Impact of RTN and Variability on RRAM-Based Neural Network

P. Freitas, Z. Chai, *W. Zhang, J. F. Zhang, J. Marsland
Department of Electronics and Electrical Engineering, Liverpool John Moores University, Liverpool L3 3AF, UK
*Corresponding author, email: w.zhang@ljmu.ac.uk

Abstract

Resistive switching memory devices can be categorized into filamentary RRAM or non-filamentary RRAM depending on the switching mechanisms. Both types of RRAM devices have been studied as novel synaptic devices in hardware neural networks. In this work, we analyze the amplitude of Random Telegraph Noise (RTN) and program-induced variabilities in both TaO_X/Ta_2O_5 filamentary and TiO_2/a-Si (a-VMCO) non-filamentary RRAM devices and evaluate their impact on the pattern recognition accuracy of neural networks. It is revealed that the non-filamentary RRAM has a tighter RTN amplitude distribution than its filamentary counterpart, and also has much lower programed-induced variability, which lead to much smaller impact on the recognition accuracy, making it a promising candidate in synaptic application.

1. Introduction

Oxide based resistive switching memory devices (RRAM) has emerged as an attractive candidate not only for the next-generation emerging memory technology [1-2], but also as synapses in large-scale artificial neural networks (ANNs) due to its natural synaptic response, simple structure, low energy consumption, and CMOS-compatible 3D integration potential [1]. There are mainly two types of transition-metal-oxide (TMO) based resistive switching devices (RRAM): the conductive filamentary type (CF) that can be implemented with a range of materials, for example, HfO_2 and Ta_2O_5 [2], etc.; and the non-filamentary type (NCF) such as TiO_2/a-Si a-VMCO [3].

In filamentary RRAM devices, variable resistance is induced by repeatable rupture and restoration of a conductive filament (CF) of nanometer scale. The large variations in read current distribution at high resistance state (HRS) is a major concern, as it deteriorates the resistive switching window and causes endurance and retention problems. This has been attributed to defects movement into/out of the constriction of the filament where only a few defects exist. The conductance of individual defect in the constriction has significant impact on the overall resistance levels, and the stochastic nature of individual defect causes large resistance variability and large read instability [4]. Non-filamentary RRAM (NCF) devices have been proposed to overcome the above problems, in which the resistance switching is controlled through the uniform modulation of the defect profile [3]. The a-VMCO RRAM device consists of two layers, in which TiO_2 serves as the switching layer and amorphous-Si as the barrier layer. The non-filamentary switching behavior is demonstrated as its resistance is inversely proportional to the area at both HRS and LRS, and the resistance distributions at both HRS and LRS in a-VMCO RRAM show smaller variations [3, 5].

Random Telegraph Noise (RTN) is the current fluctuation between discrete levels caused by electron trapping and de-trapping in defects. RTN has become a critical issue in nanoscale semiconductor devices where the impact of a single defect becomes significant [6, 7, 8]. As RRAM devices can be scaled down below 10 nm [2], RTN can significantly reduce the memory window and cause read errors in RRAM devices. It is therefore essential to evaluate the impact of RTN disturbance on the performance of RRAM-based synaptic arrays. Program-induced conductance variability in both filamentary and non-filamentary RRAM devices also need to be evaluated in synaptic applications.

In this work, we analyze the amplitude distributions of RTN and program-induced variability in both Ta_2O_5 CF RRAM and TiO_2/a-Si (a-VMCO) NCF RRAM devices. The experimental results are used to simulate their impact on the synapse arrays in a trained artificial neural network. It is revealed that the NCF RRAM has a tighter RTN amplitude distribution and smaller program-induced variability than its filamentary counterpart, leading to much less impact on pattern recognition accuracy and making it a promising candidate as synapse in neural network applications.

2. Devices and Experiments

Both types of RRAM devices were fabricated in a cross-point structure with the size of 75 nm × 75 nm and show bipolar switching characteristics (Fig. 1(a) and (b)). The Ta_2O_5 device consists of a TiN/4nm stoichiometric

Ta_2O_5/20nm nonstoichiometric TaOx/10nm TaN/TiN stack (inset of Fig. 1 (a)). The a-VMCO device has a stack of TiN/8nm amorphous-Si/8nm anatase TiO_2/TiN structure (inset of Fig. 1(b)). The detailed process parameters can be found in refs. [5,9]. All electrical tests were carried out with a Keysight B1500A analyzer. Analogue resistance levels are obtained in both devices, between 25 kΩ and 200 kΩ for Ta_2O_5, and between 1 MΩ and 7.5 MΩ for aVMCO, by incrementing the program pulse number and amplitude. The read-out is at 0.1V and 3V for Ta_2O_5 and aVMCO devices, respectively, by a read pulse width of 100 us. RTN measurement is then carried out at each R level at the read-out voltage, with a sampling time of 2 ms/point and 10,000 sampling points per resistance level for a RTN measurement period of 20 s. A 3-layer ANN was simulated using Matlab. The neural network was trained and tested with the MNIST handwritten digit database. Out of the total 60,000 images, 50,000 were used for training and the remaining 10,000 images unseen during training were used for testing.

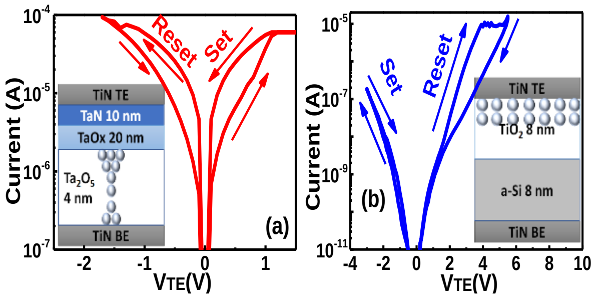

Fig.1 I-V switching curves of (a) Ta_2O_5 and (b) aVMCO devices; The insets are the schematics of the corresponding structures and the switching mechanism: the restore/rupture of a conductive filament (CF) or the areal modulation of defect distribution inside the oxide (NCF).

3. Results and discussions

3.1 RTN signals in CF and NCF RRAM devices

As shown in Fig. 2(a) and (b), the maximum relative RTN amplitude, ΔI/Iread, can be as high as ~300% in the filamentary Ta_2O_5 RRAM device, but only ~10% in the non-filamentary aVMCO. Their CDF distribution plots measured at 8 selected resistance levels are shown in Fig. 2 (c) and (d), respectively. RTN amplitude in Ta_2O_5 device spreads widely from 0.1% to 300%, whilst it is only from 1% to 10% in aVMCO. For both devices, the RTN amplitude follows the lognormal distribution. Moreover, RTN in Ta_2O_5 device has a much higher occurrence rate than in aVMCO device (not shown). The parameters of the distributions are extracted and shown in Fig. 2e & 2f, which will be used in the simulation.

This significant difference in RTN amplitude distribution and occurrence rate can be attributed to the different switching mechanisms, as shown in the insets in Fig.1: in the CF Ta_2O_5 device, the resistance switching is caused by the rupture and restoration of a conductive filament. After the reset, there are only a few defects in the constriction of the CF, and each of them is critical in current conduction, so that its trapping / detrapping leads to large RTN, and hence the higher the resistance level, the larger the RTN amplitude. In the NCF aVMCO device, resistance switching is caused by the uniform modulation of defect distribution. Resistance becomes higher when the “defect-less” region is uniformly widened. A single defect has limited contribution in conduction, hence the much smaller RTN amplitude, and much smaller occurrence rate (not shown), and the amplitude is also only slightly larger at higher resistance levels.

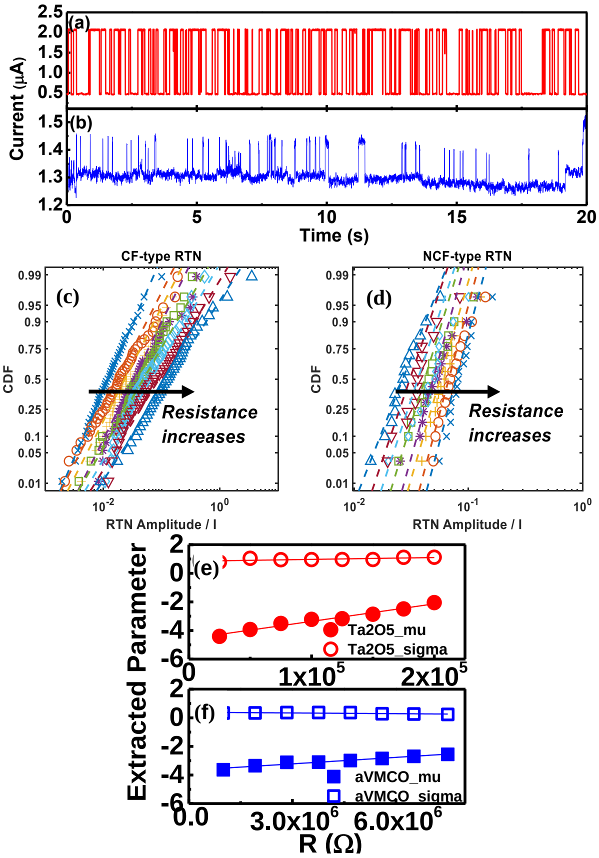

Fig. 2 (a-b) Examples of largest RTN signal in (a) Ta_2O_5 and (b) a-VMCO devices. The relative RTN amplitude can be as high as ~300% for Ta_2O_5 device, but only ~10% for a-VMCO. (c-d) CDF of relative RTN amplitude in (c) Ta_2O_5 and (d) a-VMCO devices, respectively, both following the lognormal distribution. (e-f) Extracted parameters of lognormal RTN amplitude distribution in both devices.

3.2 Program-Induced Variability in CF and NCF RRAM devices

The program-induced variability is defined as the relative variation at a target conductance level, i.e. ΔG/G induced by the programming. In Fig. 3a, four curves programmed at different constant pulse amplitudes in an NCF aVMCO RRAM device are shown as an example, where the typical exponential program kinetics are observed. A linear program approximation can be achieved in a small range on each curve, which is similar to the small signal approximation in AC circuit analysis. By applying a number of smaller identical pulses in each range and incrementing the bias in consecutive ranges, a much improved linearity in the program kinetics can be achieved during both set and reset operations, as shown in **Fig. 3b**.

The program-induced variability obtained in CF and NF devices with the linear response are compared in Fig. 3c and 3d. The program-induced variability in CF device has a wider distribution, leading to much larger variability than that in the NF device. In both devices, the distributions of the relative conductance variability are largely independent of the conductance levels. This allows the use of the observed distribution function to reproduce the variability distribution at any target conductance levels in simulation. which will be demonstrated later.

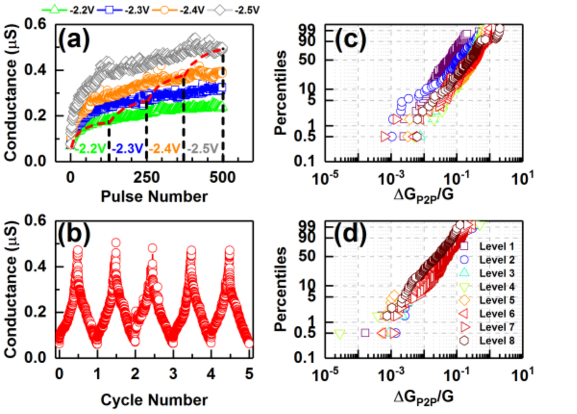

Fig.3 (a) Illustration of the program kinetics in an NCF RRAM. (b) Linear response can be achieved by applying a number of smaller identical pulses in each range and incrementing the bias in consecutive ranges. (c) Distribution of the program-induced variability at 8 selected conductance levels across the memory window in CF RRAM and (d) in NCF RRAM.

3.3 Impact of RTN on NN accuracy

The impact of RTN on the pattern recognition accuracy of RRAM based synaptic neural network is analysed first. The neural network consists of 3 layers with 30 neurons in the hidden layer, as shown in Fig. 4(a). The neural network is trained with the mini-batch gradient descent backpropagation algorithm. The accuracy after training without and with the RTN induced disturbance in both CF and NCF RRAM are statistically shown in Fig. 4(b). The change of weights in one of these procedures is visualized in Fig. 4(c), in which the weights are shown in (1) without disturbance, (2) after the CF disturbance, and (3) after the NCF disturbance. The weight differences are shown in (4) after CF disturbance and (5) after NCF disturbance.

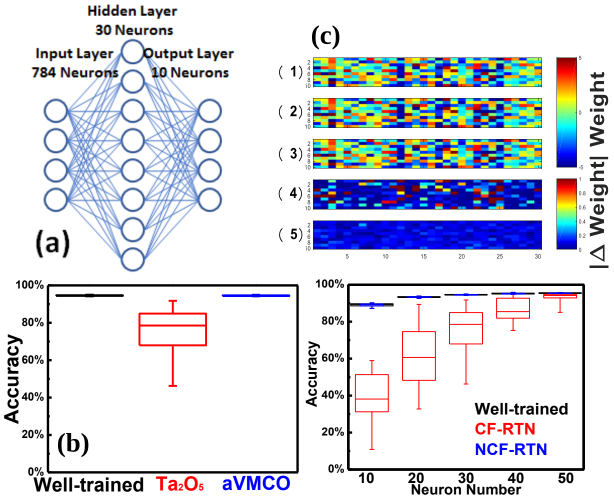

Fig. 4. (a) Schematic of the pattern recognition ANN. (b) Statistical accuracy in 50 training-disturbance procedures: Accuracy is hardly affected with the NCF disturbance, while with the CF disturbance the accuracy is severely deteriorated. (c) Visualization of weights: (1) directly after training; (2) with CF RTN disturbance; (3) with NCF RTN disturbance; (4-5) their differences to case (1), respectively. (d) Accuracy of ANN with different neuron number without and with CF and NCF RTN disturbance. ANN with NCF devices needs fewer neurons/synapse and have better accuracy.

As shown in Fig. 4(b), after the CF RTN disturbance, the average accuracy drops to ~75% with a wide repeatability distribution and its lowest is less than 50%, while after the NCF disturbance the accuracy drops negligibly only to 94% with a similar repeatability to that without disturbance, as can also be clearly seen in the weight differences shown in Fig. 4(c). This proves that the non-filamentary RRAM device has a strong advantage compared to the conventional filamentary devices in the synaptic application, due to its small RTN amplitude and low RTN occurrence rate. Furthermore, as shown in Fig.4(d), the neural network with NCF synaptic devices maintains a high accuracy of ~90% even when only 10 neurons are used in the hidden layer, whilst the accuracy drops sharply with the CF devices. NCF synaptic devices allows a much smaller ANN to achieve better accuracy due to its robust RTN resilience, therefore.

3.4 Impact of program-induced variability on NN accuracy

The impact of program-induced variability on the pattern recognition accuracy in CF and NCF RRAM devices are compared in Fig. 5. The accuracy loss caused by program-induced variability in non-filamentary RRAM device is significantly smaller than that in the filamentary RRAM device, thanks to the much smaller variability. RTN-induced accuracy loss is also shown for comparison, and it is even larger than that induced by the program-induced variability. The overall pattern recognition accuracy is limited by RTN in the NCF devices. Program-induced variability in CF RRAM and RTN induced variability in NCF RRAM are the most significant sources responsible for accuracy loss, respectively.

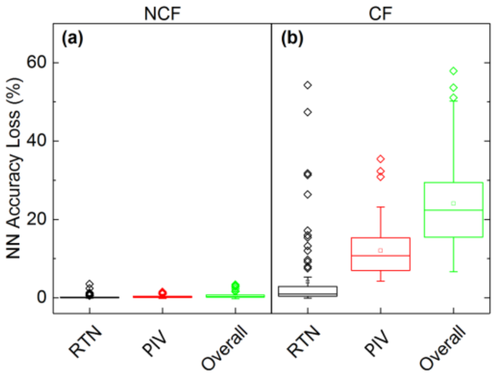

Fig. 5. Comparison of the NN accuracy loss caused by different sources of variability of (a) NCF RRAM and (b) CF RRAM, including RTN and program-induced variability (PIV) programmed with linear (LR) kinetics. NF device shows much lower variability-induced accuracy loss compared with its CF counterpart. RTN in NCF RRAM and program-induced variability in CF RRAM are the dominant sources of accuracy loss.

The difference in program-induced variability in NF and CF RRAMs can also be attributed to their different switching mechanisms. The area-dependent resistance in non-filamentary devices supports that the resistance is uniformly modulated across the lateral device area during the switching and individual defect movements are averaged out, leading to the much smaller variability in device conductance. In contrast, the area-independent resistance in conductive filamentary devices supports that the switching is controlled by the rupture and restoration of one local filament between the two electrodes. where individual defects play a significant role in the conductance change of the CF device. This translates into not only a more pronounced RTN amplitude with wider distribution, but also in higher program-induced variability. Non-filamentary RRAM demonstrates far better immunity to variability and hence better inference accuracy in HNN applications, therefore.

3.5 Conclusions

In this paper, two different variability sources are statistically measured and evaluated at different conductance levels across the memory window in both conductive-filamentary and non-filamentary RRAM devices. Based on the statistical distributions of the program-induced variability and RTN in both NF and CF RRAM devices, their impact on the pattern recognition accuracy of a RRAM-based 3-layer feedforward HNN are simulated and compared. It is revealed that NF device shows much lower variability and accuracy loss than its CF counterpart. RTN remains a major variability source in both devices. A comparison between the NF and CF switching mechanisms can explain the differences in the variability and their distributions.

Acknowledgments

This work was supported by the Engineering and Physical Science Research Council of UK under the grant nos. EP/M006727/1 and EP/S000259/1. The authors would like to thank colleagues at IMEC, Belgium, for supply of test samples used in this work and fruitful discussions.

References

[1] Wong et al, IEEE proc., 2012.
[2] Govoreanu et al, IEDM, 2011.
[3] Govoreanu, et al, VLSI Symp. Tech. Dig., 2015
[4] Degraeve et al, VLSI, 2012.
[5] Ma, et al, IEDM 2016.
[6] Chai, et al, VLSI Symp. Tech. Dig., 2016
[7] Chai, et al, IEEE, TED, 2017
[8] Kirton et al, Advances in Physics, 1989.
[9] Chai et al, IEEE EDL, 39 :955-958, 2018.