An Analog Programmable Multidimensional Radial Basis Function Based Classifier

Sheng-Yu Peng, *Student Member, IEEE*, Paul E. Hasler, *Senior Member, IEEE*, and David V. Anderson, *Senior Member, IEEE*

***Abstract*—A compact analog programmable multidimensional radial basis function (RBF)-based classifier is demonstrated. The probability distribution of each feature in the templates is modeled by a Gaussian function that is approximately realized by the bell-shaped transfer characteristics of a proposed floating-gate circuit, which we term a floating-gate bump circuit. The maximum likelihood, the mean, and the variance of the distribution are stored in floating-gate transistors and are independently programmable. By cascading these floating-gate bump circuits, the overall transfer characteristics approximate a multivariate Gaussian function with a diagonal covariance matrix. An array of these circuits constitute a compact multidimensional RBF-based classifier that can easily implement a Gaussian mixture model. When followed by a winner-take-all circuit, the RBF-based classifier forms an analog vector quantizer. We use receiver operating characteristic curves and equal error rate to evaluate the performance of our RBF-based classifier as well as a resultant analog vector quantizer. We show that the classifier performance is comparable to that of digital counterparts. The proposed approach can be at least two orders of magnitude more power efficient than the digital microprocessors at the same task.**

***Index Terms*—Analog classifier, bump circuit, floating-gate transistor, Gaussian-like analog circuit, radial basis function (RBF), vector quantizer.**

I. Motivations for Analog RBF Classifier

MULTIVARIATE Gaussian response functions can be used as building blocks in many applications including radial basis function (RBF)-based classifiers, Gaussian mixture modeling of data, and vector quantizers. This paper discusses the development of an analog Gaussian response function having a diagonal covariance matrix and demonstrates its application to vector quantization.

When followed by a winner-take-all (WTA) stage, a RBF-based classifier forms a multidimensional analog vector quantizer. A vector quantizer compares distances or similarities between an input vector and the stored templates. It classifies the input data to the most representative template. Vector quantization is a typical pattern recognition and data compression technique. Crucial issues of the vector quantizer implementation concern the storage efficiency and the computational cost of searching the best-matching template. In the past decade, efficient digital [2], [3] and analog [4]–[6] hardware vector quantizers have been developed. In general, the analog vector quantizers have been shown to be more power efficient than their digital counterparts. However, in a previous design [4], the computational efficiency is partially due to the fact that only the mean absolute distances between the input vector and the templates are compared instead of considering the possible feature distributions. To have better approximation to the Gaussian distribution, many variations of analog RBF circuits are designed [6]–[11]. Among these previous works, the simple "bump" and "anti-bump" circuits in [7] are the most classic; however, the widths of their bell-shaped transfer curves are fixed. To be able to change the width of the transfer characteristics, circuits usually become complicated. An analog RBF-based classifier can be a critical building block in an analog signal-processing front-end system for speech recognition [1]. Fig. 1 illustrates one possible application of this work as part of an analog speech recognizer that includes a band-pass-filter bank based analog Cepstrum generator, an analog RBF-based classifier, and a continuous-time hidden Markov model (HMM) block built from programmable analog waveguide stages. The input to the HMM stage could represent the RBF response directly or it could pass through a logarithmic element first. By performing analog signal processing in the front-end, not only the computational load of the subsequent digital processor can be reduced, but also the required specifications for the analog-to-digital converters can be relaxed in terms of speed, accuracy, or both. As a result, the entire system can be more

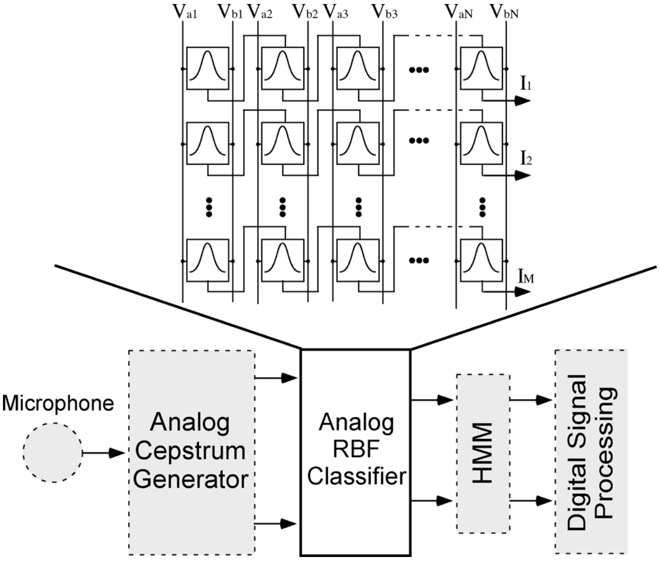


Fig. 1. Analog RBF-based classifier in an analog front-end for speech recognition. The front-end of our current speech recognition system includes a band-pass-filter bank based analog Cepstrum generator, an analog RBF-based classifier, and a continuous-time HMM. Putting the DSP stages behind the analog front-end makes the entire system more efficient.

Manuscript received July 24, 2006; revised February 22, 2007. This paper was recommended by Associate Editor A. Kuh.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-250 USA.

Digital Object Identifier 10.1109/TCSI.2007.905642

power efficient. Additionally, all of the analog RBF or vector quantization circuits reported in [6]–[11] require extra circuits to store or to periodically refresh template data. In [5], [12], and [13], floating-gate transistors are used to implement the bump and anti-bump circuits. Because the template data are stored in the form of charges on floating gates, the circuits are very compact. Particularly in [12] and [13], two adaptive versions of the floating-gate bump and anti-bump circuits are introduced to implement competitive learning. Although the bump centers in these circuits are adaptive to the mean values, the bump widths are still constant, limiting their applications. As will be shown later in this paper, our new floating-gate circuit has the potential to adapt to both the mean and the variance of the distribution.

This paper demonstrates a novel compact programmable analog RBF-based classifier that is composed of an array of two-input floating-gate bump circuits. The mean and the variance of each feature component distribution are stored in a floating-gate circuit that we term a floating-gate bump cell. These two statistical moments can be programmed accurately and independently; therefore, the stored template information can be closer to the real distributions. An array of these floating-gate bump cells can also implement Gaussian mixture models (GMMs). With a following WTA circuit, the resultant analog vector quantizer can be applied to nonuniform, as well as uniform, variance scenarios. The whole classification system is compact and can be easily scaled up.

In the next section, we present our new programmable floating-gate bump circuit, which is the most crucial element in our RBF-based classifier. In Section III, we briefly review the techniques of programming an array of floating-gate transistors. In Section IV, we illustrate the complete schematic of the floating-gate bump circuit and the architecture of a resultant analog vector quantizer. In Section V, we show the measurement results from an analog vector quantizer implementation. We evaluate the classifier performance by means of receiver operating characteristic (ROC) and equal error rate (EER). The power efficiency of the classifier is investigated in Section VI. The conclusion is drawn in the final section.

II. Programmable Floating-Gate Bump Circuit

In our classifier, the Gaussian response function is approximated by the bell-shaped transfer characteristics of a proposed floating-gate bump circuit. The height, the width, and the center of the transfer curve represent the maximum likelihood, the variance, and the mean of a distribution, respectively. The ability to program these three parameters individually empowers the classifiers to fit into different scenarios with the full use of statistic information up to the second moment. In addition, adjusting these parameters is equal to pre-scaling input signals in the analog fashion so that the circuit outputs can fall into the effective input range of the following stage. For example, in the analog vector quantizer implementation, despite the different distributions in different applications, the required precision of the following WTA circuit can remain relaxed if the input signals can be scaled properly.

The schematics of the proposed floating-gate bump circuit and its bias generation block are shown in Fig. 2. All floating-gate transistors have two input capacitances and all input capacitances are of the same size. The proposed bump circuit is composed of three parts: an inverse generation block, a conventional bump circuit [7], and in between a fully differential variable gain amplifier (VGA). The inverse generation block provides the complementary input voltages to the VGA so that the floating-gate common-mode voltage of M_{21} and M_{22} as well as the outputs of the VGA are independent of the input signal common-mode level. The width of the bell-shaped transfer curve can be adjusted by changing the VGA gain.

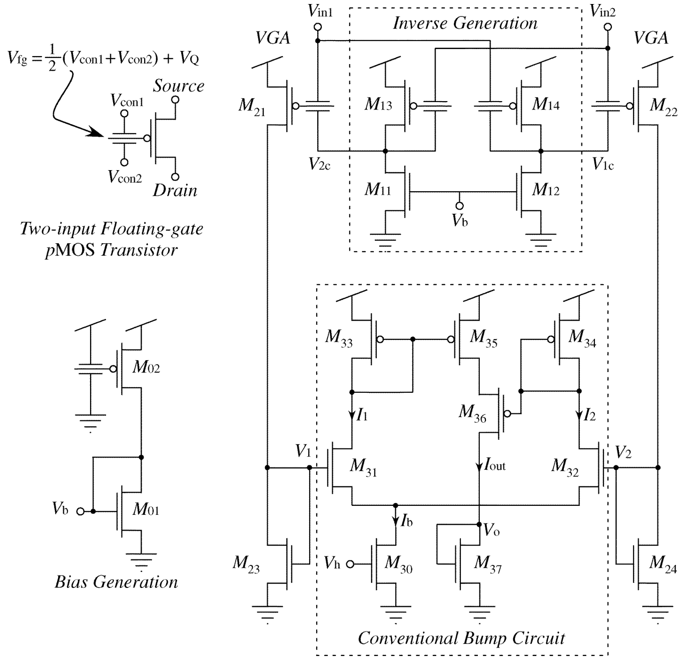


Fig. 2. Schematics of the new floating-gate bump circuit. All floating-gate transistors in the schematics have two inputs with equal weights and the floating-gate voltage can be expressed as $V_{\rm fg} = 1/2(V_{\rm con1} + V_{\rm con2}) + V_Q$, where $V_Q = Q_{\rm fg}/C_{\rm T}$, $Q_{\rm fg}$ is the charge on the floating gate and $C_{\rm T}$ is the total capacitance from the floating gate. The bump circuit is composed of an inverse generation block, a fully differential VGA, and a conventional bump circuit. The width and the center of the bell-shaped transfer function are set by the common-mode and differential charges on M_{21} and M_{22}. The height is controlled by the tail current $I_{\rm b}$. All of them are independently programmable.

The inverse generation block has two floating-gate summing amplifiers. If the charges on M_{13} and M_{14} are matched and the transistors are in saturation region, we can have

$$V_{\rm in1} + V_{\rm 1c} = V_{\rm in2} + V_{\rm 2c} = V_{\rm const} \tag{1}$$

where $V_{\rm const}$ only depends on the bias voltage $V_{\rm b}$ and the charges on M_{13} and M_{14}. If the charge on M_{02}, in the bias generation circuit, also matches that on M_{13} and M_{14}, the generated voltage, $V_{\rm b}$, provides the summing amplifiers an operating range that is one $V_{\rm DSsat}$ away from the supply rails, as shown in Fig. 3. The floating-gate voltages on M_{21} and M_{22} can be expressed as

$$\begin{aligned} {\rm V}_{\rm fg,21} &= \frac{1}{2}(V_{\rm in1} + V_{\rm const} - V_{\rm in2}) + \frac{Q_{21}}{C_{\rm T}} \\ &= \frac{1}{2}\Delta V_{\rm in} + V_{Q,\rm cm} + \frac{1}{2}V_{Q,\rm dm} \end{aligned} \tag{2}$$

$$\begin{aligned} {\rm V}_{\rm fg,22} &= \frac{1}{2}(V_{\rm in2} + V_{\rm const} - V_{\rm in1}) + \frac{Q_{22}}{C_{\rm T}} \\ &= -\frac{1}{2}\Delta V_{\rm in} + V_{Q,\rm cm} - \frac{1}{2}V_{Q,\rm dm} \end{aligned} \tag{3}$$

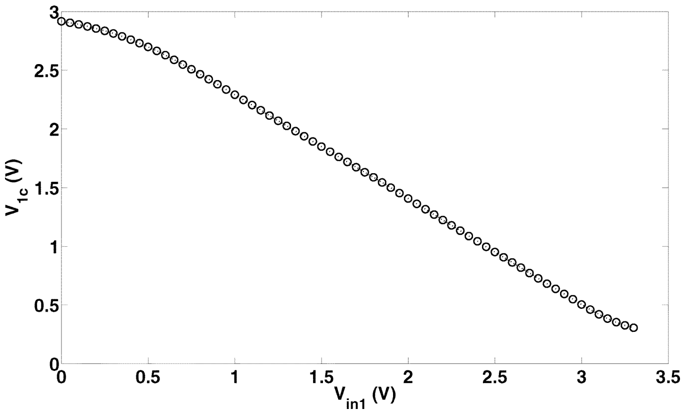


Fig. 3. Inverse generation transfer characteristics. A floating-gate summing amplifier generates a complementary input voltage. This outputs are fed to floating-gate transistors in the VGA so that the outputs of the VGA are independent of the input common-mode signals. With an appropriate bias voltage, the operating range is one $V_{\rm DSsat}$ away from supply rails.

where $\Delta V_{\rm in} = V_{\rm in1} - V_{\rm in2}$, Q_{21} and Q_{22} are the amounts of charge on M_{21} and M_{22}, respectively, $C_{\rm T}$ is the total capacitance seen from a floating gate, and

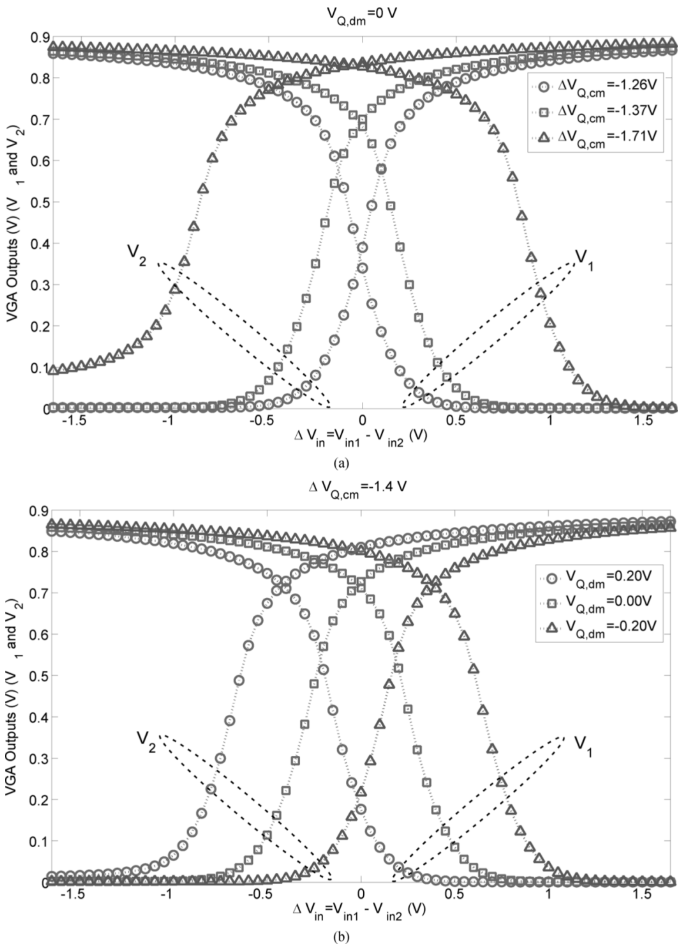
$$V_{Q,\rm cm} = \frac{1}{2}\left(\frac{Q_{21}+Q_{22}}{C_{\rm T}} + V_{\rm const}\right)$$
$$V_{Q,\rm dm} = \frac{Q_{21}-Q_{22}}{C_{\rm T}}.$$

From (2) and (3), these two floating-gate voltages do not depend on the input signal common-mode level.

The variable gain of the VGA stems from the nonlinearity of the transfer function from the floating-gate voltage, $V_{\rm fg,21}$ (or $V_{\rm fg,22}$), to the diode-connected transistor drain voltage, V_1 (or V_2). Several pairs of the transfer curves corresponding to different amounts of the charge on the floating gates are measured and are shown in Fig. 4. The measurement is taken with $V_{\rm in2}$ fixed at $V_{\rm DD}/2$ while $V_{\rm in1}$ is swept from 0 V to $V_{\rm DD}$. The value of $\Delta V_{\rm in}$ at the intersection indicates the center of the bell-shaped transfer curve. As shown in Fig. 4(a), the slopes at the intersection point varies with the common-mode charge while the value of $\Delta V_{\rm in}$ at the intersection does not. Therefore, we can program the common-mode charge to tune the width of the bell-shaped transfer characteristics without affecting the center. On the other hand, as shown in Fig. 4(b), the value of $\Delta V_{\rm in}$ at the intersection shifts as the differential charge changes, but the slopes at the intersection are invariant. Thus, by programming the differential charge, the center of the transfer function can be tuned without altering the width. Because the template information are stored in a pair of floating-gate transistors as in [12], [13], this circuit has the potential to implement adaptive learning algorithms with not only an adaptive mean but also an adaptive variance.

The detailed derivations of the relation between the VGA gain and the common-mode charge are given in the appendix . The final equation is

$$\frac{\Delta V_{\rm out}}{\Delta V_{\rm in}} \approx -\gamma\left(1 + e^{-\gamma\kappa_{\rm p}/2U_{\rm T}(V_{\rm DD}-V_{Q,\rm cm}-V_{\rm T0,p})}\right) = \eta \quad (4)$$



(a)

(b)

Fig. 4. VGA transfer characteristics. $\Delta V_{\rm in} = V_{\rm in1} - V_{\rm in2}$ and $V_{\rm in2}$ is fixed at $V_{\rm DD}/2$ and $V_{\rm in1}$ is swept from 0 V to $V_{\rm DD}$, where $V_{\rm DD}$ is 3.3 V. In the programming mode, the control gate voltages are set to be $-\Delta V_{Q,\rm cm} \mp V_{Q,\rm dm}/2$ and the floating-gate transistors are programmed to have 1 μA of current. (a) Common-mode charge on M_{21} and M_{22} are programmed to several different levels and the amount of the differential charge is fixed. (b) Differential charge on M_{21} and M_{22} are programmed to several different levels and the amount of the common-mode charge is fixed.

where $\gamma = \kappa_{\rm p}/\kappa_{\rm n}\sqrt{I_{0,\rm p}W_{\rm p}L_{\rm n}/I_{0,\rm n}L_{\rm p}W_{\rm n}}$, the subscripts "p" and "n" refer to the pMOS and nMOS transistors, respectively, I_0 is the subthreshold pre-exponential current factor, W and L are the dimensions of a transistor, κ is the subthreshold slope factor, $V_{\rm T0}$ is the threshold voltage, and $U_{\rm T}$ is the thermal voltage. Since the transfer function of the conventional bump circuit is given in [7], we can have the transfer function expression of the complete bump circuit as

$$I_{\rm out} = \frac{2I_{\rm b}}{2 + e^{\kappa\eta\Delta V_{\rm in}/U_{\rm T}} + e^{-\kappa\eta\Delta V_{\rm in}/U_{\rm T}}} \quad (5)$$

which is used to approximate a Gaussian function. By adjusting $V_{Q,\rm cm}$, the magnitude of the VGA gain increases exponentially and hence the width of bell-shaped transfer curve, which models the standard deviation of a distribution, decreases exponentially.

In Fig. 5(a), we program the common-mode charge to several different levels and measure the transfer curves with different

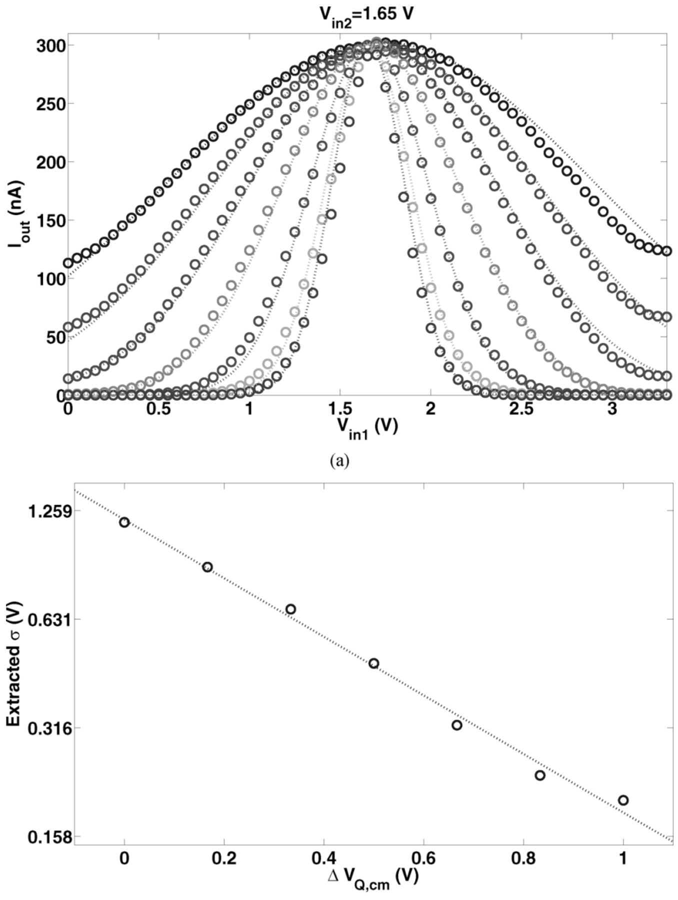


Fig. 5. Gaussian fits of the transfer curves and the width dependance. (a) Comparison of the measured 1-D bumps (circles) and the corresponding Gaussian fits (dashed lines). One of the bump input voltages is fixed at $V_{\rm DD}/2$, where $V_{\rm DD}$ is 3.3 V through the measurement. The extracted standard deviation varies 5.87 times and the mean only shifts 4.23%. The minimum achievable extracted standard deviation is 0.199 V. (b) Width and common-mode charge relation in semi-logarithmic scale. The width is characterized by the extracted σ. The shift of the programmed common-mode floating gate voltage, $\Delta V_{Q,{\rm cm}}$, represents the common-mode charge level. The dashed line is the exponential curve fit.

widths. The bell-shaped curves are compared with their correspondent Gaussian fits. In Fig. 5, the extracted standard deviation varies 5.87 times and the mean only shifts 4.23%. In the semi-logarithmic plot of Fig. 5(b), the extracted standard deviation, σ, exponentially depends on the common-mode charge as predicted by (4). The minimum achievable extracted standard deviation from our measurements is 0.199 V, which is set by the maximum gain of the VGA. If two diode-connected nMOS transistors are used as the load, the maximum VGA gain will be doubled and the minimum achievable standard deviation can be reduced by half.

A diode-connected transistor M_{37} in the bump circuit converts the output current into a voltage. By feeding this voltage to the tail transistor M_{30} in the next stage bump circuit as shown in Fig. 6, the final output current approximates a multivariate Gaussian function with a diagonal covariance matrix. Although the feature dimension can be increased by cascading more floating-gate bump cells, the bandwidth of the classifier decreases. The mismatches between the floating-gate bump circuits can be trimmed out by using floating-gate programming techniques. In Fig. 7, we show two 2-D "bumps" with different widths to approximate bivariate Gaussian functions with different standard deviations. The output currents of an array of these floating-gate bump circuits can easily be summed up to implement GMMs.

III. Programming Floating-Gate Transistor Array

How to accurately programming an array of floating-gate transistors is a critical technique in the development of our analog classifier. Fowler–Nordheim tunneling and channel hot electron (CHE) injection mechanisms are used to program charge on floating gates. The techniques of programming an array of floating-gate transistors have been detailed in many previous works [14], [15]. In this section, we will briefly review the floating-gate programming method and the way to program an array of floating-gate transistors.

Fowler–Nordheim tunneling removes electrons from the floating gates through tunneling junctions, which are schematically represented by arrowheaded capacitors shown in Fig. 8(b). Because of the poor selectivity, tunneling currents are used as the global erase. To accurately program charges on floating gates, CHE injection are employed. As detailed in [16], CHE injection current can be modeled as

$$I_{\rm inj} = I_{\rm inj0}\left(\frac{I_{\rm s}}{I_{\rm s0}}\right)^{\alpha} e^{-\Delta V_{\rm ds}/V_{\rm inj}} \tag{6}$$

where $I_{\rm s}$ is the channel current, $V_{\rm inj}$ is a device and bias dependent parameter, and α is very close to 1. Instead of using this computationally complex physical model as in [14], an empirical model proposed in [15] is used to perform floating-gate transistor characterization and algorithmic programming.

Given a short pulse of $V_{\rm ds}$ across a floating-gate device, the injection current is proportional to $\Delta I_{\rm s}/I_{\rm s0}$, where $\Delta I_{\rm s} = I_{\rm s} - I_{\rm s0}$ is the increment of the channel current. From (6), logarithmic of this ratio should be a linear function of $V_{\rm ds}$ and a nonlinear function of $\log(I_{\rm s0}/I_{\rm u})$, where $I_{\rm u}$ is an arbitrary unity current. It can be expressed as

$$\log\left(\frac{\Delta I_{\rm s}}{I_{\rm s0}}\right) = g\left(\log\left(\frac{I_{\rm s0}}{I_{\rm u}}\right)\right)V_{\rm ds} + f\left(\log\left(\frac{I_{\rm s0}}{I_{\rm u}}\right)\right) \tag{7}$$

where $g(\cdot)$ and $f(\cdot)$ are weakly linear functions when the transistor is in the subthreshold region and are nonlinear when the transistor is above threshold. In the characterization process, $V_{\rm ds}$ and $I_{\rm s0}$ are given and $\Delta I_{\rm s}$ can be measured. Thus, $g(\log(I_{\rm s0}/I_{\rm u}))$ and $f(\log(I_{\rm s0}/I_{\rm u}))$ can be regressed by high order polynomial functions. After the characterization process, we obtain the resulting polynomial regressive functions $\hat{f}(\log(I_{\rm s0}/I_{\rm u}))$ and $\hat{g}(\log(I_{\rm s0}/I_{\rm u}))$. In the programming process, with the regressive functions, the appropriate $V_{\rm ds}$ value for injection can be predicted by

$$V_{\rm ds} = \frac{\log\left(\frac{\Delta Is}{I_{\rm s0}}\right) - \hat{f}\left(\log\left(\frac{I_{\rm s0}}{I_{\rm u}}\right)\right)}{\hat{g}\left(\log\left(\frac{I_{\rm s0}}{I_{\rm u}}\right)\right)} \tag{8}$$

where $I_{\rm s0}$ is the given starting point and $I_{\rm s}$ is the target value.

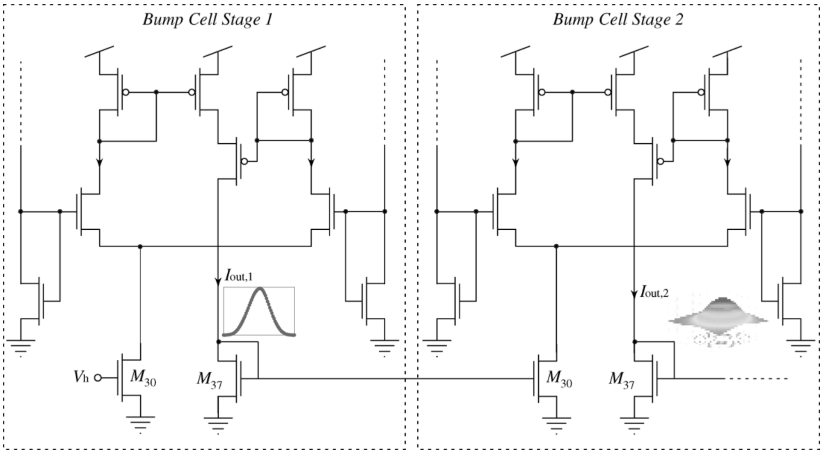
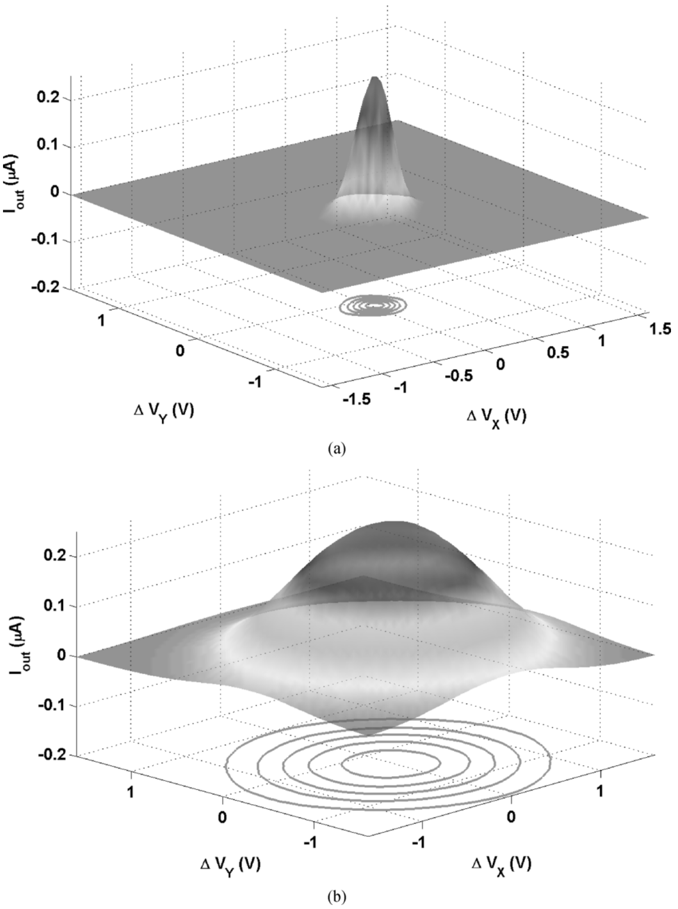


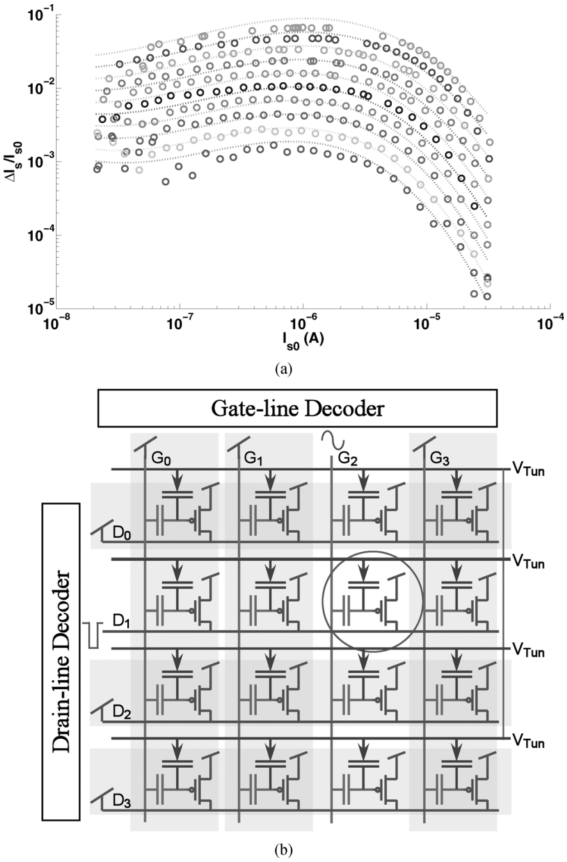
Fig. 6. Cascading bump circuits. By connecting the diode-connected output transistor to the tail transistor of the next stage bump cell, the resulting output current can approximate a multivariate Gaussian function with a diagonal covariance matrix.



(a)

(b)

Fig. 7. Measured bivariate RBFs. Measurement results from two cascading floating-gate bump circuits. $\Delta V_{\rm X}$ is the input voltage difference $\Delta V_{\rm in} = V_{\rm in1} - V_{\rm in2}$ of the first stage floating-gate bump circuit and $\Delta V_{\rm Y}$ is the input voltage difference of the second stage. In both stages, $V_{\rm in2} = V_{\rm DD}/2$. The common-mode charges are programmed to different levels to approximate bivariate Gaussian functions with different variance.



(a)

(b)

Fig. 8. Injection characterization and floating-gate array programming. (a) Measured injection characterization points (circles) and the corresponding curve fits (dashed lines). The pulsewidth is fixed at 200 μs. 10 different values of $V_{\rm ds}$ ranging from 5.6 to 6.5 V and 30 channel current levels ranging from 20 nA to 20 μA are used to obtain the curve fits for each curve. Cubic functions are used to regress the nonlinear functions $g(\cdot)$ and $f(\cdot)$ in (7). (b) Programming an array of floating-gate transistors. Drain lines and gate lines are shared in rows and in columns, respectively. By applying $V_{\rm DD}$ to unselected drain lines and gate lines, floating-gate transistors can be programmed individually. Decoders for programming are at the peripheries of the array.

The measured and the regressive results for the CHE injection characterization are compared in Fig. 8(a). Only one floating-gate transistor in the floating-gate array is used in the characterization, and the regressive functions are cubic. The measured regressive coefficient mismatches in the array are less than 10%. To avoid overshooting the target value, we always apply slightly shorter and smaller pulses of $V_{\rm ds}$ than the predicted values. Therefore, despite the mismatches and the discrepancy between the curve fits and the measured data, the current level of the floating-gate transistor approaches the target

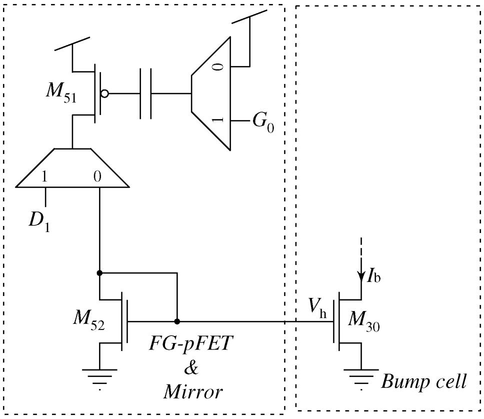


Fig. 9. Schematic of the "FG-pFET & Mirror" block. The charge on the pMOS transistor can be programmed to set the height of the bell-shaped transfer curve.

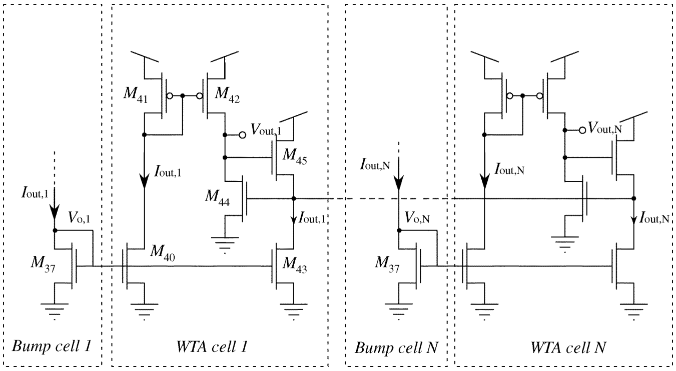


Fig. 10. Schematic of a current mode WTA circuit. Only the output voltage of the winning cell will be high to indicate the best-matching template.

value asymptotically. The precision of the programmed current level can be as accurate as 99.5%, which is consistent with other approaches [14], [15]. As presented in [17], the retention time for the charges on floating gates can last over 10 years at room temperature. Because the bump circuit is a differential structure, the center of the transfer curve would not vary with the temperature. However, its width depends on the temperature because of the U_T term in (4).

To program an array of the floating-gate bump circuits, floating-gate transistors are arranged as in Fig. 8(b) in the programming mode. Because two conditions are required for CHE injection: a channel current and a high channel-to-drain field, we can deactivate the unselected columns (or rows) by applying V_{DD} to the corresponding gate lines (or drain lines) so that there are no currents through (or no fields across) the devices for injection. In this manner, each floating-gate transistor can be isolated from others and can be programmed individually.

IV. Programmable Analog Vector Quantizer

A *"FG-pFET & Mirror"* block shown in Fig. 9 is added in front of the first bump cell to program its tail current, which sets the height of the "bump." For the analog vector quantizer implementation, the final output currents of the RBF-based classifier are duplicated and are fed into a simple current mode WTA circuit, for which the schematic is shown in Fig. 10. The output voltage of the winning cell only will be high to indicate the best-matching template.

To have the access to all drain and gate terminals of floating-gate transistors in the programming mode, multiplexers are inserted into the circuits as shown in Fig. 11. Most of the multiplexers are in the inverse generation and bias generation blocks. Because only one bias generation block is needed for the whole system, when the system is scaled up, the complexity of bias generation block does not cost. In the analog RBF-based classifier and vector quantizer, the same input voltage vector is compared with all stored templates. Therefore, the inverse generation can be shared by the same column of bump cells, each of which only includes a VGA and a conventional bump circuit. The number of inverse generation blocks is equal to the dimension of the feature space. Together with the gate-line and drain-line decoders, most of the programming overhead circuitries are at the peripheries of the floating-gate bump cell array; therefore the system can be easily scaled up and maintain high compactness. The compactness and the ease of scaling up are important issues in the implementation of an analog speech recognizer that requires more than a thousand of bump cells. The final architecture of our analog vector quantizer is shown in Fig. 12 and the circuit parameters are listed in Table I.

We use two examples to demonstrate the reconfigurability of our classifiers. Four templates are used as shown in Fig. 13. The floating-gate transistors of other unused templates are tunneled off. Four bell-shaped output currents emulate the bivariate Gaussian likelihood functions of four templates. The thick solid lines at the bottom, indicate the boundaries determined by the WTA outputs.

V. Performance of the Analog Vector Quantizer

We have fabricated an analog vector quantizer in a 0.5-μm CMOS process and the micrograph is shown in Fig. 14. We also fabricated a 16×16 highly compact low-power version of an analog vector quantizer in the 0.5-μm CMOS process occupying less than 1.5×1.5 mm^2. Some important parameters and measured results are listed in the Table II.

To measure the power consumption, we program several "bumps" with identical width and deactivate other "bumps" by tunneling their floating-gate transistors off. The power consumption is averaged over the entire 2-D input space. The slope of the curve in Fig. 15(a) indicates the average power consumption per bump cell with a specific value of width. The relation between the power consumption and the extracted standard deviation is shown in Fig. 15(b).

The VGA is the main source of the power consumption. The gain is tunable when the nMOS transistors in the VGA operate in the transition between above threshold and subthreshold regions. The width tunability can also result from the nonlinearity of the pMOS transistors when they are in transition between saturation and ohmic region. From simulation, to save the power consumed in the VGA, we can make nMOS transistors longer to reduce the above-threshold currents and raise the source voltages of M_{23} and M_{24} to reduce the headroom.

Because the RBF output current is in the nano-amp range and the bandwidth of our current preamplifier for measurement is approximately 1 kHz at that current level, we can not measure the speed of our floating-gate bump circuit directly, which is expected to be around megahertz range. We can only mea-

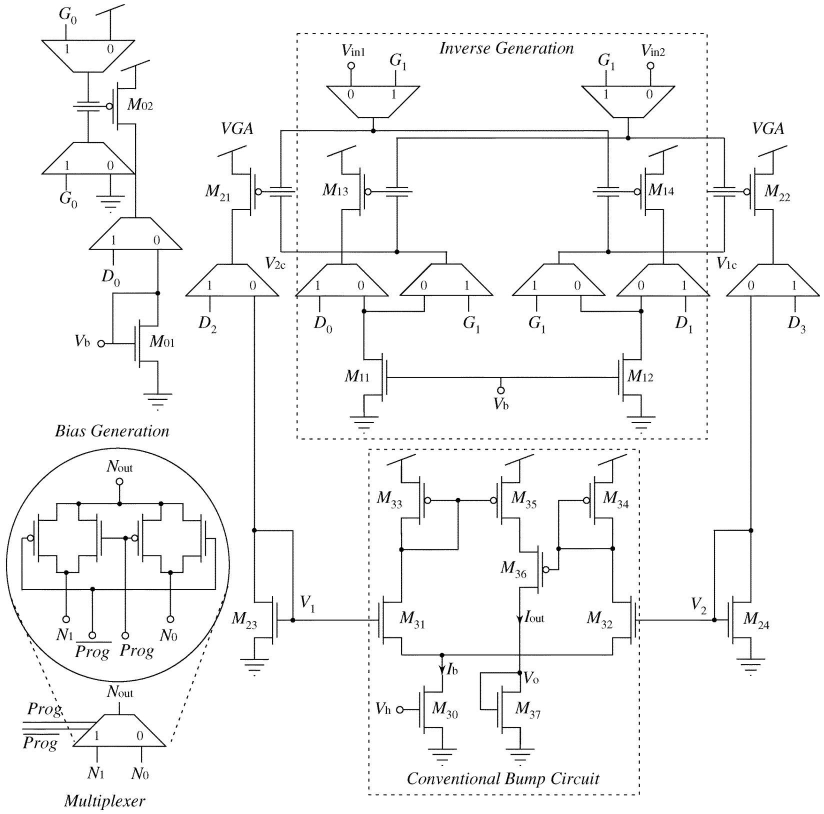


Fig. 11. Complete schematics of the floating-gate bump circuit. Multiplexers for programming are inserted into the original circuits. The "1" on the multiplexer indicates the connection in the programming mode and the "0" indicates the connection in the operating mode. The tunneling junction capacitors are not shown for simplicity. Most of the multiplexers are in bias generation and inverse generation blocks. Only two multiplexers are added in the bump cell that includes the VGA and the conventional bump circuit.

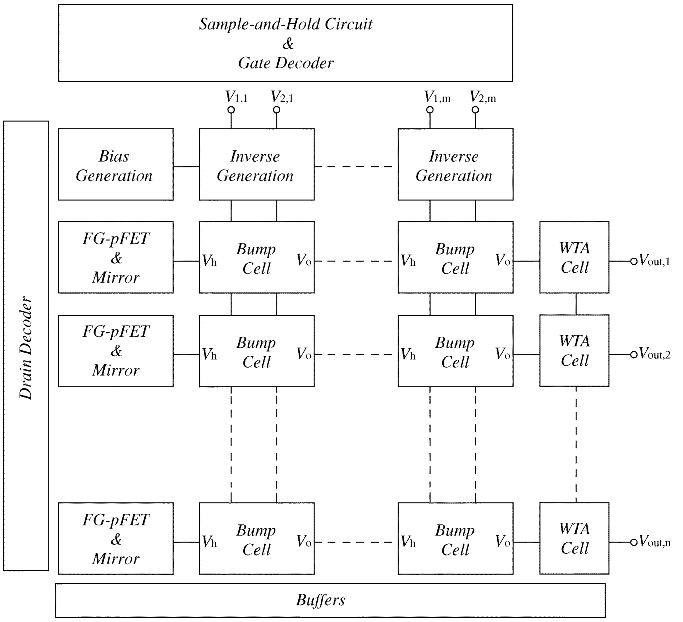


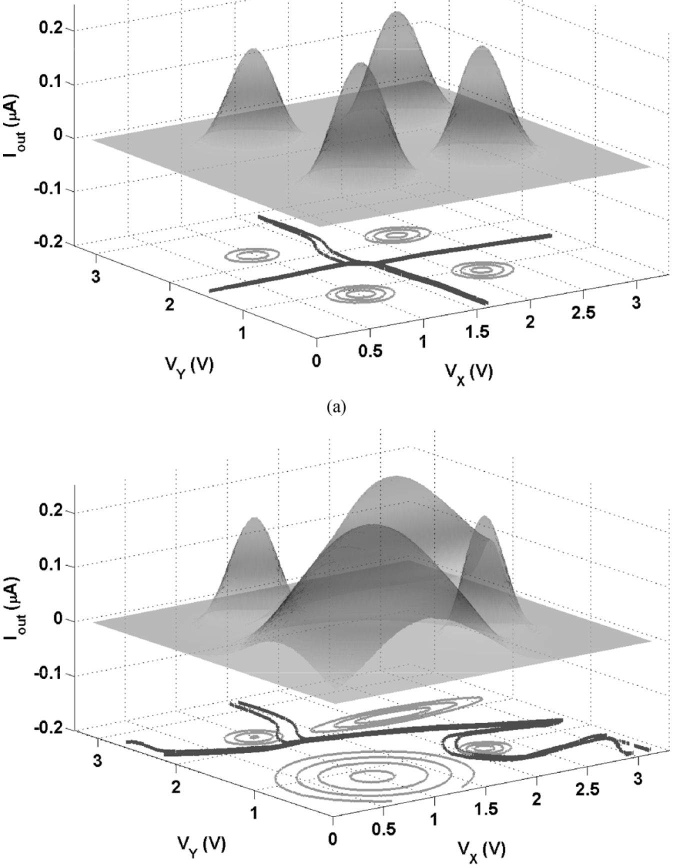
Fig. 12. Architecture of an analog vector quantizer. The core is the bump cell array followed by a WTA circuit. The main complexity from programming are at the peripheries and the system can be scaled up easily.

TABLE I
CIRCUIT PARAMETERS (μm/μm)

Bump		*FG-pFET & Mirror*	
$M_{01\sim02}, M_{11\sim14}$:	1.8/1.2	M_{41}:	1.8/1.2
$M_{21\sim22}, M_{31\sim32}$:	1.8/1.2	M_{42}:	3.6/1.2
$M_{23\sim24}$:	1.8/3.0	WTA	
$M_{30}, M_{35\sim37}$:	3.6/1.2	$M_{50\sim52}$:	1.8/1.2
$M_{33\sim34}$:	1.8/2.4	$M_{53\sim55}$:	1.8/1.5

sure the response time from the input to the WTA outputs. The measured transient response of the analog vector quantizer is shown in Fig. 16(a). One of the speed bottlenecks of the system is the inverse generation block. For a given width, the speed and the power depend on the amount of charge on M_{13} and M_{14}. With more electrons on the floating gates, the circuit can achieve higher speed but consumes more power as shown in Fig. 16(b). The steep portion of the curve implies that the inverse generation block dominates. In this region, we can increase the speed by consuming more power in the inverse generation block. The flat region in Fig. 16(b) indicates the VGA dominant region. In this region, burning more power in the inverse generation block does not improve the speed of the system. Thus, given a variance, we can program the charges on M_{13} and M_{14} so that the system operates at the knee of the curve to optimize the tradeoff between the speed and the power consumption in the inverse generation block.

Finally, we wish to evaluate the computational accuracy of the analog RBF. Since the computation method and errors are different from those of traditional digital approaches, generic comparisons of effective bit-accuracy do not make sense. Rather,



(a)

(b)

Fig. 13. Configurable classification results. The measured bump output currents (circle contours) and the WTA voltages (thick solid lines at the bottom) of four templates are superposed in a single plot. V_X and V_Y are the $V_{\rm in1}$ in the first stage and the second stage floating-gate bump circuits, respectively. Both of their $V_{\rm in2}$ terminals are fixed at $V_{\rm DD}/2$. (a) Four templates are programmed to have the same variance and evenly spaced means. (b) Four templates are programmed to have different variances with evenly spaced means.

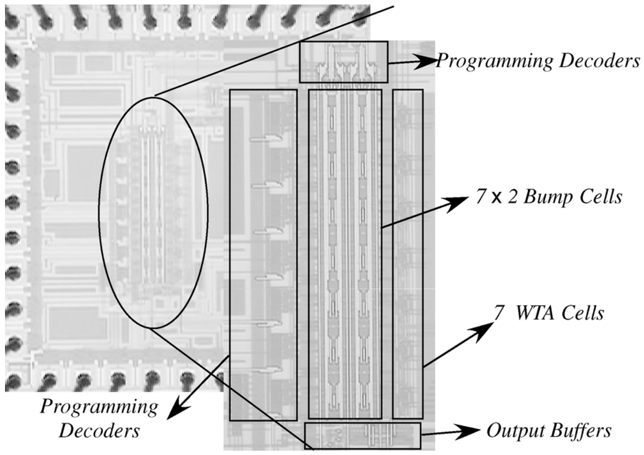
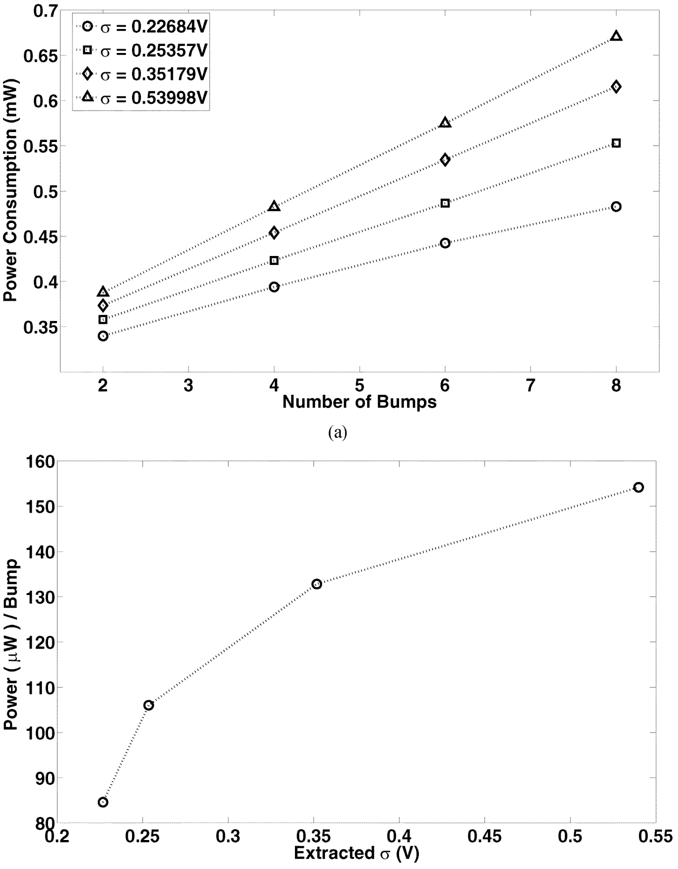


Fig. 14. Micrograph of an analog vector quantizer. A prototyped version of an analog vector quantizer is fabricated in a 0.5-μm CMOS process. It is composed of a 7 × 2 floating-gate bump cell array.

we choose to evaluate the impact of using the analog RBFs on system performance. To this end we use ROC curves and EER.

TABLE II
ANALOG VECTOR QUANTIZER PARAMETERS

Size of VQ	7(templates)×2(components)
Area/Bump Cell	$42 \times 82\,\mu\mathrm{m}^2$
Area/WTA Cell	$20 \times 35\,\mu\mathrm{m}^2$
Power Supply Rail	$V_{\rm DD} = 3.3V$
Power Consumption/Bump Cell	$90\mu W \sim 160\mu W$
Response Time	$20\mu \sim 40\mu$sec
Floating-gate Programming Accuracy	99.5%
Retention Time	10 years @ 25°C

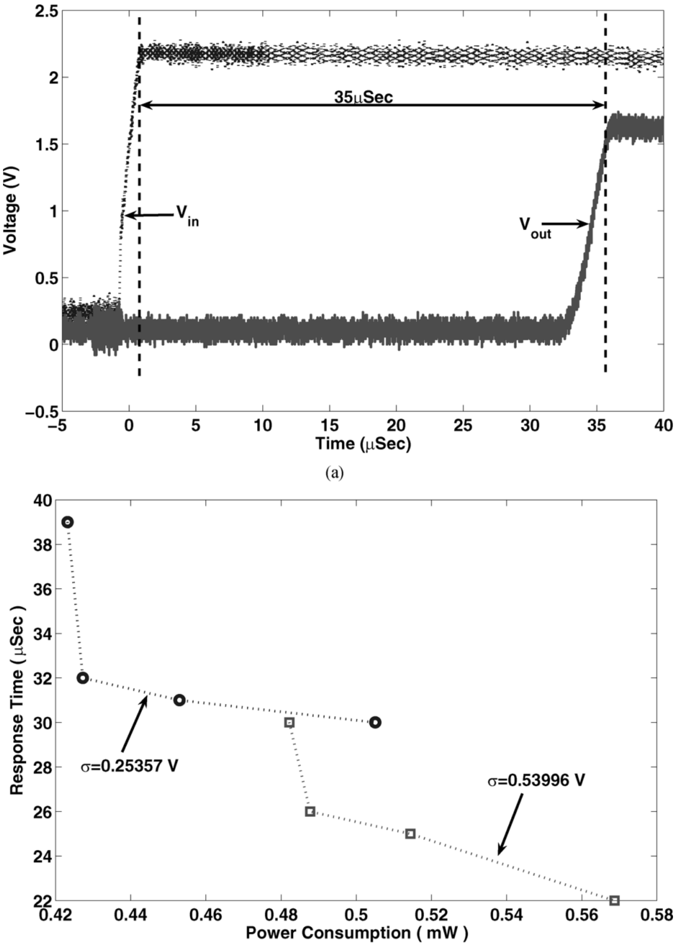


(a)

(b)

Fig. 15. Relation between the power consumption and the extracted variance. (a) Measured power consumption of the analog vector quantizer with different number of floating-gate bump cells being activated with a fixed width. The slope of the curves indicate the average power consumption per bump cell. (b) Relation between the power consumption per bump and the extracted variance of the bell-shaped transfer curve. The larger the variance is, the more the power consumption.

Two separate 2-D bumps are programmed to have the same variance with a fixed separation as shown in Fig. 17. The corresponding Gaussian fits are used as the actual probability density functions (pdfs) of two classes. Comparing these two pdfs using different thresholds renders a ROC curve of these two Gaussian distributed classes. We use it as the evaluation reference. With the knowledge of the class distributions, comparing the output currents using different thresholds generates a ROC curve for the 2-D bumps. Comparing each of the two WTA output voltages with different thresholds generates two ROC curves that characterize the classification results of the vector quantizer.



(a)

(b)

Fig. 16. Response time and speed-power tradeoff of an analog vector quantizer. (a) The response time is measured between the input step and the WTA output response. (b) Relation between the response time and the power consumption for a given bump width. The inverse generation block dominates the response time in the steep region. The VGA dominates in the flat region. Charge on M_{13} and M_{14} can be programmed to optimize the speed-power tradeoff.

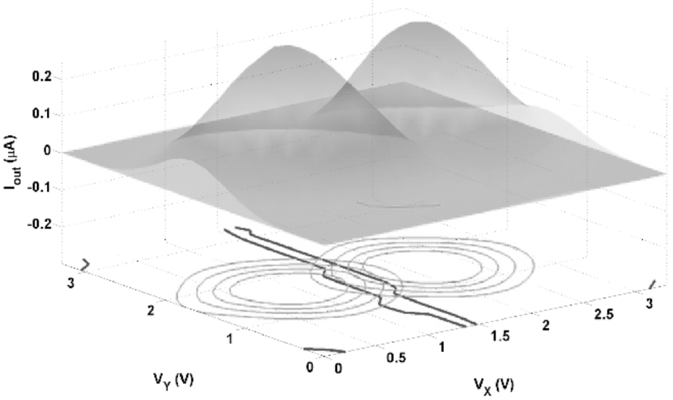
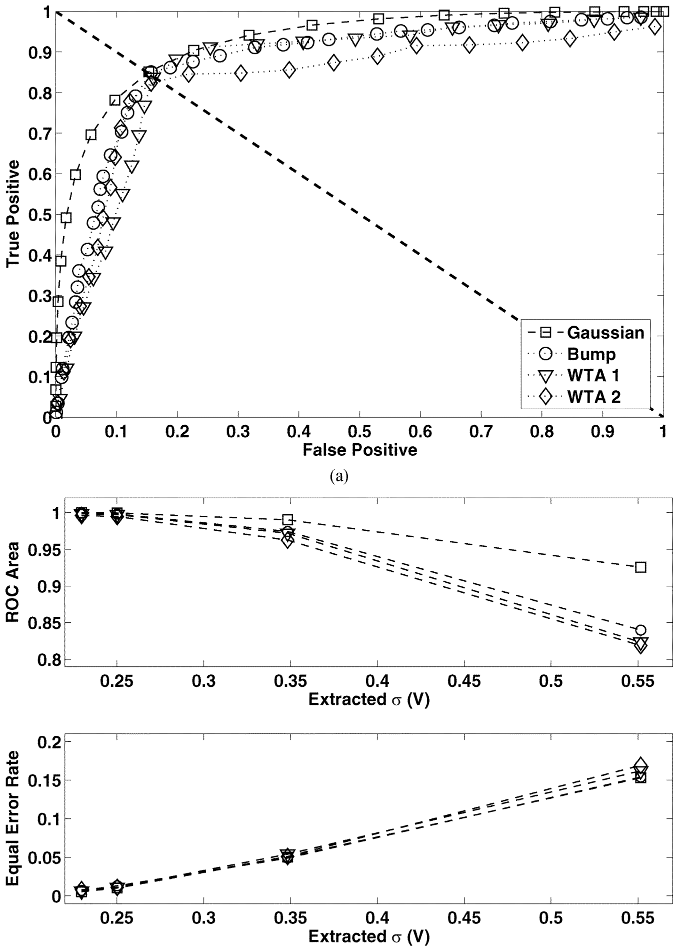


Fig. 17. Distributions of two “bumps” used to evaluate the classifier performance. In the measurements for performance evaluation, the separation of the center is kept constant but the widths of these two “bumps” varies. The measured bump output currents (circle contours) and the WTA voltages (thick solid lines at the bottom) of two templates are superposed in a single plot. V_{X} and V_{Y} are the values at the V_{in1} input terminals of the first and the second floating-gate bump circuits, respectively. The V_{in2} terminals in both stages are fixed at $V_{\mathrm{DD}}/2$.



(a)

(b)

Fig. 18. ROC and EER performance of the classifiers. The effects of different bump widths on the ROC area and the EER performance. The separation of the means of two classes is 1.2 V. The intercept plot shows the ROC curves of the Gaussian fits (squares), output currents of the 2-D bumps (circles) and WTA output voltages (triangles and diamonds) with the extracted $\sigma = 0.55$ V. The Gaussian fits are used as the actual pdf’s of the two classes and the corresponding ROC curve is used as a reference. The intersection of the ROC curve and the -45° line is the EER point, which is the usual operating point. The results show that the analog VQ is comparable to an ideal maximum-likelihood (ML) classifier.

The EER, which is the intersection of the ROC curve and the -45° line as shown in Fig. 18(a), is the usual operating point of classifiers. In Fig. 18(b), both the ROC areas and the EER are plotted to investigate the effect of the bump width on the performance. At the EER point, the performance of our RBF classifier, which uses floating-gate bump circuits to approximate Gaussian likelihood functions, is undistinguishable from that of an ideal RBF-based classifier. Despite the finite gain of the WTA circuit, the performance of the analog vector quantizer is still comparable to an ideal maximum likelihood (ML) classifier. By optimizing the precision and speed of the WTA circuit, the performance can be improved but it is beyond the scope of this paper.

VI. Power Efficiency Comparison

To compare the efficiency of our analog system with the digital signal processing (DSP) hardware, we estimate the metric of millions of multiply accumulates per second per milliwatt (MMAC/s/mW) of our classifiers. When the system is scaled up, the efficiency of the bump cells dominates the performance.

Therefore, we consider the performance of a single bump cell only.

Each Gaussian function is estimated as 10 MACs and can be evaluated by a bump cell in less than 10 μs (which is still an overestimate) with the power consumption of 120 μW or so. This is equivalent to 8.3 MMAC/s/mW. The performance of commercial low-power DSP microprocessors ranges from 1 MMAC/s/mW to 10 MMAC/s/mW and a special designed high performance DSP microprocessor in [18] is better than 50 MMAC/s/mW. If this comparison is expanded to include the WTA function, the efficiency of our analog system will improve even more relative to the digital system.

Although our power efficiency is comparable to the digital system, our classifier consumes much more power compared to other analog vector-matrix-multiplication systems [19], [20], of which efficiency ranges from 37 to 175 $\mathrm{MMAC/s/}\mu\mathrm{W}$. The reason is that the transistors M_{23} and M_{24} are operating far above threshold. By making M_{21} and M_{22} long and raising the source voltages of M_{23} and M_{24} (which is not available in the current chip), from simulation, we can easily reduce the power consumption by at least two orders of magnitude. If the WTA circuit is also optimized, it is anticipated that future ICs will be at least two to three orders of magnitude more efficient than DSP microprocessors at the same task.

VII. Conclusion

In this paper, we demonstrate a new programmable floating-gate bump circuit, of which the height, the center and the width of the bell-shaped transfer characteristics can be programmed individually. A multivariate RBF with a diagonal matrix can be realized by cascading these bump cells. Based on the new bump circuit, we build a novel compact RBF-based soft classifier and, by adding a simple current mode WTA circuit, we implement an analog vector quantizer. The performance and the efficiency of the classifiers are comparable to the digital system. With slight modifications, the overall efficiency is anticipated to be improved by at least two to three orders of magnitude better than DSP microprocessors.

Appendix

The nMOS transistors in the VGA are assumed in the transition between the above-threshold and the subthreshold regions. The pMOS transistors are assumed in the above-threshold region. Because the transfer characteristics of the two branches are symmetric, we can use the half circuit technique to analyze the VGA gain. By equating the currents flowing through the pMOS and nMOS transistors, we can have

$$\begin{aligned} I_{0,\mathrm{p}}\left(\frac{W_\mathrm{p}}{L_\mathrm{p}}\right)\frac{1}{4U_\mathrm{T}^2}\left[\kappa_\mathrm{p}(V_\mathrm{DD}-V_{\mathrm{fg},21}-V_{\mathrm{T0,p}})\right]^2 \\ = I_{0,\mathrm{n}}\left(\frac{W_\mathrm{n}}{L_\mathrm{n}}\right)\ln^2\left(1+e^{\kappa_\mathrm{n}/2U_\mathrm{T}(V_1-V_{\mathrm{T0,n}})}\right) \end{aligned} \quad (9)$$

where the subscripts of "p" and "n" refer to pMOS and nMOS transistors, respectively, I_0 is the subthreshold pre-exponential current factor, κ is the subthreshold slope factor, V_{T0} is the threshold voltage, and U_T is the thermal voltage. At the peak of the bell-shaped transfer curve, $V_{Q,\mathrm{dm}} = 0$ and

$$\begin{aligned} \mathrm{V}_{\mathrm{fg},21} &= \frac{1}{2}\Delta V_\mathrm{in} + V_{Q,\mathrm{cm}} \\ V_1 &= V_{\mathrm{out,cm}} + \frac{1}{2}\Delta V_\mathrm{out} \end{aligned}$$

where $V_{\mathrm{out,cm}} = (V_1+V_2)/2$, $\Delta V_\mathrm{out} = V_1 - V_2$. We can obtain the gain of the VGA by differentiating (9) with respect to $V_{\mathrm{fg},21}$ and have

$$\begin{aligned} \frac{\Delta V_\mathrm{out}}{\Delta V_\mathrm{in}} = \frac{dV_1}{dV_{\mathrm{fg},21}} &= -\gamma\left(1+e^{-\kappa_\mathrm{n}/2U_\mathrm{T}(V_1-V_{\mathrm{T0,n}})}\right) \\ &= \frac{-\gamma}{1-e^{-\gamma\kappa_\mathrm{p}/2U_\mathrm{T}(V_\mathrm{DD}-V_{\mathrm{fg},21}-V_{\mathrm{T0,p}})}} \\ &\approx -\gamma\left(1+e^{-\gamma\kappa_\mathrm{p}/2U_\mathrm{T}(V_\mathrm{DD}-V_{Q,\mathrm{cm}}-V_{\mathrm{T0,p}})}\right) \end{aligned} \quad (10)$$

where $\gamma = \kappa_\mathrm{p}/\kappa_\mathrm{n}\sqrt{I_{0,\mathrm{p}}W_\mathrm{p}L_\mathrm{n}/I_{0,\mathrm{n}}L_\mathrm{p}W_\mathrm{n}}$. Therefore, the gain increases approximately exponentially with the common-mode charge and, accordingly, we can expect the exponential relation between the extracted standard deviation of the transfer curve and the common-mode charge.

References

[1] P. Hasler, P. D. Smith, D. Graham, R. Ellis, and D. V. Anderson, "Analog floating-gate, on-chip auditory sensing system interfaces," *IEEE J. Sensors*, vol. 5, no. 5, pp. 1027–1034, Oct. 2005.

[2] M. Ogawa, K. Ito, and T. Shibata, "A general-purpose vector-quantization processor employing two-dimensional bit-propagating winner-take-all," in *Proc. Symp. VLSI Syst.*, Jun. 13–15, 2002, pp. 244–247.

[3] M. Bracco, S. Ridella, and R. Zunino, "Digital implementation of hierarchical vector quantization," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1072–1084, Sep. 2003.

[4] G. Cauwenberghs and V. Pedron, "A low-power CMOS analog vector quantizer," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1278–1283, Aug. 1997.

[5] P. Hasler, P. Smith, C. Duffy, C. Gordon, J. Dugger, and D. Anderson, "A floating-gate vector-quantizer," in *Proc. Midwest. Symp. Circuits Syst.*, Aug. 2002, vol. 1, pp. 196–199.

[6] T. Yamasaki and T. Shibata, "Analog soft-pattern-matching classifier using floating-gate MOS technology," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1257–1265, Sep. 2003.

[7] T. Delbruck, "'Bump' circuits for computing similarity and dissimilarity of analog voltage," in *Proc. Int. Neural Network Society*, Seattle, WA, 1991.

[8] S. S. Watkins and P. M. Chau, "A radial basis function neurocomputer implemented with analog VLSI circuits," in *Proc. Int. Joint Conf. Neural Networks*, 1992, vol. 2, pp. 607–612.

[9] J. Choi, B. J. Sheu, and J. C.-F. Chang, "A Gaussian synapse circuit for analog neural networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 2, no. 3, pp. 129–133, Mar. 1994.

[10] S.-Y. Lin, R.-J. Huang, and T.-D. Chiueh, "A tunable Gaussian/square function computation circuit for analog neural networks," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 45, no. 3, pp. 441–446, 1998.

[11] D. S. Masmoudi, A. T. Dieng, and M. Masmoudi, "A subthreshold mode programmable implementation of the Gaussian function for RBF neural networks applications," in *, Proc. 2002 IEEE Int. Symp. Intelligent Control*, Vancouver, Cananda, Oct. 2002, pp. 454–459.

[12] D. Hsu, M. Figueroa, and C. Diorio, "A silicon primitive for competitive learning," in *Proc. Conf. Neural Inf. Processing Syst.*, Dec. 2000, pp. 713–719.

[13] P. Hasler, "Continuous-time feedback in floating-gate MOS circuits," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 56–64, Jan. 2001.

[14] M. Kucic, A. Low, P. Hasler, and J. Neff, "A programmable continuous-time floating-gate Fourier processor," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 90–99, Jan. 2001.

[15] A. Bandyopadhyay, G. J. Serrano, and P. Hasler, "Adaptive algorithm using hot-electron injection for programming analog computational memory elements within 0.2% of accuracy over 3.5 decades," *IEEE J. Solid-State Circuits*, vol. 41, no. 9, pp. 2107–2114, Sep. 2006.

[16] P. Hasler and J. Dugger, "Correlation learning rule in floating-gate pFET synapses," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 65–73, Jan. 2001.

[17] V. Srinivasan, G. J. Serrano, J. Gray, and P. Hasler, "A precision cmos amplifier using floating-gates for offset cancellation," in *Proc. CICC'05*, Sep. 2005, pp. 734–737.

[18] J. Glossner, K. Chirca, M. Schulte, H. Wang, N. Nasimzada, D. Har, S. Wang, A. J. Hoane, G. Nacer, M. Moudgill, and S. Vassiliadis, "Sandblaster low power DSP," in *Prec. IEEE Custom Integr. Circuits Conf.*, Oct. 2004, pp. 575–581.

[19] R. Chawla, A. Bandyopadhyay, V. Srinivasan, and P. Hasler, "A 531-nW/MHz, 128 × 32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity," in *Prec. IEEE Custom Integr. Circuits Conf.*, Oct. 2004, pp. 651–654.

[20] R. Karakiewicz, R. Genov, A. Abbas, and G. Cauwenberghs, "175 GMACS/mW charge-mode adiabatic mixed-signal array processor," in *Proc. Symp. VLSI Syst.*, Jun. 2006.

Sheng-Yu Peng (S'02) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1995 and 1997, respectively, and the Master's degree in electrical and computer engineering from Cornell University, Ithaca, NY, in 2004. He is currently working toward the Ph.D. degree in electrical and computer engineering at Georgia Institute of Technology, Atlanta.

His research interests include interface circuits for sensors, programmable analog circuits, low power analog signal processing, and machine learning.

Paul E. Hasler (S'87–M'01–SM'03) received the B.S.E. and M.S. degrees in electrical engineering from Arizona State University, Tulsa, both in 1991, and the Ph.D. degree in computation and neural systems from California Institute of Technology, Pasadena, in 1997.

He is an Associate Professor in the School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta. His current research interests include low power electronics, mixed-signal system integrated circuits, floating-gate MOS transistors, adaptive information processing systems, "smart" interfaces for sensors, cooperative analog–digital signal processing, device physics related to submicron devices or floating-gate devices, and analog VLSI models of on-chip learning and sensory processing in neurobiology.

Dr. Hasler received the NSF CAREER Award in 2001, and the Office of Naval Research YIP award in 2002. He received the Paul Raphorst Best Paper Award, IEEE Electron Devices Society, 1997, CICC Best Student Paper Award, 2006, ISCAS Best Sensors Paper award, 2005, a Best paper award at SCI 2001.

David V. Anderson (S'92–M'98–SM'05) received the B.S and M.S. degrees from Brigham Young University, Provo, UT, in 1993 and 1994, respectively, and the Ph.D. degree from Georgia Institute of Technology (Georgia Tech) Atlanta, and 1999, respectively. He is an Associate Professor in the School of Electrical and Computer Engineering at Georgia Tech and co-director of the Center for Research in Embedded Systems Technology. His research interests include audio and psycho-acoustics, signal processing in the context of human auditory characteristics, and the real-time application of such techniques using both analog and digital hardware. He has over 90 technical publications and 5 patents/patents pending.

Dr. Anderson was awarded the National Science Foundation CAREER Award for excellence as a young educator and researcher in 2004 and is a recipient of the 2004 Presidential Early Career Awards for Scientists and Engineers (PECASE).Dr. Anderson is a senior member of the Acoustical Society of America, ASEE, and Tau Beta Pi.