Pulse-Firing
Winner-Take-All Networks

Jack L. Meador
School of Electrical Engineering and Computer Science
Washington State University
Pullman WA, 99164-2752

Abstract- **Winner-take-all (WTA) neural networks using pulse-firing processing elements are introduced. In the pulse-firing WTA (PWTA) networks described, input and activation signal shunting is controlled by one shared lateral inhibition signal. This organization yields an O(n) area complexity that is convenient for integrated circuit implementation. Appropriately specified network parameters allow for the accurate continuous evaluation of inputs using a signal representation compatible with established pulse-firing neural network implementations.**

1   Introduction

The winner-take-all (WTA) function plays a central role in competitive neural networks and is related to recurrent on-center off-surround models of natural neural systems [1-3]. Although it can be realized sequentially via pairwise comparisons, the WTA operation is more effectively realized in parallel analog circuits via a distributed network of processing elements which compare relative input magnitudes and allow only that element with the largest input (or "winner") to remain active. Parallel analog WTA realizations have been described which use Hopfield Network dynamics [4], and MOS current conveyors [5,6]. The model introduced in this paper and its electronic implementation are more like a WTA mechanism inspired by natural presynaptic inhibition feedback [7]. The new pulse-firing WTA (PWTA) model employs a unique combination of a self-shunting feedback term with output hysterisis to yield a WTA network compatible with asynchronous pulse-firing neural network implementations described variously as impulse, pulse-stream, and neural-type networks [8-10].

This paper first introduces asynchronous pulse firing processing units in Section 2. These are the basic computational units used in PWTA networks. The mathematical foundation of PWTA networks is then presented in Section 3 where the system dynamics of a general PWTA network are developed. Section 4 continues with the presentation of MOS circuit implementations. Section 5 closes with an analysis of finite circuit precision effects.

2   Asynchronous Pulse Firing Processing Units

The dynamics of the pulse firing processing units used in a PWTA network obey the following equations:

$$\kappa \dot{v} = -\alpha v + x - (\beta v + x - \gamma)g(v), \qquad v(t_0) = 0 \qquad (1)$$
$$y = g(v)$$

where v is unit activation, x is total unit input, y is unit output, and $g(\cdot)$ is the binary hysterisis function shown in Figure 1. As can be deduced from the figure $g(v)$ includes as a special case the simple threshold nonlinearity (when $V_{tl} = V_{th}$) although the specific pulse firing dynamics described here would cease to exist in that situation. Throughout this paper input x is assumed to be positive and time variant. The unit response to a constant input is a train of regularly spaced constant width pulses. The larger the input signal x, the greater the output pulse repetition rate. The parameter α establishes a first-order response to x during the input integration phase of operation as defined by the absence of an output pulse ($y = 0$). That response is shifted to one defined by $\alpha + \beta$ during the firing phase, as defined by the presence of an output pulse ($y = 1$). Since x is shunted during the output pulse period, processing element state asymptotically approaches $\epsilon = \gamma/(\alpha + \beta)$. Parameter κ uniformly scales all unit time constants. One pulse firing cycle is summarized by the integration of the input signal until v reaches V_{th}, whereupon the switches toggle, causing the discharge of v to $\min(\epsilon, V_{tl})$. Oscillation is sustained provided $\epsilon < V_{tl}$.
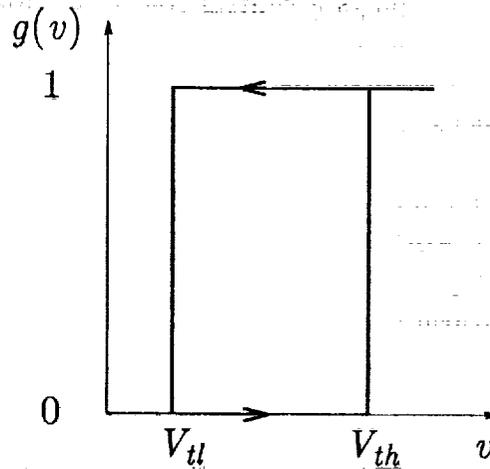
Figure 1: Output hysterisis function

3 PWTA Network Dynamics

A PWTA network combines the unit dynamics described in (2) with lateral inhibition. Lateral inhibition from a combination of unit outputs can be expressed in a form which yields network state equations similar to those of the presynaptic inhibition model described by Yuille and Grzywacz [7]:

$$\kappa \dot{v}_i = -\alpha v_i + x_i F(\mathbf{V}) - H(\mathbf{V}) \qquad (2)$$

where

$$F(\mathbf{V}) = 1 - \bigvee_k g(v_k)$$

and

$$H(\mathbf{V}) = ((\lambda - \beta)v_i + \gamma - \varsigma)g(v_i) + (\beta v_i - \gamma)\bigvee_k g(v_k)$$

with $\bigvee$ indicating the logical OR operation and $\mathbf{V}$ corresponding to the vector of unit activations. $F(\mathbf{V})$ is a binary value establishing the input inhibition which occurs while any processing element generates a pulse. It is during this "output firing phase" of the network that $H(\mathbf{V})$ controls the processing elements such that a winning unit activation decays at a different rate to a different equilibrium than that of a losing unit. The model parameters allow for the independent adjustment of winning and losing unit decay rates and asymptotes. Properly chosen parameter values guarantee WTA function independent of initial system state without the need for external synchronization [11].

All units contribute to a shared lateral inhibition signal identically. The winning output is indicated by the dominance of the first unit activation to reach V_{th}: since it establishes the synchronized re-initialization of all units, it is the only one to fire. In general, the winning unit is determined by a combination of initial network activation state and input magnitude. With appropriately chosen parameters, however the reset state establishes initial conditions which make the winning unit decision independent of initial state and dependent exclusively upon the x_i inputs.

For the winning unit to exactly correspond to the one having the largest input, it is important that initial condition independence be maintained. To guarantee this independence in the PWTA network described by (2) parameters are chosen such that all units reset to an identical initial condition.

All activations in the PWTA of (2) will reset to near-identical initial states if parameters are chosen such that $v_i^* < V_{tl}$, $v_k^* = V_{tl}$ and $\beta \ll \lambda$ [11]. During the output firing phase, these values cause the losing units to approach V_{tl} well before the winning unit does. When the winning unit reaches V_{tl}, it terminates the output firing phase and all activations cease to decay. Theoretically, the losing units only asymptotically converge to V_{tl} while the winning unit converges via a truncated exponential. Even though this is mathematically imprecise, in a practical sense it can be assumed that β is chosen large enough with respect to λ for losing units to converge to within the limits of finite precision hardware well before the firing phase ends.

A geometric interpretation of ideal PWTA network operation with constant inputs is illustrated in Figure 2. Each loop in the state diagram corresponds to one firing cycle. Unit activations are reset to V_{tl} at state S_0 in the diagram. The input integration phase (1 and 3 in the figure) begins at S_0 and terminates when the V_{th} threshold is reached. That is followed by the output firing phase (2 and 4 in the figure) during which unit activations
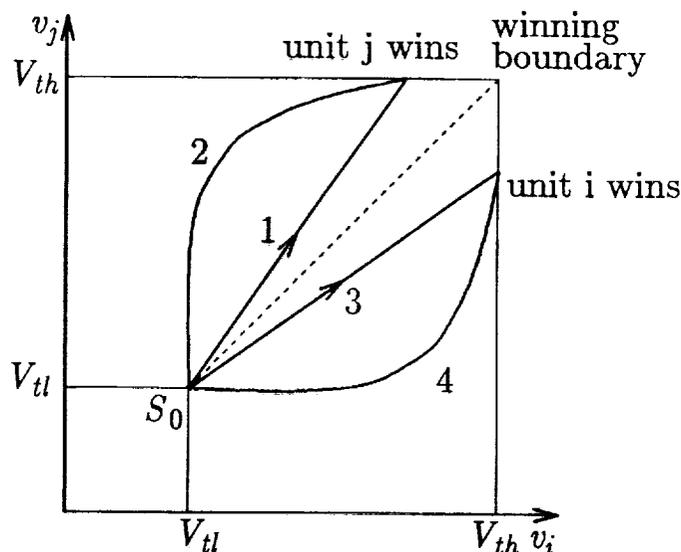
Figure 2: Ideal PWTA network operation

decay. In the figure, trajectory 1-2 corresponds to the path followed when unit j wins, and trajectory 3-4 to when unit i wins. Which unit wins is determined by the one which first reaches V_{th}. That in turn is determined by the state trajectory during the input integration phase. With constant inputs, that trajectory is linear with slope proportional to the quotient of the input signal magnitudes. A "winning boundary" for constant inputs can be identified by unit slope (dashed line in the figure). This geometric interpretation of PWTA operation will prove useful in a later discussion of finite precision effects.

Figure 3 illustrates the operation of a 2-unit PWTA network in response to a smooth transition between two inputs. In this example, input signals x_1 and x_2 move from 0 to 1 and from 1 to 0 respectively, crossing at t=10. The parameters chosen for this simulation are $V_{tl} = 1$, $V_{th} = 4$, $\kappa = 0.1$ $\alpha = 0.1$, $\beta = 1.2$, $\gamma = 1.3$, and $\lambda = 0.6$. The activation state space diagram for this simulation is shown in Figure 4. It can be seen how the reset state with these parameters assures input order preservation.

4 CMOS Circuit Implementation

By way of introduction to the CMOS PWTA network, a CMOS implementation of a pulse firing processing element shall first be considered. Figure 5 shows the circuit for an impulse neural circuit as described previously in [8]. For simplicity, $\gamma = 0$, and α is for practical purposes nonexistent by virtue of the low leakage currents exhibited in MOS technology. C_κ includes not only the ideal capacitance of a poly-1 capacitor, but also stray wiring capacitance and the input capacitance exhibited by the Schmitt trigger G. The Schmitt trigger provides high voltage gain at the threshold voltages V_{tl} and V_{th}, with positive feedback from the output establishing the active threshold. The Schmitt trigger output can be expressed in terms of the hysterisis function g of (2) as $G(v) = V_{DD}g(v)$. β corresponds to the channel conductance of M2 which operates in the active region when an output pulse is generated. Further details regarding the operation of this circuit are
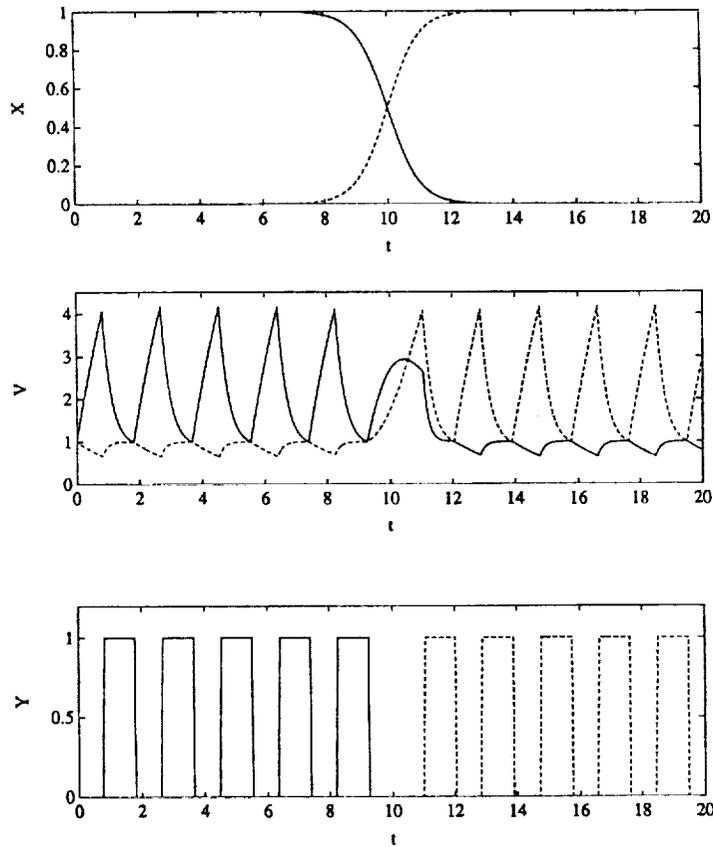
Figure 3: Inputs, state variables and output of a 2-unit PWTA network
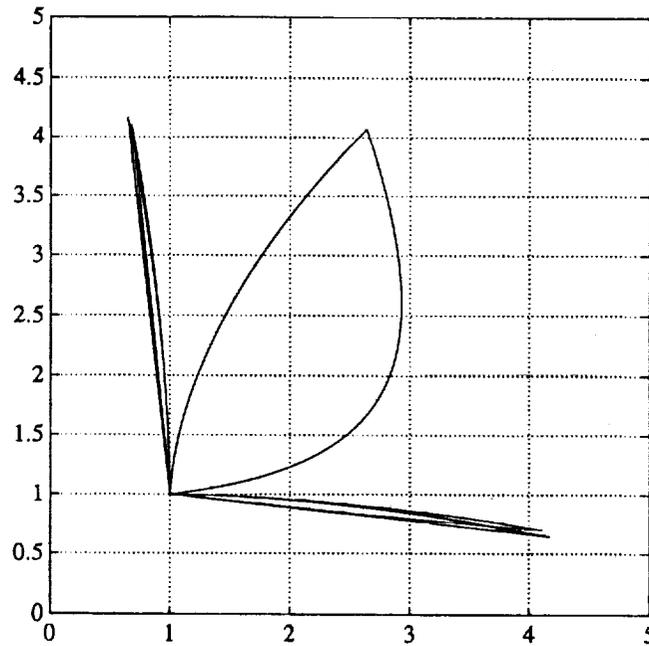

Figure 4: State trajectory of a 2-unit PWTA network simulation

Figure 5: CMOS implemantation of a pulse firing processing unit

provided in [8].

A CMOS implementation of a PWTA cell is shown in Figure 6. The basic elements of the CMOS impulse circuit are augmented by additional MOSFETs which establish various parameters associated with the ideal model. Variations of this circuit having reduced transistor counts are also possible [11]. The circuit of Figure 6 simply represents the most general CMOS PWTA implementation consistent with the Equation (3) definition.

Two local signals and one global signal control circuit operation in accordance with current network state. The local signals G_i and $\overline{G_i}$ indicate that unit i is the winning unit when $G_i = V_{DD}$ and $\overline{G_i} = 0V$. These signals select the local firing response. Both the true and complemented global lateral inhibition signal F and $\overline{F}$ are derived by a pseudo-NMOS NOR gate and a CMOS inverter consisting of transistors M11 through M14 in the diagram. These signals are distributed on two wires between all cells of the WTA network. NOR pulldown transistors (M11) are distributed across all cells while there need only exist a single pullup transistor (M12) and single inverter (M13, M14). When any unit in the network initiates a pulse, F becomes active, causing all units to enter the output firing phase.

Transistor M1 disconnects input current x_i during the output firing phase, causing it to be shunted into the parallel capacitance of some input circuit (not shown – see [8] for further details). Also during the firing phase, transistors M2, M7, M6, and M10 conduct, allowing some combination of the currents I_1 through I_4 to flow.

The circuit branch consisting of M2 through M4 establishes a current which corresponds to the constant $I_1 = \gamma$. Similarly, branch M7 through M9 establishes a constant current corresponding to $I_3 = \zeta$. Ignoring the nonlinear component of channel conductance, the branch consisting of transistor M10 establishes a current corresponding to $I_4 = \lambda v_i$ and the M5, M6 branch a current analogous to $(\beta - \lambda)v_i$. As with the circuit of Figure 5, α is assumed to be negligible.

During the output phase, the signals G and *overlineG* control the unit response. If the unit is a winner, then $G = V_{DD}$, $\overline{G} = 0V$, and branch currents I_3 and I_4 are allowed

to flow. This establishes a winning unit response where the unit activation asymptotically decays to ζ/λ, but is truncated at V_{tl} when the winning unit terminates the output firing phase. If the unit is a loser, then $G = 0V$, $\overline{G} = V_{DD}$, and branch currents I_1, I_2, and I_3 are allowed to flow. This establishes a losing unit response, where the unit activation asymptotically approaches V_{tl} at a much faster rate than it would otherwise as the winning unit.

5   Finite Precision Effects

Thus far, the effects of finite parameter precision have been ignored. Intra-cell parameter variation will contribute to deviations from the ideal performance previously described. This section focuses upon the parametric variations which will have the greatest effect upon PWTA performance that also are the most likely to occur in contemporary CMOS fabrication processes.

The overall function of a PWTA network is to select the input signal having the greatest magnitude. Inspection of Figure 2 reveals that there are two potential error sources which can interfere with that function. These are errors in the determination of the initial network state, S_0 and deviations in the position of the winning boundary. These variations effectively give an unfair advantage to some processing units, sometimes allowing units to fire even thought their inputs are not necessarily the largest. Fortunately, it can be shown that this occurs only when two inputs have very nearly the same magnitude. Units having input signals which are "clearly" not the largest will remain quiescent. The definition of "clearly" is expressed as a hysterisis deadband which naturally occurs around the winning boundary. This hysterisis arises directly from parametric variation.

For the remainder of this section only constant inputs will be considered. This allows for the analysis of parameter precision effects while the network is in a steady-state operating condition. Figure 2 illustrates network operation under ideal conditions when the critical parameters V_{tl}, V_{th}, κ, γ and β are assumed identical across all units. Under these conditions, unit i wins if

$$\frac{dv_i}{dv_j} > 1 \tag{14}$$

with unit j winning otherwise. $\frac{dv_i}{dv_j} = 1$ corresponds to the ideal winning boundary (dashed line) of Figure 2.

Variations in the scaling constant κ yield an inaccurate winning boundary definition. κ is determined in the CMOS implementation by the MOS capacitor C_κ in the previous circuit diagram. Variations in capacitor geometry will lead to inter-unit κ variation and subsequently give those units having a smaller κ an advantage in the race toward V_{th}. Recognizing that

$$\frac{dv_i}{dv_j} = \frac{\kappa_i \dot{v}_i}{\kappa_j \dot{v}_j} \tag{15}$$

yields the decision rule that unit i wins if

$$\frac{dv_i}{dv_j} > \frac{\kappa_j}{\kappa_i} \qquad (16)$$

This rule reduces to the ideal one of (14) when $\kappa_i = \kappa_j$ Although all units are initialized to the same state at S_0, the decision boundary is shifted such that one unit is favored over another.

Variation of C_κ alone leads to finite precision for the winning boundary. Parameter variations which affect the initial network state and the unit firing thresholds lead to more complex hysterisis effects at the winning boundary. Firing thresholds, as defined by V_{th} in the Schmitt trigger are typically determined by device geometries. The initial network state is determined in part by V_{tl} which is also dependent upon Schmitt device geometries. The other part of initial network state is determined by the geometries of M2-M6 and M10 of Figure 6 (corresponding to β, γ and λ in the ideal equations). Not only do such variations change the slope of the winning boundary, but the slope is also dependent upon the winning unit as well. Under these conditions, the current winner is favored by the initial states such that a new winner must have a significantly larger input than the present one. These effects can be used to extend the decision rule expressed by (16) to one where unit i wins if

$$\frac{dv_i}{dv_j} > \frac{\kappa_j}{\kappa_i} \frac{V_{thi} - V_{tli}}{V_{thj} - V_{0j}} \qquad (17)$$

where V_{0i} and V_{0j} are determined by the combined variations of M2-M6 and M10 between units i and j. It can be easily verified that this decision rule reduces to the ideal case when there is no inter-unit variation. The effect this has on overall PWTA function is to introduce a hysterisis deadband which only affects close decisions.

6   Conclusion

An ideal linear model has been used to establish a general basis for PWTA function. This model improves an earlier one based upon presynaptic inhibition in two ways. The new model uses O(n) interconnect for lateral inhibition and does not require an external reset signal since it is fully asynchronous. Furthermore, it provides information regarding how strong a winning input is – a feature not always found in winner-take-all networks. Model parameters can be chosen to guarantee ideal winner-take-all function given precise parameter specifications. The model is also fully compatible with previously established asynchronous pulse firing analog neural ICs.

CMOS PWTA circuits have also been presented. These circuits necessarily deviate from the ideal linear model, but they can be designed to exhibit similar behavior simply by accounting for the nonlinear characteristics of the electronic devices they employ. Non-ideal effects arising from practical implementation considerations have also been addressed.

Figure 6: A genral CMOS PWTA cell

Parametric variation between pulse-firing processing units leads to a finite decision accuracy and the potential existence of a hysterisis deadband.

References

[1] J. Hertz, A. Krogh, and R.G. Palmer, Introduction to the Theory of Neural Computation, 217-228, Addison Wesley, 1991.

[2] S. Grossberg, Contour Enhancement, Short Term Memory, and Constancies in Reverberating Neural Networks, *Studies in Applied Mathematics*, Vol. 12, 213-257, MIT Press, 1973.

[3] S. Grossberg, Nonlinear Neural Networks: Principles, Mechanisms, and Architectures, *Neural Networks*, Vol. 1, 17-61, Pergamon Press, 1988.

[4] E. Majani, R. Erlanson, and Y. Abu-Mostafa, On The K-Winners-Take-All Network, *Progress in Neural Information Processing Systems*, Vol. 1, 635-642, Morgan Kaufman, 1988.

[5] J. Lazzaro, et al., Winner-Take-All Networks of O(n) Complexity, *Progress in Neural Information Processing Systems*, Vol. 1, 703-711, Morgan Kaufman, 1988.

[6] A. Andreou, et al., Current-Mode Subthreshold MOS Circuits for Analog VLSI Neural Systems, *IEEE Trans. on Neural Networks*, Vol. 2, 205-213, IEEE Press, 1991.

[7] A. L. Yuille, and N. Grzywacz, A Winner-Take-All Mechanism Based on Presynaptic Inhibition Feedback, *Neural Computation*, Vol. 1, 335-347, MIT Press, 1989.

[8] J. Meador, et al., Programmable Impulse Neural Circuits, *IEEE Trans. on Neural Networks*, Vol. 2, 101-109, IEEE Press, 1991.

[9] A. F. Murray, et al., Pulse-Stream VLSI Neural Networks Mixing Analog and Digital Techniques, *IEEE Trans. on Neural Networks*, Vol. 2, 193-204, IEEE Press, 1991.

[10] G. Moon, et al., Analysis and Operation of a Neural-Type Cell (NTC), *Proc. IEEE ISCAS*, pp. 2332-2334, IEEE Press, 1991.

[11] J. Meador, Pulse Firing Winner-Take-All Networks, *Internal Manuscript, School of Electrical Engineering and Computer Science*, Washington State University, Pullman WA., June, 1991.