Circuits for Multi-level Neuron Nonlinearities[1]

Jen-dong Yuh and Robert W. Newcomb

Microsystems Laboratory
Department of Electrical Engineering
University of Maryland
College Park, MD 20742

Abstract: In this paper, we present some circuits for realizing multi-level nonlinearities. These nonlinearities can be used for circuit implementations of multi-level neuron networks. As examples, we show how to use a tanh(·) function to build different Multi-level nonlinearities. Circuit simulations are included using SPICE3 with level 2 MOSIS BiCMOS process parameters to show that these multi-level nonlinearities function.

1. Introduction

In the literature, most neural networks are considered as "two state" neural networks since the nonlinearities they use are Hard limiters, saturating linear elements, or sigmoidal functions [1][2]. In [3] Banzhaf did some simulations involving multi-state neural associative memory. However, he did not show what nonlinearity was used and how to implement that nonlinearity. In [4] Si and Michel did analysis and synthesis of discrete-time neural networks wit multi-level threshold functions. Nevertheless, no realization is mentioned in their paper. Since we think multi-level neural networks are with good potential in terms of image processing applications, multi-valued logic applications, and reducing weights and still achieving desired goals for neural networks, we present two kinds of circuits to realize multi-level nonlinearities. One is a BiCMOS circuit and the other one is an all MOS circuit. These circuits can be readily applied to the realization of multi-level neural networks.

2. Multi-level Nonlinearities

We introduce a multi-level nonlinearity as

$$M(x) = \sum_{j=1}^{b} a_j f_j(x\text{-}th_j) \tag{1}$$

Where, the multi-level nonlinearity M(x) consists of b, $b \geq 1$, basis functions weighted by real nonnegative coefficients a_j and with thresholds th_j. These basis functions $f_j(\cdot)$'s are chosen among any of a number of monotonically nondecreasing step-type functions. Therefore, M(x) is a monotonically nondecreasing function. As an example, for basis functions, the hyperbolic tangent functions tanh(λx) are shown in Fig.1(a) with $\lambda = \{2, 3, 7\}$. With increasing values of λ, the function behaves more like a hard-limiter. This is a popular neuron nonlinearity in the literature and considered as a "two state" function, since its output has only one sharp transition region. Beyond this sharp transition region, we denote the upper region as logic "1" and the other as logic "0". By substituting $f_j(.)$ with tanh(λx) in equation (1), we express a tanh(·) multi-level nonlinearity as

$$M_{tanh}(x) = \sum_{j=1}^{b} a_j \tanh(\lambda(x\text{-}th_j)) \tag{1a}$$

[1] Research supported by NSF Grant MIP 891122 and ONR Grant N00014-90-J-1114

0-7803-0559-0/92 $3.00 © 1992 IEEE

As an example, $M_{tanh}(x)$ with b = 3, $a_1 = a_2 = a_3 = 1$, $th_1 = -1$, $th_2 = 0$, $th_3 = 1$ and $\lambda = 2$ is shown in Fig.1(b)

3. BiCMOS Circuit Realization of M_{tanh}

To start, we build a tanh(x) Nonlinearity Building Block (NBB) with the BiCMOS circuit shown in Fig. 2(a). In analyzing this circuit, we assume Q1 and Q2 are matched. For simplicity, assume that channel modulation effects are negligible, in which case we describe the current output in Fig.2(a) as [5, pp445]

$$I_o = -(I_{c1} - I_{c2}) = \alpha I_{EE} \tanh((x-th)/2V_T) \qquad (2a)$$

where α is the bipolar transistor forward current gain, I_{EE} is the value of the bias current sink for the emitter coupled pair shown in Fig.2(a), V_T is the thermal voltage, and th is the threshold setting given as a differential pair input. Here, I_{EE} is set by a cascode current mirror as shown in Fig. 2(a) and described by [5,pp346],

$$I_{EE} = I_{dsr1} \frac{W_4/L_4}{W_{r1}/L_{r1}} \qquad (2b)$$

where I_{dsr1} is the transistor M_{r1} drain current, and W_4/L_4 & W_{r1}/L_{r1} are width/length ratios of M_4 and M_{r1} respectively. By connecting in parallel the outputs of b NBBs with different threshold points as shown in Fig. 2(b) and using different I_{EE}'s, we get the following voltage input, x, current output, $M_{Bi}(x)$, multi-level nonlinearity

$$M_{Bi}(x) = \alpha [I_{EE1} \tanh(\frac{x - th_1}{2V_T}) + I_{EE2} \tanh(\frac{x - th_2}{2V_T}) + \ldots + I_{EEb-1} \tanh(\frac{x - th_{b-1}}{2V_T})] \qquad (3)$$

Note that, on comparing with (1a), $\lambda = 1/(2V_T)$ is obtained. Here, to control the nonlinearity we scale the I_{EE}'s, and also adjust the thresholds. These two parameters are key for programming a multi-level nonlinearity.

3.1 Circuit simulation for M_{tanh}

A SPICE3e1 simulation for the BiCMOS circuit of Fig.2(a) for different bias I_{EE}, is shown in Fig. 3(a). Simulations are performed with Level 2 parameters obtained from MOSIS for a 2-μ, n-well, double-poly, analog, BiCMOS process [run N15S of June 18th, 1991]. This circuit uses cascode current mirrors and a cascode current sink (for the I_{EE} bias) to reduce the channel modulation effects. This simulation shows us that we can adjust a BiCMOS multi-level nonlinearity output by choosing appropriate I_{EE}. A circuit simulation for a BiCMOS multi-level nonlinearity is show in figure 3(b) with b = 3. For a BiCMOS NBB circuit, the λ of $\tanh(\lambda x)$ is a constant value at a fixed temperature.

4. CMOS Realization of M_{tanh}-like nonlinearity

To obtain variable λ while also using only MOS processing, we now turn to a CMOS realization. This circuit is that of Fig.2(a) except that, Q_1 and Q_2 are replaced by two NMOS transistors denoted by M_1 and M_2. This NBB circuit gives us a similar result as the previous one. However, the analysis is not as straightforward as the previous one since circuit equations used for a MOS differential pair can not be described by a single equation for all operation regions. For analyzing the circuit, assume that M_1 and M_2 are in the saturated mode of operation and channel

modulation effects are negligible, then the drain currents of M_1 and M_2 are described in the following [5,pp.145]

$$I_{d1} = \mu_n \frac{C_{ox} W_1}{2 L_1} (V_{gs1} - V_{th})^2, \quad I_{d2} = \mu_n \frac{C_{ox} W_2}{2 L_2} (V_{gs2} - V_{th})^2 \tag{4}$$

Here, μ_n is the electron mobility, C_{ox} is the gate capacitance per μ^2 unit of an input MOS transistor, V_{gs} is the gate to source voltage of a MOS transistor, and V_{th} is the threshold voltage of an input MOS transistor. In Fig. 2(a), $I_{EE} = I_{d1} + I_{d2}$, and $V_{gs1} - V_{gs2} = (x\text{-th})$. Assume that M_1 and M_2 geometry sizes are equal, $L_1/W_1 = L_2/W_2$. By solving equations (4) in terms of $(I_{d1} - I_{d2})$ and (x-th), we have the following [6,pp706]

$$I_o' = I_{d1} - I_{d2} = \frac{\mu_n C_{ox} W_1}{2 L_1} (x - th) \sqrt{\frac{2 I_{EE}}{(\mu_n C_{ox}/2)} \left(\frac{L_1}{W_1}\right) - (x - th)^2} \tag{5a}$$

The above equation is valid only when both M_1 and M_2 are operated in the saturation region which is when the following condition holds [6,pp706]

$$|(x - th)| \le \sqrt{\frac{I_{EE}}{(\mu_n C_{ox}/2)} \frac{L_1}{W_1}} \tag{5b}$$

where $|\cdot|$ denotes the absolute value. Therefore, the excursion of the differential input voltage that is required to turn off one of the input transistors is a function of transistor W/L ratios and the bias current I_{EE}. By decreasing the I_{EE}, increasing the W/L ratios of input transistors, the active range for two input transistors can be reduced. This effect is similar to that due to the λ of the function $\tanh(\lambda x)$, i.e. increasing λ gives a sharper transition output for a hyperbolic tangent function. The source coupled pair displays a saturating behavior when the differential input voltage exceeds a certain value given by equation (5b). As a result, a large differential voltage will effectively turn off either one of the source pair transistors. Let λ' be the effective λ for the CMOS realization, then λ' is a function of W_1/L_1 and I_{EE} as suggested by equation 5(b). We derive λ' by taking the fact that $\tanh(\lambda' x) \approx \lambda' x$ when $\lambda' x \ll 1$. Then, from equation 5(a), let $I_o'/I_{EE} \approx \lambda'(x\text{-th})$, ignoring the $(x\text{-th})^2$ term in the square root since it is small when $\lambda' x \ll 1$. We obtain the following

$$\lambda' = \sqrt{\frac{\mu_n C_{ox}}{I_{EE}} \left(\frac{W_1}{L_1}\right)} \tag{6a}$$

Then an all MOS multi-level nonlinearity is approximated by the following equation

$$M_{mos}(x) \approx I_{EE1} \tanh(\lambda_1'(x - th_1)) + I_{EE2} \tanh(\lambda_2'(x - th_2)) + \ldots + I_{EEb} \tanh(\lambda_{b-1}'(x - th_b)) \tag{6b}$$

As compared with equation (3), (6) is not a direct function of the thermal voltage V_T. We next show that circuit simulations support equation (6).

4.1 Circuit Simulation for M_{tanh}-like nonlinearity

A simulation of the all MOS modification of Fig.2(a) is shown in Fig. 4(a) for the same I_{EE} values used in Fig.3(a). This simulation shows us that we can adjust the all MOS multi-level nonlinearity output by choosing appropriate I_{EE}. Circuit simulation for an all MOS multi-level nonlinearity, b=3, is shown in Fig. 4(b). A simulation using a variety of sizes of input transistors for an all MOS NBB is shown in Fig. 5(a). Comparisons between curves of Fig.5(a) and $\tanh(\lambda x)$ with $\lambda = \{2,3,7\}$ is shown in Fig.5(b) to illustrate how close the characteristic transfer function of an all MOS NBB is to a hyperbolic tangent function. Therefore, to choose an appropriate input transistor W_1/L_1 ratio of an all MOS NBB is similar to choosing a $\lambda' \approx \lambda$ of the function $\tanh(\lambda x)$.

5. Discussion and Conclusions

We have demonstrated two multi-level neuron nonlinear circuits using BiCMOS and all MOS technologies, respectively. Simulations for these two circuits using level 2 BiCMOS process parameters from MOSIS are shown in this paper in which case these two circuits can be built on the same chip with BiCMOS technology. A comparison is given between the BiCMOS circuit and the all MOS circuit for a multi-level neuron. Channel modulation effects are investigated in these simulations and precision multi-level nonlinear circuits are designed by using cascode current mirrors and cascode current sinks for the NBBs. These multi-level neurons can be programmed by their I_{EE}'s and threshold voltage th's. By organizing b-1 NBBs, we can obtain suitable b neuron output levels. Since we take advantage of the current output of a NBB, it is relatively easy to augment a b-level one into a b+1-level neuron. If an output voltage is desired, a BiCMOS current controlled voltage source can be inserted. Therefore, these multi-level nonlinear circuits can be readily applied to the realization of multi-level neural networks.

REFERENCES

[1] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, Vol. 4, April 1987, pp. 4-21

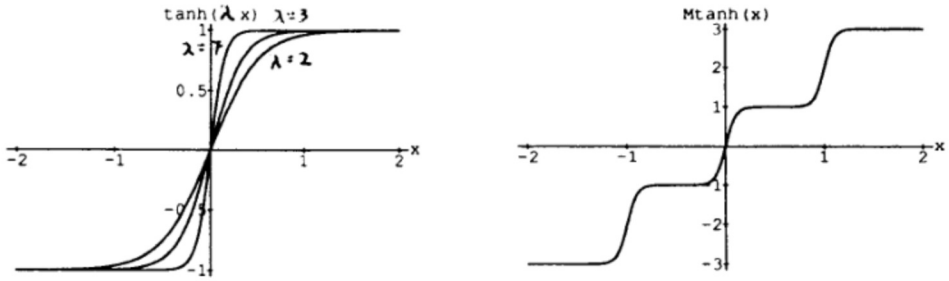
[2] A. N. Michel, and J. A. Farrel, "Associative Memories Via Artificial Neural Networks," *IEEE Control Systems Magazine*, Vol. 10, April, 1990, pp. 6-17

[3] W. Banzhaf, "A Network of Multistate Units Capable of Associative Memory and Pattern Classification," *Physica D*, Vol 34, March 1989, pp. 418-426

[4] J. Si and A.N.Michel, "Analysis And Synthesis Of Discrete-Time Neural Networks With Multi-Level Threshold-Functions," Proceedings of the IEEE International Symposium on Circuits and Systems, The Westin Stamford and Westin Plaza, Singapore, June 1991, pp. 1461-1464.

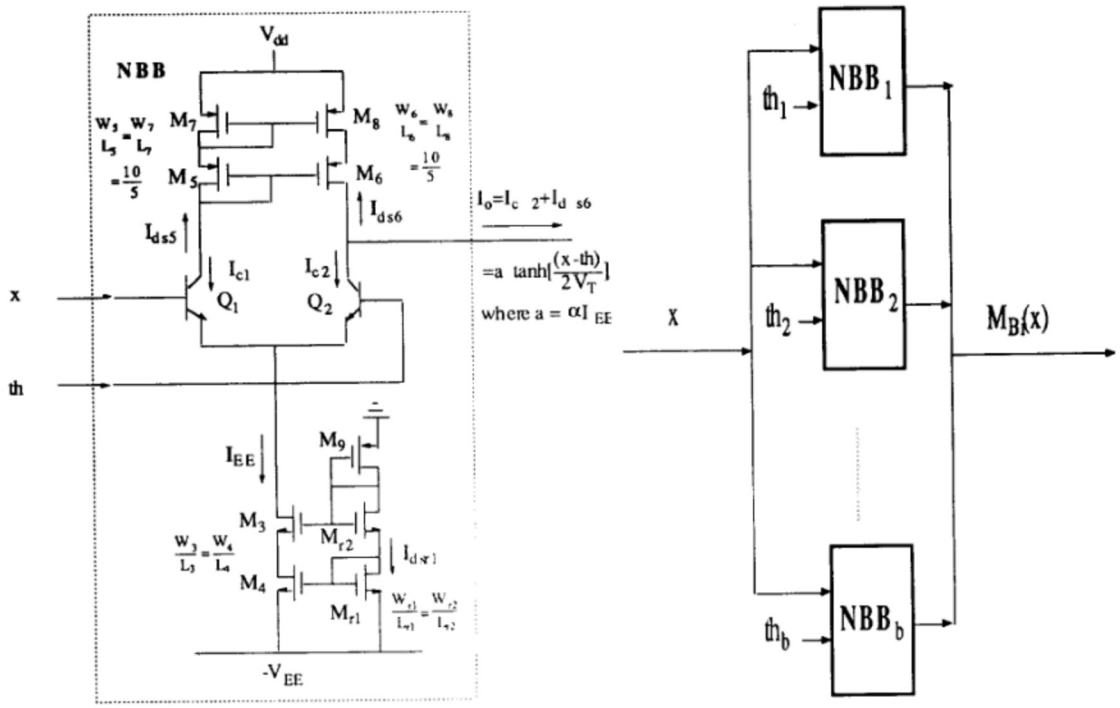
[5] R. L. Geiger, P. E. Allen, N. R. Strader, "VLSI design techniques for analog and digital circuits," McGraw-Hill Inc., New York, 1990, pp. 144-150,318-372, 431-449.

[6] P. R. Gray, R. G. Meyer, "Analysis and Design of Analog Integrated Circuits," 2nd edition, John Wiley & Sons, Inc., New York, 1984, pp.705-709



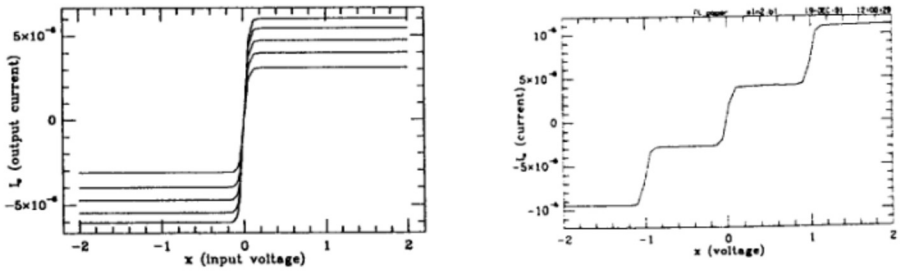
(a) (b)

Figure 1: (a) A hyperbolic tangent function tanh(λx) with different λ, λ = {2,3,7}; (b) A multi-level tanh(λx) nonlinearity with b = 3, λ = 10.



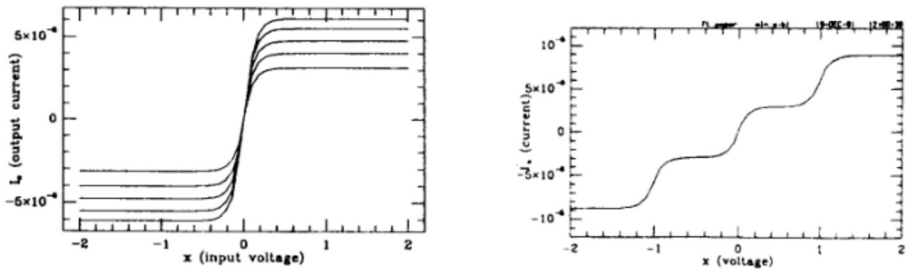
(a) (b)

Figure 2: (a) A BiCMOS tanh(·) Nonlinearity Building Block(NBB); (b) A multi-level nonlinearity with b NBBs.



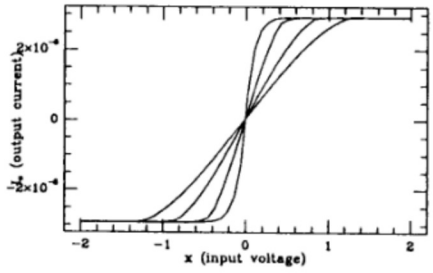
(a) (b)

Fig. 3: (a) A BiCMOS NBB circuit simulation for different bias I_{EE}'s, taking the values 3.1(smallest curve), 4, 4.8, 5.5, to 6μA; (b) Simulation of a BiCMOS 4-level Nonlinearity (b=3)

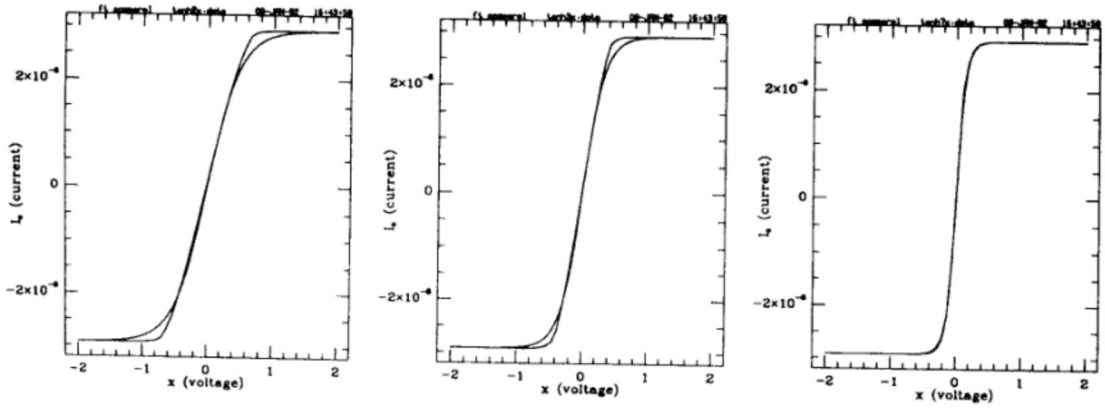


(a) (b)

Figure 4:(a) An all MOS NBB circuit simulation for different bias I_{EE}'s, ranging from 3.1, 4, 4.8, 5.5, to 6μA(largest curve); (b) Simulation of a BiCMOS multi-level Nonlinearity.



(a)



$\lambda = \lambda' = 2$ $\lambda = \lambda' = 3$ $\lambda = \lambda' = 7$

(b)

Fig. 5: (a) An all MOS NBB circuit simulation with a variety of W_1/L_1 ratios for input transistors; W_1/L_1 = {5/60, 5/30, 5/10, 25/5[sharpest curve]}; (b) Comparisons between output curves of an all MOS NBB and tanh(λx) with $\lambda' = \lambda = \{2,3,7\}$ sharper curves for λ'.