Factorial Hidden Markov Model analysis of Random Telegraph Noise in Resistive Random Access Memories

Francesco Maria Puglisi^{*1} and **Paolo Pavan***, Non-members

ABSTRACT

This paper presents a new technique to analyze the characteristics of multi-level random telegraph noise (RTN). RTN is defined as an abrupt switching of either the current or the voltage between discrete values as a result of trapping/de-trapping activity. RTN signal properties are deduced exploiting a factorial hidden Markov model (FHMM). The proposed method considers the measured multi-level RTN as a superposition of many two-levels RTNs, each represented by a Markov chain and associated to a single trap, and it is used to retrieve the statistical properties of each chain. These properties (i.e. dwell times and amplitude) are directly related to physical properties of each trap.

Keywords: RTN, Multi-level, FHMM, Trapping, Noise.

1. INTRODUCTION

Random telegraph noise (RTN) is usually found in metal-oxide-semiconductor field-effect transitors (MOSFETs) and in other novel devices (e.g. Resistive Random Access Memeories, RRAM [1-3]) as an abrupt and random change of either the voltage or the current between discrete levels. This can result in unpredictable deviation of key parameters (e.g. threshold voltage of MOSFETs, read current in RRAMs) from their expected values. Currently, RTN is becoming a challenging issue limiting the full industrial exploitation of RRAM concepts and their reliability since its effects are expected to be even more severe in nanoscale devices. Even though the physical mechanisms responsible for RTN have not been completely assessed, it is commonly accepted that it is the result of capture and emission processes of charge carriers in/from defect centers acting as traps [4]. Fig. 1 shows our interpretation of the mechanism leading to both two-levels (a) and multi-level (b) RTN in metal-oxide-based RRAMs in High Resistance State (HRS) along with experimental time series [5,6]. Our previous works [5,6] exploited the color-coded time-lag plots to investigate the nature of RTN in RRAMs, showing that multi-level RTN can be seen as a superposition of many two-levels RTNs. Formerly, we used hidden Markov model [7] (HMM) to investigate RTN. However, HMM can be used to extract the discrete levels in the current but cannot be used directly to extract the amplitude of each two-levels fluctuation in multi-level RTN.

In this paper, we propose a more refined implementation of the HMM which is best suited to solve for the statistical properties of multi-level RTN caused by multiple traps. Retrieving traps parameters is of utmost importance to gain a deeper understanding of the phsyical mechanisms leading to RTN since they are strictly related to the physical properties of the defect centers, such as their positions and energies in the oxide layer and their relaxation energy (taking into account the structural lattice relaxation occuring during the trapping and detrapping of charge carriers) [4]. Traps parameters are estimated using a factorial hidden Markov model [8,9] (FHMM) approach: the the proposed method is self-consistent (the number of active traps is automatically determined) and its implementation can be parallelized, leading to better performances with respect to other methods such as Markov chain Monte Carlo-based techniques [10]. This paper is organized as follows: the mathematical description is given in Section 2, underlining the limitations of HMM approach when analyzing multi-level RTN and proposing FHMM; in Section 3 we report results and discussion. Conclusions follow.

2. STATISTICS BACKGROUND

The capture/emission process of charge carriers in/from traps into the barrier, Fig. 1, can be described by a Hidden Markov Model, i.e. a Markov (memoryless) process with unobserved (hidden) states. Whereas in simple Markov models the state of the system at each instant of time is directly visible to the observer, in HMM the output of the system is directly visible at each instant of time while the state of the system is hidden, even though the output strictly depends on the state. Each state is characterizied by a probability distribution over all the possible values assumed by the output, statistically linking the sequence of observations (output) to the sequence of hidden states. Moreover, each state is associated to a set of transition probabilities (one per each state) defining how likely is for the system,

Manuscript received on January 15, 2014 ; revised on February 2, 2014.

*The authors are with with DIEF - Università di Modena e Reggio Emilia, Italy, E-mail: francescomaria.puglisi@unimore.it[1]

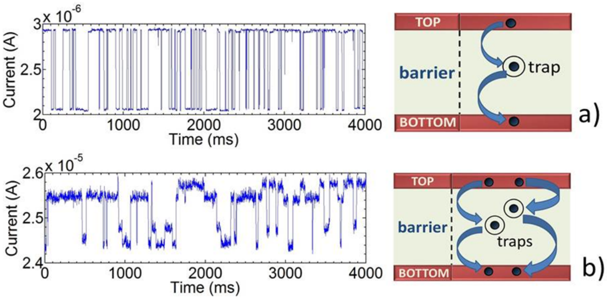

Fig.1: *a) Experimental two-levels RTN and its simplified physical mechanism involving one trap only. b) Experimental multi-level RTN and its simplified physical mechanism involving two (many) traps. Black spots are the charge carriers and the black holes are active traps. Metal-oxide-based RRAM conduction in HRS is modelled as trap-assisted tunneling via the traps in the barrier, leading to 2-levels (a) or multi-level RTN (b).*

being in a given state at a given instant of time, to switch to another of the possible states (including the same state) at the successive instant of time.

2.1 The probabilistic model of HMM

In HMM, a sequence of observations $\{Yt\}\ t = 1...T$ is modeled by specifying a probabilistic relation between the observations and a set of hidden (unknown a priori) states St through a Markov transition structure linking the states. In this framework the state is represented by a random variable assuming one out of N values at each instant of time. The HMM approach relies on two conditional independence assumptions:

1) S_t only depends on S_{t-1} (known as the first-order Markov or "memory-less" property)

2) Y_t is independent of all other observations $Y_1, ..., Y_{t-1}, Y_{t+1}, ..., Y_T$ given S_t.

The joint probability for the state sequence and observations can be formalized as:

$$\begin{aligned} P(S_t|y_t) = P_{init} \cdot \prod_{t=2}^{T} P(S_t|St-1) \cdot P(Y_t|S_t) \\ P_{init} = P(S_1) \cdot P(Y_1|S_1) \end{aligned} \tag{1}$$

A schematic representation of the HMM is given in Fig. 2, where the Markov property is evidenced. According to the formalism used by Rabiner in [7], an HMM is completely defined as a 5-tuple (N, M, A, B, π). N is the number of hidden states, S, in the model (i.e. the number of discrete current levels to be found in RTN); since observations assume discrete values, M is defined as the number of distinct observable symbols (i.e. the possible current values assumed by RTN). A is an N-by-N matrix defining the transition probabilities among states and B is a N-by-M matrix defining the observation probability of each observable symbol in each hidden state; π is a vector defining the initial state probability distribution [7]. The inference problem in this model consists in finding the most likely set of probability of hidden states given the observations. This is achieved through a maximum likelihood estimate of the HMM parameters given the observations using the forward-backward algorithm [7]. Then the most likely sequence of hidden states representing the dynamics of the observations can be achieved via the "Viterbi" algorithm, a dynamic programming paradigm. As a result, HMM analysis can efficiently estimate the discrete current levels and the best sequence of states representing RTN data, as shown in Fig. 3(a).

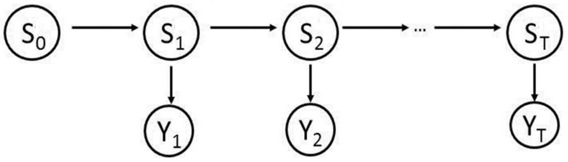

Fig.2: *Graphical representation of an HMM. At each instant of time t, each output Yt is related only to the current state of the Markov chain defining the model.*

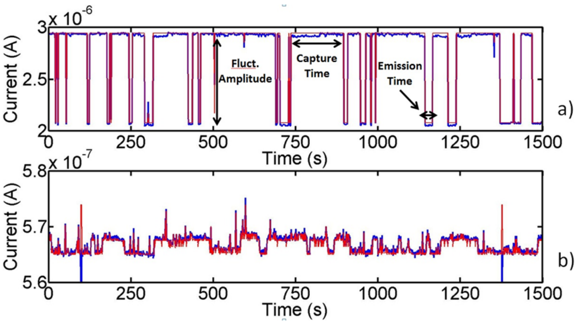

Fig.3: *a) Experimental two-levels RTN and HMM fitting. The hidden levels and most likely state sequence are correctly retrieved. The distinctive features of the two-levels RTN (amplitude of the fluctuation and average capture/emission times) are evidenced. b) Experimental multi-level RTN and HMM fiting. Though succesful in characterizing the hidden states and their durations, HMM output is unsufficient to achieve a comprehensive characterization of all the traps leading to the observed noise.*

2.2 The limitations of HMM for multi-level RTN analysis

Even though being effective in capturing the Markov dynamics of RTN, HMM is nevertheless unsuitable to comprehensively characterize multi-level RTN, although is still valid to fully describe a two-levels RTN (for which $N = 2$). Indeed, the distinctive feature of a two-levels RTN fluctuation, related to the physical properties of the associated trap, are the amplitude of the fluctuation and the average capture/emission times, see Fig. 3(a). While dealing with a two-levels fluctuation, the output of the HMM analysis is sufficient to extrapolate all the characteristic features of the RTN data: the amplitude of the fluctuation is simply given by the difference between the two hidden states while the average capture and emission times can be extracted by averaging the duration of each state. In the case of $multi-level\ RTN$, a superposition of many two-levels RTNs each generated by the contribution of a single trap [5,6], a complete characterization would be achieved only by defining all the distinctive features of every trap contributing to the observed RTN. Unfortunately this result cannot be achieved with the HMM approach: Fig. 3(b) reports an experimental multi-level RTN (generated by many traps) along with the HMM output. Though the HMM analysis is correctly defining the hidden states of the multi-level RTN and their most likely sequence, it is generally impossible to separately define the amplitudes of fluctuations and capture/emission times for each single trap contributing to the RTN. This implies that even though the characterization of the RTN signal is achieved, it is impossible to retrieve the distinctive features of each trap contributing to the observed noise. In this paper we show how this limitation can be overcome by using a more refined HMM-based concept, namely the FHMM.

The FHMM [8] extends the HMM potential by considering the hidden state as a collection of K state variables, instead of a single random variable, each potentially assuming one out of N values at each instant of time (i.e. K different and parallel Markov chains). This results in a space state having a dimension of N^K. If no constraints are applied to the model, it can potentially take into account all the possible interdependencies between the K Markov chains, resulting in a high computational burden. However, a natural approach consists in assuming that each of the K Markov chains evolves independently from the other chains, resulting in a significant reduction of the problem complexity. This can be formalized as:

$$P(S_t|S_{t-1}) = \prod_{k=1}^{K} P(S_t^k|S_{t-1}^k) \tag{2}$$

This is also the most suitable representation of a multi-level RTN, seen as a superposition of many two-levels RTNs [5,6], each associated to a single trap. This assumption also constraints each Markov chain state to assume only one out of two values at each instant of time (which is N=2). A graphical representation of the FHMM concept is given in Fig. 4: S_t^m represents the state of the $m-th$ chain at time t, while Y_t represents the output of the whole process (i.e. the expected value of the multi-level RTN) at time t. The inference problem is solved by using the expectation-maximization algorithm, an iterative method for finding maximum likelihood estimates of parameters in statistical models depending on hidden variables. The iteration alternates between an expectation step, calculating the expectation of the likelihood evaluated using the current estimate for the model parameters, and a maximization step, computing the parameters maximizing the expected likelihood found on the previous step. These estimates can be used to determine the probability distribution of the hidden variables in the next iteration. This approach allows decomposing the multi-level RTN into a superposition of two-levels RTNs: since the output of the FHMM is a collection of two-levels fluctuations, it is now possible to separately retrieve the distinctive features of each trap contributing to the observed multi-level RTN.

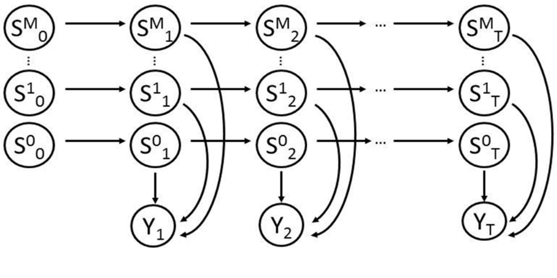

Fig.4: *Graphical representation of an FHMM. At each instant of time t, each output Yt is related to the superposition of the states of M independent and parallel Markov chains.*

2.3 Implementation issues and self-consistency

The implementation of either HMM or FHMM suffers from a trade-off between the computational burden and the fitting accuracy. Indeed, a more complex model (higher number of hidden states in HMM or higher number of Markov chains in FHMM) results in higher time-to-solution. Regrettably, as in HMM the number of hidden states is an input parameter for the model, so the number of parallel Markov chains (i.e. the number of traps contributing to the observed RTN) should be estimated beforehand in FHMM. This theoretically requires an a priori estimation of the number of traps contributing to the RTN. However this issue can be solved by feeding the algorithm a reasonably large number of expected traps (though resulting in a more time-consuming algorithm): the chains related to traps which are unnecessary to match the input RTN will be characterized by negligible amplitude of the fluctuation and can easily be discarded after the analysis. The advantage of the FHMM over HMM is evident even in this aspect: a too large estimation of the number of hidden states in HMM can cause the algorithm to be forced to identify more hidden levels than the effective number, resulting in an erratic signal characterization. Instead in the FHMM approach, using a large number of parallel chains is not affecting the goodness of fitting. This results in the FHMM approach to be self-consistent and extremely accurate.

3. RESULTS AND DISCUSSION

The proposed method has been tested using mathematically generated RTN data simulating the activity of three traps with additive gaussian noise. The number of expected traps was intentionally set to five, a reasonably high value, in order to check the algorithm capability of automatically determining the number of active traps contributing to the observed RTN. Results are summarized in Tab. 1 and plotted in Fig. 5. Fig.5 (a) shows the extremely accurate matching of the FHMM output and the input multi-level RTN. Moreover, input data are correctly separated in three two-levels RTNs time series with remarkable accuracy, Fig.5 (b-d). Noticeably, since only three traps were necessary to match the input signal, the algorithm assigned negligible amplitude to two out of five chains (see Tab. 1), confirming its self-consistency. The algorithm was also successfully applied to a real experimental time series, see Fig. 6. Results are summarized in Tab. 2.

Other statistical machine learning methods can solve this problem using Markov chain Monte Carlo (MCMC) approaches, the simplest of which is Gibbs sampling [10]. Although this technique is guaranteed to converge to the real probability distribution of the data [8,10], the whole space defined by unknown variables in the model has to be sampled, resulting in an extreme computational burden as a consequence of the so-called curse of dimensionality. Moreover, Gibbs sampling technique is intrinsically non parallelizable since every step is based on the result of the previous one, preventing an efficient implementation from being possible. Conversely, the FHMM approach takes advantage of parallel computing: since the time-to-solution is strictly dependent on the initial guess of the model variables (namely A, B and π, according to the formalism used by Rabiner it is possible to run parallel instances of the FHMM algorithm on different cores and then choose the result maximizing the likelihood. Furthermore, as shown in [8], approximated inference techniques can be used to speed-up the FHMM routine: the expectation step of the expectation-maximization algorithm discussed in Section 2.3 can be replaced with either a Gibbs sampling approach or a variation inference method at the cost of lower accuracy. Nevertheless the speed-up gain is strictly dependent on the model complexity and, in our specific case, the exact expectation has been found to be the best choice in terms of trade-off between a good accuracy and a reasonable time-to-solution. [7]), it is possible to run parallel instances of the FHMM algorithm on different cores and then choose the result maximizing the likelihood. Furthermore, as shown in [8], approximated inference techniques can be used to speed-up the FHMM routine: the expectation step of the expectation-maximization algorithm discussed in Section 2.3 can be replaced with either a Gibbs sampling approach or a variation inference method at the cost of lower accuracy. Nevertheless the speed-up gain is strictly dependent on the model complexity and, in our specific case, the exact expectation has been found to be the best choice in terms of trade-off between a good accuracy and a reasonable time-to-solution.

Table 1: *FHMM Output for Generated Data.*

Trap Nr.	Amplitude (Generated)	Amplitude (FHMM)	Amplitude % of Trap 1
1	2	1.994	100.00%
2	1	1.004	50.35%
3	5	4.998	250.65%
4	-	0.024 (discarded)	1.20%
5	-	0.002 (discarded)	0.10%

Table 2: *FHMM Output for Experimental Data.*

Trap Nr.	Amplitude (FHMM)	Amplitude (% of Trap 1)
1	1.78×10^{-8}	100.00%
2	9.87×10^{-9}	55.45%
3	4.1×10^{-11} (discarded)	0.23%
4	3.4×10^{-11} (discarded)	0.19%
5	7.6×10^{-12} (discarded)	0.04%

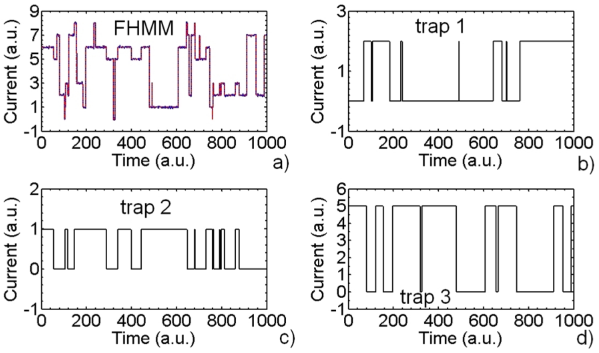

Fig.5: *a) An eight-levels RTN generated by three traps with superimposed additive gaussian noise (blue curve) and FHMM fitting (red curve). b, c, d) Amplitudes of fluctuations and state sequences for all traps are easily found and traps characteristics can be inferred.*

4. CONCLUSIONS

In this paper we proposed the FHMM approach to achieve a full and comprehensive characterization of multi-level RTN, resulting from the activity of multiple traps. HMM limitations in characterizing the multi-level RTN were underlined and the novel FHMM approach has been used to solve for the statistical properties of each trap contributing to multi-

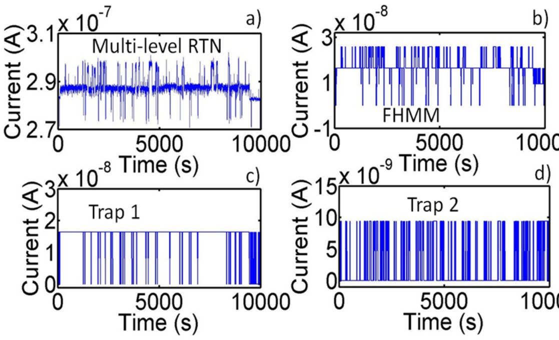

Fig.6: *a) An experimental multi-level RTN and b) FHMM fitting. c, d) Amplitudes of fluctuations and state sequences for all traps are easily found and traps characteristics can be inferred.*

level RTN. As a result, a complex RTN has been separated in multiple two-levels RTNs, allowing inferring the distinctive features of each of the traps leading to the observed RTN. This is of crucial relevance for studying the trap-assisted conduction in novel devices as the inferred trap properties are directly linked to their physical properties. This method is comprehensive and self-consistent, and has been succesfully tested with both experimental and mathematically generated data. Moreover it can take advantage of parallel computing on distributed cores, resulting in a consistent speed-up.

References

[1] F. M. Puglisi et al., "An empirical model for RRAM resistance in low- and high-resistance states," *IEEE Electron. Device Lett.*, vol.34, no.3, pp. 387-389, Mar. 2013.

[2] F. M. Puglisi et al., "A compact model of hafnium-oxide-based resistive random access memory," *Proc. Int. Conf. IC Design Technology*, 2013, pp. 85-88.

[3] D. Veksler et al., "Random telegraph noise (RTN) in scaled RRAM devices," *Proc. IEEE Int. Reliability Physics Symp.*, 2013, pp. MY.10.1-MY.10.4.

[4] L. Vandelli et al., "A physical model of the temperature dependence of the current through SiO_2/HfO_2 stacks," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 2878-2887, Sept. 2011.

[5] F. M. Puglisi et al., "Random telegraph signal noise properties of HfOx RRAM in high resistive state," *Proc. European Solid-State Device Research Conf.*, 2012, pp. 274-277.

[6] F. M. Puglisi et al., "RTS noise characterization of HfOx RRAM in high resistive state," *Solid-State Electron.*, vol. 84, pp. 160-166, Jun. 2013.

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no.2, Feb. 1989, pp. 257-285.

[8] Z. Ghahramani et al., "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245-273, Nov./Dec. 1997.

[9] F. M. Puglisi, P. Pavan, "RTN analysis with FHMM as a tool for multi-trap characterization in HfOX RRAM," *Proc. IEEE Int. Conf. Electron. Devices Solid-State Circuits.* 2013 , pp. 1-2.

[10] G. Casella et al., "Explaining the Gibbs Sampler" *American Statistician*, vol. 46, no. 3, pp. 167-174, Aug. 1992.

Francesco M. Puglisi was born in Cosenza, Italy, in 1987. He received the B.S. and M.S. degrees in electronic engineering summa cum laude in 2008 and 2010, respectively, from Università della Calabria, Rende, Italy. He is currently working toward the Ph.D. degree at Università di Modena e Reggio Emilia, Modena, Italy, in the Dipartimento di Ingegneria "Enzo Ferrari" since 2012. His work focuses on characterization of emerging non-volatile memories, especially RRAMs, and complex noise analysis, particularly RTN, and both compact and physics-based modeling. Mr. Puglisi is currently serving as a reviewer for IEEE Transactions on Device and Materials Reliability. He was awarded with the "E. Loizzo Memorial Award" for being the best Master student of the engineering faculty at Università della Calabria during the 2010-2012 timespan. He was the recipient of the Best Student Paper Award at the IEEE ICICDT 2013 Conference.

Paolo Pavan was born in Italy in 1964. He graduated in Electrical Engineering at the University of Padova, Italy, in 1990 working on latch-up and hot-electron degradation phenomena in MOS devices. In 1991 he started his PhD program studying impact ionization phenomena in advanced bipolar transistor and received his PhD in 1994 from the same University. From 1992 to 1994 he was at the University of California at Berkeley where he studied radiation effects on MOS devices and circuits. In 1998 he worked with Saifun Semiconductors, in Israel, on the development of NROM, a new nonvolatile memory device. His research interests are in the characterization and modeling of Flash memory cells and on the development of new nonvolatile cells and, more recently, in the development of safety critical and wireless applications for automotive electronics. His activity is strongly connected to companies and start-ups in the hi-tech business. He partecipates to many research projects, national and european. In 2002-2003 he was in the IEDM Technical Committee "CMOS and Interconnect Reliability," and Chair it in 2004. European Co-Chair of IEDM 2005 and 2006. He has been Chairman of the Technical Committee "Nonvolatile and Programmable Device Reliability" for ESREF 2002, and Guest Editor of the IEEE Transactions on Device and Material Reliability Special Issue on Nonvolatile Memories in Sept. 2004. He is in the Technical Committee of VLSI-TSA from 2006 to 2010. He is in the Technical Committe of ESREF 2012 and IRPS 2014; in ESSDERC from 2012, and Technical Program Chair of ESSDERC 2014. Prof. Pavan has been the President of the Italian University Nano Electronics Team (IU.NET) Consortium, Bologna, Italy from 2005 to 2011. He authored and co-authored many technical and invited papers, one book and two chapters in edited books; he gave seminars and short courses at international conferences and schools. He is currently Professor of Electronics at the University of Modena and Reggio Emilia, Italy, where he acted as deputy dean for his College and Department and he has been member of the Academic Senate. He is currently Dean of the Electronics Engineering Program.