Robust Symmetric Multiplication for Programmable Analog VLSI Array Processing

C. Domínguez-Matas, R. Carmona-Galán, F. J. Sánchez-Fernández, A. Rodríguez-Vázquez
Instituto de Microelectrónica de Sevilla-CNM-CSIC
Campus de la Universidad de Sevilla. Avda. Reina Mercedes s/n,
41012-Sevilla, Spain. E-mail: cmanuel@imse.cnm.es

Abstract— **This paper presents an electrically programmable analog multiplier. The circuit performs the multiplication between an input variable and an electrically selectable scaling factor. The multiplier is divided in several blocks: a linearized transconductor, binary weighted current mirrors and a differential to single-ended current adder. This paper shows the advantages introduced using a linearized OTA-based multiplier. The circuit presented renders higher linearity and symmetry in the output current than a previously reported single-transistor multiplier. Its inclusion in an array processor based on CNN allows for a more accurate implementation of the processing model and a more robust weight distribution scheme than those found in previous designs.**

I. INTRODUCTION

Real-time image processing is an extremely demanding computing task that can easily exceeds the capabilities of a conventional serialized digital signal processing scheme. The extraordinary amount of data contained in the visual stimuli is difficult to handle by conventional microprocessors. They usually do it at the expense of large physical profile and considerable energy consumption. This should not be a problem when dealing with machine vision applications in industrial environments, but it is certainly a drawback when trying to migrate automatic vision to different scenarios. In applications like robotic vision [1], sensor networks for ambient intelligence [2] and retinal prosthesis for the blind [3], for example, power efficient computation and the use of the simplest and the least hardware possible are mandatory.

One way to avoid the signal processing bottlenecks inherent to a serialized scheme —camera plus A/D converter plus digital processor—, is to convey an important fraction of the computing facilities to the focal plane. With this, the architecture of the system adapts to the nature of the stimuli [4]. This characteristic is quite common in biological sensory organs, in which high performance is obtained by exploiting a high parallelism [5]. For the efficient implementation of array processing in VLSI, analog and mixed-signal circuits represent a good alternative. On one side, A/D conversion at the pixel level is avoided. On the other, for the moderate accuracy required in sensory applications, analog functional blocks occupy less area and consume less power. Many neuromorphic models render multi-dimensional signal processing as a result of the evolution of the network dynamics, described by a set of coupled reaction-diffusion equations. In the most of the developments emulating the dynamic processing capabilities of biological retinas, the cooperative behavior between the different processing nodes of the array is hard-coded into the network architecture [6]. In our work, we have tried to maintain a reasonable level of programmability while trying to enhance the robustness of the implementation of the multiplier. Also to solve the limitations found in less linear and less symmetrical multiplying blocks. Based on this multiplier, we have developed a programmable 3-layer Cellular Neural Network [7] (CNN) that constitutes the central element of a vision chip for fast and efficient focal-plane image processing.

II. CNN PROCESSING UNIT

The dynamic behaviour of this network is described in terms of the input ($\mathbf{u}_k$), state ($\mathbf{x}_k$) and output ($\mathbf{y}_k$) variables. Each layer, k, of the array follows the evolution law expressed in:

$$\tau_k \frac{d\mathbf{x}_k(t)}{dt} = -\mathbf{g}[\mathbf{x}_k(t)] + \sum_n [\mathbf{A}_{kn} \otimes \mathbf{y}_n + \mathbf{B}_{kn} \otimes \mathbf{u}_n] + \mathbf{z}_k \qquad (1)$$

The symbol ⊗ stands for the linear convolution between the so-called feedback and feedforward templates, with the output and input matrices of layer, n, where n can be 1, 2 or 3. If the full-signal-range CNN model is applied [8], the output and state variables can be identified. The operators, responsible for multiplying the state (or input) variable by the programmable weight, are referred to as synapses or synaptic blocks. They are basically four quadrants multipliers in which linearity with the state (or input) variable and a symmetric characteristic are strongly desired. The effect of varying the programmed weights is to modify the network dynamics, and thus, changing the type of processing realized by the array. The losses term and the activation function are also those of the FSR CNN model [8].

1-4244-0395-2/06/$20.00 ©2006 IEEE.

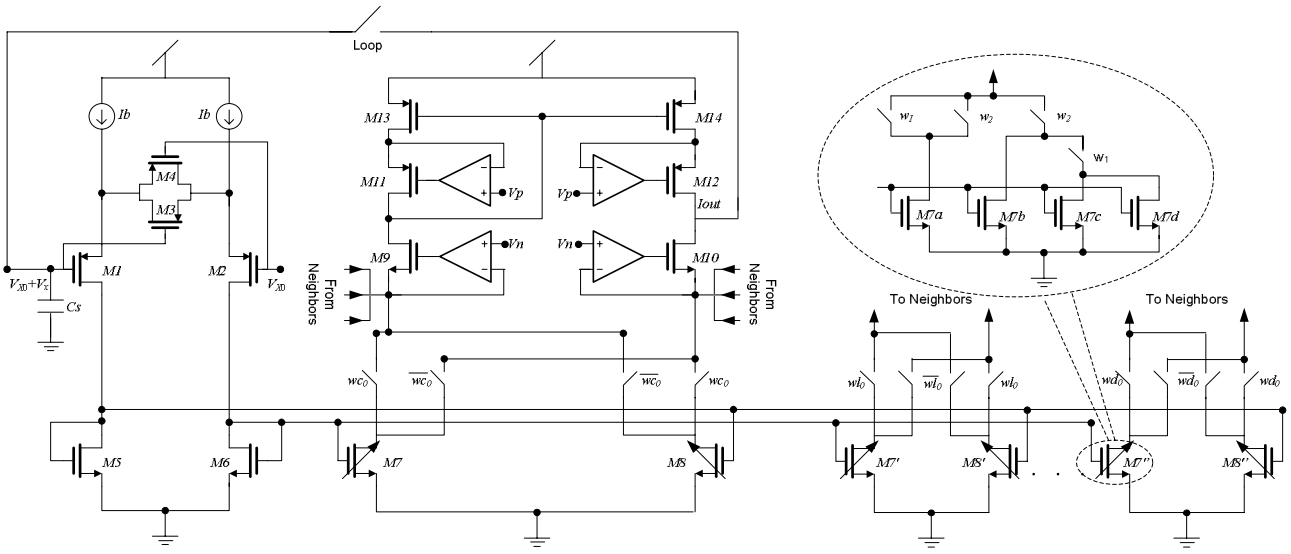

Figure 1. The schematic of synapse based on linearized OTA

The physical realization of the elementary processing unit of the CNN starts with the selection of the appropriate format for signal representation. Therefore, input, output and state variables are chosen to be matrices of voltages: $\mathbf{V}_u$, $\mathbf{V}_y$ and $\mathbf{V}_x$. On the other side, signal addition can be easily realized in the form of currents. Then, the summands in the second member of Eq. (1) can be represented by currents. This sum of currents will be integrated in the state capacitor to obtain the instantaneous state variable voltage:

$$C_k \frac{d\mathbf{V}_{x,k}(t)}{dt} = -\mathbf{G}_g[\mathbf{V}_{x,k}(t)] + \sum_n [\mathbf{G}_{A,kn} \otimes \mathbf{V}_{y,n} + \mathbf{G}_{B,kn} \otimes \mathbf{V}_{u,n}] + \mathbf{I}_{z,k} \quad (5)$$

Here, the elements of the feedback and feedforward templates, $\mathbf{A}_{kn}(i,j)$ and $\mathbf{B}_{kn}(i,j)$, are now linear transconductances, $\mathbf{G}_{A,kn}(i,j)$ and $\mathbf{G}_{B,kn}(i,j)$, that render the current contributions of the neighbours. Thus, the synaptic block is a transconductor whose output current is proportional, in the ideal case, to the product of the state (or input) variable and the weight. The accuracy of these terms is very important to accomplish a correct operation of the network, since the synapse errors are integrated in the state capacitor. The double transformation implicit in Eq. (5), V-I and then I-V, allows for a compact realization of the processing node, achieving higher cell densities. The main linearity concerns are found in the V-I conversion. In this design, we have employed a linearized OTA in order to generate the unitary current contribution. Though the elementary transconductor achieving V-I conversion has a larger number of transistors than the single-transistor synapse in [9], advantages in linearity with the state (or input) variable and symmetry of the V-I characteristic justify its use. In addition, the supporting circuitry can be simplified resulting in a more robust implementation finally without a serious area penalty (see table in Sect. IV).

III. SYNAPSE BASED ON A LINEARIZED OTA

The circuit in Fig.1 constitutes the core of the elementary dynamic processor. Operating in closed loop (when the switch controlled by signal Loop is on), it implements the evolution law described by Eq. (5). The weighted V-I conversion of the state voltage is carried at several stages. The single-to-differential V-to-I conversion is realized by a linearized transconductor (at the left). Currents are scaled and replicated by programmable mirrors to generate the contributions towards the neighborhood (center and right of the figure). These contributions are added to self-feedback and integrated in the state capacitor, when the loop is closed (center).

The transconductor responsible of transforming the single-ended state capacitor voltage (V_x-V_{X0}) into a differential current is a source degenerated differential pair with diode-connected loads. It is based on the linearized CMOS differential transconductance amplifier described by Krummenacher in [10]. The output current of the OTA, when the degeneration transistors operate in triode region, is given by:

$$I_{out} = \frac{\sqrt{2\beta_1 I_b}}{1 + \frac{\beta_1}{4\beta_3}} V_x \sqrt{1 - \frac{\beta_1 V_{in}^2}{(1 + \frac{\beta_1}{4\beta_3})^2 I_b}} \quad (7)$$

M1-M2, and M3-M4 are matched pairs. If the degeneration transistors move into saturation, transfer function changes to:

$$I_{out} = \left[\frac{V_{in}\sqrt{\beta_1(4a-2)} + \sqrt{(8a-2)I_b - \beta_1 V_{in}^2}}{4a-1} \right]^2 \quad (9)$$

The best linearity performance is obtained when the ratio β_1/β_3 is around 7 and the transistors are biased with reduced current density [10]. In these conditions the deviation from the average transconductance can be below 1%. The operation of this circuit alone is inherently symmetric if working in fully-differential mode, representing an enhancement from what have been achieved by previous implementations. This symmetry though is broken by using a single-ended input voltage, but still the resulting V-I characteristic maintains symmetry levels beyond those of other implementations.

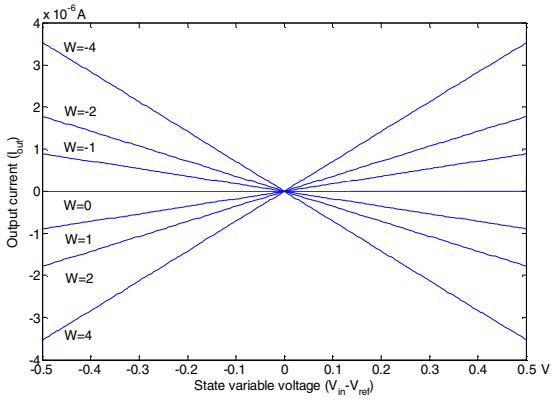

Figure 2. Output current vs. state voltage of the OTA-based synapse

The implementation of the weights is based on geometrical relations. This has the advantage of being less influenced by process parameter variations both inter- and intra-die. It has also the drawback of only permitting a discrete set of weight values (–4, –2, –1, 0, 1, 2 and 4). Opposite-sign contributions are obtained by crossing the wires conveying the currents, thus achieving a symmetric operation by architecture. Finally, the sum of all the currents coming from the neighborhood is injected into the target state capacitor. But before that, differential to single-ended current conversion is realized with the help of a current mirror. It is important to mention that the achievable output resistances using self-biased or externally biased Cascode current mirrors are not sufficient to achieve a minute error in the copied current. Hence, gain boosting of the Cascode devices is needed. The accuracy of the current replication in this mirror is crucial for achieving the required linearity and symmetry in the V-I characteristic.

The multiplier in Fig.1 has been designed in a 0.35µm CMOS technology. In order to characterize its behavior, simulations have been carried out in open loop with a 1V range for the input signal, null contributions from the neighbors, and for different weight values, using HSPICE transistor models level 49. Fig.2 shows the output current of the multiplier corresponding to the central element of the weight matrices, i. e. the self-feedback or self-feedforward contribution, vs. the input voltage. The output current has a high linearity. The transconductance relative error in large signal is below 0.7%. Concerning the symmetry of the characteristic, the difference of the output currents corresponding to weights with the same absolute value but opposite sign is zero on average because offset cancellation, with the typical deviation being 2% of the absolute value of the individual currents.

IV. Comparison with Single-Transistor Synapse

The single-transistor four-quadrant multiplier, described in [9], is based on the linear dependence between I_{ds} and V_{gs}, when the MOS transistor is biased in the ohmic region in strong inversion. Fig.4 shows the output current versus input voltage (I_{out} vs. V_x) of the PMOS single-transistor synapse. Transistor sizes are W=1µm and L=13µm. It has been designed in same 0.35µm CMOS technology employed in the previous case. V_{X0} is chosen to be 2.7V, to maximize the available voltage range for the weights, varying here from 2.3 to 3.1V, while using a 0.8V range for the state variable, to avoid going off the ohmic region. The maximum weight differential voltage corresponds to 4 by definition, and the other are proportional to the weights being implemented. The output current presents an important asymmetry respect to the origin. The maximum difference in the output currents corresponding to weights with the same absolute value but opposite sign is now 9%, mainly due to mobility degradation, since the transconductance depends on V_x. In this occasion, device mismatch is not as important.

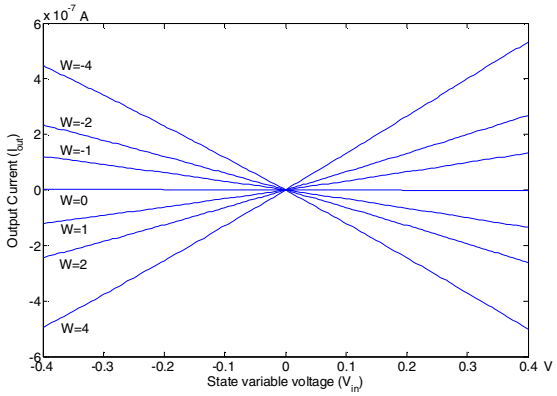

Figure 3. Output current of the 1-T synapse

If a comparison is established between both multipliers, there are some advantages in using the linearized OTA:

- The linearity of the multiplier increases, since the transconductance relative error is one order of magnitude lower in the OTA-based multiplier than in the synapse based on single-transistor.
- The OTA-based synapse is less affected by the variations of the process parameters than the single-transistor synapse. This is because the output current of the latter depends lineally on process parameters, whereas the output current of the OTA-based synapse depends on geometrical relations among transistors, whose scattering is better controlled.

- The reproducibility of the algorithm in a system built upon OTA-based multipliers is enhanced, since the sensitivity to process variations is low for this multiplier and so is the sensitivity of the weights. Because of that, the same weight code represents the same network dynamics even for different samples of the chip, undergoing process parameter changes.
- The weight signals distribution across the array is simplified in the OTA-based synapse. In a network employing single-transistor synapses, a DC current consumption is needed since these signals are broadcasted through low-impedance nodes. When the OTA synapse is employed, the weights are distributed to high impedance nodes only requiring transient current consumption.

On the other side, there are some disadvantages in the use of the synapse based on linearized OTA:

- With the intention of achieving a compact implementation of the OTA-based synapse, transistors with small sizes have been used. These devices are more sensitive to mismatch what originates undesirable effects in the multiplier. This is not observed, in principle, in the single-transistor synapse because it does not rely in device matching to operate, though mismatch can affect to the previously referred weight uniformity through the array.

Concerning response times, the transconductances of both alternatives are of the same order (2.5 and 3.5μA/V), thus, for the same state capacitor, we will have similar time constants. Weight signals are not intended to change during network evolution. In what refers to power consumption, the elementary 1-T synapse consumes 1.075μA, while the unitary current in the other case is 2.4μA. In both designs, the most of the power is demanded by the rest of the circuitry supporting the network dynamics, adding up to 300μA in both chips, the one based on the 1T-synapse [11] and that using the linearized OTA. The following table permits to establish a comparison on the final area requirements:

	CACE1k	CACE2	CACE2[1]	CACE2[2]
Total area	34200	30976	44251	37046
CNN layers	18196	12385	17692	17692
Memories	4928	2652	3789	1894
Photo-sensor	0	1118	1597	0
3rd layer	0	1404	2006	0
Synapses	2112	3240	4629	4629
Wiring	8964	10177	14539	14539

All the quantities expressed in $(\mu m)^2$.

1. CACE2 scaled up to a 0.5μm process
2. CACE2 w/o 3rd layer, sensor and half of the memories

V. CONCLUSIONS

In this paper an electrically programmable analog multiplier suitable to be included in a massively-parallel array processor is proposed. If we compare the results obtained with multiplier based on the linearized OTA with the results of the 1-T multiplier, we can conclude that the proposed architecture reduces the linearity error (from a transconductance error of a 6% to a 0.7% now) and increases the symmetry of the characteristic. The deviation in the output currents corresponding to opposite-sign weights is reduced to one fourth. In the OTA-based case the mismatch effects dominate and the single-transistor case the inherent asymmetry is the more important source of deviation. The weight implementation is now more robust and the design does not take much more area, an extra 8%. The inclusion of this multiplier in an array processor improves the accuracy in the implementation of the processing model and the network dynamics. A 32x32 CNN array processor has been designed in a 0.35μm CMOS technology. The prototype has already been fabricated and now it is being tested. We expect to present some experimental results of the synapse and its influence in the operation of the complete system in the conference.

ACKNOWLEDGEMENTS

This work has been supported by the Office of Naval Research through ONR contract N-00014- 02-1-0884 and by the Spanish MCyT through TIC2003-09817-C02-01.

REFERENCES

[1] T. Makimoto, T. T. Doi, "Chip Technologies for Entertainment Robots - Present and Future". *Int. Electron Devices Meeting*, pp. 9-16, Dec. 2002.

[2] E. Aarts, R. Roovers, "IC Design Challenges for Ambient Intelligence", *Design, Automation and Test in Europe (DATE)*, pp. 2-7, March 2003.

[3] Eyal Margalit et al. "Retinal Prosthesis for the Blind", *Survey of Ophthalmology*, Vol. 47, No. 4, pp. 335- 356, July-August 2002.

[4] C. Diorio, D. Hsu, M. Figueroa, "Adaptive CMOS: From Biological Inspiration to Systems-on-a-Chip". *Proceedings of the IEEE*, Vol. 90, No. 3, pp. 345-357, March 2002.

[5] D. H. Hubel, *Eye, Brain and Vision*. Scientific American Library, No. 22. W. H. Freeman and Co., New York, 1995.

[6] S. Kameda, T. Yagi, "An Analog VLSI Chip Emulating Sustained and Trans. Response Channels of the Vertebrate Retina". IEEE Transactions on Neural Networks, Vol. 14, No. 5, pp. 1405-1412, Sept. 2003

[7] L. O. Chua, T. Roska, "The CNN Paradigm", IEEE Transactions on Circuits and Systems-I, Vol. 40, No. 3, pp. 147-156, March 1993.

[8] S. Espejo, R. Carmona, R. Domínguez and A. Rodríguez, "A VLSI Oriented Continuous-Time CNN Model". Int. J. of Circuit Theory and Apps., Vol. 24, No. 3, pp. 341-356, May-June 1996.

[9] R. Domínguez, S. Espejo, A. Rodríguez and R. Carmona, "Four-Quadrant 1-T Synapse for High-Density CNN Implementations". Proc. 5th Int. W. CNNs and Apps., pp. 243-248, London, UK, April 1998.

[10] F. Krummenacher and N. Joehl, "A 4-MHz CMOS Continuous-Time Filter with On-Chip Automatic Tuning". IEEE J. of Solid-State Circuits, Vol. 23, pp. 750-758, June 1988.

[11] R. Carmona et al., "2nd-Order Neural Core for Bioinspired Focal-Plane Dynamic Image Processing in CMOS". IEEE Trans. Circuits and Systems—I, 51 (5) pp.913-925, May 2004.