1998 Fifth IEEE International Workshop on Cellular Neural Networks and their Applications, London, England, 14-17 April 1998

Four-Quadrant One-Transistor-Synapse for High-Density CNN Implementations

R. Domínguez-Castro, A. Rodríguez-Vázquez, S. Espejo, R. Carmona

Inst. de Microelectrónica de Sevilla - Centro Nac. de Microelectrónica - C.S.I.C. - Universidad de Sevilla,
Edf. CICA, C/Tarfia s/n, 41012-Sevilla, SPAIN.
Phone: +34 5 4239923; FAX: +34 5 4231832; email: rafael@imse.cnm.es

ABSTRACT: *This paper presents a linear, four-quadrants, electrically-programmable, one-transistor synapse strategy applicable to the implementation of general massively-parallel analog processors in CMOS technology. It is specially suited for translationally-invariant processing arrays with local connectivity, and results in a significant reduction in area occupation and power dissipation of the basic processing units. This allows higher integration densities and therefore, permits the integration of larger arrays on a single chip.*

1. Introduction

The most important trend in the electronic implementation of CNNs is the maximization of the number of elementary processing units that can be placed in a single chip. This must be combined with the achievement of an acceptable accuracy for the system parameters. The first trend imposes a strong commitment in the design of the cell circuitry: area-efficiency. A second but also important objective is power-economy. Unfortunately, in analog processing circuits, area and power consumption of individual devices are directly related to accuracy [1]. Therefore, the selection of a circuit strategy as simple as possible for the prescribed processing function is crucial.

This paper proposes a one-transistor, four-quadrants, electrically-programmable, linear synapse for the implementation of massively-parallel analog-array processors in CMOS technologies. The proposal is easily extensible to general artificial neural networks.

2. Analog-Array-Processors Elementary Units

Each processing unit (or *cell*) in a massively-parallel analog-array processor can be characterized by an interconnection pattern and by an specific processing function. These characteristics are often considered invariant from cell to cell, resulting in an additional simplification of the electronic implementation. This property, commonly referred to as spatial invariance or uniformity, will be assumed without loss of generality.

A common characteristic of massively-parallel analog processing algorithms is the *local* computation (within each cell c) of a weighted aggregation of contributions from the cells in its neighborhood,

$$y_c = \sum_{i=1}^{N} A_i x_i \tag{1}$$

The aggregated signal y_c is then used as input to a *processing-block* which realizes some specific function, and generates an output x_c representative of the cell state. In turn, this output constitutes the (unscaled) contribution of the cell to its neighbors. This general cell architecture is illustrated in Figure 1. The output x_i of each neighbor is weighted by a coefficient A_i, which is independent of the particular receptor cell c under the assumed spatial uniformity. Except for specific purpose systems it is generally required that the scaling coefficients (or *weights*) A_i be electrically programmable, for versatility reasons.

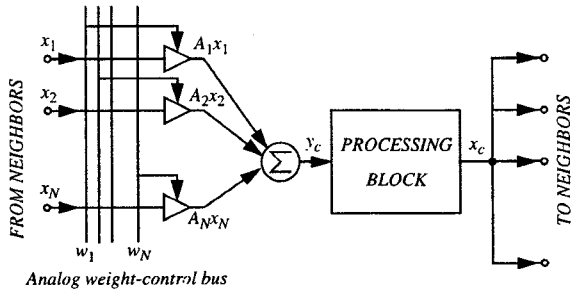

Figure 1. General architecture of an analog-processing-array elementary processing unit (cell).

At a system level, the required number of signal-scaling circuit-blocks (or *synapses*) can be computed as N times the number of cells. The associated area and power consumption, as well as the obvious effects of the synapse accuracy on the overall performance of the system, renders the selection of the synapse circuitry a crucial issue in the design of integrated analog array processors.

3. Electrically-Programmable Synapses

Electrically programmable synapses must be driven by the *input signal* x_i, and by a *weight signal* w_i used to program the scaling coefficient A_i. Under the assumed spatial uniformity, only N different weight values will coexist throughout the array. Therefore, a reduced number (N) of *global nodes* (common to all cells in the network) satisfy the programming-related routing requirements of the array, if the programming signals w_i are codified as voltages. Since every cell output x_c is transmitted to N synapses located within its N neighboring cells, it is also appropriate that synapse input signals x_i be codified as voltages. Finally, because scaled signals must be added at the input of each cell's processing block, it is convenient that synapse output

0-7803-4867-2/98/$10.00©1998 IEEE

signals $A_i x_i$ be given in current form, eliminating the need of a dedicated summing circuit.

Although the required functionality of a programmable analog synapse may suggest the use of linear analog multipliers, there are some specific circumstances common to almost every analog-array processing algorithm which expand the set of selectable circuit blocks.

First note that while the synapses output current is expected to be linear with the input signal x_i, it is not required to be linear with the weight signal w_i, whose function is simply to allow weight variations in some prescribed range. Therefore, function $A_i(w_i)$ may be nonlinear in general.

Second, in almost every analog-array processing algorithm, the weight values are invariant during processing. Therefore, the dynamic response with respect to the weight signal is of little concern and, more important, after setting the weight values, any error or deviation from the ideal behavior independent of the input signal x_i (but in general dependent on w_i) may be cancelled using autozeroing, before performing the processing function.

This is a good practice in general because the offset of the aggregated signal is given by the addition of the output-current offset of the N synapses driving each cell. Indeed, this is often the dominant error source of this class of systems.

In the last years, synapse circuits based on MOS transistors operating in their ohmic region have been employed by several authors [2], [3]. The choice is based on a combined estimation of several performance figures (including area occupation, accuracy, linearity, programming weight range, signal-range, and power efficiency) which predicts important advantages as compared to other classes of synapse circuits based on the quadratic law of MOS transistors in saturation, or the exponential law of bipolar transistors and MOS transistors in weak inversion.

Regardless the family, practically all synapse circuits employ differential or fully differential architectures to achieve four-quadrants behavior and also for linearity reasons. Typical examples include those synapses based on differential pairs, like the Gilbert multiplier [4], and also the synapses employed in [2] and [3]. In addition to the larger complexity of the synapse, this usually forces the use of differential or fully differential architectures in the processing block as well, thus resulting in a substantial increase in area occupation.

In the following section we propose a one-transistor, four-quadrants, electrically-programmable synapse circuit with single-ended architecture.

4. A One-Transistor Synapse Circuit

The DC current of an MOS transistor operating in its strong-inversion ohmic region can be described by the following well-known first-order approximation

$$I_{DS} = \beta\left(V_{GS} - V_T(V_{SB}) - \frac{V_{DS}}{2}\right)V_{DS} \tag{2}$$

where

$$\beta = \mu C_{ox}\frac{W}{L} \tag{3}$$

$$V_T(V_{SB}) = V_{T0} + \gamma\left(\sqrt{V_{SB} - \phi_B} - \sqrt{\phi_B}\right) \tag{4}$$

and every symbols has its well established meaning in MOS literature.

Eq. (2) predicts an *incrementally* linear relation between I_{DS} and V_{GS}, and an *approximately* linear dependence with V_{DS} for $V_{DS} \ll 2[V_{GS} - V_T(V_{SB})]$. These considerations have been widely exploited for many applications, including MOS implementations of active RC filters [5], analog multipliers for RF communication circuits [6] and also synapse circuits for massively-parallel analog processing systems [2], [3] using differential architectures.

The use of a single transistor to implement an electrically programmable synapse with signals represented by single-ended voltages requires that one of the diffusion terminals be set to a fixed voltage level. The gate and the other diffusion terminals can then be employed as input points, while the output is obtained from the current flowing out of the fixed-voltage diffusion terminal. This is conceptually illustrated in Figure 2a in which a nullator and a DC voltage source represent the ideally null-impedance input terminal of the processing block in Figure 1. Such a virtual-reference level is required because the output impedance of the synapse is low due to its operation in the ohmic region. The input impedance at the diffusion input is also low, while that at the gate input is high. Two alternatives can then be considered: using the gate terminal for x_i and the diffusion terminal for w_i, or the other way around. This has implications on the output-impedance requirements for either the cell processing-block or the voltage sources driving the analog-weight control bus, but the major decision factor is related to linearity.

Equation (2) is valid only for $V_{DS} \geq 0$ and therefore, the use of the notation introduced in Figure 2a requires an independent consideration of the two possible cases $V_A \geq V_L$ and $V_A \leq V_L$. Still, simple analysis results in the following combined expression for I_N, valid in either of the two cases,

$$I_N = \beta(V_A - V_L)V_G - \beta\left(\hat{V}_T + \frac{V_A + V_L}{2}\right)(V_A - V_L) \tag{5}$$

with

$$\hat{V}_T = \begin{cases} V_{T0} + \gamma\left(\sqrt{V_L - \phi_B} - \sqrt{\phi_B}\right) & ; V_A \geq V_L \\ V_{T0} + \gamma\left(\sqrt{V_A - \phi_B} - \sqrt{\phi_B}\right) & ; V_A \leq V_L \end{cases} \tag{6}$$

Note that the second summand in (5) *is independent of* V_G and that the first summand is linear with V_G. Since we need a

linear behavior with respect to one of the inputs (x_i) and we can eliminate any systematic offset in a previous step, it seems straight forward that we can chose $x_i \equiv V_G$ and $w_i \equiv V_A$, as shown in Figure 2a.

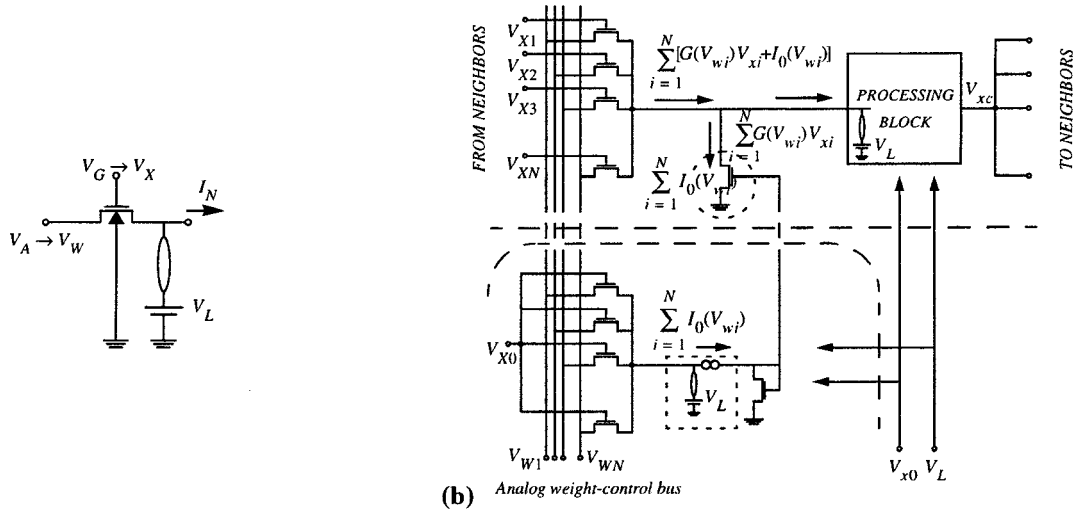

Figure 2. a) One transistor synapse concept, b) Cell (top) and peripheral (bottom) circuitry for independent-terms substraction.

There is still another issue related to the obtention of a four-quadrants behavior. While the two possibilities $V_A \geq V_L$ and $V_A \leq V_L$ provide double sign capability for the weight, V_G must always be positive. Therefore we must select a sufficiently high reference level on the gate voltage to act as the zero level for x_i. Let us define:

$$V_X = V_x + V_{x0} \Rightarrow V_x = V_X - V_{x0} \equiv x_i \tag{7}$$

and in the same manner, referring the weight voltage to V_L, this is, selecting $V_{w0} = V_L$, we have

$$V_W = V_w + V_L \Rightarrow V_w = V_W - V_L \equiv w_i \tag{8}$$

Using this new notation, equations (5) and (6) can be rewritten respectively as:

$$I_N = \beta V_w V_x + \beta V_w \left(V_{x0} - \hat{V}_T - V_L - \frac{V_w}{2} \right) \tag{9}$$

$$\hat{V}_T = \begin{cases} V_{T0} + \gamma\left(\sqrt{V_L - \phi_B} - \sqrt{\phi_B} \right) & ;V_w \geq 0 \\ V_{T0} + \gamma\left(\sqrt{V_w + V_L - \phi_B} - \sqrt{\phi_B} \right) & ;V_w \leq 0 \end{cases} \tag{10}$$

where both V_x and V_w can be either positive or negative and still, the first summand in (9) is linear with V_x and the second one is independent of V_x. We define the weight and the output offset of the synapse, respectively, as

$$G(V_w) = \beta V_w \tag{11}$$

$$I_0(V_w) = G(V_w)\left(V_{x0} - \hat{V}_T - V_L - \frac{V_w}{2} \right) \tag{12}$$

This allows (9) to be written in the following form

$$I_N(V_w, V_x) = G(V_w)V_x + I_0(V_w) \tag{13}$$

Let us now assume that we can eliminate the term $I_0(V_w)$, Then, we can define

$$I_N = I_n + I_0 \Rightarrow I_n = I_N - I_0 \equiv A_i x_i \tag{14}$$

and rewrite (13) as

$$I_n(V_w, V_x) = G(V_w)V_x \tag{15}$$

which is the equation of a four-quadrants, electrically-programmable, linear analog synapse. This relies only on the *separated* dependencies shown in (13), and not on the specific forms of $G(V_w)$ and $I_0(V_w)$, a fact that will become relevant for the consideration of second order effects in a latter section

5. Cell and Control Circuitry

In order to preserve the high area efficiency provided by the one-transistor synapses, the circuitry employed within each cell for the elimination of the N second summands (one per synapse) should be as simple as possible. This is made easier by the fact that the cell circuitry itself computes the *sum* of these summands, which can therefore be eliminated altogether.

Under the assumed spatial invariance of the weight signals, common to most analog-array processing systems, the term to

be eliminated in each cell is also spatially invariant. Therefore, we can reproduce its value in a small circuitry, shared by all the cells in the network and placed at the periphery of the cell array, and substract it at the input nodes of each cell processing block. This can be achieved, without additional cost in the cell circuitry, through proper *weight-dependent biasing* of the processing-block input stage.

As an example, Figure 2b describes a simple technique. The top part of the figure represents the cell circuitry (identical to Figure 1 using one-transistor synapses) which includes a *class-A* input stage as part of the processing-block input circuitry (bias current source is not shown for simplicity). The circled transistor performs the substraction of the N second summands in (13). The lower part of the figure describes the required biasing devices, placed at the periphery of the cell-array and shared by the whole network.

This technique relies on the use of *matched* current-conveyors [4] at the peripheral circuitry and at the input node of the cells processing-block. Their electronic implementation can be shown to be highly efficient in terms of area and power consumption [3].

It might be argued that large-distance mismatch effects could result in *cancellation errors* of the independent terms. However, as mentioned earlier, it is generally convenient to employ autozeroing techniques. Such autozeroing, which can be easily implemented with area-efficient current memories [7] in the cases considered (current-output synapses), would eliminate the cancellation error as well. Indeed, the use of autozeroing may render unnecessary the proposed substraction circuitry, since it could be used to eliminate the complete sum of independent terms rather than their remaining error. Although combining high-signal-ranges and high-absolute-accuracy is often difficult, the recently proposed class of S^2I current memories [8] may provide a good solution.

6. Operation Limits and Second Order Effects

One fundamental limitation to the operation range of the proposed synapse is imposed by the ohmic region limits of the MOS transistor, which can be approximated by

$$V_{GS} \geq V_{DS} + V_T(V_{SB}) \tag{16}$$

Substitution of the previously employed notation in this equation yields the following *lower* limit for the gate voltage

$$V_X \geq \begin{cases} V_L + V_T(V_W) & ;V_W \leq V_L \\ V_W + V_T(V_L) & ;V_W \geq V_L \end{cases} \tag{17}$$

Except for this limit, no other restrictions exist on the proposed synapse, on the basis of the first order model considered. It can be shown that there is only one second order effect, *mobility degradation*, which represents a relevant deviation from the functional dependence expressed in (13). Other second order effects affect only to the precise form of $G(V_w)$ and $I_0(V_w)$, something irrelevant for our discussion.

Mobility degradation models predict a reduction in the effective carriers mobility (μ in equation (3)) with transversal (normal to channel surface) electric field, something that affects our present discussion because the transversal electric field depends on the gate voltage and thus, the first summand will not be linear with V_x. Although the widely accepted *simple* model for mobility degradation [9].

$$\mu = \frac{\mu_0}{1 + \theta (V_{GS} - V_T(V_{SB}))} \tag{18}$$

predicts a continuous reduction of the effective mobility starting just above $V_{GS} = V_T(V_{SB})$, the fact is that in most technologies, there is an appreciable (~2.0v) V_{GS} range above $V_T(V_{SB})$ within which mobility reduction is negligible. Furthermore, some higher level models accounting for mobility degradation employ a specific parameter to define a field threshold below which no mobility degradation occurs (*UCRIT* in SPICE level 2 [10]). Regardless the continuous or thresholded modelling of mobility degradation, we can always define a *maximum effective gate voltage*

$$V_{GE\mathrm{MAX}} = [V_{GS} - V_T(V_{SB})]_{\mathrm{MAX}} \tag{19}$$

below which any reasonable linearity requirements are satisfied. The operation of the synapse must be restricted to this range.

Performing the appropriate substitutions in (19) yields the following *upper* limit for the gate voltage,

$$V_X \leq \begin{cases} V_W + V_T(V_W) + V_{GE\mathrm{MAX}} & ;V_W \leq V_L \\ V_L + V_T(V_L) + V_{GE\mathrm{MAX}} & ;V_W \geq V_L \end{cases} \tag{20}$$

The selection of V_L and V_{x0} must be made based on the limits imposed by (17) and (20). In turn, this will result in an upper limit for the allowed signal ranges of V_w and V_x. Equation (17) can be rewritten as,

$$\begin{aligned} V_x + V_{x0} &\geq V_w + V_L + V_T(V_L) \quad ;V_w \geq 0 \\ V_x + V_{x0} &\geq V_L + V_T(V_w + V_L) \quad ;V_w \leq 0 \end{aligned} \tag{21}$$

Let us denote the signal ranges of V_w and V_x by $|V_w| \leq V_{w\max}$ and $|V_x| \leq V_{x\max}$, respectively. In the above equation, the worst-case limit for V_{x0} is given by the first inequality when $V_w = V_{w\max}$ and $V_x = -V_{x\max}$, which yields,

$$V_{x0} \geq V_L + V_T(V_L) + V_{w\max} + V_{x\max} \tag{22}$$

Regarding V_L, its value must be sufficiently high to provide room for the minimum V_w value and also for some possible loss of voltage range due to the limited output swing of the circuits generating the analog weight control signals, which we denote as V_{Wmin}. This is,

$$V_L \geq V_{wmax} + V_{Wmin} \tag{23}$$

Because an upper limit for the voltage ranges exist due to mobility degradation, and also because we are interested in maximizing the signal ranges and/or allowing a reduced power supply operation, we will select the minimum allowed value for V_L. Substituting $V_L = V_{wmax} + V_{Wmin}$ in (22) results in a minimum value for V_{x0},

$$V_{x0} \geq V_{Wmin} + V_T(V_{wmax} + V_{Wmin}) + 2V_{wmax} + V_{xmax} \tag{24}$$

Again, we select V_{x0} as its minimum allowed value. The resulting maximum value for V_X is then given by $V_{Xmax} = V_{x0} + V_{xmax}$, this is,

$$V_{Xmax} = V_{Wmin} + V_T(V_{wmax} + V_{Wmin}) + 2V_{wmax} + 2V_{xmax} \tag{25}$$

and the worst case mobility degradation limit will be imposed by the first inequality in (20), when V_W is minimum, this is, $V_W = V_L - V_{wmax} = V_{Wmin}$, which

yields,

$$V_{Xmax} \leq V_{Wmin} + V_T(V_{Wmin}) + V_{GEMAX} \tag{26}$$

Using (25) into (26), yields,

$$V_{wmax} + V_{xmax} \leq \frac{1}{2}V_{GEMAX} - \frac{1}{2}[V_T(V_{wmax} + V_{Wmin}) - V_T(V_{Wmin})] \tag{27}$$

For moderate linearity requirements, the right hand side of the above equation takes values in the range of one volt for typical CMOS technologies, which for $V_{wmax} = V_{xmax}$ provides a peak-to-peak signal range of about one volt for both V_x and V_w.

Figure 3 illustrates the above discussion and shows the voltage distribution selected for a particular technology: a standard n-well, 0.8μm CMOS process available through EUROPRACTICE. With small changes, these values should be valid for most typical CMOS technologies. Note that the minimum power supply level should be at least slightly above $V_{Xmax} = 3.4v$ to pre-

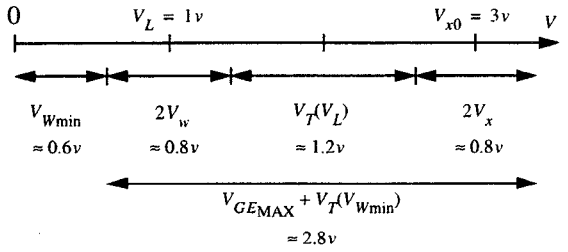

Figure 3. Voltage range distribution for synapse operation.

vent the possible loss of voltage range due to the limited output swing of the processing block. Still, an optimization of the output swing of both the analog weight control drivers and the output stage of the processing block should allow the operation of the proposed synapse with power supply levels in the range of 3.3v, with similar signal swings for V_x and V_w.

Figure 4 provides an additional insight into the selection of V_L and V_{x0} values and the associated signal ranges. It shows the allowed operation region, delimited by (17) and (20) in the $V_X \cdot V_W$ plane, within which a squared range (under the assumption $V_{wmax} = V_{xmax}$) for signals V_x and V_w, centered around (V_L, V_{x0}), must be defined. The graphs correspond to $V_L = 1v$ and the specific parameters of the technology being employed. Note that although apparently, an appreciable increase in signal ranges could be obtained by increasing the values of V_L and V_{x0}, this is not true in general because the limits imposed by (17) and (20) will also shift with V_L. In view of (27), the increase would be small. On the other hand, it would require a larger power supply.

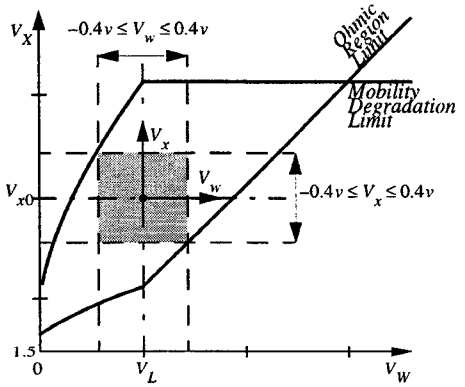

Figure 4. Signal ranges delimited by ohmic region and mobility degradation limits.

7. Results

In this section we will illustrate the behavior of the proposed synapse in a specific n-well, 0.8μm CMOS technology available through EUROPRACTICE. Figure 5 contains HSPICE level 2 simulated transfer characteristics of the proposed synapse. Transistors sizes are $W = 6\mu m$ and $L = 24\mu m$. These geometries correspond to reasonable sizes in a practical application, in which a *low aspect ratio* serves to the purposes of having reasonable current levels through the analog weight-control lines, as well as moderate power dissipation in the chip. Large channel areas (relative to technology resolution) are required for matching considerations [1]. Still, low resolution technologies are highly convenient for matching considerations and also because most of the cell area is usually dedicated to contacts, routing, and active region separations.

Figure 5a reflects the total transistor current I_N versus the total gate voltage V_X, for different values of the weight signal voltage V_w. The value of V_L is $1.0v$. Values of V_w, relative to V_L, range from - 0.4v (lower trace) to + 0.4v (upper trace) in 50mv increments. The mobility degradation effects are clearly visible at the right side, while those related to the pinch-off region can be observed at left side, specially for positive values of V_w. The region around $V_{x0} = 3.0v$ reflects the behavior predicted by (13). Figure 5b shows the result obtained after substracting the independent term $I_0(V_w)$, by means of the circuitry described in Figure 2b. Wort-case $(V_w = 0.4v)$ total harmonic distortion (THD) is below 0.05% at 1Hz and 0.7% at 1MHz. Similar results are obtained from the p-channel version of the circuitry.

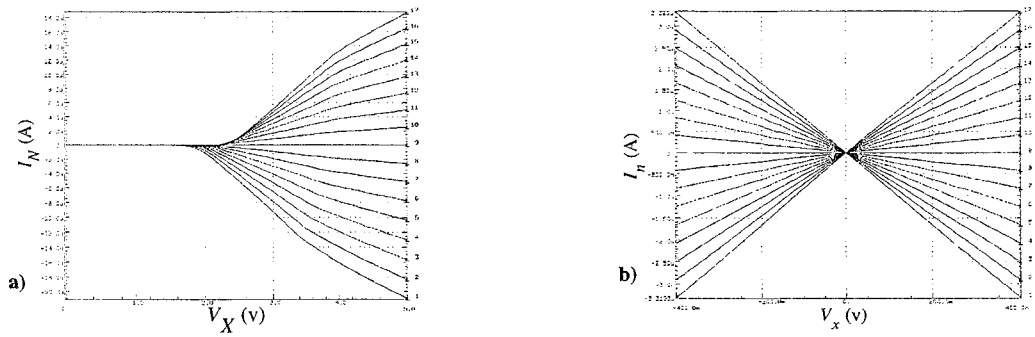

Figure 5. HSPICE level-2 simulations of the proposed synapse: a) Current I_N in Figure 2 versus gate voltage V_X for different values of V_w. b) Current $I_n = I_N - I_0$ obtained using the circuitry described in Figure 2b.

8. Conclusions

This paper has proposed and discussed an electrically programmable, one-transistor, four-quadrants linear synapse strategy for massively-parallel analog array processing systems, based on MOS operation in the triode region. Signal ranges for both the input and the weight signal are in the range of 1vpp, and total harmonic distortion is below 0.7% at 1MHz. Operation from reduced power supplies of about 3.3v seems feasible. The proposed synapse circuit results in a substantial reduction in area and power consumption of the basic cell, as compared to traditionally employed synapses based on differential or fully-differential architectures. This allows the realization of array processors with a larger number of units in the same chip.

Acknowledgment: This work has been funded by spanish CICYT under contract TIC96-1392-C02-02 (SIVA: Active Vision Integrated System).

9. References

[1] M.J.M Pelgrom, A.C.J. Duinmaijer and A.P.G. Welbers: "Matching Properties of MOS Transistors". *IEEE J. Solid-State Circuits*, Vol. 24, pp 1433-1440, October 1989.

[2] P. Kinget and M. Steyaert, "An Analog Parallel array Processor for Real-Time Sensor Signal Processing", *1996 Int. Solid State Circuits Conference*, paper 6.1, 1996.

[3] S. Espejo, A. Rodríguez-Vázquez, R. Carmona and R. Domínguez-Castro: "A 0.8μm CMOS Programmable Analog-Array-Processing Vision-Chip with Local Logic and Image-Memory". 1996 European Solid State Circuits Conference, pp. 276-279. Neuchâtel, September 1996.

[4] C. Toumazou, F.J. Lidgey, D.G. Haigh (Eds.): "Analog IC Design: the Current-Mode Approach", Peter Peregrinus, 1990.

[5] Y.P. Tsividis: "Integrated Continuous-Time Filter-Design -- An Overview". *IEEE J. Solid-State Circuits,* Vol. 29, pp 166-176, march 1994.

[6] B. Song: "CMOS RF Circuits for Data Communication Applications", *IEEE J. Solid-State Circuits*, Vol.21, pp 310-317, April 1986.

[7] C. Toumazou, J.B. Hughes, N.C. Battersby (Eds.): "Switched-Currents an Analog Technique for Digital Technology", Peter Peregrinus, 1993.

[8] J.B. Hughes and K.W. Moulding: "S^2I: A Switched-Current Technique for High Performance". Electronic Letters. Vol.29, No. 16, pp. 1400-1401, August 1993.

[9] Y.P. Tsividis, *"Operation and Modeling of the MOS Transistors"*. New York: McGraw-Hill, 1987.

[10] P Antognetti, G. Massobrio (Eds.): "Semiconductor Device Modeling with SPICE", McGraw-Hill, 1988.