HMM classifier using biophysically based CMOS dendrites for wordspotting

Suma George, *Graduate Student Member, IEEE* and Paul Hasler, *Senior Member, IEEE*
Georgia Institute of Technology,Atlanta,GA
suma.george@gatech.edu, phasler@ece.gatech.edu

***Abstract*—We explore the co-relations between Neural systems, CMOS transistors and Hidden Markov Models(HMM). We have built a computational model, implementing an HMM classifier that was built using biophysically based CMOS dendrites for wordspotting. The system was implemented on a reconfigurable analog platform. The system thus realized, was found to have high computational efficiency. We discuss the implications of such a computational model. We will also discuss how analog systems can effectively model biological systems, considering benefits both in terms of cost and power dissipation.**

We have built a YES/NO wordspotter system, modeled on an HMM classifier using CMOS dendrites. Wordspotting is the detection of specific words in unconstrained speech [1]. The objective was to build computational models using circuits that are biologically inspired. Dendrites have been known to perform computations like coincidence detection [2]. It has been shown mathematically, that dendrites are similar to a continuous-time HMM [3]. We implemented the system on a reconfigurable analog platform, the RASP 2.8a [4]. We will also show experimental results for the same. We will further discuss the advantages of such a system in terms of computational efficiency and the broader impact of such modeling.

HMM models are a popular choice for speech recognition systems. They have been known to be highly accurate. However, there is still no solution for wordspotting in unconstrained speech [5], [6]. Now even though digital systems have greater accuracy than analog systems; analog systems have lower power consumption. This is closer to how biological systems function. Also, speech is analog in nature. Thus for certain applications especially implantable devices, an analog system is preferred [7]. Previously analog systems were not used much as they were neither programmable nor reconfigurable. However now that we have programmable/reconfigurable analog systems, building larger bio-inspired systems has become a reality.

In section 1 we will overview the inter-relation between the fields of Neural systems, CMOS transistors and HMMs. In section 2 we will discuss the HMM classifier model and discuss the experimental results seen. In section 3 we will give a brief overview of the tools used. In section 4 we will compare the computational efficiency of the analog HMM classifier. In section 5 we will talk about the broader impact of this hypothesis and future directions in this research.

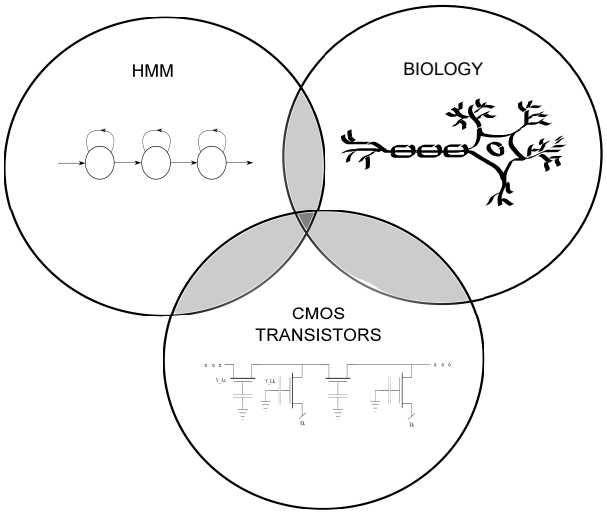

Fig. 1. The diagram depicts the intersection between the fields of Neuro-biology, Hidden Markov Models and CMOS transistors. We have demonstrated in the past how we can build reconfigurable dendrites using programmable analog techniques. We have also shown how such a dendritic network can be used to build an HMM classifier which is typically used for speech recognition systems. Thus it is reasonable to believe that one can compare a HMM network with a group of cortical cells. The co-relations between these two areas is significant for many applications such as low-power implantable devices to aid hearing.

I. Neuro-biology, CMOS transistors and HMM networks

It is an established fact that biological processes can be emulated using silicon devices. Neuromorphic engineers have modeled the channels, synapses, dendrites etc. using CMOS transistors [2], [8], [9]. These circuits aid our understanding of their biological counterparts. A bigger agenda in understanding biological processes is not just studying the elements themselves, but also understanding the intricate relationships they share [9]. This we hope, will help us understand how computation takes place in neural systems and build similar systems. Thus, an important part of this process is not only to build circuits that emulate individual biological elements but build computational systems using these circuits. Studies have shown that dendrites act as computational sub-units that contribute to overall computation of the neural network [10], [11]. It is then imperative that we build computational models using dendrites or say a network of dendrites. One such computational model is an HMM classifier used for speech recognition [3], [12].

978-1-4577-1470-2/11/$26.00 ©2011 IEEE

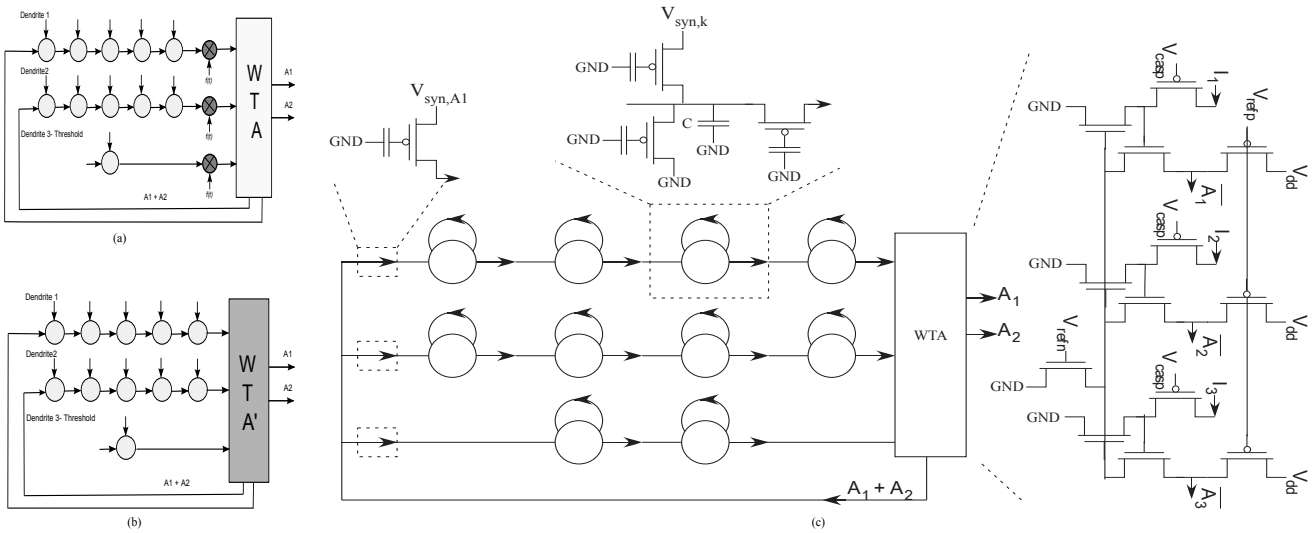

Fig. 2. HMM Classifier Block Diagram (a) The classifier structure with the normalization factor multiplied, $f(x) = exp{-t/\tau}$ (b) The classifier structure after normalization(c) Detailed structure of the HMM classifier using reconfigurable floating-gate devices. There are three main structures here : The dendrite branches, the Winner-Take-All (WTA) circuit and the supporting circuitry. The dendrite branch consists of a 5-stage dendrite for the both the branches representing YES and NO; and a single stage dendrite 3 to set the the threshold current. Each dendrite has synaptic inputs at each node, which represent the phonemes of the word detected. Inputs represent, voltage applied at source of the synapse, $V_{syn,k}$When the line output of exceeds the threshold limit i.e. if a YES/NO is detected then the threshold loses. The supporting circuitry consists of a VMM structure and a floating-gate pFET that acts as a synaptic input at the start of the line(V_{syn_A1}). It also acts as reset function once a word is detected. Portion of image reprinted from [8]

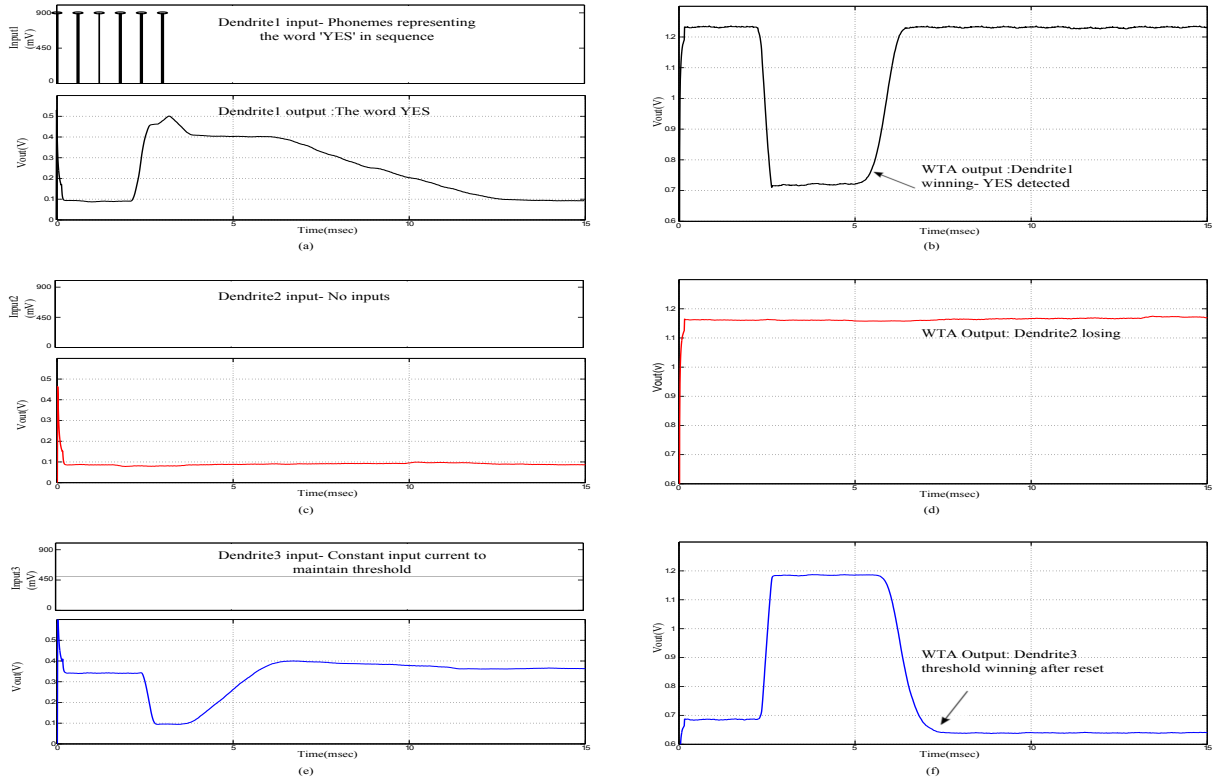

Fig. 3. Experimental results for the YES/NO classifier system. The results shown are for the case when the word YES is detected by the system. (a) Inputs into the nodes of the the dendrites and the line output for the first dendrite (b) Corresponding WTA output to it. A low value signifies that it is winning. (c) The line input and output for the second dendrite. (d) Corresponding WTA output (e) The line output for the third dendrite (f) Its corresponding WTA output. The third dendrite acts as a threshold parameter. The amplitude of the word detected on a particular line needs to be higher than the threshold to win

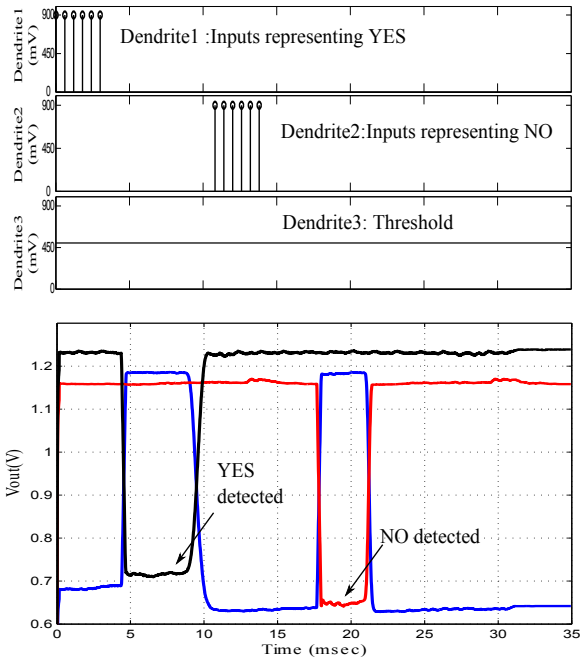

Fig. 4. Experimental results for the classifier system when a sequence of words is detected. First dendrite wins when the word YES is detected and then the second dendrite wins when, NO is detected. At all other times the threshold dendrite is winning.

Studies have shown how a discrete-time HMM can be represented as a wave-propagating PDE that is continuous in time and space [12]. This can be compared to analog diffuser circuits.

$$\underbrace{\tau\frac{\partial\varphi\left(x,t\right)}{\partial t}}_{\substack{\text{state}\\\text{element}}}+\underbrace{\left(\frac{1}{b\left(x,t\right)}-1\right)\varphi\left(x,t\right)}_{\substack{\text{decay}\\\text{term}}}+\underbrace{a\left(x\right)\Delta\frac{\partial\varphi\left(x,t\right)}{\partial x}}_{\substack{\text{wave}\\\text{propagation}}}=0 \tag{1}$$

Compare this to an RC delay line, based on which we model dendrites.

$$\underbrace{R_xC_i\frac{dV\left(x,t\right)}{dt}}_{\substack{\text{state}\\\text{element}}}+\underbrace{R_xG_iV\left(x,t\right)}_{\substack{\text{decay}\\\text{term}}}-\underbrace{\left(\Delta_x\right)^2\frac{d^2V\left(x,t\right)}{dx^2}}_{\substack{\text{wave}\\\text{propagation}}}=0 \tag{2}$$

From Eq. 1 and Eq.2 we can see the similarities in a continuous-time HMM and an RC delay line [12]. The CMOS dendrites are modeled as an RC delay line only using CMOS transistors instead of the resistances. The above equations establish a co-relation between a continuous-time HMM model and a CMOS dendrite. Typically for speech recognition, short segments of the speech signals are analyzed and then the information is integrated for the entire word [5], [7]. The probability distribution b_i, represents the estimate if a symbol (short segment of speech/phoneme) was produced by a state i. This acts as the input to the HMM state machine. Now, to group these symbols we calculate the likelihood of every state which is given by $\phi_i(t)$. This gives us an estimate of the likelihood that the particular state, was the end-state in a path of states that models the input signals [7]. Now a continuous-time Hidden Markov Model with a left-to-right topology has the following update rule,

$$\phi_i(t) = b_i(t)((1-a_i)\phi_i(t-\tau) + a_i\phi_{i-1}(t-\tau)) \tag{3}$$

where, $b_i(t)$ is the input probability of symbol in state i and $\phi_i(t)$ is the likelihood of a state i for time t and a_i is the transition probability between adjacent states. In a speech recognition model, the states are represented by the phonemes. It is interesting to note here, that even though the state sequence of such systems is implied; in continuous-time HMM one can't determine when the transition from states takes place. This is the reason why they are called '*Hidden*' Markov Models although the state sequence has a Markovian structure [5]. Considering we have established the similarities between biology and silicon devices; and between continuous-time HMM and CMOS dendrites, we can postulate there is some inter-relation between HMMs and neural systems

II. HMM CLASSIFIER USING CMOS DENDRITES

We have built a YES/NO wordspotter using CMOS dendrites with synaptic inputs, a WTA circuit and supporting circuitry. Fig. shows a complete block diagram of the system. The EPSP inputs represent the probability distribution $b_i(t)$ or confidence level for a symbol for a particular state i. The EPSP inputs account for the normalization of the output. The update rule for the HMM is similar to coincidence detection in dendrites. Also to ensure a left-right topology, we can model the dendrites with their diameter increasing from the distal to the proximal end. This is typically seen in biological dendrites. It can be modeled in a CMOS dendrite by increasing the axial conductance from left-right for the model. The WTA circuit models the neuron soma and inhibitory inter-neurons. The winning output of the WTA is akin to an action potential. In terms of classification too, the WTA output signifies if a 'word' has been detected. Our results have demonstrated that, such a system looks similar to an HMM state machine for a word/pattern. We have based our comparisons to the analog HMM wordspotter built by Lazzaro et all [7]. We can postulate from these discussions that there are some similarities in computation done by HMM networks and a network of dendrites. The results are shown in Fig. for a single word and for continuous detection of words in Fig. 4

III. METHODS

We used the Reconfigurable Analog Signal Processor (RASP) 2.8a [4], which is a Field Programmable Analog Array(FPAA) for our experiments. We used biophysically based CMOS dendrites representing an HMM state sequence, a WTA circuit representing the soma and supporting circuitry consisting of a VMM structure and transistor synapses , to build the YES/NO wordspotter. The dendrites were implemented using Floating-Gate(FG) pFETS. This enabled us to build a dense structure that was reconfigurable. We have also previously built a simulink dendrite model with other analog blocks like the WTA block and synapses [2], [13]. The system blocks were designed using MATLAB Simulink.

TABLE I
COMPARING COMPUTATIONAL EFFICIENCY OF DIGITAL, ANALOG AND BIOLOGICAL SYSTEMS

Computing Type	Efficiency	Energy/MAC
Digital (DSP)	$<$ 10MMAC/mW	$>$ 100pJ
Analog SP (VMM)	10MMAC/μW	100fJ
Neural Process	$>$ 10MMAC/pW	$<$ 0.1aJ

TABLE II
COMPARING COMPUTATIONAL EFFICIENCY DEPENDING ON LOAD CAPACITANCE

Process	Capacitance	V_{dd}	Energy/MAC
Analog	$1pF$	2.4V	$12.00fJ/MAC$
Analog	$10fF$	2.4V	$0.12fJ/MAC$
Biological	$1fF$	$200mV$	$0.02fJ/MAC$

A tool, sim2spice [13], was then used to generate a SPICE netlist that was used to program the analog hardware using GRASPER [14].

IV. WORDSPOTTING COMPUTATIONAL EFFICIENCY

A major advantage that analog systems have over digital systems is computational efficiency. This can be seen in Table I. The unit used to compare computational efficiency is Multiply ACcumulates per second. The energy efficiency at a given node of the system, depends on the bias currents, supply voltage and also the node capacitance. For a single node of an HMM classifier, we have 2 MAC/sample. Assuming $\tau \sim delay$, which at a given node is approximately $1ms$ Thus,

$$Energy/MAC = \frac{1}{2}V_{dd}(V_{rest} - E_k)C \tag{4}$$

where, V_{dd} is the supply voltage, V_{rest} and E_k the internal and external potentials of the leak channel. From the equation it is evident that major factors contributing to energy efficiency in this case is the node capacitance. As we scale down the process used, this value will reduce. Currently the node capacitance on the chip we used was $1pF$. If we further scale down the process used, this number will also reduce. This effectively means higher computational efficiency. A decrease to $10fF$ itself will give us an improvement of 2 orders of magnitude as seen in Table II.

V. FUTURE DIRECTIONS

The broader impact of such a system is two-fold. First, this system is an example of a computational model using bio-inspired circuits. Secondly the system proposes a computationally efficient solution for speech-recognition systems using analog VLSI systems. As we scale down the process, we can get more efficient and denser systems. We can also address how synaptic learning can be implemented and classification systems be trained. It is evident from the computational efficiency discussions, that clearly analog systems are the way to go for higher computational efficiency and lower costs. This calls for greater effort to build such systems. This classifier structure has a scalable modular architecture. We currently have built architectures that will enable building larger systems; further details of which are beyond the scope of this paper. Reconfigurable/programmable analog systems open a wide range of possibilities in demonstrating biological processing and also for signal processing problems. This will not only enhance our understanding of biological processes but will also help us design more efficient systems that have tremendous computational abilities.

VI. CONCLUSION

We have discussed the similarities between the field of Neuro-biology, silicon devices and HMMs. We have thus postulated similarities between Hidden Markov Models and Neural Systems. This work also demonstrates a computational model using bio-inspired CMOS dendrites. We built an HMM classifier that was used for wordspotting. This The computational efficiency of this system was found to be higher than digital implementations. This technology is attractive especially for implantable devices.

REFERENCES

[1] R. P. Lippmann, E. Chang, and C. R. Jankowski, "Wordspotter training using figure-of-merit back-propagation," *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 389–392, 1994.

[2] S. Nease, S. George, P. Halser, and S. Koziol, "Modeling and Implementation of Voltage-Mode CMOS Dendrites on a Reconfigurable Analog Platform," *IEEE Transactions on Biomedical Circuits and Systems*, In Press.

[3] P. Hasler, S. Koziol, E. Farquhar, and A. Basu, "Transistor Channel Dendrites implementing HMM classifiers," *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, vol. 1, pp. 3359 – 3362, 2007.

[4] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C. Twigg, and P. Hasler, "A Floating-Gate-Based Field Programmable Analog Array," *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 1781–1794, 2010.

[5] B. H. Juang and L. R. Rabiner, "Hidden Markov Models for Speech Recognition," *Technometrics*, vol. 33, pp. 251–272, 1991.

[6] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77 (2), pp. 257–28, 1989.

[7] J. Lazzaro, J. Wawrzynek, and R. Lippmann, "A Micropower Analog VLSI HMM State Decoder for Wordspotting," *Advances in Neural Information Processing Systems 9*, vol. M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, Massachusetts: MIT Press, p. 727733, 1996.

[8] E. Farquhar, D. Abramson, and P.Hasler, "A Reconfigurable Bidirectional Active 2 Dimensional Dendrite Model," *IEEE International Symposium on Circuits and Systems*, vol. 1, pp. 313–316, 2004.

[9] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[10] A. Polsky, B. W. Mel, and J. Schiller, "Computational subunits in thin dendrites of pyramidal cells," *Nature Neuroscience*, vol. 7, pp. 621–627, 2004.

[11] Y. Wang and S.-C. Liu, "Input evoked nonlinearities in silicon dendritic circuits," *IEEE International Symposium on Circuits and Systems*, vol. 1, pp. 2894 – 2897, 2009.

[12] P. Hasler, P. Smith, D. Anderson, and E. Farquhar, "A Neuromorphic IC Connection Between Cortical Dendritic Processing and HMM Classification," *IEEE 11th Digital Signal Processing and 2nd Signal Processing Education Workshop*, pp. 334–337, 2004.

[13] C. Schlottmann, C. Petre, and P. Hasler, "A High-Level Simulink-Based Tool for FPAA Configuration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. Issue:99, pp. 1–1, 2010.

[14] F. Baskaya, D. Anderson, P. Hasler, and S. K. Lim, "A generic reconfigurable array specification and programming environment," *Circuit Theory and Design, 2009. ECCTD 2009. European Conference on*, vol. 1, pp. 619–622, Aug. 2009.