neuroscience, the first stage with time constant ≈52.5 ms is considered short-term and the second stage with time constant ≈92.5 s (i.e., >1000x longer) is considered long-term. We note that even though the absolute values of the short-term and long-term characteristic time constant are different from those reported in biological synapses (e.g.,

tens of milliseconds to a few minutes for short-term and minutes to hours for long-term), the separation of the two time scales that differ by more than three orders of magnitude is evident in the memristor, and that circuits based on memristors can potentially operate at higher clock frequency (e.g., kHz or higher compared to ~Hz in biological systems) to utilize the different dynamics in the two regimes.

3.3.1.2 Nonlinear Response: Saturation Effect

Another property of the memristor device is the nonlinear response to programming as shown in the figure below.

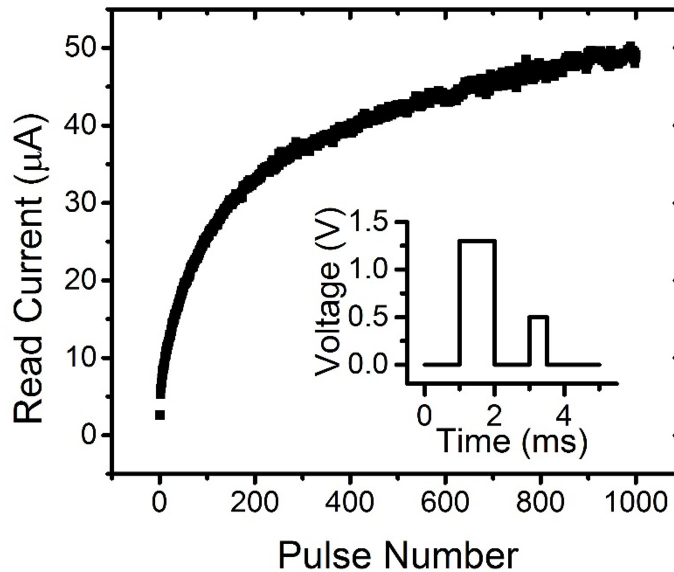

Figure 3-2: Conductance saturation by repeated stimulations. Inset: the programming waveform with a write pulse (1.3 V, 1 ms) followed by a read pulse (0.5 V, 500 us).

Here 1000 consecutive write pulses (1.3 V, 1 ms) were applied to the device at a repetition frequency of 50 Hz, and the device conductance was measured by a small read pulse (0.5 V, 500 μs) after each write pulse (as shown in the inset). The read current increases quickly following the first few write pulses but the rate of increase slows down as the device conductance increases with an increasing number of write pulses. The loss of programming capability at high conductance states has been hypothesized to be caused by the exhaustion of the supply of readily-available oxygen vacancies in the switching layer.

3.3.1.3 The Effect of Short-term Activity on Long-term Plasticity

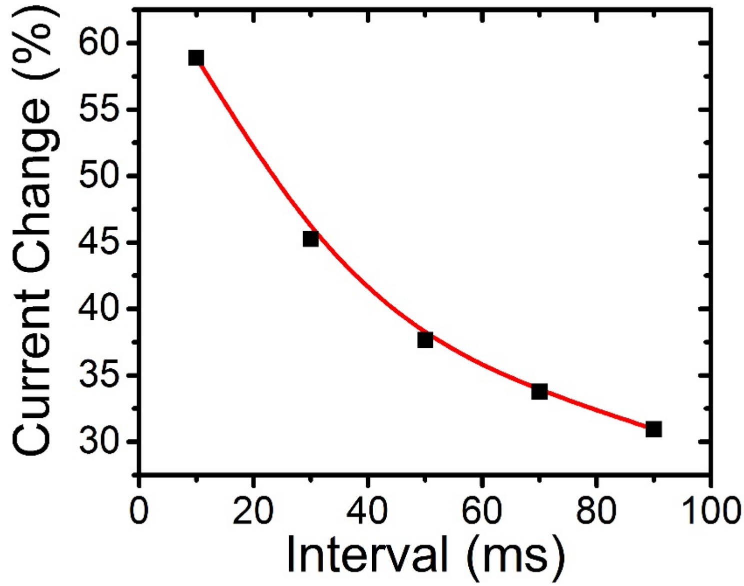

Figure 3-3: Memristor conductance change by programming pulse pairs with difference intervals. The change, represented by read current difference, was measured long (5 s) after the application of a pair of programming pulses.

The effects and interplay of the short- and long-term dynamics in the WO_x memristors were studied in several experiments. In one experiment, the long-term device conductance increase (measured 5 s after the application of the programming pulse pairs) was monitored as a function of the time interval between two programming pulses. As shown in Figure 3-3, the long-term plasticity shows a clear dependence on the short-term, intra-pair delay between the two pulses. Longer intervals led to smaller conductance increase, consistent with a short-term, vanishing effect induced by the first pulse, whose effect is experienced by the second pulse and leads to different long-term effects.

3.3.2 Equations Describing Memristor Dynamics

Considering these observed behaviors, the memristor can be modeled with two state variables derived from generic memristor to capture the ionic dynamics:

$$I = (1 - w_c) * \alpha * [1 - exp(-\beta V)] + w_c * \gamma * sinh(\eta V) \qquad (3\text{-}8)$$

$$\frac{dw_m}{dt} = \lambda_m M(w_m, V)\, sinh(\rho_m |V|) - \frac{w_m - w_{m0}}{\tau_m^*(w_m)} \qquad (3\text{-}9)$$

$$\frac{dw_c}{dt} = \lambda_c\, M(w_c, V) exp(\epsilon w_m)\, sinh(\rho_c V) - \frac{w_c - w_{c0}}{\tau_c^*(w_m)} \quad (3\text{-}10)$$

Here, Equation (3-8) is the current-voltage equation determined by the state variable w_c, which represents the effective area of the conductive region, as discussed in previous studies[4,10]. Equation (3-9) and (3-10) are the dynamic equations of the two state variables w_m and w_c in which the first term describes the effect of the stimulation voltage on oxygen vacancy mobility, while the second term describes the effect of decay with effective time constants (τ_m^* and τ_c^*). Specifically, the dynamics Equation (3-9) for state variable w_m, which represents the effective mobility of V_Os, shows that V_Os become more mobile with stimulation since more V_Os are driven out of equilibrium, and the mobility enhancement fades after stimulation is removed[8]. Additionally, w_m affects how w_c changes with stimulation through the factor $exp(\epsilon w_m)$ in Equation (3-10). The decay of w_c , which shows two time periods, may also be affected by the V_O mobility so the effective decay time constant τ_c^* is considered as a function of w_m too, as shown in Equation (3-12), to capture and reproduce the two-stage decay. In this sense, the memristor can be considered as a "second-order" memristor[2,11] as discussed at the beginning of this chapter. In this kind of memristor, the state variable w_c that directly controls the device current-voltage characteristics is modulated by another state variable w_m.In particular, even though the enhancement of w_m is mostly short-term with an effective time constant of only tens of milliseconds, it can (indirectly through Equation (3-10)) have long-term effects on the device conductance.

The value of the device-specific parameters and the choice of the "window function" $M(w, V)$ that reflects the nonlinear state-dependent programming capability and the effective time constant functions $\tau_m^*(w_m)$ and $\tau_c^*(w_m)$ are listed below.

$$\frac{1}{\tau_m^*(w_m)} = \frac{1}{\tau_s} \quad (3\text{-}11)$$

$$\frac{1}{\tau_c^*(w_m)} = \frac{1}{\tau_l} + \frac{\sigma \cdot w_m}{\tau_s} \quad (3\text{-}12)$$

$$M(w, V) = \begin{cases} 1 - \exp\left(-\frac{w_{max} - w}{0.0001}\right) & if\ V \geq 0 \\ 1 - \exp\left(-\frac{w - w_{min}}{0.0001}\right) & if\ V < 0 \end{cases} \quad (3\text{-}13)$$

Parameter	Value	Parameter	Value
α	1.5e-6	**β**	4
γ	3.2e-6	**κ**	5
λ_m	1e-6	**λ_c**	1e-6
ρ_m	15.5±2.5	**ρ_c**	14±2
w_{c0}	0-0.3	**W_{m0}**	0.001
τ_s	0.0025	**τ_l**	298
σ	0.25	**ϵ**	15

Table 3-1: The values of parameters used in the simulation.

3.3.3 SPICE Code

The memristor model discussed here can be directly implemented in a circuit simulator (SPICE). Note that in Table 3-1 the values of the two time constants are different from the fitting results because in the simulation normal exponential decay form is used while in the data fitting, a stretched exponential form is used.

The SPICE code for memristor model and simulation are shown below.

```
****************** LTspice code for metal oxide memristors*****************
*Parameters:
*alpha is prefactor for Schottky barrier
*beta is exponent prefactor for Schottky barrier
*gamma is prefactor for tunneling
*delta is exponent prefactor for tunneling
 **************************************************************************
```

```
.SUBCKT memristor 1 2 params:

+ alpha=1.5e-6 beta=4 gamma=3.2e-6 delta=5 wmax=1 wmin=0

*State variable:

.param lambda=1e-6 rhoc=14.5 rhom=17 taul=298 taus=0.0025 epsilon=15 sigma=0.25

.param cc={1}

.param cm={1}

Cpvar1 c 0 {cc}

Cpvar2 m 0 {cm}

*rate equation considering the diffusion effect

Gc 0 c value={trunc1(V(1,2),cc*V(c))*(lambda*exp(epsilon*cc*V(c))*sinh(rhoc*V(1,2)))-
(cc*V(c)-0.001)*(1/taul+sigma*cm*V(m)/taus)}

Gm 0 m value={trunc2(V(1,2),cm*V(m))*(lambda*sinh(rhom*abs(V(1,2))))-(cm*V(m)-
0.001)*(cm*V(m)/taus)}

.ic V(c) = 0.001

.ic V(m) = 0.001

*******************************************************

*auxiliary functions to limit the range of w

.func sign2(var) {(sgn(var)+1)/2}

.func trunc1(var1,var2) {sign2(var1)*sign2(wmax-var2)*(1-exp(-(wmax-var2)/0.0001))+sign2(-
var1)*sign2(var2-wmin)*(1-exp(-(var2-wmin)/0.0001))}

.func trunc2(var1,var2) {sign2(var1)*sign2(wmax-var2)*(1-exp(-(wmax-var2)/0.0001))+sign2(-
var1)*sign2(var2-wmin)*(1-exp(-(var2-wmin)/0.0001))}

***********************************************************************

*Output:
```

```
Gw 1 2 value={(1-cc*V(c))*alpha*(1-exp(-beta*V(1,2)))+(cc*V(c))*gamma*sinh(delta*V(1,2))}
.ENDS memristor
```

3.3.4 Simulation Results

This memristor model based on two state variables can quantitatively capture the ionic dynamics and describe the device response over a large range of programming conditions. Below we summarize simulation results obtained from the 2nd order memristor model using parameters listed in Table 3-1.

First, we show that the two-stage decay phenomenon can be well-reproduced through simulation, as shown below.

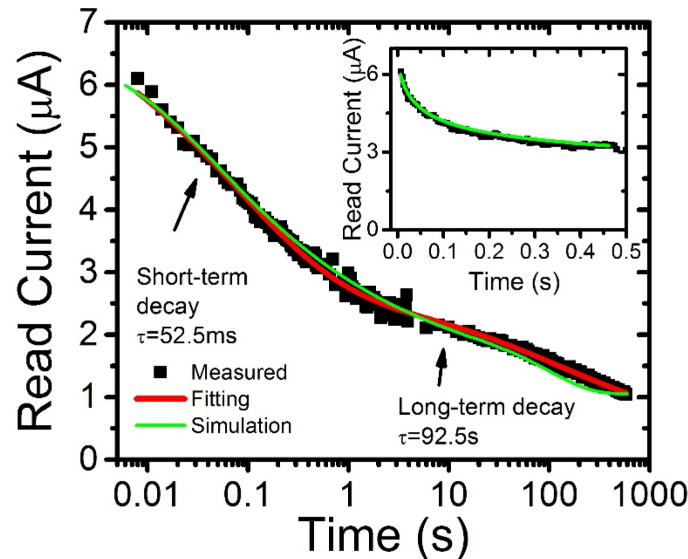

Figure 3-4: Simulation of the two-stage decay using the 2nd order WO_x memristor model.

Both AC and DC dynamic responses of the device can be also captured, as shown below.

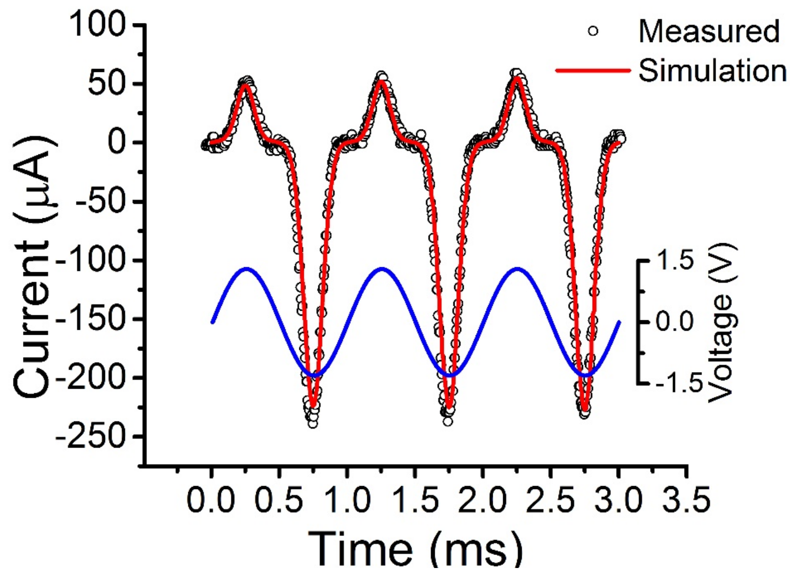

Figure 3-5: AC sweep experimental and simulation results. Inset: AC sweep waveform applied to the device.

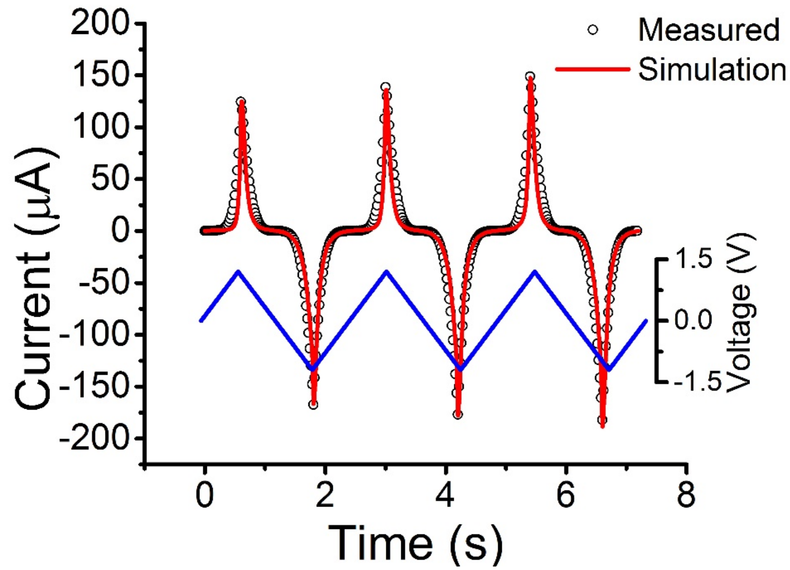

Figure 3-6: DC sweep experimental and simulation results. Inset: DC sweep waveform applied to the device.

The effect of short-term dynamics on long-term plasticity is also captured by the model, shown in Figure 3-7 and Figure 3-8. A large intra-pair interval between two pulses will lead to the weakening of the enhancement effect on long-term plasticity induced by the short-term dynamics after the first pulse, thus reducing the overall conductance increase after the second pulse.

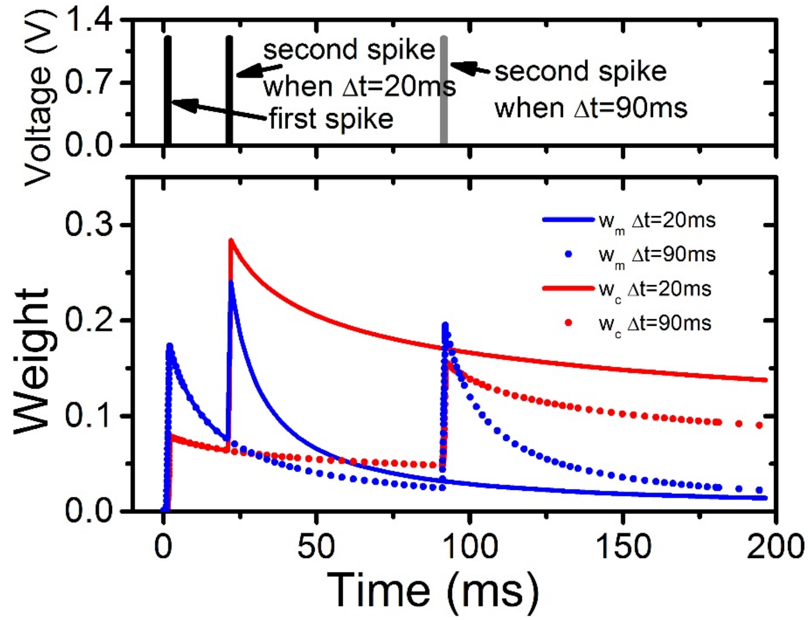

Figure 3-7: Simulation results showing the evolution of short-term and long-term state variables. After the stimulation (the pulse pair) is removed, they will decay with different time constants and be affected by the intra-pair interval between two pusles.

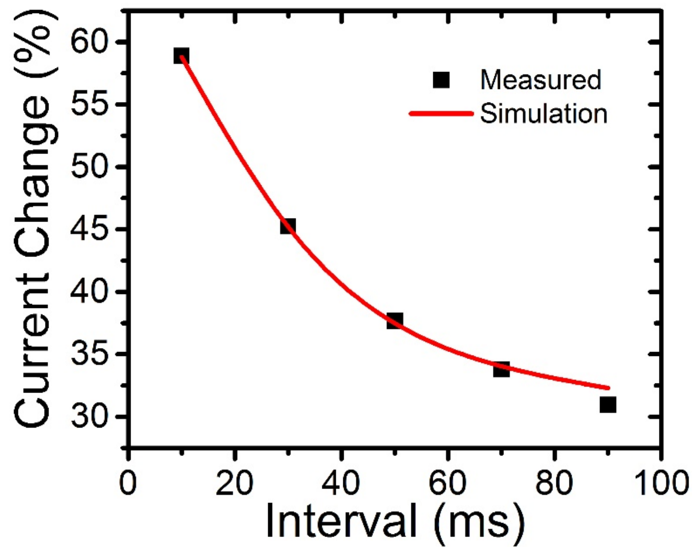

Figure 3-8: Simulation results showing the effect of short-term dynamics on long-term plasticity.

The success of the 2nd order memristor model to describe the device dynamics allows design and implementation of rate- and timing-dependent plasticity in memristor-based hardware in a bio-realistic fashion, as will be discussed below.

3.4 Ta_2O_5-TaO_x (nonvolatile) Memristor as 2nd Order Memristor

The 2nd order memristor effect is not restricted to volatile memristor effects. For example, for the commonly studied Ta_2O_5-TaO_x-based memristor, we found that the temperature can be used as the 2nd order state variable that can in turn modulate the first state variable (conduction filament geometry)[11].

Specifically, when a voltage pulse is applied then removed, the local temperature inside the switching layer, especially at the conduction filament gap, will increase then spontaneously decay therefore affecting the programming ability of stimulation afterwards.

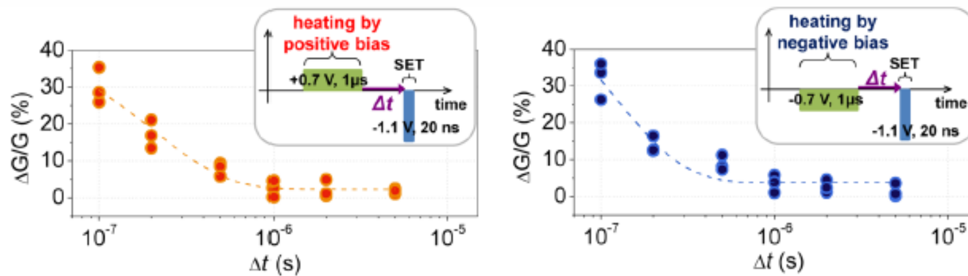

Figure 3-9: Temporal effect of heat on weight change in Ta_2O_5-TaO_x memristor. The elevated temperature by a heating pulse will affect the effectiveness of weight change caused by the subsequent programming pulse, depending on the interval between the two pulses.

The 2nd order effect caused by internal heating was verified by a series of carefully designed experiments. In a typical experiment, a pulse pair is designed as the stimulation which consists of two pulses: the first pulse (+0.7 V, 1 μs) has a small amplitude well below both the RESET and the SET threshold voltage so the device conductance could not be changed. However, it has a long pulse width so that it lasts long enough to cause sufficient heat accumulation and raise the internal temperature in the switching layer, especially in the conduction filament gap. This pulse is termed the "heating pulse". The other pulse in the pulse pair (-1.1 V, 20 ns), the so-called "programming pulse", has a larger amplitude than the SET threshold voltage, but the width is extremely short so it alone is not sufficient to change the device conductance. When these two pulses are applied consecutively and the intra-pair interval is short enough, there could be significant conductance change as shown in Figure 3-9. As the interval getting shorter, especially around hundreds of nanoseconds, the effect becomes more significant. Additionally, substituting the heating pulse with a pulse with negative amplitude (-

0.7 V, 1 μs) leads to similar conductance changes, consistent with the fact that the effect from the first pulse is Joule heating and not polarity-dependent.

Overall these results show that the internal temperature, increased by the first heating pulse and decays after the pulse removal, affects the effects of the subsequent programming pulse. The strong dependence of the conductance change on the intra-pair interval suggests the elevated temperature decays very fast, probably due to heat dissipation through metal top electrode, with a time constant around hundreds of nanoseconds.

Similar to the WO_x memristor case, the experimental results obtained from the Ta_2O_5-TaO_x memristor can be explained by a 2nd order memristor model using the local temperature as the 2nd order state variable.

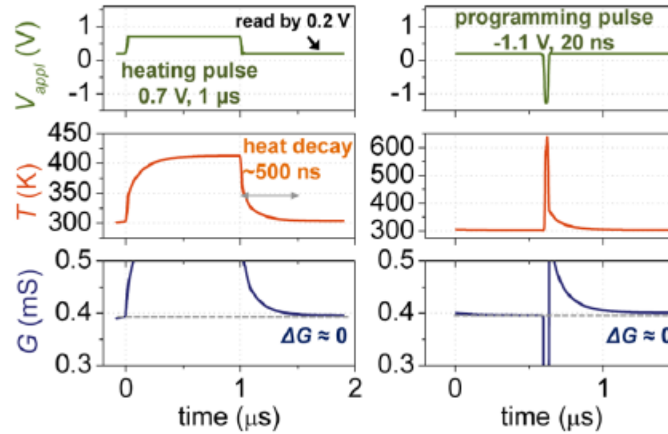

Figure 3-10: Simulated transient temperature evolution of the device (measured at the edge of filament) during and after the application of only heating pulse or programming pulse.

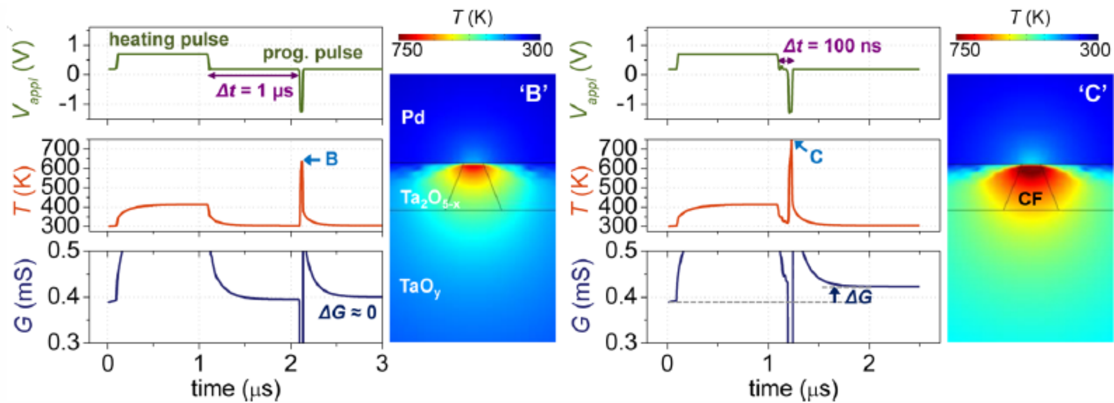

Figure 3-11: Simulated transient temperature evolution during and after the application of a heating pulse followed by a programming pulse with the internal of 1 μs (left) and 100 ns (right).

The simulation results verify the heat accumulation when adequate pulse is applied and the heat dissipation in a very fast but still noticeable speed. When a heating pulse is applied, Joule heating is generated at the conduction filaments and the temperature rises to more than 400 K. After the pulse is removed the temperature decays with a time constant around 500 ns, as shown in Figure 3-10.The relatively long decay time is due to the relatively low thermal conductivity values of the oxide and metal electrode used, which allows for slower heat dissipation and sheds light upon future device performance improvement. Addtionally, the SiO_2 substrate acts as a heat insulator which further slows down heat dissipation. The conductance is not changed due to the low amplitude of the heating pulse. If a progamming pulse is applied alone, little conductance change is observed due to the extremely short pulse width. However if the programming pulse is applied shortly after the heating pulse, higher temperature is obtained (e.g. $T > 750$ K when $\Delta t = 100$ ns) during the second programming pulse since the elevated temperature created during heating pulse has not returned to its resting value. The elevated temperature then enables fast V_O migration and results in detectable conductance change during the programming pulse, as shown in Figure 3-11(right). On the other hand when the intra-pair pulse interval is long, the tempertaure has decayed to the ambient level before the programming pulse and the internal temperature during the programming pulse is kept low, leading to negligible conductance change (Figure 3-11(left)).

The simulation and experimental results confirm the 2nd order memristor hypothesis, that is 1) Joule heating generated by the application of voltage pulses lead to a temporary temperature increase; 2) temperature exhibits short-term dynamics and will decay spontaneoulsy when the pulse is removed, with a decay time constant ~500 ns for our devices and 3) the temporal summation of the thermal effect can occur and lead to an elevated device temperature that is higher than produced by a single pulse alone, and subsequently enable device conductance modulation, the extent of which depends on the relative timing of the pulses.

We will utilize those second-order dynamics, both caused by enhanced oxygen vacancy mobility and elevated internal temperature, to implement synaptic functions in a bio-realistic fashion, as will be discussed in the next chapter.

3.5 Conclusion

Originated from basic memristor model, 2nd order switching models are proposed for both WO_x memristor and Ta_2O_5-TaO_x memristor to quantitively capture both the short-term and the long-term switching dynamics, which are observed in electrical measuremens. Apart from previously introduced 1st order state variable, which can be the effective conductive area, 2nd order state variables, i.e., oxygen vacancy mobility in WO_x memristor and internal temperature in Ta_2O_5-TaO_x memristor, are introduced to reflect the effect of short-term dynamics on long-term plasticity. Simulations based on those 2nd order models can reproduce the experimental results very well.

Reference

1. Chua, L. O. Memristor-the missing circuit element. *Circuit Theory, IEEE Trans.* **18,** 507–519 (1971).
2. Pershin, Y. V. & Di Ventra, M. Neuromorphic, Digital, and Quantum Computation With Memory Circuit Elements. *Proc. IEEE* **100,** 2071–2080 (2012).
3. Du, C., Ma, W., Chang, T., Sheridan, P. & Lu, W. D. Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics. *Adv. Funct. Mater.* **25,** 4290–4299 (2015).
4. Chang, T. *et al.* Synaptic behaviors and modeling of a metal oxide memristive device. *Appl. Phys. A Mater. Sci. Process.* **102,** 857–863 (2011).
5. Wang, Z. Q. *et al.* Synaptic Learning and Memory Functions Achieved Using Oxygen Ion Migration/Diffusion in an Amorphous InGaZnO Memristor. *Adv. Funct. Mater.* **22,** 2759–2765 (2012).
6. Yang, R. *et al.* On-Demand Nanodevice with Electrical and Neuromorphic Multifunction Realized by Local Ion Migration. *ACS Nanoano* **6,** 9515–9521 (2012).
7. Chen, Y. Y. *et al.* Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current. *Int. Electron Devices Meet. 2013* 10.1.1-10.1.4 (2013). doi:10.1109/IEDM.2013.6724598

8. Nian, Y. B., Strozier, J., Wu, N. J., Chen, X. & Ignatiev, A. Evidence for an Oxygen Diffusion Model for the Electric Pulse Induced Resistance Change Effect in Transition-Metal Oxides. *Phys. Rev. Lett.* **98,** 146403 (2007).

9. Valov, I. *et al.* Nanobatteries in redox-based resistive switches require extension of memristor theory. *Nat. Commun.* **4,** 1771 (2013).

10. Strachan, J. P. *et al.* State Dynamics and Modeling of Tantalum Oxide Memristors. *Electron Devices, IEEE Trans.* **60,** 2194–2202 (2013).

11. Kim, S. *et al.* Experimental Demonstration of a Second-Order Memristor and Its Ability to Biorealistically Implement Synaptic Plasticity. *Nano Lett.* **15,** 2203–2211 (2015).

Chapter 4

Bio-realistic Implementation of Synaptic Functions Using Internal Ionic Dynamics of Memristors

During systematic studies on biological synaptic behaviors, diverse synaptic plasticity effects, including rate- and timing-based synaptic functions have been observed[1–3]. Neuroscientists have been trying to reconcile the experimental results and the different models created to explain specific synaptic functions. Among those, a unified model based on calcium ion (Ca^{2+}) concentration dynamics was shown to offer plausible explanation at the molecular level and was able to account for many different synaptic behaviors[4,5]. In this model, as has been confirmed in many previous biological studies, the calcium ion concentration in both the presynaptic neuron and the postsynaptic neuron will surge when an action potential arrives, then decreases following an exponential trend[1,6]. Specially the calcium dynamics is described by a first-order linear differential equation:

$$\frac{d[Ca(t)]}{dt} = I_{NMDA}(t) - \frac{[Ca(t)]}{\tau_{Ca}} \qquad (4\text{-}1)$$

The relative interval between action potentials then determines the cumulative calcium ion concentration, which in turn affects the activity of the receptors for neural transmitters (e.g. NMDAR) in the synapse that lead to synaptic plasticity.

We note that Equation (4-1) is similar to the equations describing the 2nd order state variable dynamics in metal oxide memristors. Specially, for WO_x memristor, the oxygen vacancy mobility decreases after the simulation is removed and can be modeled as an exponential decay. For Ta_2O_5-TaO_x memristor, the temperature decay by heat dissipation after stimulation is removed is also described as an exponential process. Similar to the biological synapse, the short-term dynamics of these 2nd order state variables in turn determines the dynamics of the 1st order

state variable (e.g. conduction filament geometry) which in turn determines the device conductance (synaptic weight). Therefore, by emulating Ca^{2+} dynamics in the 2nd order memristors, it becomes possible to mimic different synaptic functions naturally, in a bio-realistic fashion without having to manually adjust pulse parameters for different plasticity effects[7,8].

Below we discuss how the 2nd order memristors can be utilized to implement different synaptic functions, for both short-term effects and long-term effects, and for both rate- and timing-dependent plasticities.

4.1 Short-term Synaptic Behaviors

4.1.1 Paired-pulse Facilitation

Paired-pulse facilitation (PPF) is an important short-term phenomenon extensively discussed in neuroscience studies[1]. PPF states that when two excitatory presynaptic spikes are applied successively, the second spike will generate a larger excitatory postsynaptic current (EPSC) than the first pulse does. Additionally, the amplitude of EPSC caused by the second pulse is determined by the time interval between the two pulses and a larger interval will lead to a smaller EPSC amplitude enhancement as shown in Figure 4-1.

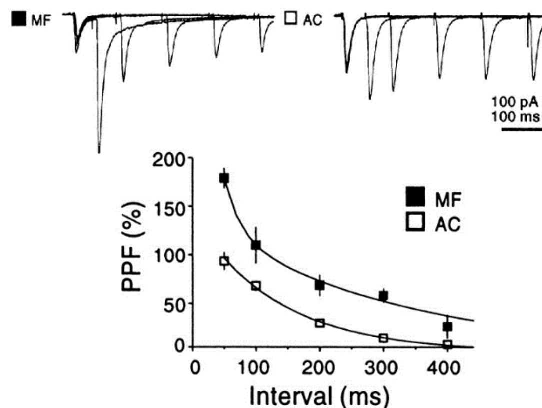

Figure 4-1: Results of PPF. The results were obtained from mossy fiber (MF) and assoc/com (AC) synapses[2]. Top: EPSC obtained from paired pulses with different intervals. The EPSC by the second pulse is enhanced and the enhancement is weaker with longer intervals between the pulses. Bottom: PPF ratio as a function of pulse interval. The ratio was defined as $(p_2-p_1)/p_1$, where p_1 and p_2 are the amplitude of the EPSCs evoked by the first and second pulse, respectively.

The PPF effect is believed to be caused by the residual Ca^{2+} concentration in the presynaptic neuron induced by the first spike which enhances the overall Ca^{2+} level and the

resulting EPSC generated by the second spike. Due to the exponential decay of the residual Ca^{2+} caused by the first spike, the effect naturally becomes weaker when the interval between the two spikes increases[1].

We obtain similar effects in the WO_x memristor when applying two identical, non-overlapping pulses as shown in Figure 4-2.

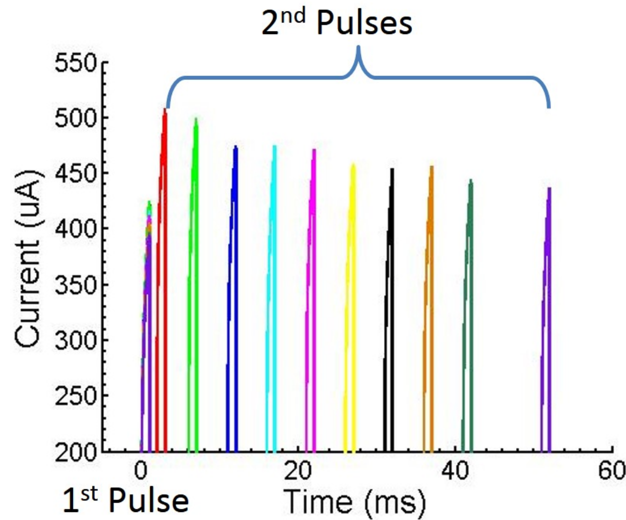

Figure 4-2: PPF effect obtained in WO_x memristor. Two paired pulses (1.4 V, 1 ms) are applied to the device at ten different intervals and the current during all pulses is recorded. The second pulse produces an enhanced response in all cases, and increasing the interval leads to a decrease in the enhancement.

First, we find that the peak current during the second pulse is indeed higher than that during the first pulse which is similar to the PPF effect. Analogous to the residual effects of Ca^{2+} in biological synapses, the enhancement in programming current (determined by the state variable w_c) observed in the second pulse can be explained by the residual effect of oxygen vacancy mobility enhancement (represented by the state variable w_m) from the first pulse. If the second pulse is applied before w_m has decayed to its resting value, w_m during the second pulse, which starts from a higher initial value when the pulse is applied, is larger than during the first pulse and so is w_c (and its increase), therefore an enhanced current spike will be obtained.

Second, with increasing interval between the two pulses, the enhancement in reduced. This can also be easily explained as w_m gradually decays toward its resting value so the difference of w_m during the two pulses becomes smaller with longer interval, leading to a smaller w_c (therefore the peak current) enhancement during the second pulse.

The PPF effect observed in memristor can be better illustrated by measuring the device conductance immediately after the first pulse (p_1) and the second pulse (p_2), and calculating the conductance change ratio as $(p_2-p_1)/p_1$, similar to the biological method, as shown in Figure 4-3.

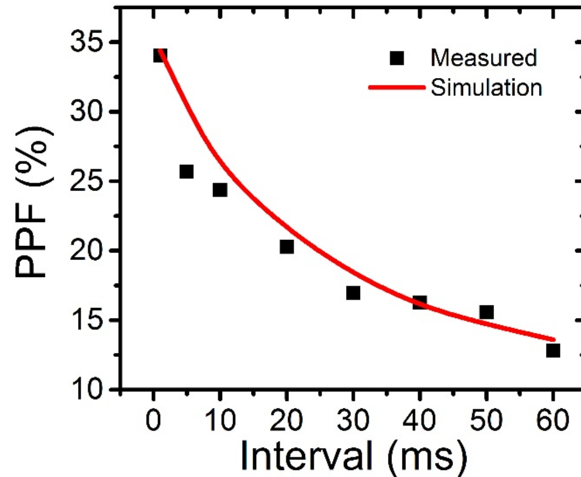

Figure 4-3: PPF ratio for different pulse intervals obtained from WO_x memristor. The results (squares: experimental data, line: simulation results by the second order WO_x memristor model) show similar trend to biological experimental results.

The dependence of the conductance enhancement on the pulse interval again shows a similar trend to that observed in biological synapses (Figure 4-1). A larger interval will lead to a smaller conductance enhancement and this could be directly explained from the perspective of the enhancement and decay of w_m as discussed above. Specifically, the change in device conductance as a function of pulse interval can be quantitatively explained through simulations based on the device model, shown as the solid line in Figure 4-3.

4.1.2 Frequency-dependent Weight Change

As an extension of PPF, if more than two excitatory presynaptic spikes are applied to the synapse, the amplitude of the resulting EPSC will continue increasing gradually. The extent of the synaptic weight change should depend on the frequency of stimuli, which is inversely related to the interval between each pulse. To verify that the internal dynamics can naturally lead to similar frequency-dependent weight change, we applied ten continuous write pulses (1.25 V, 1 ms) with different frequencies to WO_x memristor and monitored the current during each pulse and calculated the increase by comparing the current of the last pulse to that of the first pulse.

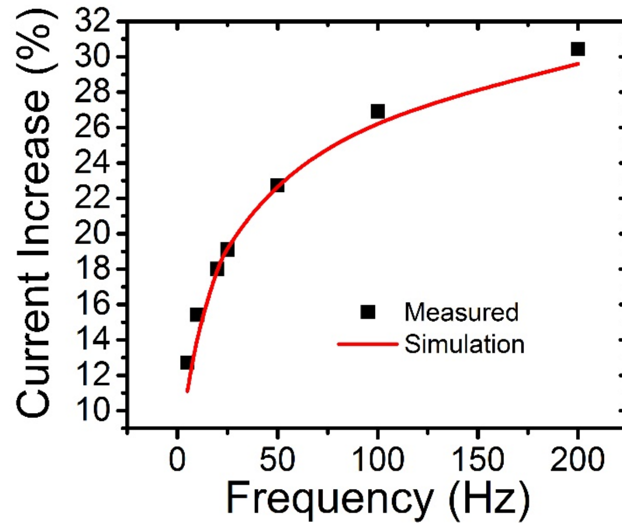

Figure 4-4: Frequency-dependent weight change in WO_x memristor. Ten write pulses (1.25 V, 1 ms) with different frequencies are applied to the device and the currents of the last and the first pulse are compared to calculate the increase. Higher stimulation frequency leads to larger conductance enhancement. Squares: experimental data. Line: simulation results from the second order WO_x memristor model.

As shown in Figure 4-4, the device current is indeed found to increase gradually and more interestingly but not surprisingly, at different rates depending on the stimulation frequency. A clear trend in the potentiation effect with respect to the stimulation frequency can be observed: as the stimulation frequency increases, the increase in current is more significant. This frequency dependent weight change in WO_x memritor can be readily explained using the V_O dynamics following the residual Ca^{2+} concentration model, as pulse trains with higher frequency means smaller intervals between each pulse to allow w_m decay and result in more effective accumulation of the V_O (analogous to the more enhanced Ca^{2+} concentration in presynaptic neuron in biological synapses), as already explained in the PPF experiments. Again, the experimental data can be quantitatively explained by the second order WO_x memristor model (solid line in Figure 4-4).

4.2 Long-term Synaptic Behaviors

The synaptic behaviors discussed in Section 4.1 are mostly believed to be caused by the Ca^{2+} concentration surge and decay in the presynaptic neuron therefore the effects are short-term[1]. However the more important synaptic behaviors, relating to memory and functionality of

brain, are long-term[9]. The long-term synaptic plasticity is mostly believed to be related to the Ca^{2+} concentration change in the postsynaptic neuron and involve many complex dynamics. Here we show that some important long-term synaptic behaviors can be implemented by metal oxide memristors and more importantly, in a bio-realistic way.

4.2.1 Spike-timing Dependent Plasticity

Spike-timing dependent plasticity (STDP) is a very important synaptic behavior. In STDP, the synapse is subjected to repetitive pre- and postsynaptic spike trains, and the relative timing of the pre- and postsynaptic spikes determines whether the synaptic weight will be potentiated or depressed and by how much[3]. If presynaptic spike arrives before postsynaptic spike (pre-post pair), the pairs will cause potentiation while a reversed sequence (post-pre pair) will cause depression. Moreover, a larger time interval between the pre- and postsynaptic spikes will lead to smaller weight modification.

In previous studies, STDP was implemented by applying carefully designed, overlapped waveforms on memristors, such that the relative timing of the spikes was converted into the amount of overlap of the pre- and postsynaptic signals, allowing the device to respond accordingly[10], as shown in Figure 4-5.

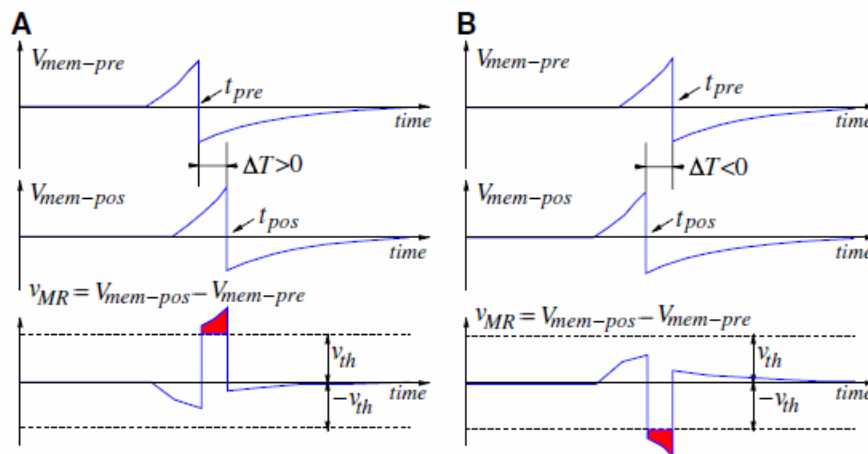

Figure 4-5: Waveforms with overlap for STDP. Pre- and postsynaptic membrane potential waveform for the situations of positive Δt (A) and negative Δt (B) where $\Delta t = t_{post} - t_{pre}$. Voltage V_{MR} is the difference between the postsynaptic membrane voltage $V_{mem-pos}$ and the presynaptic membrane voltage $V_{mem-pre}$. Overlap of two waveforms generates the effective programming signal (red area) which is above the threshold.

Instead of relying on external factors that keep the timing information, in neurobiology, the relative timing information between the pulses is natively embedded, e.g., by the natural decay of Ca^{2+} level which provides an internal timing mechanism[4,11]. Here we demonstrate that STDP behavior can be achieved in memristors with similarly simple, non-overlapping pre- and postsynaptic spike pairs. Similar to the case of biological synapses, STDP is achieved naturally since the relative timing information is encoded internally through the V_O dynamics of WO_x memristor and heat dynamics of Ta_2O_5-TaO_x memristor.

4.2.1.1 STDP Achieved by WO_x Memristor

The pulse pair we used contains a negative erase pulse (-1.1 V, 1 ms) representing the effect of a presynaptic spike and a positive write pulse (+1.1 V, 1 ms) representing that of a postsynaptic spike, both applied at the postsynaptic side. This configuration is equivalent to applying identical, positive pulse on both presynaptic and postsynaptic sides of the device as shown in Figure 4-6.

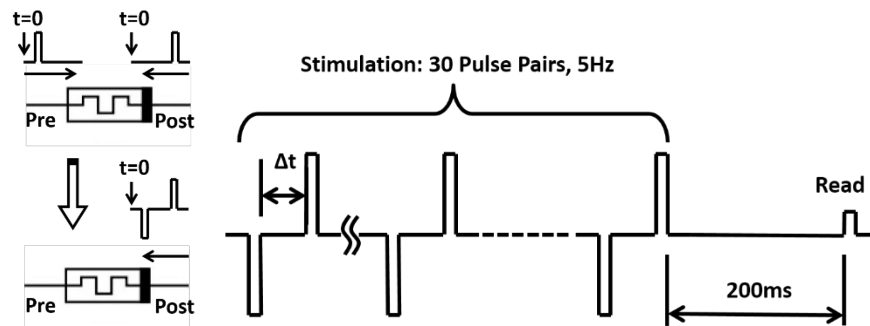

Figure 4-6: Stimulation protocol for STDP on WO_x memristor. Left: experimental setup, a pre-post pair consisting of identical spikes is equivalent to a negative/positive pulse pair applied on the postsynaptic side. Right: the pre-post programming protocol including 30 pulse pairs (-1.1 V, 1 ms/1.1 V, 1 ms) applied at 5 Hz for stimulation, followed by a read pulse 200 ms after stimulation. Post-pre pairs are applied similarly.

Before each test, the device was stimulated with the same pulse train consisting of ten positive pulses (1.2 V, 1 ms, 200 Hz). In each test, 30 pulse pairs of either positive-negative pulse pair (+/- pair, representing the post-pre spike condition) or negative-positive (-/+ pair, representing the pre-post spike condition) were then applied at 5 Hz repetition frequency, as shown in Figure 4-6 (right). The device conductance was measured 0.2 s after the last pair and compared to that of the reference value, which was measured at identical time but without

applying pre-post or post-pre pairs. The device was then brought back to the same starting condition and the experiment was repeated for different pulse pair configurations.

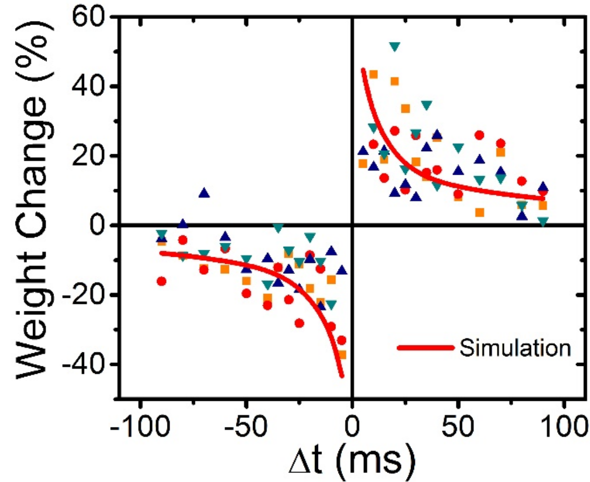

Figure 4-7: STDP implemented in WO_x memristor. The weight of memristor changes as a function of the relative timing between the pre- and postsynaptic pulses. $\Delta t = t_{post} - t_{pre}$. Symbols: experimental results obtained from four different tests. Solid lines: simulation data from the second order WO_x memristor model.

As shown in Figure 4-7, for pre-post condition ($\Delta t > 0$), the memristor conductance (weight) increases while for post-pre condition ($\Delta t < 0$), the memristor conductance decreases. In other words, even though symmetric presynaptic (negative) and postsynaptic (positive) pulses were applied with identical amplitude and pulse width, their effects do not cancel each other and the net effect is found to be strongly dominated by the effect of the second pulse. This observation can be understood again with the second order WO_x memristor model, as illustrated in Figure 4-8.

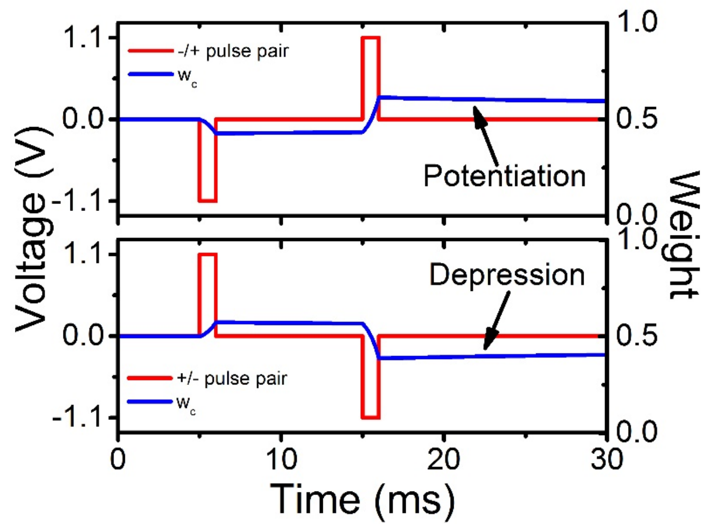

Figure 4-8: Illustration of weight change under different pulse pair timing by simulation based on WO_x memristor model. The dynamics of only w_c within one pulse pair is shown for clarity. The second pulse in the pair has a large effect on w_c change due to residue enhanced w_m from the first pulse, and can cause either potentiation or depression depending on the relative timing between the pre- and postsynaptic pulses.

Here the state variable w_c after a negatvie-positive pulse pair shows a net increase, since the write effect of the second pulse will be stronger due to the larger residue value of w_m, which is enhanced by the first pulse (Figure 4-8, upper panel). Therefore, the negative-positive pair leads to a net increase of the state variable w_c and a net increase of device conductance.

Moreover, since the enhancement effect of the second pulse is caused by the residue of the increased state variable w_m, the amplitude of the enhancement is dependent on the relative timing (Δt) of the first (which enhances w_m) and the second pulse (which utilizes this enhancement) inside the pulse pair as shown in Figure 4-9.

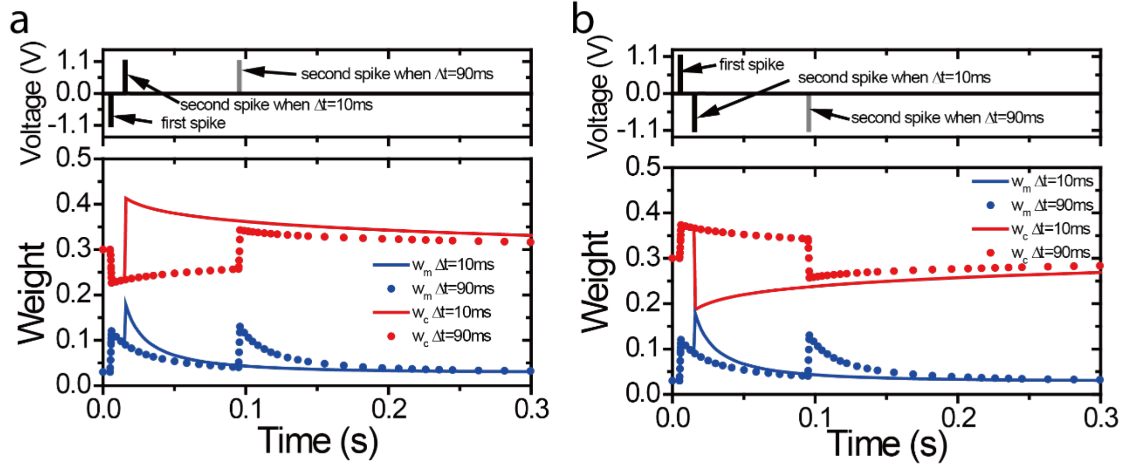

Figure 4-9: Illustration of state variable dynamics for different Δt by simulation based on WO_x memristor model. a) The case of pre-post pulse pair. b) The case of post-pre pulse pair. With longer interval Δt, w_m decays to a smaller value, leading to less significant w_c change by the second pulse.

After the first pulse, the enhanced state variable w_m decays following a characteristic time constant around tens of milliseconds, so a larger Δt between the two pulses leads to a smaller residual w_m and subsequently a smaller change in the state variable w_c during the second pulse, and a smaller measured conductance change. As a result, the accumulating net weight change of 30 pulse pairs, which is dominated by the second pulse in each pair, shows an inverse relationship with Δt. Indeed, an effect analogous to STDP was clearly observed, with larger relative timing between two pulses (larger Δt) resulting in smaller conductance change[3] and vice versa, as shown in Figure 4-7.

Here again the state variable w_m plays the role of the (postsynaptic) Ca^{2+} concentration and provides an intrinsic timing mechanism, and in turn affects the plasticity of the weight-determining state variable w_c.

Similar to other experiments discussed earlier, the second order memristor model can quantitatively capture the STDP behavior (Figure 4-7) and explain the mechanisms from known physical processes with controlled internal dynamics (Figure 4-8, Figure 4-9).

4.2.1.2 STDP Achieved by Ta_2O_5-TaO_x Memristor

Similarly, STDP can be achieved in Ta_2O_5-TaO_x memristor since the short-term temperature dynamics can also provide an intrinsic timing mechanism and enables timing-

dependent long-term weight changes. The spikes used to implement STDP are shown in Figure 4-10(a), (b), where the presynaptic (postsynaptic) spike consists of a programming element: 1.6 V (1.1 V) with 20 ns duration and a heating element: 0.7 V with 1 μs duration, applied to the TE (BE) respectively. Equivalently, the postsynaptic spike corresponds to a programming element of -1.1 V with 20 ns duration and a heating element of -0.7 V with 1 μs applied to the TE, as shown in Figure 4-10(c).

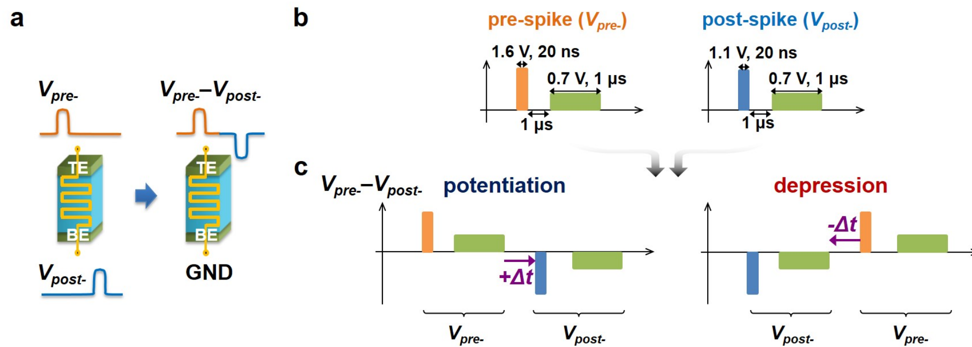

Figure 4-10: Stimulation protocol for STDP on Ta_2O_5- TaO_x memristor. a) Experimental setup: a pair of spikes (V_{pre} -V_{post}) applied to TE is equivalent to a pair of spikes V_{pre} and V_{post} applied to the pre- and postsynaptic side, respectively. b) Pre- and postsynaptic spikes used for STDP implementation. Each spike consists of a programming pulse (1.6 V, 20 ns for presynaptic spike and 1.1 V, 20 ns for postsynaptic spike) and a heating pulse (0.7 V, 1 μs). c) Equivalent pulses applied to the TE of the device, for $\Delta t > 0$ and $\Delta t < 0$.

Similar to earlier discussions, each programming and heating element alone cannot modulate the device conductance. However, when the presynaptic spike reaches the device earlier than the postsynaptic spike (i.e., in the case of $\Delta t > 0$), the postsynaptic spike will be effected by the temporal heating effect from the presynaptic spike. The elevated temperature during the second spike causes the effect of second spike (the postsynaptic spike in this case, which with a negative programming pulse will increase the device conductance) to be stronger than that of the first spike (the presynaptic spike, which with a positive programming pulse will decrease the device conductance) and thus the overall effect will be dominated by the second spike and an overall increase in device conductance (potentiation) is obtained. In the opposite case ($\Delta t < 0$), a decrease in device conductance (depression) can be obtained by similar arguments. Indeed, STDP results measured in the Ta_2O_5-TaO_x memristor using the non-overlapped, spike-pairing protocols, where the conductance was measured with a small (0.2 V) read voltage after the application of the spikes, are shown in Figure 4-11.

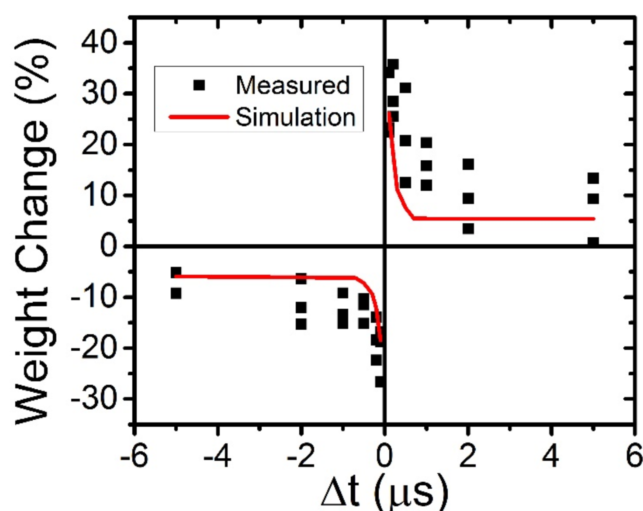

Figure 4-11: STDP implemented in Ta_2O_5-TaO_x memristor. The weight of memristor changes as a function of the relative timing between the pre- and postsynaptic spikes, i.e., Δt. Squares: experimental results. Solid lines: simulation results from the second order Ta_2O_5-TaO_x memristor model.

Significantly, an effect analogous to STDP observed in biology is clearly observed, with the sign of the conductance (weight) change determined by the sign of Δt and the amplitude of that change determined by the value of Δt, with larger relative timing between the pre- and postsynaptic spikes (larger Δt) resulting in smaller conductance change and vice versa.

Here again, the internal temperature plays a similar role of the (postsynaptic) Ca^{2+} concentration and provides an intrinsic mechanism to encode the spike timing and activity information, which in turn causes the first order state variable, which determines the device conductance, to change accordingly. Similar to the WO_x memristor case, the experiment results can be quantitatively captured by the second order model, shown as red curves in Figure 4-11.

4.2.2 Frequency-dependent Long-term Weight Change

As discussed above, we have demonstrated that the Ca^{2+}-like short-term dynamics of the 2nd order state variable can have strong effects on the modulation of the 1st order (determining synaptic weight) state variable, leading to (long term) synaptic plasticity effects that are controlled by short-term temporal properties, e.g. spiking timing in STDP experiments.

One more example is the frequency-dependent plasticity observed in the Ta_2O_5-TaO_x memristor. As has been discussed previously, the rise and decay of the local temperature offers

an intrinsic timing mechanism, and different (short-term) temporal patterns of the input spikes can lead to different long-term conductance changes. To test this concept, we applied a series of stimuli with different frequencies to the device. Each stimulus is a pulse pair consisting of a large but narrow programming pulse (-1.1 V, 20 ns) and a small but long heating pulse (-0.7 V, 1 μs), and is considered together as a single spike here as shown in the top panel in Figure 4-12. The frequency is determined by controlling the inter-spike time interval Δt.

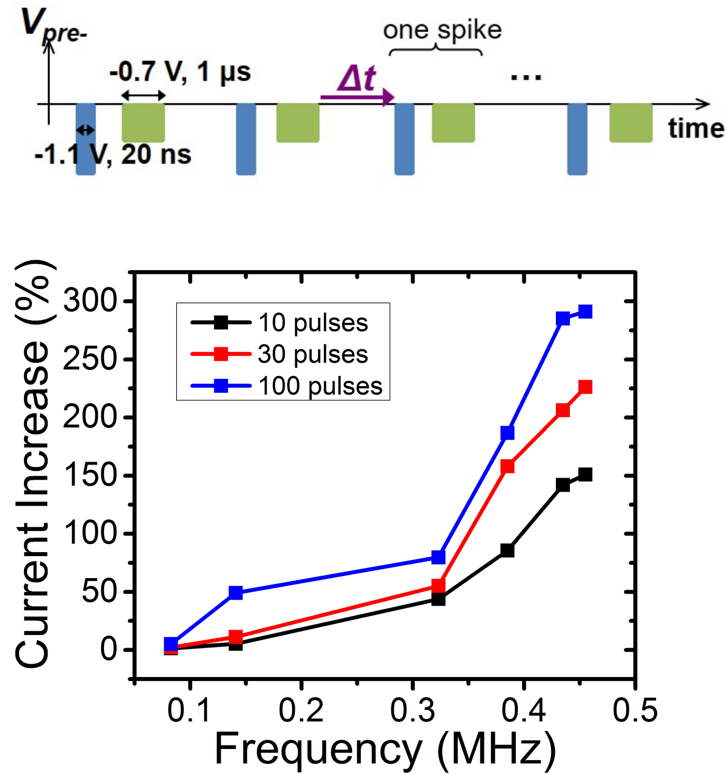

Figure 4-12: Frequency-dependent long-term weight change of Ta_2O_5-TaO_x memristor. Top: the spikes applied. Each spike contains a programming pulse (-1.1 V, 20 ns) and a heating pulse (-0.7 V, 1 us). The interval Δt between the spikes are tuned to obtain different frequencies. Bottom: conductance (represented by the current measured by read pulses) change by stimulations with different frequencies and spike numbers. Higher frequency and more pulses lead to stronger weight increase.

At each frequency, an increase in device conductance, i.e. potentiation, is generally observed as the number of spikes is increased. More importantly, when the stimulation frequency is low (e.g. 0.14 MHz for Δt=5 μs), the potentiation effect is weak and no potentiation is observed for frequency below 0.1 MHz (Δt=8 μs), while higher stimulation frequencies (e.g. 0.43 MHz for Δt=200 ns and 0.45 MHz for Δt=100 ns) lead to stronger potentiation, as shown in Figure 4-12. This obvious frequency dependence of the conductance modulation can be readily explained by the heat accumulation/dissipation dynamics based on second order memristor

model. Since spike trains with higher frequency means shorter interval between spikes, more pronounced heat summation and higher device temperature, caused by the heating pulse in one spike, is experienced during the following programming pulse of the next spike, leading to a more significant conductance change.

4.3 Metaplasticity

As has been discussed and demonstrated above, neural activity can generate persistent forms of synaptic plasticity, such as long-term potentiation (LTP) and long-term depression (LTD), which are used to retain and process information in activated networks of neurons. However, there must exist some mechanisms to prevent the saturation of LTP or LTD. Apart from various intercellular signaling molecules directly regulating the degree of LTP and LTD[12], there exists a different regulation that persists across time. Here stimulations at certain time can affect neurons or synapses such that their ability to exhibit LTP or LTD is changed after a later bout of activity. This form of plasticity regulation, i.e. plasticity of plasticity, is termed metaplasticity[13].

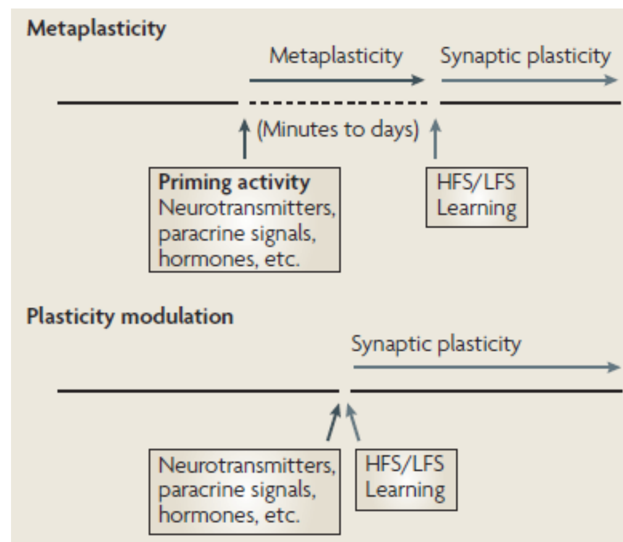

Figure 4-13: The standard paradigm of metaplasticity. An episode of priming activity at one point in time is applied and no weight change is observed. Then a subsequent event evokes synaptic plasticity. A change in neural function as a result of priming alters the response to the subsequent plasticity-inducing event.

As shown in Figure 4-13, the metaplasticity states that the historical activity, or termed priming activity, may induce synaptic weight change or may not, but there must be some change in neural function as a result of the priming that persists after the termination or washout of the priming stimulus that alters the response to a subsequent plasticity-inducing event. For example, weak prior stimulation (only leading to short-term potentiation, STP) that does not cause a long-lasting change in the synaptic efficacy was found to inhibit subsequent induction of LTP in the CA1 region of the hippocampus[14]. Therefore, the 'meta' part reflects a higher order of plasticity, that is, the plasticity of plasticity. Metaplasticity is important to the learning and memory of the brain, since it functions as an internal modulator that dynamically regulates the synaptic plasticity according to its previous activity and maintains the synaptic efficacy within a range. Studying the metaplasticity in memristor based artificial neuromorphic system can help explore how the network adaptively evolves with activities.

Steady implementation of metaplasticity in memristor requires the existence of a state variable that can be adjusted not only by the historical activity but also determines the future exhibition of plasticity. Here we show that metaplasticity can be implemented by our metal oxide memristor.

4.3.1 Experience-dependent Plasticity in WO_x Memristor

An important instance of metaplasticity is the experience-dependent plasticity. For example, according to the theory of Bienenstock, Cooper and Munro (BCM)[15], the synapse can exhibit either potentiation (synaptic weight strengthening) or depression (synaptic weight weakening) even when subjected to the same spike trains. In other words, not only the amplitude but also the sign of synaptic weight change depends on the present stimulation conditions as well as the stimulation history. Specifically, Bear et al. found that high frequency stimulation normally leads to potentiation and low frequency stimulation normally leads to depression, and there exists a threshold frequency at which the synaptic weight can be maintained[16]. Additionally, the threshold frequency will also shift accordingly, depending on the experience of the synaptic activities[16]. For example, after a period of increased synaptic activity, the threshold will slide to right (higher frequency), promoting synaptic depression such that spike trains that previously caused potentiation may now be below the threshold frequency and will cause

depression instead. Similarly, after a period of decreased activity, the threshold will slide to left, promoting synaptic potentiation and it will be easier to enhance synaptic weight with lower frequency spikes. The sliding threshold effect from Bear's study[16] on visual cortex is reproduced in Figure 4-14.

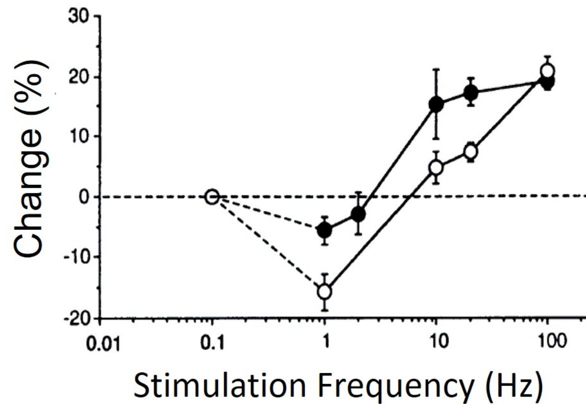

Figure 4-14: Experience-dependent synaptic weight change of biological synapse. The relative weight change as a function of stimulation frequency was obtained in rat visual cortex for two different cases. Low stimulation frequency results in depression and high stimulation frequency results in potentiation, and the threshold moves to lower frequency under light-deprived condition (filled symbols) compared to the normal condition (open symbols).

In our experiment, we applied a series of pulse trains, each consisting of five identical programming pulses (1 V, 1 ms) with different frequencies and recorded the memristor conductance change (represented by the current produced by each pulse) as shown in Figure 4-15.

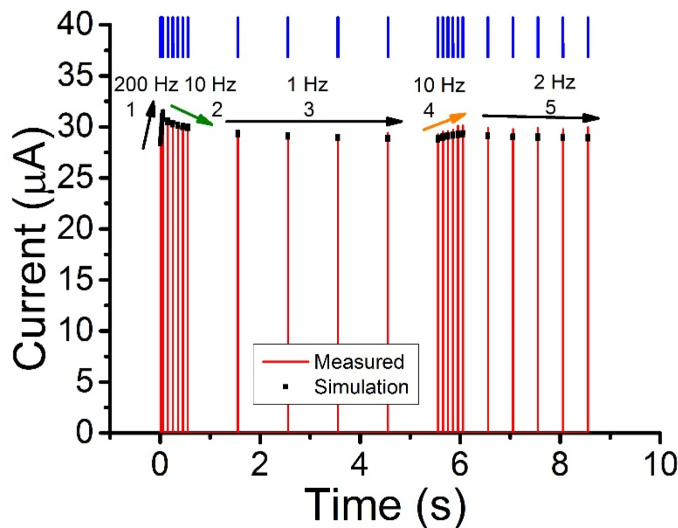

Figure 4-15: Experience-dependent weigh change of WO_x memristor. Consecutive programming pulse trains (1 V, 1 ms, blue lines) at different frequencies were applied, with five pulses for each frequency. The 10 Hz pulse train caused current decrease in step 2 following strong 200 Hz stimulation in step 1, but current increase in step 4 following weak 1 Hz stimulation in step 3. Black squares: simulation results from the second order WO_x memristor model.

In step 1, the first pulse train with a 200 Hz stimulation frequency was applied and resulted in an increase in current through the memristor. Subsequently, in step 2, a 10 Hz pulse train caused the memristor current to drop. On the other hand, following the 1 Hz pulse train in step 3, the same 10 Hz pulse train in step 4 caused an increase in memristor current instead. The sign reversal with respect to current change at the same 10 Hz stimulation condition shows that the effect of the stimulation on a memristor can also be dependent on previous activity.

This behavior observed in WO_x memristor can be explained by the V_O dynamics of the memristor, as the conductance change is determined by the competition of the effects of the stimulation pulse and the decay of the state variables w_c and w_m. The experimentally observed experience-dependent behaviors can be fully reproduced by simulation based on second order WO_x memristor model (black squares, Figure 4-15). Briefly, during the experiment, the 200 Hz pulse train drove w_m to a high value which leads to small effective time constant τ_m^* and τ_c^* and enhanced the decay of w_c and w_m in Equations (3-9) and (3-10). As a result, the subsequent 10 Hz pulse train was not sufficient to overcome the fast w_c decay to increase the memristor conductance anymore so an overall conductance drop was observed. On the contrary, after the 1 Hz pulse train, w_m had fully relaxed so the decay of w_c and w_m had slowed down significantly. As a result, the same 10 Hz pulse train afterwards was enough to bring the conductance up. In other words, the same device can experience either conductance increase or decrease at a given stimulation condition, depending on previous activities.

With this understanding, we performed an experiment analogous to that of Bear et al.[16] (Figure 4-14). In this study, we first experienced the device to one of three levels of activities by the application of ten pulses at either 10, 20 or 50 Hz, then five write pulses (1.2 V, 1 ms) with different repetition frequencies were applied and the net current (before and after the application of the write pulses) were recorded and the change was calculated. The experiment was repeated by fully relaxing the device to the resting state, and the change in current by the five write pulses was plotted against the stimulation frequency of the write pulses for the three cases, as shown in Figure 4-16.

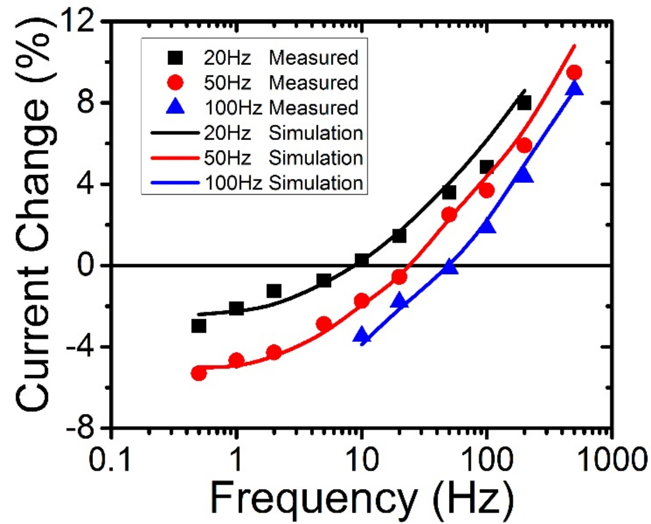

Figure 4-16: Experience-dependent weight change by different stimulation strength on WO_x memristor. The current changes as a function of the stimulation frequency after the memristor has been experienced to three different levels of activities (10, 20 and 50 Hz stimulation). Pulse trains consisting of five pulses (1.2 V, 1 ms) with different repetition frequencies were used to program the memristor. Black squares, red circles, blue triangles represent experimental data and the solid lines are simulation results by the second order WO_x memristor model.

A low stimulation frequency in general leads to conductance decrease (negative change) due to the previous activities and a high frequency in general leads to conductance increase. Moreover, the threshold frequency at which the net conductance change is zero is observed to depend on the previous activities as well, as evidenced by the shift in the three curves corresponding to the three levels of previous activities the device has been subjected to. The threshold frequency will slide to the right (higher frequency) when previous activity is stronger (50 Hz stimulation), similar to the theory in neurobiology, specifically the BCM theory. On the contrary, after experiencing weak previous activity (10 Hz), the threshold frequency will slide to the left (lower frequency).

These behaviors can be fully explained by the internal memristor dynamics using w_m and w_c as state variables, as evidenced by the quantitative agreement between experimental data and simulation results shown in Figure 4-16.

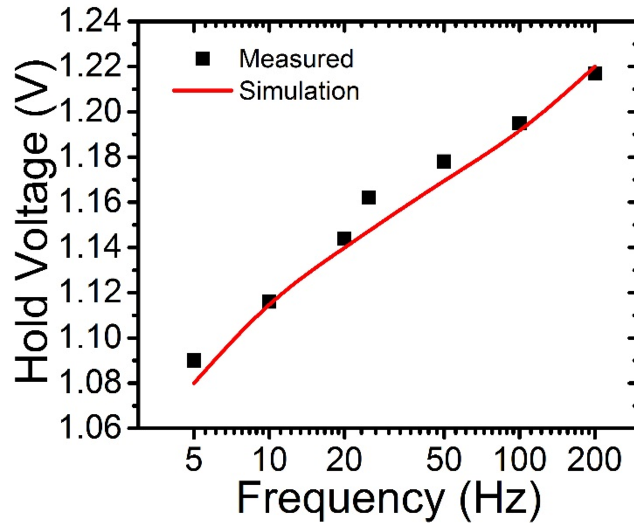

Figure 4-17: Experience-dependent plasticity as shown by the hold voltage of WO_x memristor. The device was stimulated by pulse trains with the same repetition frequency of 50 Hz but different amplitudes as different previous activities then the voltage by which the weigh could be maintained was recorded. Black squares: experimental data. Solid line: simulation results by the second order WO_x memristor model.

The sliding threshold effect can be reflected as either a change in threshold frequency at the same stimulation amplitude, as observed in neurobiological studies and shown in Figure 4-16, or a change in threshold amplitude at the same stimulation frequency. Both effects may be relevant for memristor devices and can be used in hardware-based neuromorphic systems. The sliding threshold amplitude effect is demonstrated in Figure 4-17. In this study, the device again was first subjected to different levels of activities, then a pulse train consisting of five pulses with a fixed frequency (50 Hz) but different amplitudes was applied and the amplitude at which the device conductance can be maintained was recorded. As can be seen in Figure 4-17, to maintain the device conductance at a given stimulation frequency, the threshold amplitude shifts to higher voltages with stronger previous activities (higher frequency in Figure 4-17). This behavior can be again quantitatively captured by the memristor model (solid line, Figure 4-17).

4.3.2 Metaplasticity in Ta_2O_5-TaO_x Memristor

As has been discussed in Chapter 2, the Ta_2O_5-TaO_x memristor made by ebeam lithography can demonstrate well-controlled complimentary resistive switching (CRS) behavior, which can be used to phenomenologically implement metaplasticity.

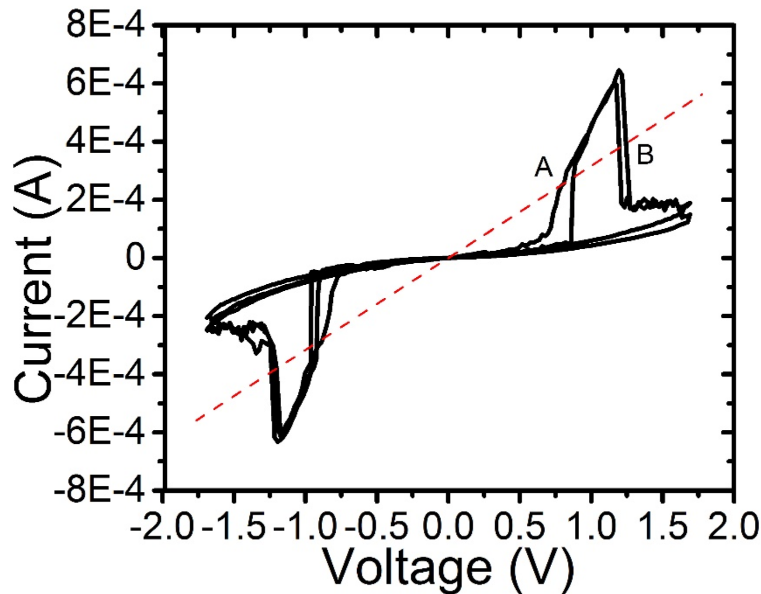

Figure 4-18: CRS of Ta_2O_5-TaO_x memristor.

As shown in Figure 4-18, after experiencing different DC sweep voltage ranges, which can be regard as the priming activity in metaplasticity, the device can be switched to two states A and B. While the conductances of states A and B, analogous to the synaptic weight, are of the same value (~0.326 mS), the response of the device to subsequent stimulations are dramatically different, for example with A leading to LTP while B leading to LTD when faced with identical large positive voltage. This effect is similar to the most common paradigm of metaplasticity, and can be explained by the different conduction filament profiles in the two cases, i.e. the location of the depletion gap at the opposite sides of the device in cases A and B, leading to conductance increase in A and conductance decrease in B when programmed with identical pulses[17].

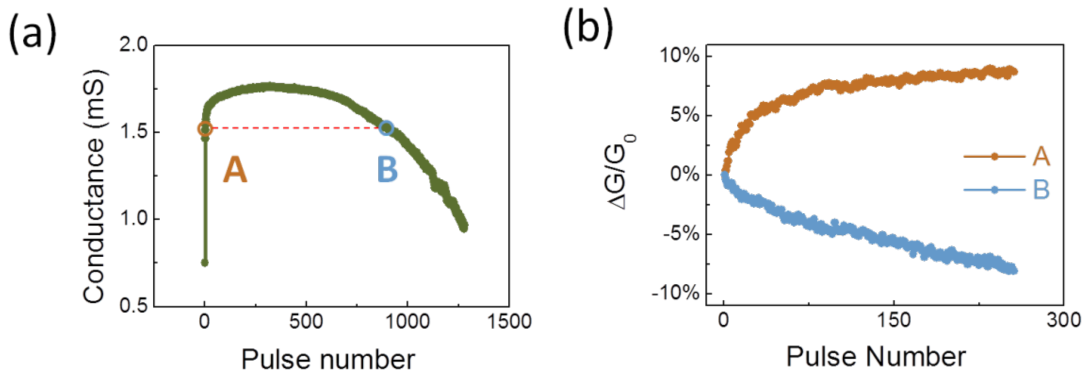

Figure 4-19: Metaplasticity observed in Ta_2O_5-TaO_x memristor. (a) Evolution of the device conductance with the number of applied programming pulses (1.2 V, 0.5 µs) staring from HRS. (b) The conductance change with the number of subsequently applied programming pulses (1.1 V, 0.5 µs) after the device was switched into states A and B.

This hypothesis was verified experimentally, shown in Figure 4-19. Here a series of programming pulses, acting as the priming activity, was applied to a device initially in HRS and the conductance evolved with the number of pulses applied (Figure 4-19(a)). The device was switched into states A and B, both of which have same conductance (~1.5 mS), meaning the priming activity does not cause measurable weight difference. However, when subjected to identical, subsequent programming pulses (1.1 V, 0.5 µs), LTP was observed in case A while LTD was observed in case B (Figure 4-19(b)), meaning the plasticity had indeed been altered by the priming activity. This result suggests the pulse-number-dependent metaplasticity in the Ta_2O_5-TaO_x memristor, similar to the spike-number dependent metaplasticity observed in the hippocampal slices of rats[18].

Similar effects were also observed by using pulse trains with different frequencies and pulse amplitudes as the priming activities (results to be published), and again suggest that memristors are much more than a programmable resistor characterized by a single weight state variable, but can exhibit interesting dynamics that can be used to emulate biological synapses efficiently.

4.4 Conclusion

Based on the understanding of internal ionic dynamics of WO_x memristor and Ta_2O_5-TaO_x memristor, the dynamics of 2nd order state variables are found to play a similar role of Ca^{2+} concentration dynamics in biological synapse that results in many different synaptic functions and can be utilized to implement them. Short-term synaptic behaviors, including paired pulse facilitation, rate dependent plasticity, long-term synaptic plasticities, including spike-timing dependent plasticity, frequency-dependent long-term weight change, and metaplasticity, including experience-dependent plasticity, are implemented in either WO_x memristor or Ta_2O_5-TaO_x memristor, and more importantly, in a natural and bio-realistic fashion.

Reference

1. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64,** 355–405 (2002).

2. Salin, P. A., Scanziani, M., Malenka, R. C. & Nicoll, R. A. Distinct short-term plasticity at two excitatory synapses in the hippocampus. *Proc. Natl. Acad. Sci.* **93,** 13304–13309 (1996).

3. Bi, G. & Poo, M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18,** 10464–10472 (1998).

4. Shouval, H. Z., Bear, M. F. & Cooper, L. N. A unified model of NMDA receptor-dependent bidirectional synaptic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 10831–10836 (2002).

5. Rachmuth, G., Shouval, H. Z., Bear, M. F. & Poon, C.-S. A biophysically-based neuromorphic model of spike rate- and timing-dependent plasticity. *Proc. Natl. Acad. Sci. U. S. A.* **108,** E1266-74 (2011).

6. Zucker, R. Calcium-and activity-dependent synaptic plasticity. *Curr. Opin. Neurobiol.* **9,** 305–313 (1999).

7. Kim, S. *et al.* Experimental Demonstration of a Second-Order Memristor and Its Ability to Biorealistically Implement Synaptic Plasticity. *Nano Lett.* **15,** 2203–2211 (2015).

8. Du, C., Ma, W., Chang, T., Sheridan, P. & Lu, W. D. Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics. *Adv. Funct. Mater.* **25,** 4290–4299 (2015).

9. Bliss, T. V. P. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361,** 31–39 (1993).

10. Zamarreño-Ramos, C. *et al.* On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. *Front. Neurosci.* **5,** 26 (2011).

11. Graupner, M. & Brunel, N. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proc. Natl. Acad. Sci. U. S.*

A. **109,** 3991–3996 (2012).

12. Yang, S.-N., Tang, Y.-G. & Zucker, R. S. Selective Induction of LTP and LTD by Postsynaptic [Ca2+]i Elevation. *J. Neurophysiol.* **81,** 781–787 (1999).

13. Abraham, W. C. Metaplasticity: tuning synapses and networks for plasticity. *Nat. Rev. Neurosci.* **9,** 387–399 (2008).

14. Huang, Y., Colino, A., Selig, D. & Malenka, R. The influence of prior synaptic activity on the induction of long-term potentiation. *Science (80-.).* **255,** 730–733 (1992).

15. Bienenstock, E. L., Cooper, L. N. & Munro, P. W. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* **2,** 32–48 (1982).

16. Kirkwood, A., Rioult, M. G. & Bear, M. F. Experience-dependent modification of synaptic plasticity in visual cortex. *Nature* **381,** 526–528 (1996).

17. Yang, Y., Sheridan, P. & Lu, W. Complementary resistive switching in tantalum oxide-based resistive memory devices. *Appl. Phys. Lett.* **100,** 1–5 (2012).

18. Mockett, B., Coussens, C. & Abraham, W. C. NMDA receptor-mediated metaplasticity during the induction of long-term depression by low-frequency stimulation. *Eur. J. Neurosci.* **15,** 1819–1826 (2002).

Chapter 5

Memristor Array for Sparse Coding

In previous chapters, we have demonstrated that at the single device level, memristors can emulate synaptic functions by storing the analog synaptic weights and implementing synaptic learning rules[1–6]. When constructed into a crossbar form, memristor networks can implement certain matrix operation, especially the dot product, very easily as discussed in Chapter 1 and offer the desired density and connectivity that are required for hardware implementation of neuromorphic computing systems[7–9]. Recently, memristor arrays and phase change memory devices have been used as artificial neural networks to perform tasks such as feature extraction and pattern recognition[10–14]. Here we experimentally demonstrate a sparse coding algorithm implemented by a memristor crossbar network, and show that the memristor network can be used to perform applications such as natural image analysis using learned dictionaries.

5.1 Sparse Coding

Sparse representation of information provides a powerful method to perform feature extraction on high-dimensional data, and is of broad interest for applications in signal processing, machine vision, object recognition and neurobiology[15,16]. Sparse coding is also believed to be a key mechanism by which biological neural systems can efficiently process complex, large amount of sensory data while consuming very little power[17,18].

Sparse representation reduces the complexity of the input signals and enables more efficient processing and storage, as well as improved feature extraction and pattern recognition functions[15,16]. Briefly, given a signal x, which may be a vector (*e.g.* representing the pixel values in an image patch), and a dictionary of features D, the goal of sparse coding is to represent x as a

linear combination of features from D using a sparse set of coefficients a, while minimizing the number of features used. A schematic of the sparse coding concept is shown in Figure 5-1, where an input (e.g. the image patch of a clock) is represented by a few features selected from a large dictionary[16,17].

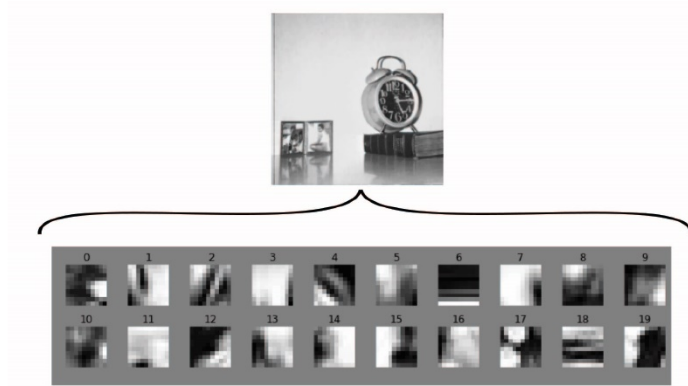

Figure 5-1: Schematic of the sparse coding concept. An input (e.g., the image patch of a clock) can be decomposed into and represented with a minimal number of dictionary elements.

5.2 Sparse Coding Algorithm

The Locally Competitive Algorithm (LCA)[19] is a sparse coding algorithm that uses a dictionary of feature vectors (represented as synaptic weights in a neural network) to transform a vector of input signal into a relatively small number of output coefficients, which can be used as a compressed form of the input for image compression or object recognition. Different from simple feed-forward neural networks, LCA describes a dynamical system where neurons compete with each other in proportion to the similarity of their respective receptive fields (the collection of synaptic weights entering a neuron) so that a more optimal representation, out of many possible representations, can be obtained.

The neuron dynamics during LCA analysis can be summarized by the following equations:

$$\frac{du}{dt} = \frac{1}{\tau}\left(-u + x^T D - a(D^T D - I_n)\right) \tag{5-1}$$

$$a = \begin{cases} u \text{ if } u > \lambda \\ 0 \text{ otherwise} \end{cases} \tag{5-2}$$

where u is the neuron's membrane potential, τ is a time constant, x is the signal to be encoded, D is the dictionary of features, a is the neuron activity (whose non-zero elements form the sparse code), and I_n is the $n \times n$ identity matrix. During LCA analysis, each neuron i integrates its input $x^T D$, leakage $-u$, and inhibition $a(D^T D - I_n)$ terms and updates its membrane potential u_i (Equation (5-1)). Specifically, the input to neuron i results from the signal x scaled by the weights D_{ji} connected to the neuron (second term in Equation (5-1)). To this regard, the collection of the synaptic weights D_{ji} associated with neuron i, corresponding to a feature column of D, is also referred to as the receptive field of neuron i, analogous to the receptive fields of biological neurons in the visual cortex[17,20]. Here the input $x^T D$ increases the neuron's membrane potential, by an amount proportional to the similarity between the input and the neuron's receptive field (through the vector-matrix dot-product), while the inhibition term decreases the neuron's membrane potential, particularly from other neurons with similar receptive fields. If and only if u_i reaches above a threshold (set by parameter λ), neuron i will produce an output $a_i = u_i$, otherwise the neuron's activity a_i is kept at zero through the thresholding function (Equation (5-2)).

5.3 Memristor Network for Sparse Coding

The memristor network is particularly suitable for implementing neuromorphic algorithms such as LCA since the matrix-vector dot-product operations can be performed through a single read operation in the memristor array[8], as already discussed in Chapter 1.

Equation (5-1) can be rewritten as:

$$\frac{du}{dt} = \frac{1}{\tau}(-u + (x - \hat{x})^T D + a) \tag{5-3}$$

By doing so, the matrix-matrix operation $D^T D$ in Equation (5-1) is reduced to two sequential matrix-vector dot-product operations (one used to calculate $\hat{x} = Da^T$ and the other used to calculate the contribution from the updated input $(x - \hat{x})^T D$).

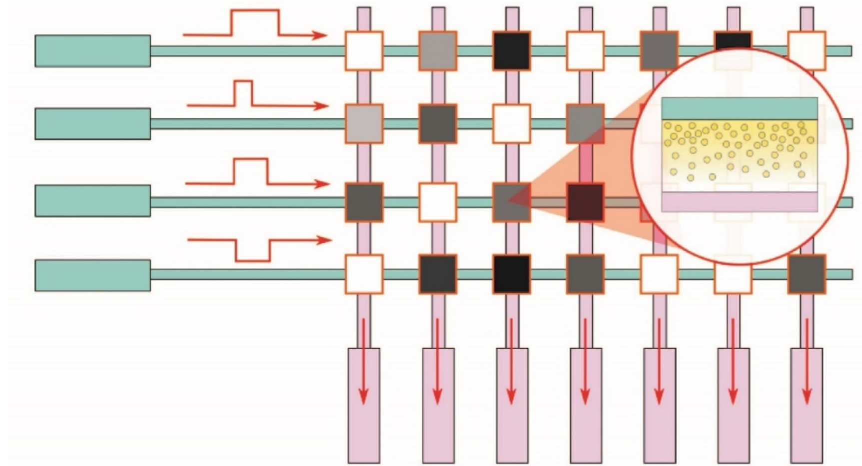

Figure 5-2: Schematic of memristor crossbar based computing. A memristor is formed at each crosspoint and can be programmed to different conductance states (represented as grayscale color).

We have mapped the LCA algorithm into a memristor crossbar network, schematically shown in Figure 5-2. In this implementation, x is an m-element column vector applied to the rows of the memristor crossbar (cyan pads on the left), with each element corresponding to an input element (e.g. intensity of a grayscale pixel in an image patch). It is implemented by read pulses with a fixed amplitude but variable width proportional to the pixel intensity. The dictionary D is an m×n matrix directly mapped element-wise into the memristor crossbar with each column j storing the corresponding synaptic weights, an m-element vector, for one element in the dictionary. As a result, the total charge Q_{ij} passed by a memristor at crosspoint (i, j) is linearly proportional to the product of the pixel intensity x_i and the conductance D_{ij} of the memristor $Q_{ij} = x_i \mathrm{D}_{ij}$, and the charge passed by all memristors sharing column j is summed via Kirchhoff's current law $Q_j = \sum_i x_i D_{ij} = x^T D_j$, thus achieving the desired vector-vector dot-product *in physics*. Since the dot-product of vectors measures how close the input vector is matched with the stored vector, the ability to implement this operation in a single read process allows the memristor network to conveniently and efficiently perform this important pattern matching task.

In this implementation, each column is connected to a leaky-integrator output neuron (pink pads at the bottom), such that the total charge accumulated at neuron j is proportional to the dot-product of the input x with the neuron's receptive field D_j. Afterwards the neuron activity is obtained through Equation (5-2), and the neuron activity coefficients form a row vector a, where

the jth element of a represents the activity of the jth neuron. After feeding input x to the network and allowing the network to stabilize through lateral inhibition, a reconstruction of x can be obtained as Da^T, and in a sparse representation only a few elements in a are non-zero while the other neurons' activities are suppressed to be precisely zero.

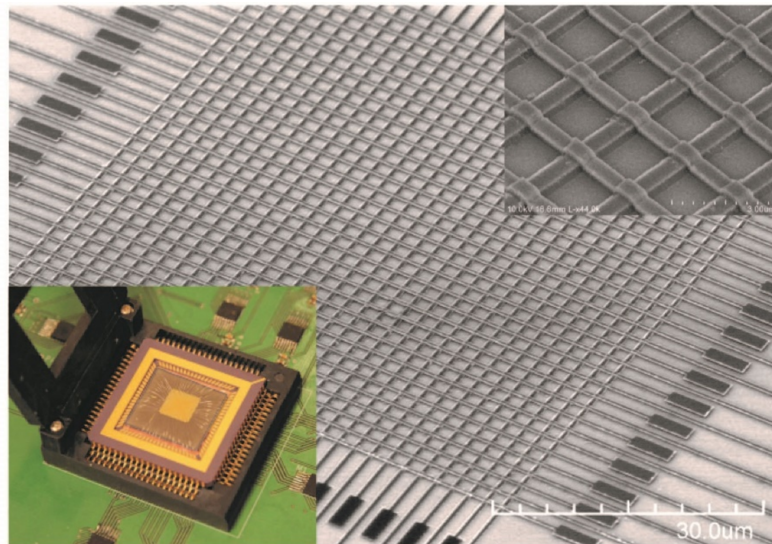

Figure 5-3: Memristor crossbar network for sparse coding. A scanning electron microscope (SEM) image of a fabricated memristor array used in this study is shown. Upper right inset shows a magnified SEM image of the crossbar. Lower left inset shows the memristor chip integrated on the testing board after wire-bonding.

The hardware system used in our study is based on a 32x32 memristor crossbar array, with a memristor formed at each intersection in the crossbar (Figure 5-3). The WO_x memristor devices are fabricated following the previously developed procedures[21] discussed in Chapter 2. After fabrication, the chip of memristor crossbar array is wire-bonded and integrated on a custom-build testing board. The original input, such as an image, is fed to the rows of the memristor array and the columns of the array are connected to output neurons. The memristor network performs critical pattern matching and neuron inhibition operations to obtain a sparse, optimal representation of the input. After the memristor network stabilizes, the re-constructed image can be obtained based on the (sparse) output neuron activities and the features stored in the crossbar array.

5.4 Sparse Coding Experiment Results

Figure 5-4 shows an example of encoding an image composed of horizontal and vertical bars using the procedures given above. The dictionary shown in Figure 5-4(a) contains 20 features with each feature consisting of 25 weights. If the input images are restricted to only combinations of horizontal and vertical bars, the input dimensionality is reduced to 9. Therefore, the dictionary is larger than the input space to form an over-complete dictionary set. In this experiment, a 25x20 sub-array was used out of the 32x32 memristor array. The 20 features were written into the 20 columns and the inputs were fed into the 25 rows. An input signal, shown in Figure 5-4(b) and consisting of a combination of 3 bars, is used as a test input and the final reconstruction is shown in Figure 5-4(b). It can be correctly reconstructed with neurons 8 and 16, which are the two neurons with membrane potential larger than the threshold after certain iterations, as shown in Figure 5-4(c). The network not only correctly reconstructed the input image, but more interestingly, picked the more efficient solution – a solution based on neurons 8 and 16, over another solution based on neurons 1, 4 and 8. This result emphasizes the sparsity of the coding.

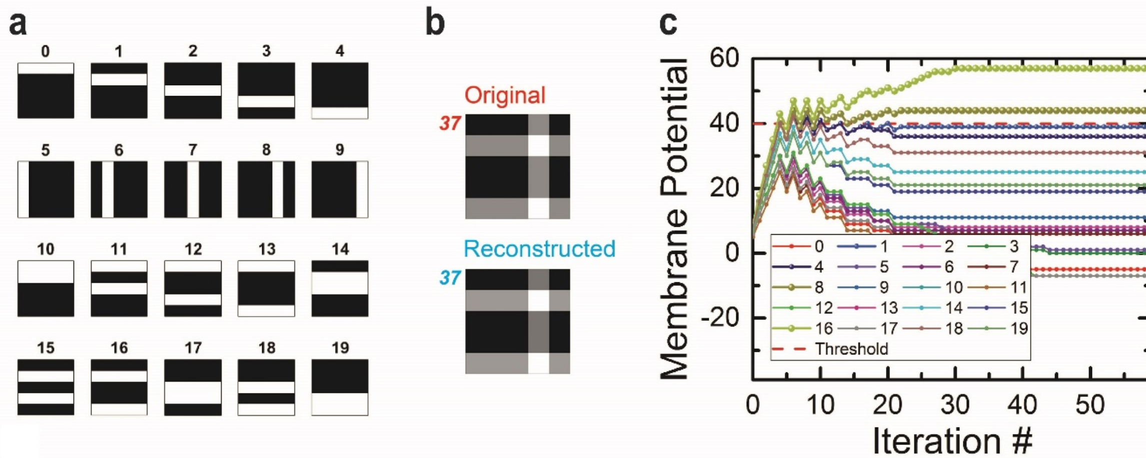

Figure 5-4: Sparse coding results by WO_x memristor crossbar array. a) Dictionary elements base on horizontal and vertical bars. b) The original image to be encoded and the reconstructed image. c) Membrane potentials of the neurons as a function of iteration number during LCA analysis.

5.5 Conclusion

Utilizing the merits of memristor crossbar network, i.e., easy implementation of dot product and weight storage/modulation, an important neuromorphic algorithm, LCA, is implemented experimentally on a fabricated WO_x memristor crossbar array. The algorithm can be used to code and reconstruct patterns and images with sparsity.

Apart from encoding the bar patterns shown above, we also demonstrated that the memristor array can be used to experimentally code and reconstruct natural images using sparse coding algorithm and a learned dictionary (results to be published). The dictionary elements were obtained offline using a realistic memristor model and algorithm based on winner-take-all (WTA) approach and Oja's learning rule. The obtained dictionary elements were programmed into a physical 16x32 crossbar array. Using the trained dictionary, we successfully preformed reconstruction of 120x120 pixel grayscale images using the 16x32 memristor crossbar array.

Reference

1. Jo, S. H. *et al.* Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10,** 1297–1301 (2010).
2. Pershin, Y. V. & Di Ventra, M. Experimental demonstration of associative memory with memristive neural networks. *Neural Networks* **23,** 881–886 (2010).
3. Du, C., Ma, W., Chang, T., Sheridan, P. & Lu, W. D. Biorealistic Implementation of Synaptic Functions with Oxide Memristors through Internal Ionic Dynamics. *Adv. Funct. Mater.* **25,** 4290–4299 (2015).
4. Kuzum, D., Jeyasingh, R. G. D., Lee, B. & Wong, H.-S. P. Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing. *Nano Lett.* **12,** 2179–2186 (2012).
5. Ohno, T. *et al.* Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nat. Mater.* **10,** 591–5 (2011).
6. Kim, S. *et al.* Experimental Demonstration of a Second-Order Memristor and Its Ability to Biorealistically Implement Synaptic Plasticity. *Nano Lett.* **15,** 2203–2211 (2015).

7. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nat. Nanotechnol.* **8,** 13–24 (2013).

8. Sheridan, P. M., Du, C. & Lu, W. D. Feature Extraction Using Memristor Networks. *IEEE Trans. Neural Networks Learn. Syst.* 1–10 (2015). doi:10.1109/TNNLS.2015.2482220

9. Legenstein, R. Nanoscale connections for brain-like circuits. *Nature* **521,** 37–38 (2015).

10. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nat. Commun.* **4,** 2072 (2013).

11. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521,** 61–64 (2015).

12. Burr, G. W. *et al.* Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. *IEEE Trans. Electron Devices* **62,** 3498–3507 (2015).

13. Sheridan, P. M., Du, C. & Lu, W. D. Feature Extraction Using Memristor Networks. *IEEE Trans. Neural Networks Learn. Syst.* **pp,** 1–10 (2015).

14. Sheridan, P., Ma, W. & Lu, W. Pattern recognition with memristor networks. *Proc. - IEEE Int. Symp. Circuits Syst.* 1078–1081 (2014). doi:10.1109/ISCAS.2014.6865326

15. Wright, J. *et al.* Sparse representation for computer vision and pattern recognition. *Proc. IEEE* **98,** 1031–1044 (2010).

16. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.* **37,** 3311–3325 (1997).

17. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).

18. Vinje, W. E. & Gallant, J. L. Natural Vision Sparse Coding and Decorrelation in Primary Visual Cortex During Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. **287,** 1273–1277 (2000).

19. Serre, T., Wolf, L. & Poggio, T. Object recognition with features inspired by visual cortex. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2,** 994–1000 (2005).

20. Hubel, B. Y. D. H. *et al.* Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160,** 106–154 (1962).

21. Chang, T. *et al.* Synaptic behaviors and modeling of a metal oxide memristive device. *Appl. Phys. A Mater. Sci. Process.* **102,** 857–863 (2011).

Chapter 6

Memristor Array for Reservoir Computing

Beyond using memristors to store weights and perform vector-matrix dot-products in place for applications discussed in Chapter 5, the internal dynamics discussed in Chapter 4 can be used to perform other computing tasks natively. One interesting and important application is reservoir computing, where the internal dynamics of memristors can be utilized for efficient temporal information processing.

6.1 Reservoir Computing

The concept of reservoir computing was developed to solve problems in implementing recurrent neural networks (RNNs). RNN is a class of artificial neural network in which connections between nodes form a directed cycle, creating an internal state of the network which allows it to exhibit dynamic temporal behaviors[1].

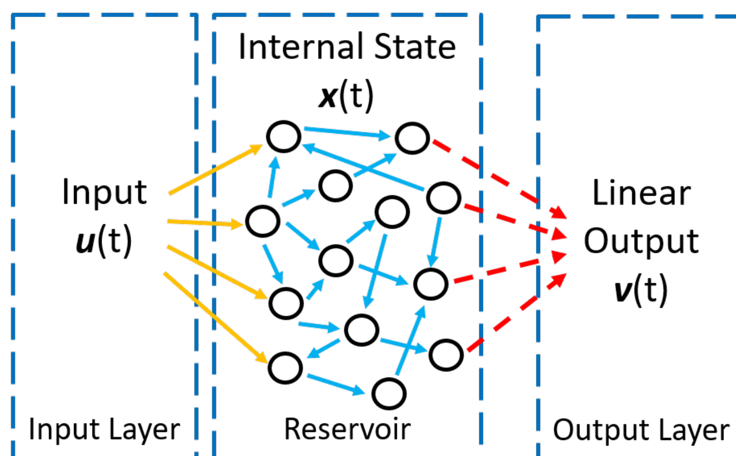

Figure 6-1: Reservoir computing.

Although RNNs are theoretically very powerful tools for solving complex temporal machine learning tasks, several factors still hinder the large scale deployment of RNNs in practical applications including factors such as not many learning rules exist and most that do exist suffer from slow convergence rates and require intense computation during the training of the network. To solve these problems, two concepts or structures were proposed independently. Echo State Network (ESN), proposed by Jaeger[2] in 2001, and Liquid State Machine (LSM), proposed by Maass[3] in 2002, described an improved constructive learning algorithm for RNN. These two structures were later unified by Verstraeten[4] in 2007 as Reservoir Computing (RC) since they have a similar structure as shown in Figure 6-1: the whole system is separated into two parts: the first part, connecting to the input, will evolve dynamically with the temporal input signal, is called *reservoir* and the weight of the connections between all nodes will not be trained; while the second part, called *readout function* (red arrows in the picture), reads the state of the reservoir and generates the desired output. The most significant difference of RC systems compared with previous RNNs is that during training only the weight of connections between the reservoir and readout will be trained, while the connections inside the reservoir remain unmodified. In this way, it is much easier to perform training since for RNNs it is extremely difficult to find an optimal training algorithm of the network and it is usually computationally complex.

Essentially, the reservoir conducts a pre-processing of the input by projecting it to a high-dimension space, therefore converting initially linearly inseparable input signals into linearly separable signals after this nonlinear transformation. Afterwards, a simple readout, normally a single layer linear readout, can generate the desired output after training.

6.2 Liquid State Machine

Before we discuss the implementation of reservoir computing by memristor, it is worth mentioning the differences between ESN and LSM:

1. Origins: ESN was proposed for machine learning and to solve engineering problems, while LSM was proposed for computational neuroscience, initially aiming to investigate the microcircuit of cortex.

2. Nodes: ESN, mostly using simple sigmoidal neuron, is easier to implement and run simulation. While for LSM, as it is derived from biological neural networks, in most cases Leaky Integrate and Fire (LIF) neurons are used and dynamic synapses are used for the connections. Therefore, LSM is more complex but potentially more computationally powerful.

3. Reservoir types: although both set very loose restrictions on the reservoir, ESN uses recurrent neural networks so there are loops inside the network. Especially, it requires a feedback from the output to the reservoir to obtain the so-called "echo state property". LSM, on the other hand, sets a much looser requirement on the reservoir. It even does not need to be recurrent, meaning feed-forward networks without any loops can also constitute a liquid state machine as long as the reservoir obeys a quite unrestrictive property, which will be discussed later.

4. Input signals: the difference between ESN and LSM with respect to the inputs, which could be spike trains, is that the former focuses on the firing rate as the nodes are simple sigmoidal neurons while the latter emphasizes the timing of each spike. As the spike timing is very important for our memristor device, we will also focus on the timing of input spikes as in the LSM case.

Therefore we focus on LSM type RC for two reasons:

1. The reservoir does not need to be recurrent so it can be implemented by a simple memristor array.

2. WO_x memristor has been intensively investigated to implement several dynamic synaptic functions while in LSM the dynamic synapses are commonly used in the reservoir as the connections between nodes or even constitute the reservoir itself.

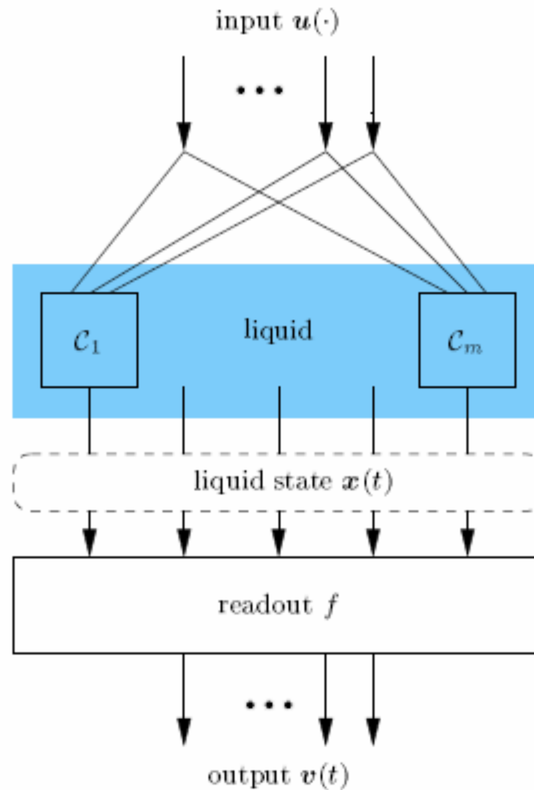

*Figure 6-2: Schematic of a Liquid State Machine. The system consists of two parts: the liquid, which will generate an internal state **x**(t) according to the history of input **u**(t), and a readout function, which can generate a desired output **v**(t) after training.*

The structure of a Liquid State Machine is shown in Figure 6-2. The input ***u(t)***, which is a temporal signal, is fed into the reservoir. The reservoir, or the liquid, will response to the input by changing its internal state ***x(t)***. Then the readout function *f* obtains the internal state of the liquid and generates the output ***v(t)***, which should be as close to the desired output ***d(t)*** as possible, after training the weight of connections between the liquid and the readout.

The readout function only linearly maps the liquid state ***x(t)*** at time t to output ***v(t)**=f(**x(t)**)* so it is memory-less. Then how could the system obtain information at t'<t as a filter? The answer is that the liquid state is determined not only by the input currently applied but also those in a certain period in the past. Therefore, the liquid itself must have short-term memory. In fact, it has been mathematically proved that the Liquid State Machine should have two very unrestrictive properties to obtain universal computation power for time-varying inputs[3]:

1) The liquid satisfies the point-wise separation property. This means that all output-relevant differences in the preceding part of two input series $u_1()$ and $u_2()$ (before time t) are reflected in the corresponding liquid state $x_1()$ and $x_2()$ which are *separable*.

2) The readout function satisfies the approximation property. This means the readout function can map the current liquid state to the desired current output with required accuracy.

Bearing these two properties in mind, it has been shown that a Liquid State Machine, which can approximate any time invariant filter with fading memory with arbitrary precision, can be constructed with the following simple protocol[5]:

i. Choose a suitable liquid
ii. Record liquid states in response to inputs
iii. Train a readout function with input-output samples

6.3 WO_x Memristor Based Synapses as the Liquid

With the above mentioned protocol, a Liquid State Machine, in which the liquid is built directly using dynamic synapses[6] has been demonstrated to implement an arbitrary finite state machine.

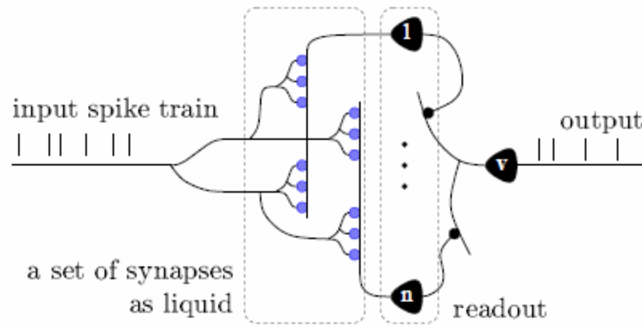

Figure 6-3: Schematic of a Liquid State Machine with synapses as the liquid.

The structure of this Liquid State Machine is shown in Figure 6-3. The synapses constituting the liquid in this case are expected to show very similar behaviors to what we have already observed in our WO_x memristor, as shown in Figure 6-4. It is obvious that the synapse has a fading memory, similar to the conductance decay of WO_x memristor (Figure 3-1), and responds differently according to the timing of spikes. In the WO_x memristor, this effect is caused by the nonlinear conductance change by multiple pulses (Figure 3-2) and the pulse timing dependent conductance change (Figure 4-2).

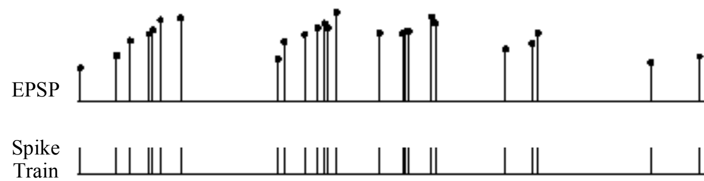

Figure 6-4: Expected synapse's response to a spike train. The EPSPs generated through synapse are different depending on the timing of the spikes.

To test this effect, a pulse train composed of write pulses having the same amplitude (1.4 V, 500 μs) but different relative timing (intervals between the pulses), was applied to the device and the response of the memristor, which was represented by current measured by a small read pulse (0.6 V, 500 μs) following each write pulse, was recorded and shown in Figure 6-5.

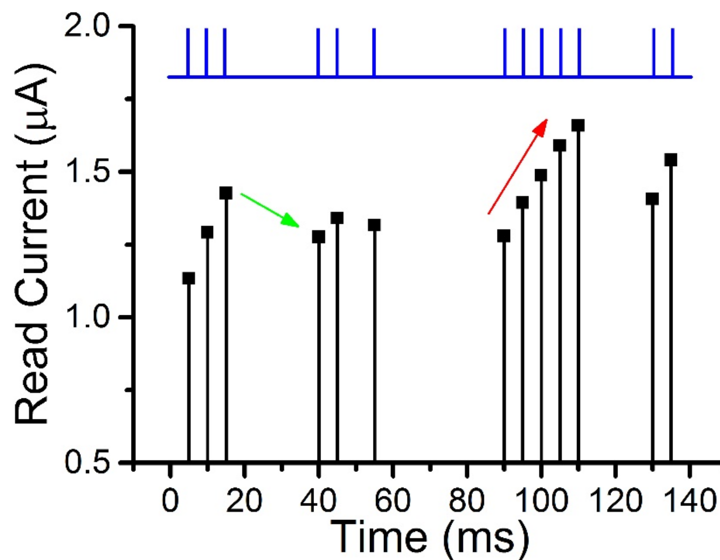

Figure 6-5: Memristor's temporal response to a pulse train. Write pulses (1.4 V, 500 μs) with different timing (blue lines) were applied and the response, represented by current measured by a small read pulse (0.6 V, 500 μs) after each write pulse is recorded. A temporal response is observed.

The response of the memristor is indeed very similar to the synapse in a LSM in at least two aspects: 1) if multiple pulses are applied with short intervals, the response will gradually increase as indicated by the red arrow in Figure 6-5, showing a cumulative effect, 2) if there is a certain period without any stimulation, then the response to the next stimulation will be weaker as indicated by the green arrow in Figure 6-5, showing the decay effect. Therefore, we think that

by adopting the Liquid State Machine concept and structure, a reservoir computing system can be constructed using the WO_x memristors.

6.4 System and Task Design

Apart from the dynamics of WO_x memristor, which has been mentioned and compared to that of synapses used in LSM, it is important to design a suitable LSM for specific applications. The task should also be carefully selected and designed as to show the potential of using memristors as the liquid. In previous research, finite state machine[6], speech/word recognition[7], movement predication[8] and many other nonlinear temporal applications using LSM have been demonstrated. Here we start from a digit recognition task and in the future we hope to demonstrate more powerful and universal computation abilities of WO_x memristor based LSMs.

6.5 LSM for Simple Digit Recognition

Here a LSM is designed for a very specific task, that is, to recognize the digit from an input image, as shown in Figure 6-6.

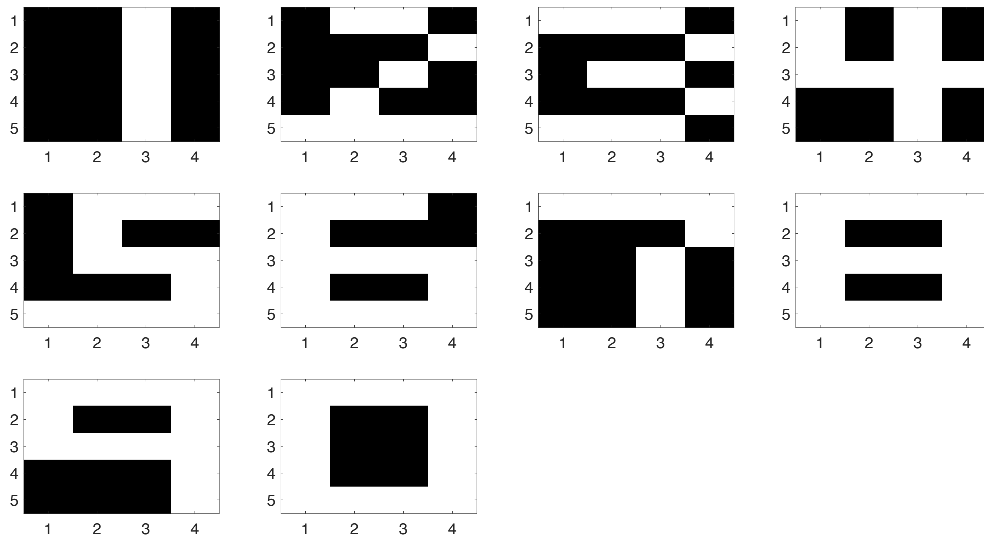

Figure 6-6: Simple digit images. Each digit image contains twenty pixels, either black or white.

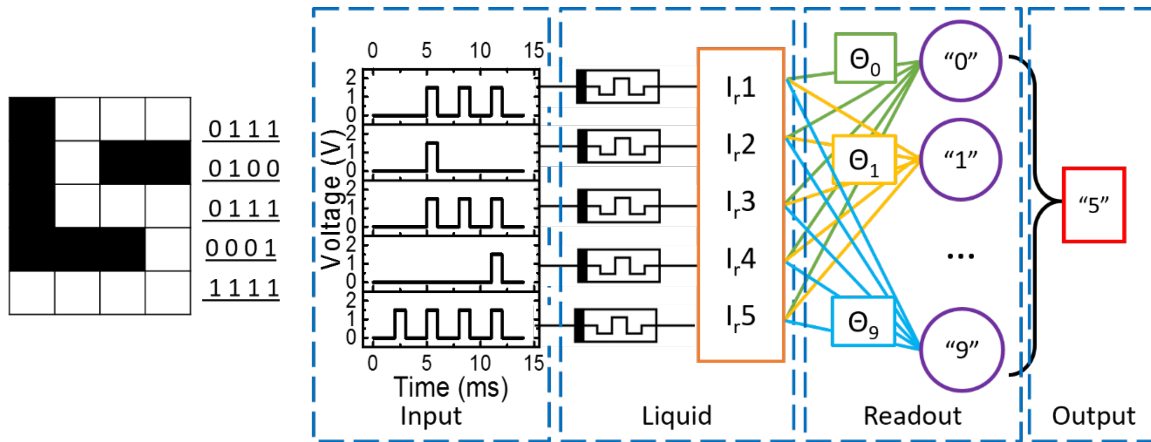

Figure 6-7: LSM for simple digit recognition. Left: digit "5" as an example. Right: the LSM containing the inputs (pulse trains transformed from the image), the liquid (consisting of 5 memristors) and the readout function (a network with 10 output neurons).

The LSM is shown in Figure 6-7, taking digit "5" as an example. The digit is shown as an image with 20 pixels, either black or white. Then the image is separated into 5 rows, each row containing 4 pixels. The pixels are transformed into either a write pulse (1.5 V, 1 ms) for white pixel, or no pulse (or pulse with amplitude of 0 V) for black pixel. Therefore, the spatial information of the image for digit "5", which is the locations of white pixels in each row, is transformed into temporal information, i.e., a pulse train with pulses applied with different timing. From the 5 rows, 5 pulse trains are obtained and fed into 5 memristors. Here the 5 memristors constitute the liquid and as discussed above will respond differently to each pulse train according to the relative timing of the pulses within the pulse train. When a write pulse is applied, the state of the memristor will be changed (conductance increase) and if multiple pulses are applied with short interval the cumulative effect will increase the conductance gradually, while if there is no stimulation, the state (conductance) will decay towards its resting state, i.e., the initial state before any pulse is applied. Therefore, the final state of the memristor after applying a pulse train corresponds to the result of a non-linear transformation of the temporal information of that pulse train, and if these states are different for different input patterns the patterns become linearly separable and can be identified through the readout function. Here the states of the 5 memristors are obtained by applying a read pulse at the end of each pulse train, and the measured read currents reflect the internal state of the liquid. After supervised training of the simple readout function, the desired output, which is the correct recognition of digit "5" can be obtained. More specifically, here the readout function is a 5x10 network, with the read

currents from the 5 memristors in the liquid as the input, and the 10 output neurons representing the 10 digits as the output. Each output is connected to the memristors through 5 weights (totaling 50 weights for the 10 digits). During the classification task, the output from the 10 output neurons are calculated by obtaining the dot product of the input (read currents from the 5 memristors in the liquid) with the 5 weights for each digit, and the digit with the maximum value is selected as the recognition result.

A significant advantage of using a LSM for digit recognition task is that fewer weights need to be trained. A normal neural network for this task will have 20 inputs as there are 20 pixels. If there is no hidden layer, i.e., the inputs are directly connected to the output, then from 20 inputs to 10 outputs, there are 200 weights in total that need to be trained and that number will quickly increase if one or more hidden layers are used. By using a LSM, the spatial information is encoded in the temporal domain so a smaller network (e.g. a 5x10 network) can be used for readout and only 50 weights need to be trained. Essentially, we are trading speed (the time needed to input pulse train) for space (memory to store 50 weights instead of 200 or more) and the computation complexity of training (fewer weights need to be trained).

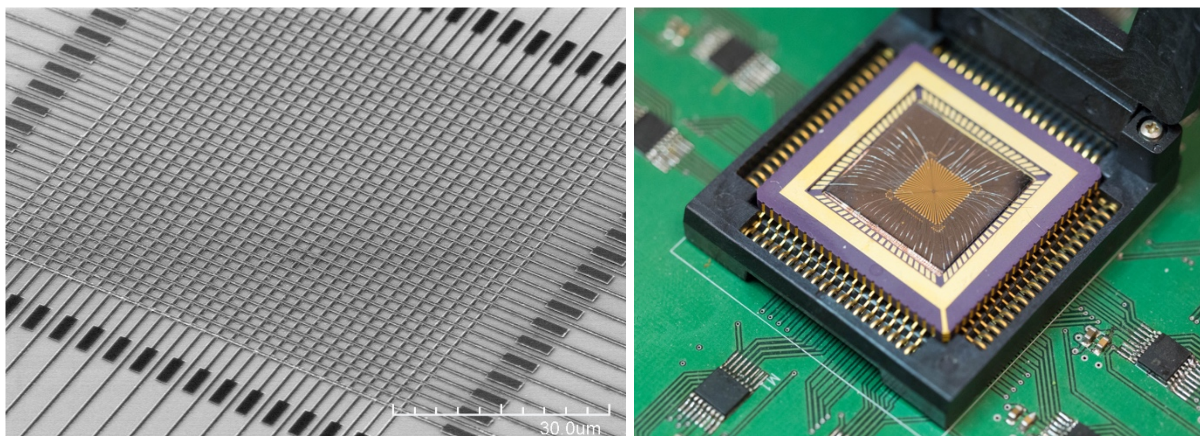

Figure 6-8: Experimental setup for LSM. Left: 32x32 WO_x memristor array fabricated. 5 cells from the array are used as the liquid. Right: the chip containing the memristor array is wire bonded to a chip carrier and integrated on a customized board for testing.

The experimental setup is shown in Figure 6-8. A 32 by 32 WO_x memristor array is fabricated with 500nm line width. The chip is wire bonded to a chip carrier and mounted on a customized board for testing.

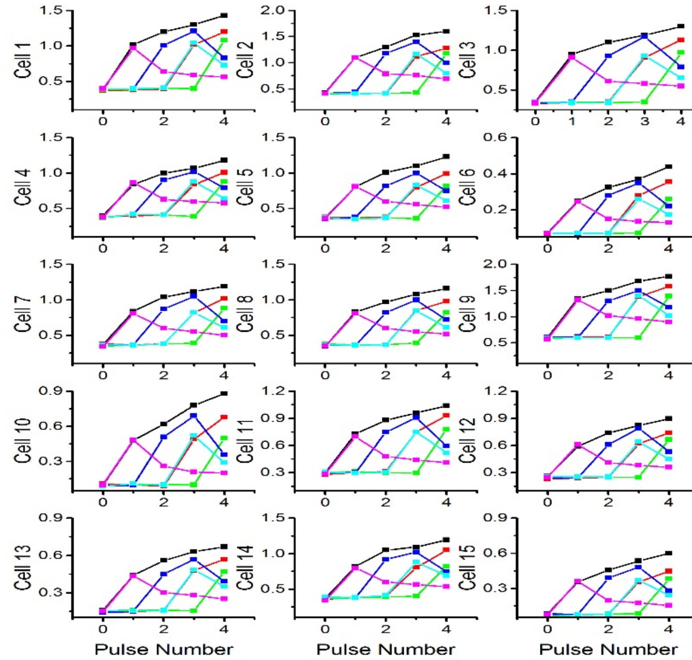

Figure 6-9: Memristors' variation and response to six pulse trains. The response of 15 cells from the array to 6 different pulse trains are shown, indicating 1) variations between cells which are manageable and 2) similarity of the temporal response of each cell and clear separation of those different inputs by the final cell conductances.

15 cells were chosen from the array for LSM studies. The response of these cells to 6 different pulse trains are shown in Figure 6-9. Although there are some variations among the cells, especially the absolute read current levels, all cells show the same trend when subjected to these inputs and the read current after the forth pulse can be well separated for different inputs (pulse trains) in all devices.

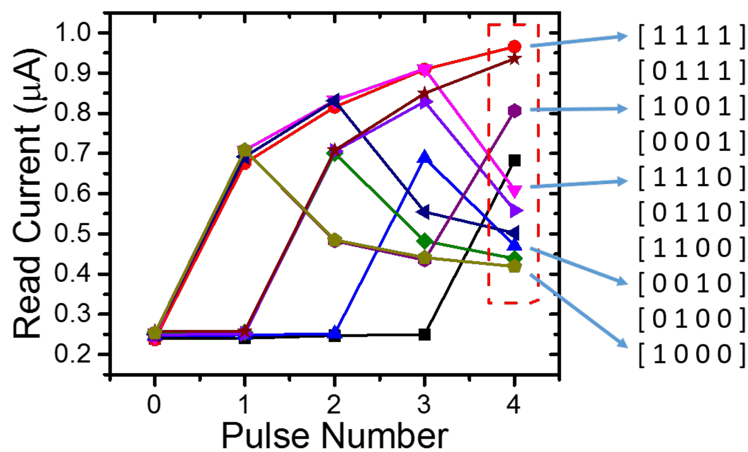

Figure 6-10: Memristor's response to ten pulse trains. Ten pulse trains corresponding to ten different row pixel arrangements for the ten digit images were input to a memristor and the read currents after the forth pulse show ten different levels that can be well separated.

For the ten digits represented by the 4x5 images, there are overall ten different pixel arrangements along each row direction, corresponding to ten different possible pulse trains as input. We tested the memristor's response to these pulse trains, shown in Figure 6-10, and a different read current after the forth pulse was obtained for each pulse train, indicating that the memristor can separate those ten different inputs well. This is more clearly shown in Figure 6-11. Here the internal states of the liquid, represented by the combination of the read currents from all 5 memristors in the liquid (with each memristor corresponding to a row of the input image), are significantly different when the liquid is subjected to the 10 different digit inputs, verifying the liquid can clearly separate those ten digit images.

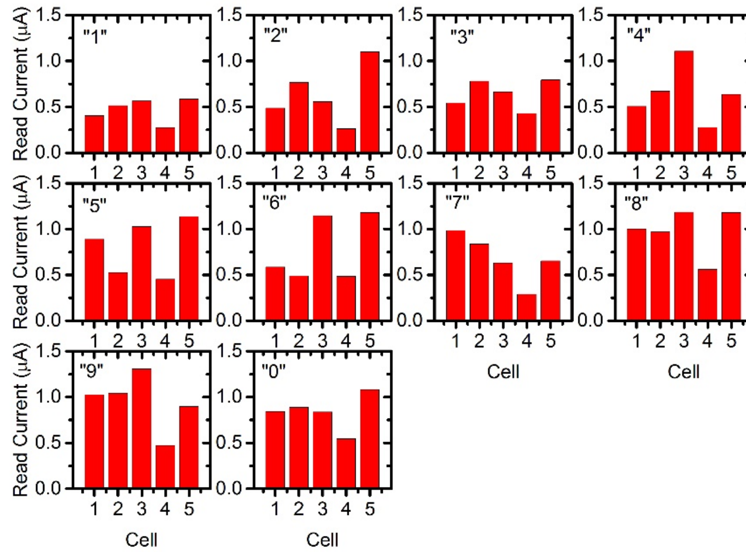

Figure 6-11: Liquid's internal states after subjected to the ten digit inputs. The read currents of the 5 memristors were recorded as the internal state of the liquid and significant differences can be observed.

After the liquid states for all the inputs were obtained, the readout function was trained using logistic regression which is commonly used for classification in machine learning.

Suppose the liquid state is x, a vector containing 5 elements, i.e., the 5 read currents, and for each output neuron in the readout network there is a set of weights θ, which is also a vector with 5 elements. Then the hypothesis function is

$$h_\theta(x) = g(\theta^T \cdot x) \quad (6\text{-}1)$$

$$g(z) = \frac{1}{1+e^{-z}} \quad (6\text{-}2)$$

The cost function is

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}log\left(h_\theta\left(x^{(i)}\right)\right) - \left(1 - y^{(i)}\right)log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] \quad (6\text{-}3)$$

where m is the number of samples, $y^{(i)}$ is the desired output for input $x^{(i)}$.

The gradient descent is

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)} \quad (6\text{-}4)$$

The training of the weights is achieved by Matlab 2016b using function *fmincg()*, which is provided by Jason Rebello as a logistic regression with regularization and commonly used to classify hand written digits[9].

After training, any example from the original ten digit images can be recognized with 100% accuracy after passing the input image through the memristor-based LSM and the readout network (each image tested 10 times). This unambiguously demonstrates the memristor array based LSM's ability to encode the spatial information in temporal domain and to process the temporal information due to the internal dynamics of the memristor.

To more clearly demonstrate the effect of short-term decay, or the temporal information processing ability of the liquid, two distorted digit images for “2” and “3” were generated by adding noise to the original training samples, as shown in Figure 6-12. A close inspection will reveal that the number of pulses in the pulse trains for these two digits are the same for each row (i.e., 2, 1, 2, 1, 3 pulses for row 1 to 5). The only difference is the relative timing of the pulses for the last two rows. As expected, the liquid states corresponding to the last two rows of the two noisy digit images, as shown in Figure 6-12, are significantly different, therefore enabling the liquid to clearly separate these two different inputs and allowing the system to successfully recognize these inputs as digit “2” and “3” after the readout network.

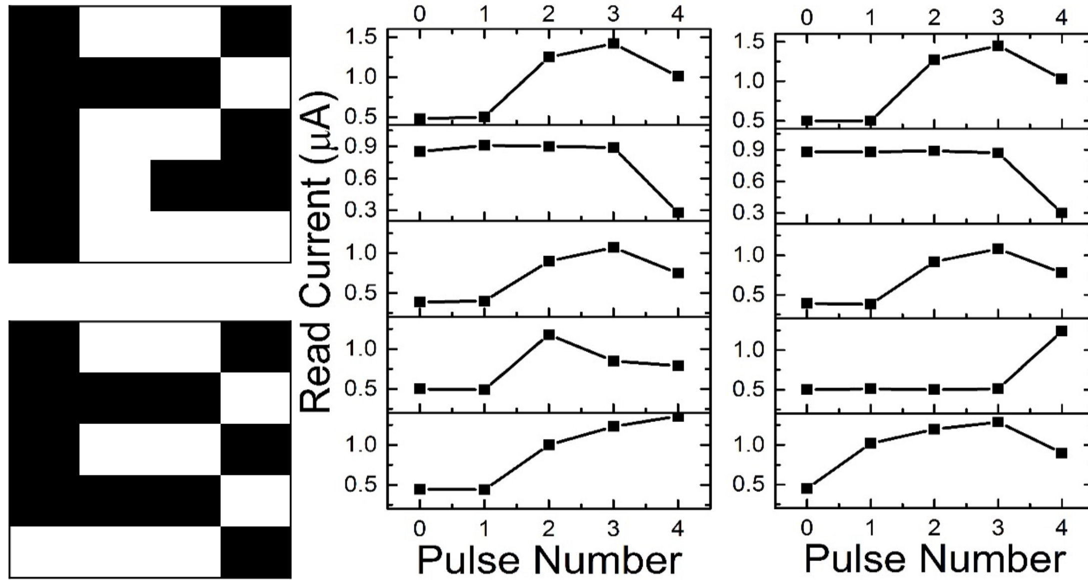

Figure 6-12: The effect of decay in separating different digit images. Noisy digit "2" and "3" images were generated by adding noise to the original data. The corresponding liquid states for these two inputs are shown. The responses to the last two rows are significantly different, enabling the liquid to distinguish these two digits.

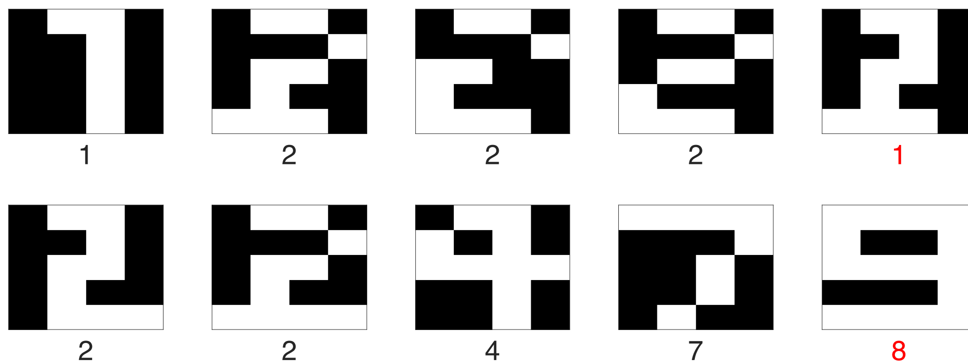

Figure 6-13: Recognition of noisy digits. Noisy digit images were generated by adding noise to the original training samples. The recognition results are shown below each digit image. Most of the noisy digits can still be successfully recognized until too much noise was added, as in the last two cases.

After adding additional noise to the original training samples, the digit images can still be recognized correctly by the system as shown in Figure 6-13. However, if too much noise was added, as in the last two examples shown in the figure, the system will not be able to recognize them, as the liquid states will be too close to other digits thus it becomes very difficult to distinguish those inputs. However, it could be argued that in these two cases, the noisy "2" can indeed be alternatively considered as a noisy "1", while the noisy "9" can in fact be considered as a noisy "8" (with a missing pixel) instead.

6.6 LSM for Handwritten Digit Recognition

The Liquid State Machine was then tested with a more complex but real task, that is, recognition of handwritten digits. The data set, MNIST database (Mixed National Institute of Standards and Technology database) is a large database that is commonly used for training and testing in the field of machine learning. The database was created by "remixing" the digit samples written by high school students and employees of the United States Census Bureau, and consists of 60000 training samples and 10000 test samples. Some of the samples are shown in Figure 6-14.

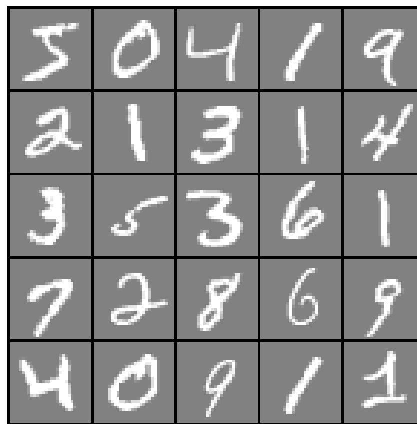

Figure 6-14: Samples from the MNIST database.

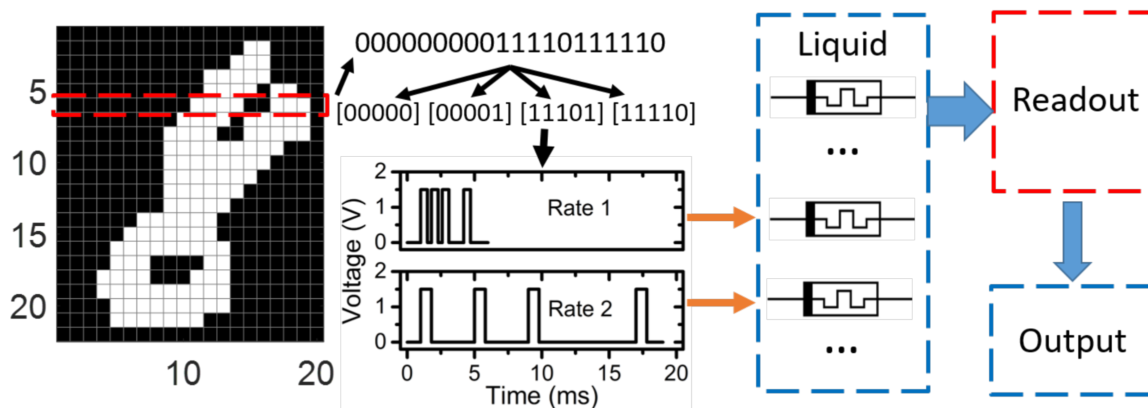

Figure 6-15: LSM for handwritten digit recognition. Image of the digit was preprocessed and transformed into pulse trains. Then pulse trains with different temporal patterns were input to the liquid with different rates. With a trained readout, the recognition results will be obtained.

Figure 6-15 shows the preprocessing of the digit images and the LSM. The original grayscale image was first transformed into a binary-pixel image. The unused boarder area was removed by reducing the original 28-by-28-pixel image into 22-by-20. For each row, there are

now 20 pixels. If the entire row is used as one input pulse train then in theory, there could be 2^{20} different input patterns which may be too difficult for one memristor to distinguish. Therefore, each row should be divided into several sections with each section containing fewer pixels. For the example shown in Figure 6-15, one row was divided into 4 sections and each section contained 5 pixels.

Another strategy to increase the ability of the memristor-based liquid to separate the inputs is to apply pulse trains with different rates/frequencies. If the frequency is very high (compared to the decay time constant of the memristor), the increased conductance by each pulse will not decay much since the interval between pulses will be short. As a result, the number of pules in the input will be the dominant factor to determine the final memristor conductance (the liquid state) due to the cumulative effects of the conductance increase. In the other extreme, if the frequency is very low, the memristor has enough time to decay so the relative timing between pulses will play a more significant role and pulses applied later will also have a stronger effect than pulses applied earlier. These types of effects allow the memristor-based liquid to perform different non-linear transformations of the temporal information in the input and allow different inputs to be linearly separated by reading the liquid states.

With these considerations, each pulse train, converted from each section, was applied with two different rates to the memristor based liquid, as shown in Figure 6-15. From a system point of view, by doubling the total number of input cells, the responses of two memristors fed with pulse trains with the same pattern will capture different features of the input thus improving the separation ability of the liquid. To associate the different liquid states with the digits, the readout network was trained using logistic regression discussed earlier. During training, the liquid states of training samples were recorded and weights in the readout networks were updated following the procedure used in the simple digit recognition task. After training, another set of samples, not in the training set, were used to test the recognition accuracy.

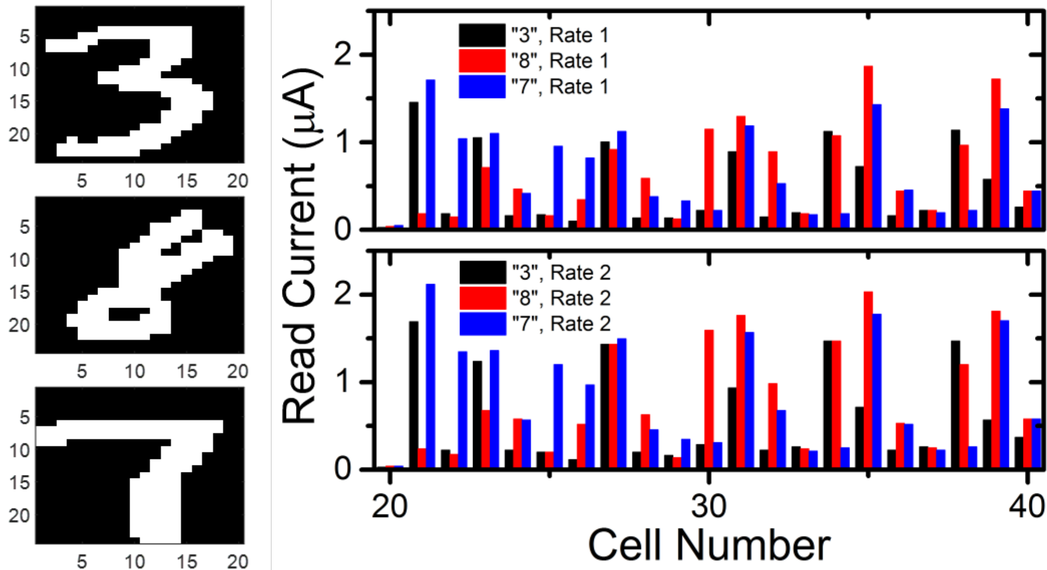

Figure 6-16: Liquid states for three MNIST digit images. Results corresponding to digits "3", "8" and "7" at two different input rates are shown, and significant differences between the digits can be observed.

Figure 6-16 shows the liquid states corresponding to digits "3", "8", "7" at two different input rates, demonstrating that significant difference can be achieved in the liquid states to allow effective separation of the inputs and subsequent classification in the readout network.

The performance of the memristor-based LSM to analyze handwritten digits were tested through both experiment and simulation studies. The results are shown in Table 6-1. From simulation, a system with 96 input cells (24 rows, each row has 2 sections and each section is input at 2 rates) can already achieve more than 90% recognition accuracy. As more inputs are used, a system with 288 inputs (24 rows, 4 sections, 3 rates) can achieve close to 92% accuracy, which is very good for a neural network with no hidden layer (as the single layer readout network needs to be trained) as previously up to 88% accuracy was achieved by one layer neural network. In experiment, 14000 samples from the training sample set and 2000 samples from the test set were used and 88.1% accuracy were obtained from the system, using 4 sections and 2 input rates. Note that not all the samples from the MNIST database were used in the experiment to prevent the device wearout.

	Data Preprocessing	Sample Number	Sections of Each Row	Rates	Recognition Accuracy (%)
Simulation	Reduce to 24 by 24	Train: 60000 Test: 10000	4	3	91.8
				2	91.5

				1	88.7
			3	3	91.6
				2	90.6
			2	2	90.2
Experiment	Reduce to 22 by 20	Train: 14000 Test: 2000	4	2	88.1
				1	85.6

Table 6-1: Experimental and simulation results of handwritten digit recognition by memristor-based LSM.

The results also verified that input with more than one rate indeed improves the recognition accuracy as it improves the system's ability to process temporal information. Also by dividing each row into more sections, each pulse train will have fewer number of pulses, resulting in a better separation by the liquid.

6.7 Discussion

In those two tasks, the spatial information of an image, i.e., the location of pixels in each row, was transformed into temporal information, i.e., the timing of different pulses, and by responding differently to different temporal patterns in the pulse trains, the system can separate different images (digits) as inputs. By using a pulse train containing multiple pulses instead of processing each single pixel in the space domain, we essentially trade speed, that is, the time needed to process the whole pulse train, for space, i.e., the number of inputs and the memory to store weights, and computational power, i.e., fewer weights need to be trained. Fundamentally the LSM performs a non-linear transformation of the inputs and allows the inputs to be separated, and subsequently classified through a readout function based on a simple neural network without hidden layer.

It should be mentioned that the system is not optimized for the handwritten digit recognition task so the performance could still be improved further. First, information from the original data has already been partially lost during the preprocessing, such as transforming to binary data and trimming off borders. Second, the pulse amplitude, width and rates could still be finely tuned for optimum results. More importantly, while normal neural networks aim to extract features across the image from several rows through training, the LSM presented here only processes each row separately and independently. A quick solution, by scanning the digit also in

vertical direction and inputting each column to the liquid to allow relations between the rows to be processed as well, can improve the recognition accuracy to 95.2%.

Additionally, the experimental studies are limited by device wearout. As shown in Figure 6-17, the read current corresponding to consecutive pulses before and after handwritten digit recognition are plotted, and degradation is clearly observed as 1) current level drop and 2) failure to increase conductance by the 3rd pulse after the device was extensively programmed. As a result, the internal states of the liquid will no longer correctly represent different inputs with device degradation, finally becomes unable to distinguish and recognize different digits.

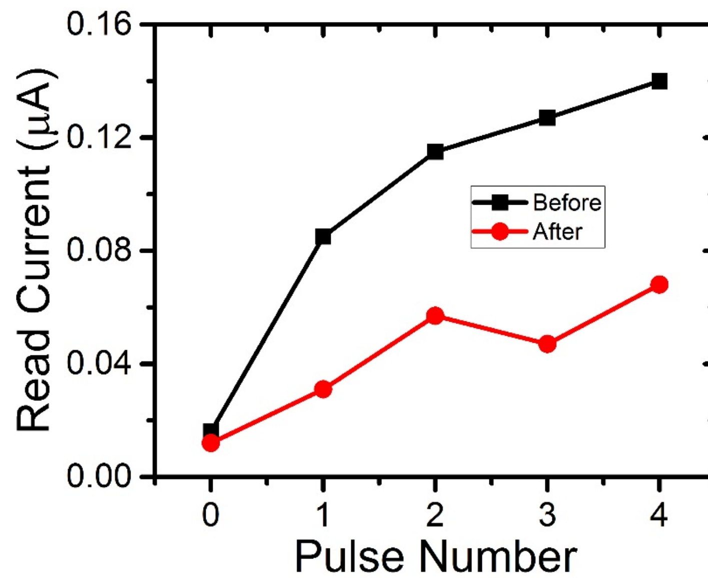

Figure 6-17: Device degradation during digit recognition task. The read currents corresponding to four consecutive pulses before and after handwritten digit recognition task are shown. An obvious degradation is observed as 1) current level drop, 2) failure of conductance increase by 3rd pulse.

In summary, even though the system is not fully optimized for this specific task, we have already successfully demonstrated the ability of memristor-based system to process temporal information and illustrated a way of trading processing speed for less memory and computational power. The memristor-based LSM approach may be more preferred than conventional neural networks for certain applications, e.g., those do not require fast processing speed but have limited resources for memory and computation power.

6.8 Conclusion

Neural networks have been intensively investigated as a new computing architecture for many years. To reduce the computation complexity required for training the whole network, Liquid State Machine, one instance of Reservoir Computing, is proposed as an improved neural network structure, in which only the readout function, i.e. the connections from the liquid to the output, is trained, and can process temporal information due to the short-term memory of the liquid. WO_x memristor, with its internal ionic dynamics, has been used to implement the liquid in LSM. Two tasks, starting from simple digit recognition to more complex handwritten digit recognition, are demonstrated on memritor-based LSM with reasonably good recognition accuracy due to memristor's ability of processing temporal information, both by simulation and experiment on real memristor array.

Reference

1. Wikipedia. Recurrent neural network. Available at: https://en.wikipedia.org/wiki/Recurrent_neural_network.
2. Jaeger, H. *The 'echo state' approach to analysing and training recurrent neural networks. GMD Report, German National Research Center for Information Technology* (2001). doi:citeulike-article-id:9635932
3. Maass, W., Natschlager, T. & Markram, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput* **14,** 2531–2560 (2002).
4. Verstraeten, D., Schrauwen, B., D'Haene, M. & Stroobandt, D. An experimental unification of reservoir computing methods. *Neural Networks* **20,** 391–403 (2007).
5. Maass, W., Markram, H. & Natschläger, T. The‘ liquid computer': A novel strategy for real-time computing on time series. *Spec. Issue Found. Inf. Process. Telemat.* **8,** 39–43 (2002).
6. Natschläger, T. & Maass, W. Spiking neurons and the induction of finite state machines. *Theor. Comput. Sci.* **287,** 251–265 (2002).

7. Verstraeten, D., Schrauwen, B., Stroobandt, D. & Van Campenhout, J. Isolated word recognition with the Liquid State Machine: A case study. *Inf. Process. Lett.* **95,** 521–528 (2005).

8. Burgsteiner, H., Kröll, M., Leopold, A. & Steinbauer, G. Movement prediction from real-world images using a liquid state machine. *Appl. Intell.* **26,** 99–109 (2007).

9. Rebello, J. fmincg() from File Exchange - MATLAB Central. (2013). Available at: https://www.mathworks.com/matlabcentral/fileexchange/42770-logistic-regression-with-regularization-used-to-classify-hand-written-digits/content/Logistic Regression with regularisation/fmincg.m.

Chapter 7

Memristor Performance Improvement

In previous chapters, we discussed the rich dynamics of the internal states of memristors, and how these effects can enable the memristor network to perform real-world applications including sparse coding and handwritten digit recognition, even with memristor arrays that can be readily produced by research labs. However, several challenges still remain before memristor-based neural networks can be commercially used to tackle more complex problems with the desired performance metrics. Below we discuss the challenges and attempts to address them in this study.

7.1 Single Device Performance Improvement

For analog type applications, especially neuromorphic applications, several characteristics are preferred:

1) Better weight update linearity. Many algorithms require the device weight to be updated linearly in response to stimulations during training. For our metal oxide memristors using programming voltage pulses as the stimulation, we prefer the conductance update follows with the number of pulses as linearly as possible. This allows an accurate mapping of the learning rules to the memristor network during training.

2) Larger on/off ratio. This is necessary since with a larger on/off (high/low conductance) ratio, different middle(interstitial) conductance states are more reliably obtained, enabling a larger read margin and more reliable operations.

3) More intermediate conductance states. More intermediate states allow better incremental weight updates. This can probably be achieved from strict control of the stimulation parameter tuning rather than from additional device performance improvements, if a device with

a good linearity and large on/off ratio can be achieved, since weaker programming pulses could increase the conductance with very small steps.

4) Control of the internal dynamics parameters. If we want to use the second order effects to process temporal patterns, better control of those parameters is required.

For the WO_x memristor, the decay of the second order state variable, which represent the ion/vacancy mobility, is most crucial for the memristor dynamics. Therefore, the decay time constant is the most important parameter. However, ion mobility, or the decay of enhanced ion mobility after stimulation is difficult to control, since the physical process is still not well understood and requires further study.

For the Ta_2O_5-TaO_x memristor, the second order state variable, representing the internal temperature which can be enhanced by the heat generated during stimulation and decay afterwards, significantly affects the switching dynamics. As the timing is encoded in the heat decay, we want to control the heat decay time constant. This can be achieved by engineering the thermal conductance of the electrode and switching material, as well as through optimized device structure to tailor the competition between heat generation and dissipation.

Below we first discuss a few attempts to improve the performance of the Ta_2O_5-TaO_x devices, whose switching layers can be modulated during the film deposition process. Then we discuss a few issues during the WO_x device fabrication and solutions.

7.1.1 TaO_x Memristor Performance Improvement

7.1.1.1 Better Linearity During Weight Update

An intrinsic idea about getting better linearity during weight update is to achieve a more uniform oxygen vacancy distribution instead of separated V_O-rich/V_O-deficient layers, therefore the two main driving forces, i.e., drift under electric field and diffusion along V_O concentration gradient, can be more evenly distributed and gradual resistive switching can be achieved without abrupt filament formation/rupture processes.

Following this hypothesis, a device based on a single TaO_x layer was fabricated, by removing the Ta_2O_5 layer from the previous device structure. The single layer device indeed can also be switched, following a much weaker forming process compared with the conventional

device. The weak forming suggests that there is no thick insulator gap between the TE and the TaO_x switching layer, unlike the Ta_2O_5-TaO_x device. A more gradual conductance increase/decrease, as shown in Figure 7-1, can indeed be observed, meaning a better linearity compared to previous results.

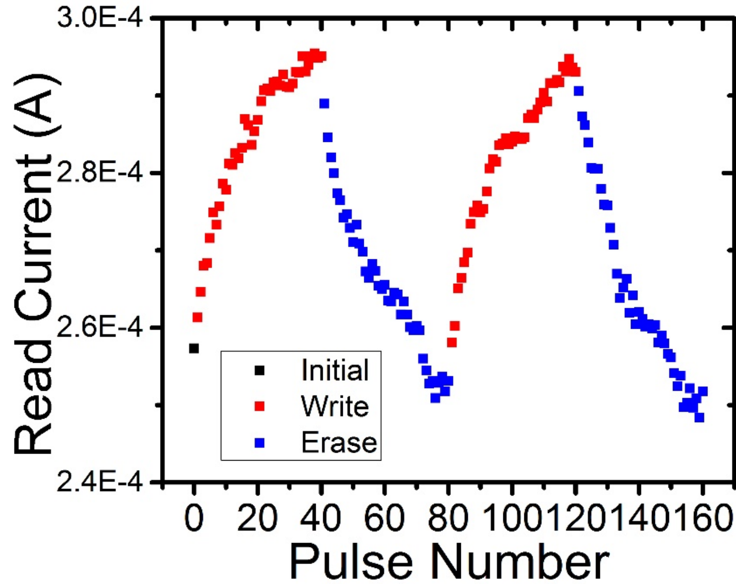

Figure 7-1: Better linearity in a TaO_x single layer memristor. 40 write pulses (-0.9 V, 100 ns) and 40 erase pulses (0.95 V, 100 ns) were applied and the read current was recorded by a read pulse (0.3 V, 50 µs) after each programming pulse.

7.1.1.2 Heat Decay Time Constant Modulation

The heat decay time constant observed in the conventional Ta_2O_5-TaO_x device is typically around 500 ns, which may be too short (for applications where power is more important than speed). Here we aim to achieve a longer heat decay time constant. The heat dissipation equation is

$$\mu C_p \frac{\partial T}{\partial t} - \nabla \cdot \kappa_{th} \nabla T = \gamma \cdot \sigma |\nabla \psi|^2 \qquad (7\text{-}1)$$

where $\mu, C_p, \kappa_{th}, \sigma$ denote the density, specific heat capacity, thermal conductivity and electrical conductivity, respectively and γ is a fitting parameter. It shows that the mass and thermal conductivity of the TE strongly affect the heat dissipation process. Therefore, we could explore different materials as the TE to modulate the heat decay time constant.

Initial attempts of using Pt, which is twice heavier than Pd, has been tried, and the results are shown in Figure 7-2, although more extensive studies are needed to reliably extract the decay time constant.

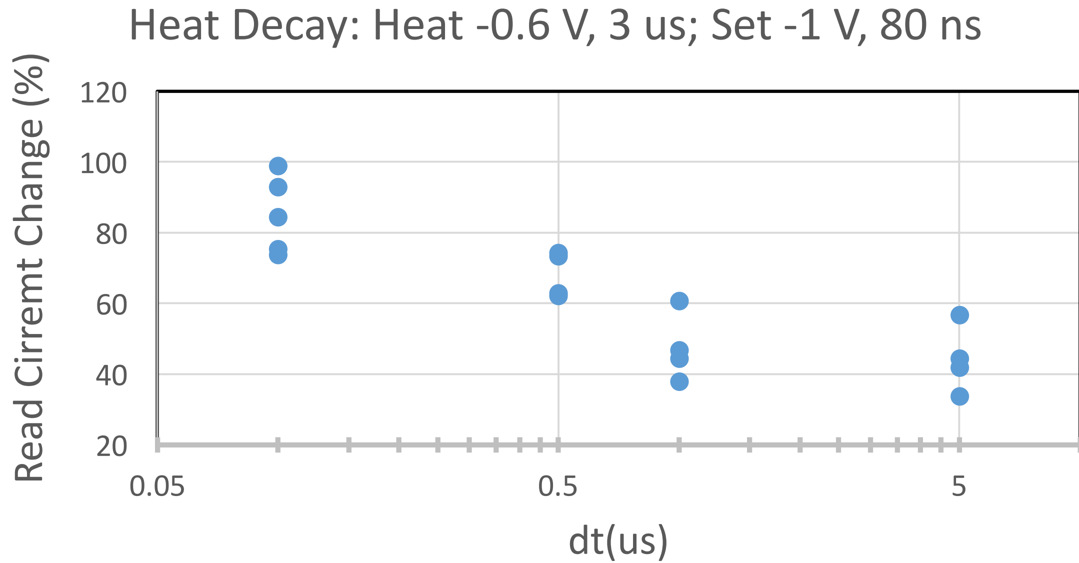

Figure 7-2: Heat decay measured in Ta_2O_5-TaO_x memristor with Pt TE. A heating pulse (-0.6 V, 3 μs) was applied, followed by a SET pulse (-1 V, 80ns) with time interval dt, and the conductance change was recorded as the read current increase. From the results, the heat decay time constant seems to be still shorter than 1 μs.

ITO, which has a much smaller thermal conductivity, has also been tried as the TE. However, it is found that the device with ITO TE is very hard to switch. One possible reason is that the conductance of ITO is too low therefore there is not enough voltage drop on the switching layer although additional studies are still needed.

7.1.2 WO_x Memristor Fabrication Process Optimization

Beyond fundamental device characterizations, we also optimized the processes during the WO_x device fabrication to improve the performance. Attempts on WO_x memristor process optimization are discussed below.

7.1.2.1 Spacer for Top Electrode

Normally two issues exist for the conventional WO_x memristor structure. The first one, as shown in Figure 7-3, is the residues along sidewalls of the bottom electrode after oxidation. We are not quite sure about the chemical nature of the residue. A possible explanation is that some residue could be deposited near the W BEs during the W BE etching step. The residue was then oxidized during the W oxidation process and became expanded and visible.

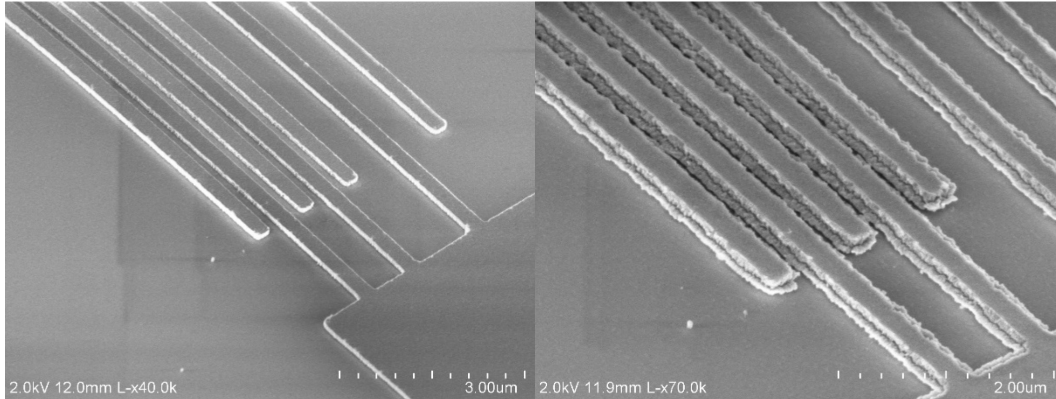

Figure 7-3: Issue of residues along the sidewalls of WO_x memristors. Left: W bottom electrodes with 200 nm line width before oxidation, very clean. Right: W bottom electrodes after oxidation, there are residuals along the sidewalls.

Another issue is the breakage of TEs at the crossing points, which can be amplified by the former issue, although fundamentally it is due to the very thick bottom electrodes after oxidation (approximately 100 nm) together with the bad step coverage of evaporation deposited TE materials, as shown in Figure 7-4.

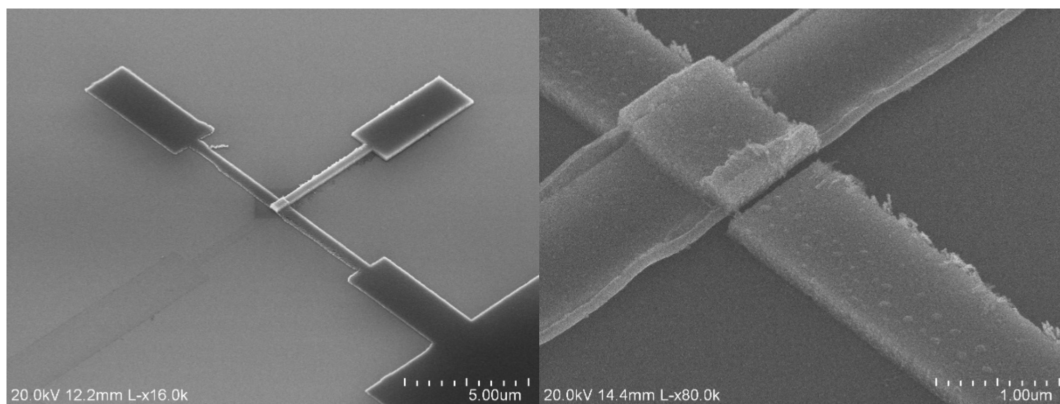

Figure 7-4: Broken TE of WO_x memristor without spacer. Due to the height of the oxide/BE stack, which could be around 100 nm, and the bad step coverage of TE materials. The problem is excerbated by the residue (ridges) around the BE.

To solve this issue, we adopted an improved design, inspired by the spacer structure in CMOS technology. As shown in Figure 7-5, after the W BE was fabricated, a thick layer of SiO_2 is globally deposited by PECVD. Since the deposition is isotropic, the film can cover the edge well. Then the SiO_2 layer was etched back vertically by RIE (reactive ion etching). Due to the nature of this anisotropic etching, the film was removed with equal thickness along the vertical direction, leaving a "spacer" structure self-aligned along the sidewalls of the BE, as shown in Figure 7-6. The width of the spacer is roughly the same as the thickness of the deposited film. The spacer suppresses the ridges formed around the BE and also creates the sloped sidewalls (Figure 7-6) that allows reliable step coverage by the TE.

Figure 7-5: Schematic of spacer formation. Left: after isotropic deposition of SiO_2. Right: after anisotropic etching of SiO_2. The spacer is formed self-aligned along the sidewall of BE.

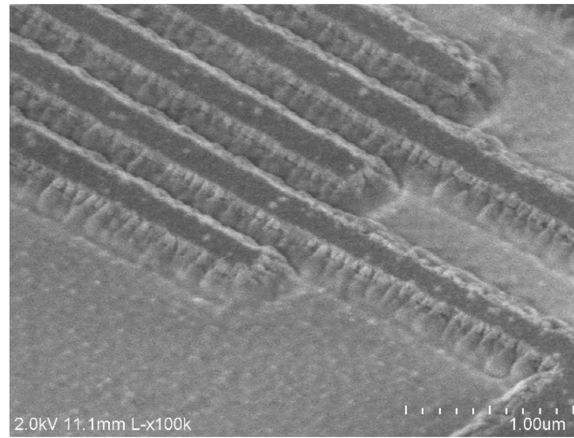

Figure 7-6: SEM image of self-aligned spacer formed along the sidewall of BE.

7.1.2.2 Recessed Structure

Another option to address the step coverage issue would be using a recessed bottom electrode. The basic idea is to bury the bottom electrode, together with the oxide layer, into the substrate, therefore the surface will be flat when the top electrode is deposited. This is inspired

by the flat surface design from crossbars fabricated using nano-imprint technology. There are two obvious advantages of using this technology: 1) the surface will be flat, eliminating the broken TE problem as there is no step height thus no stress, 2) the reaction area is ideally restricted to only the surface of oxide where it contacts the TE, as the BE is buried in the substrate so the sidewall is passivated by surrounding oxide.

Starting from a substrate with a thick enough SiO_2 layer, we first pattern the BE and etch the SiO_2 with appropriate depth, using photoresist as a mask. After that, with resist still on the substrate, W is deposited and lifted-off to form the BE. The subsequent steps are the same as the conventional crossbar structure fabrication while removing the processes for the spacer.

This technology is still under investigation, as the biggest problem associated with the recessed structure is the oxidation conditions. As the contact area with oxygen environment is limited to only the top surface of W during RTA, the WO_x film quality and thickness can be quite different from previous results, therefore careful calibrations are required to ensure the performance of final fab-out devices.

7.2 Memristor Array Performance Improvement

Although we have fabricated small scale WO_x memristor arrays and used them for proof-of-concept neuromorphic applications, additional optimizations at the array level are still required. Currently we can already fabricate WO_x memristor array with line width of 500 nm and pitch of 4 µm as shown in Figure 7-7, with device yield of 100%. The size can be further scaled for several reasons: 1) smaller size cells require smaller programming and read currents, lowering the total power consumption of the chip; 2) reducing pitch size can decrease the series resistance of both TEs and BEs, which causes device variability issue and voltage dividing effect. We have tried cells with 200 nm line width and 500 nm pitch, the yield is high but not 100%. There are a few fabrication processes and parameters to be optimized, such as:

1. Layout design, for example adding redundancy structures at the edges of the array
2. Ebeam lithography dose calibration for desired line width
3. Spacer thickness and etching recipe adjustment

4. Oxidation recipe optimization

5. BE and TE film thickness adjustment

By further optimization of the fabrication methods and parameters for each step of WO_x memristor array, which is an on-going project, we hope to achieve a better yield for larger scale networks as well. As for now, out of a 32-by-32 array, the 28-by-28 sub-array in the center as shown in Figure 7-8 can operate successfully for 200 nm line width.

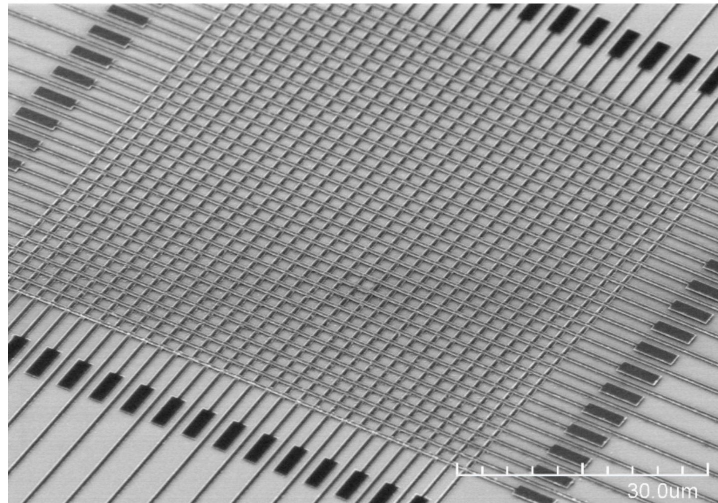

Figure 7-7: 32x32 WO_x memristor array with 500 nm line width and 4 μm pitch. The yield is 100% after fabrication process optimization.

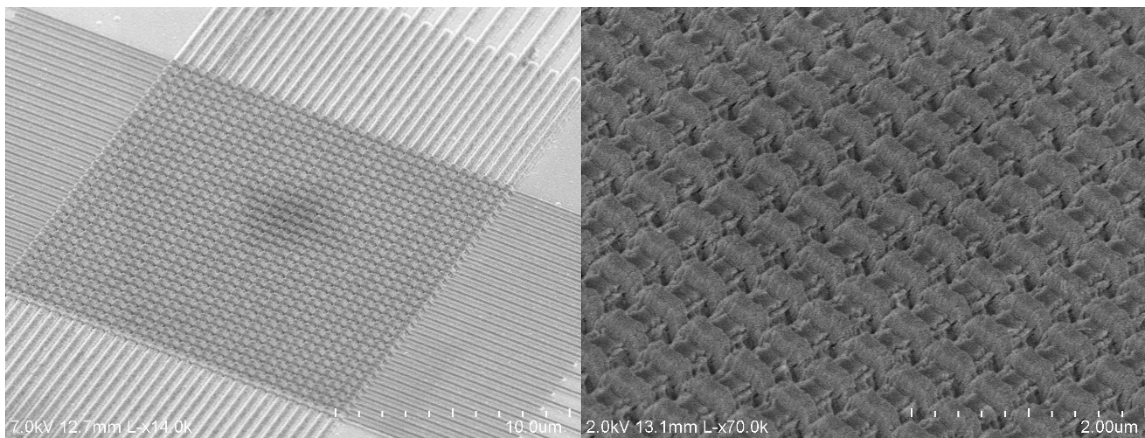

Figure 7-8: 32x32 WO_x memristor array with 200 nm line width and 500 nm pitch. The 28x28 sub-array out of 32x32 array works properly.

7.3 Transition of WO_x Memristor from Analog to Digital Type Switching

During the early stage of memristor research, the application was generally aimed for next generation memory, usually referred to as ReRAM (resistive random access memory), due to its high-density and non-volatility. Memristive devices of this type typically have two resistance states: high resistance state (HRS) and low resistance state (LRS). The device can be switched between these two states by applying certain stimulation, for example voltage pulse above a threshold amplitude. The switching from one state to another happens dramatically, with very few intermediate states that are either unstable or non-recognizable between the two stable states (HRS and LRS). Therefore, the device resembles a digital memory and the way its resistance state switches is recognized as digital type switching and we categorize this type of nonvolatile memristor as digital type memristor.

Analog type memristor, in comparison, has a more gradual resistance change as shown in Figure 2-5. It usually has more stable "intermediate states" that resembles the gradual synaptic weight change in biology.

Additionally, we observed that there could be a transition between digital and analog switching in the WO_x memristor with thinner switching layer and more oxygen vacancies compared with normal analog WO_x memristor devices. The transition can be triggered in two ways:

1) Applying a strong DC sweep with amplitude exceeding a specific threshold

2) Applying a pulse train with sufficient pulse amplitude and pulse number

Here we show results from the second approach as it better resembles processes in biology and is more likely to be implemented in a neuromorphic circuit.

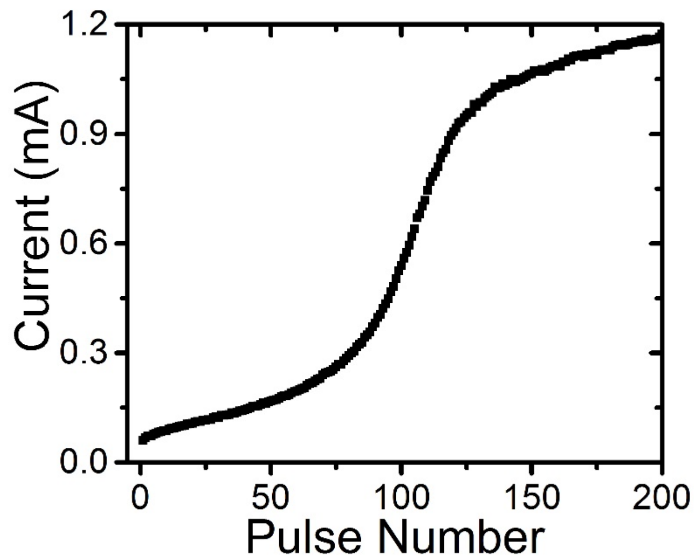

Figure 7-9: Transition of WO_x memristor from analog to digital type switching by multiple pulses. After applying around 100 programming pulses (1.3 V, 100 µs), there was a sharp current increase, indicating a transition.

The transition was demonstrated by applying 200 positive programming pulses (+1.3 V, 100 µs) and the current was recorded and shown in Figure 7-9. For the first tens of pulses the current showed a gradual increase which means the device was still working in analog regime. From around 80th pulse to 120th pulse, there was a sharp current jump and after that the current gradually increased again. The behaviors of the WO_x memristor before and after applying these pulses are also very different, indicating a transition has happened.

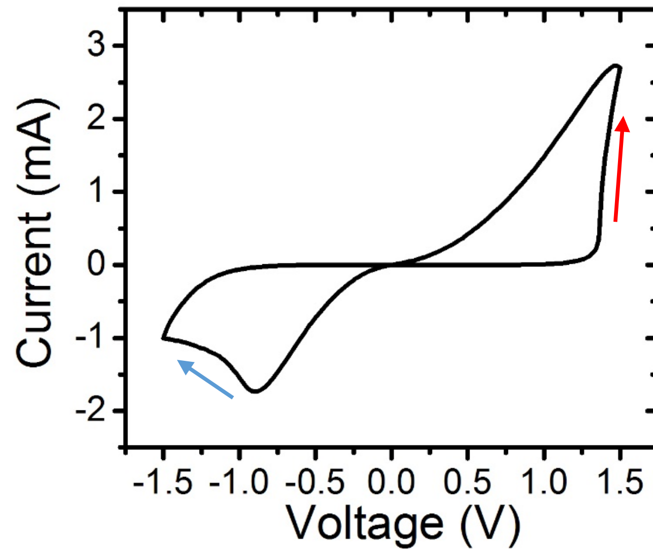

Figure 7-10: Full range DC sweep after transition. A typical nonvolatile memristor (or ReRAM) I-V characteristic was observed. Abrupt SET (red arrow) and RESET (blue arrow) were observed.

A DC sweep on devices after the transition resulted in a typical digital type switching characteristic as shown in Figure 7-10, in which abrupt SET (red arrow) and RESET (blue arrow) processes were observed. For comparison, a DC sweep on WO_x memristor before transition was shown in Figure 7-11, without abrupt SET or RESET process. These results would suggest that after the transition, resistive switching in WO_x memristor is likely dominated by the formation and rupture of very localized conduction filaments inside the switching layer, similar to the case of a typical ReRAM; while for devices before the transition, rather than conduction filaments, an interface-type resistive switching with a gradually moving front between the conductive region and resistive region, likely drives the resistance changes[1], as has been discussed in Chapter 2.

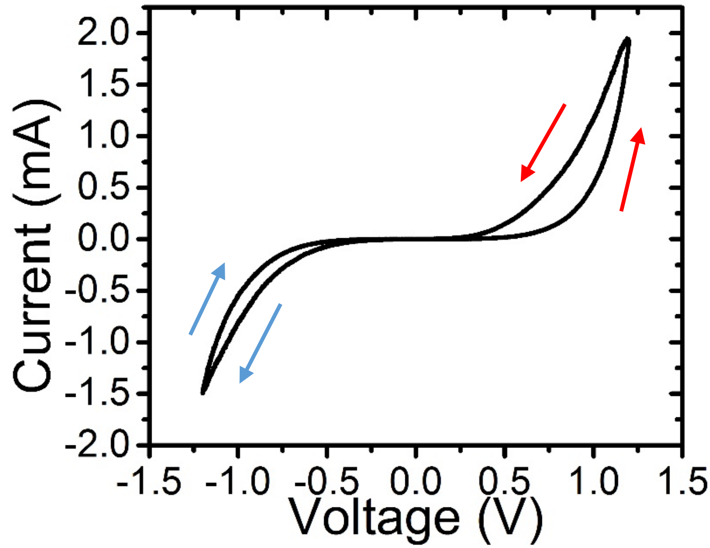

Figure 7-11: DC sweep before transition. Gradual resistance change, i.e. resistance increase with positive voltage sweep (red arrows) and decrease with negative voltage sweep (blue arrows) were observed.

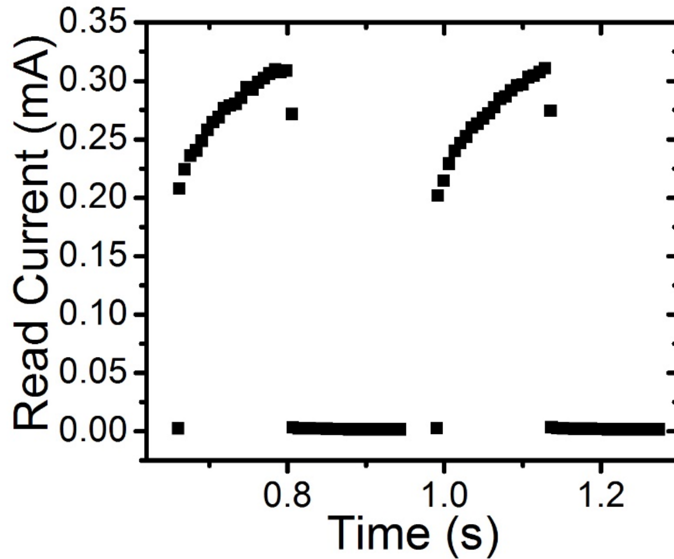

Figure 7-12: Abrupt resistance state change by pulses after transition. 20 write pulse (+1.5 V, 100 µs) and 20 erase pulses (-1.5 V, 100 µs) were applied. Abrupt state change was observed.

Similar effects were observed from pulse tests. Figure 7-12 shows measurement results by applying pulse trains consisting of 20 write pulses (+1.5 V, 100 µs) and 20 erase pulses (-1.5 V, 100 µs). Read currents were recorded with read pulses (0.5 V, 1 ms) after each write or erase pulse. Very abrupt SET and RESET were observed. A single positive pulse can SET the device to LRS due to the formation of conduction filaments, and a single erase pulse is enough to

RESET the device to HRS with almost no intermediate states. The gradual increase in the LRS during subsequent write pulses is likely caused by the widening of the filaments.

Since conduction filaments were formed inside the switching layer during the digital type switching, the retention should be improved as the filaments are typically more stable than a partially doped conductive region.

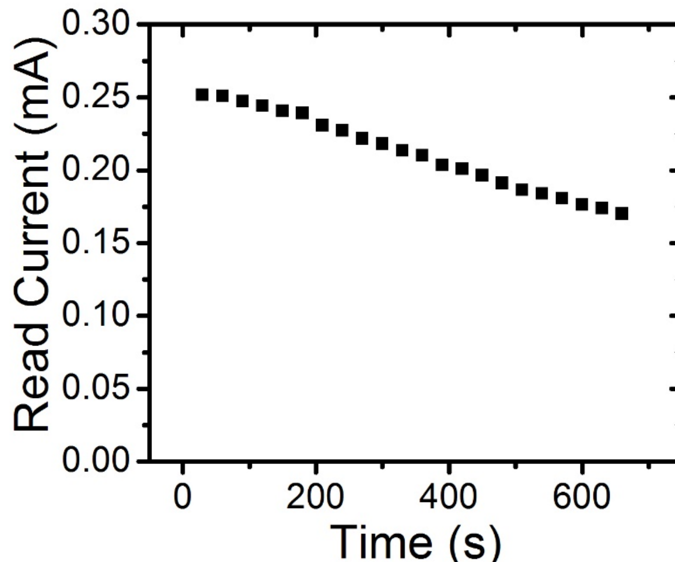

Figure 7-13: Retention test after transition. After SET to the LRS with a DC sweep (up to +1.5 V), the state can be maintained much longer than in analog type device.

As shown in Figure 7-13, retention tests after the transition showed a much longer decay time scale (> 10 minutes) than that of the analog type devices before the transition (e.g. ~ seconds in Figure 3-1).

It should be mentioned that this type of analog-to-digital transition is probably not reversible. There could be some structural changes inside the oxide, leading to the formation of conduction filaments and the entirely different switching behaviors.

7.4 Conclusion

Although some synaptic functions and neuromorphic applications have been implemented or demonstrated by memristors or memristor networks, the performance of memristors still need to be improved. Single device performance improvement has been explored

by optimizing the fabrication processes and employing new materials and designs for both WO_x and TaO_x memristors. For neuromorphic applications, the requirement for large scale memristor networks with higher yield demands further optimization of fabrication methods. By regulating the working region, that is, a transition from analog type switching to digital type switching, WO_x memristor is found to show a more abrupt resistance change (but more linear subsequent conductance increase after being switched to LRS) with better retention, which may be preferred and more suitable for certain neuromorphic applications in future research.

Reference

1. Chang, T. *et al.* Synaptic behaviors and modeling of a metal oxide memristive device. *Appl. Phys. A Mater. Sci. Process.* **102,** 857–863 (2011).

Chapter 8

Future Works

In the previous chapters, we discussed the advances in WO_x and TaO_x based memristors and prototype applications using memristor networks. Several challenges remain to be resolved, requiring more thorough investigations in the future to fully utilize memristors for neuromorphic applications.

8.1 Device Performance Improvement

8.1.1 TaO_x Memristor Analog Behavior Improvement

As discussed in Chapter 7, two aspects of TaO_x memristor analog behaviors still need to be further improved. First, the heat decay time constant still needs to be increased to a more reasonable range for low-power applicable neuromorphic applications. The top electrode materials could be explored further and new structures, for example, a heat insulting layer or a buffer could be added to prevent rapid heat dissipation. Second, a more linear conductance change, with large enough on/off ratio are required for controlled weight updates during learning. As oxygen vacancy movement under electric field, assisted by the elevated temperature, is the main mechanism for conductance change, the solution could exist in but not limited to, 1) optimizing the ion hopping process during oxygen vacancy migration, 2) creating layered oxygen vacancy gradient by depositing multiple TaO_x layers under different oxygen partial pressures, 3) increasing the lateral movement of oxygen vacancies during switching. All these concepts need further investigation.

8.1.2 WO_x Memristor Fabrication Methods Optimization

As discussed in Chapter 7, there are several fabrication options to further improve the performance of the WO_x memristor and the crossbar array. The critical consideration is still to control the quality of the oxide, limit the switching location, and resolve the issue associated with the height growth of vertical structure. Further research will be needed to extensively test these concepts.

8.2 Memristor Array for Reservoir Computing

8.2.1 Memristor Array Optimization

As has already been discussed in Chapter 6, the simple 1-by-n memristor array is not optimized for handwritten digit recognition. The recognition accuracy should be further improved by tuning the pulse parameters, adding more input cells by different image scanning methods.

However, the performance should not be improved significantly due to the intrinsic disadvantage of the proposed memristor array, that is, each memristor is performing information processing separately, without interaction or communication between each other.

One solution could be introducing loops inside the memristor array. For example, some terminals of the 32x32 array can be connected randomly to create loops inside the array. However, while the loops in neural networks are formed by connecting neurons so the signal can be transmitted without loss or even with amplification, the loops formed by directly connecting memristors(synapses) will consume power and weaken the strength of signal along the path, causing the dissipation of information in the network. A more thorough investigation is required in the future.

8.2.2 Other Neuromorphic Applications

The Liquid State Machine is a great application utilizing memristor's dynamics for temporal information processing. In this research, the system has been used for handwritten digit

recognition. However, handwritten digit recognition is probably not the best application to utilize memristor's temporal information processing capabilities.

A more suitable, or benchmark-like task would be the implementation of non-linear systems through reservoir computing.

For example, in a 2nd order non-linear system[1], the input is

$$u(k) = random([0,\ 0.5]) \tag{8-1}$$

and the output is

$$y(k+1) = 0.4y(k) + 0.4y(k)y(k-1) + 0.6u^3(k) + 0.1 \tag{8-2}$$

As can be seen from the equation, *y(k+1)* depends on *y(k)* and *y(k-1)*, so this is a 2nd order non-linear system.

The Liquid State Machine can be used to implement this system efficiently. The input *u(k)*, after linear amplification and shift (since the original range is 0 to 0.5 so we need to transform it to voltage pulses with a range of 0.5 V-1.5 V as this is the normal range for write pulse of the memristors), is applied on a memristor and the corresponding state (read current) is recorded. For a certain input pulse train, the read current for each pulse is determined by not only the current input *u(k)* but also the previously applied pulses within a certain period, as the effect of earlier pulses will decay and eventually vanish. Therefore, the system has some short-term memory and appears to be very suitable for this type of tasks such as mapping Equation (8-2) to a physical system. The read current, reflecting the internal state of the liquid, will be used to generate the desired output *y(k+1*) through a trained readout network, similar to the approach discussed in Chapter 6. Moreover, the same input could be fed to different cells and the variations among different cells, which is normally considered detrimental for other large-scale applications that require good device uniformity, can be utilized to obtain different responses from different memristors, thus further improving the network's ability to separate the inputs.

More research needs to be performed with a focus on this type of applications to fully explore the potential of using the internal dynamics of memristors for efficient information processing in the temporal domain.

Reference

1. Atiya, A. F. & Parlos, A. G. New results on recurrent network training: unifying the algorithms and accelerating convergence. *IEEE Trans. Neural Networks* **11,** 697–709 (2000).