A Programmable Multi-Dimensional Analog Radial-Basis-Function-Based Classifier

Sheng-Yu Peng, Paul E. Hasler, and David V. Anderson

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250

Abstract. A compact analog programmable multidimensional radial-basis-function (RBF)-based classifier is demonstrated in this chapter. 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 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 multi-dimensional 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. Receiver operating characteristic curves and equal error rate are used to evaluate the performance of the RBF-based classifier as well as a resultant analog vector quantizer. It is shown 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.

1 Motivations for Analog RBF Classifier

The aggressive scaling of silicon technologies has led to transistors and many sensors becoming faster and smaller. The trend toward integrating sensors, interface circuits, and microprocessors into a single package or into a single chip is more and more prevalent. Fig. 1**A** illustrates the block diagram of a typical microsystem, which receives analog inputs via sensors and performs classification, decision-making, or, in a more general term, information-refinement tasks in the digital domain. Although fabrication and packaging technologies enable an unprecedented number of components to be packed into a small volume, the accompanying power density can be higher than ever, which has become one of the bottle-neck factors in the microsystem development. If the information-refinement tasks can be performed in the analog domain with less power consumption, the specifications for the analog-to-digital-converters, which are usually power-hungry, can be relaxed. In some cases, analog-to-digital conversion can

Please use the following format when citing this chapter:

Peng, S.-Y., Hasler, P.E. and Anderson, D.V., 2009, in IFIP International Federation for Information Processing, Volume 291; *VLSI-SoC: Advanced Topics on Systems on a Chip*; eds. R. Reis, V. Mooney, P. Hasler; (Boston: Springer), pp. 33–52.

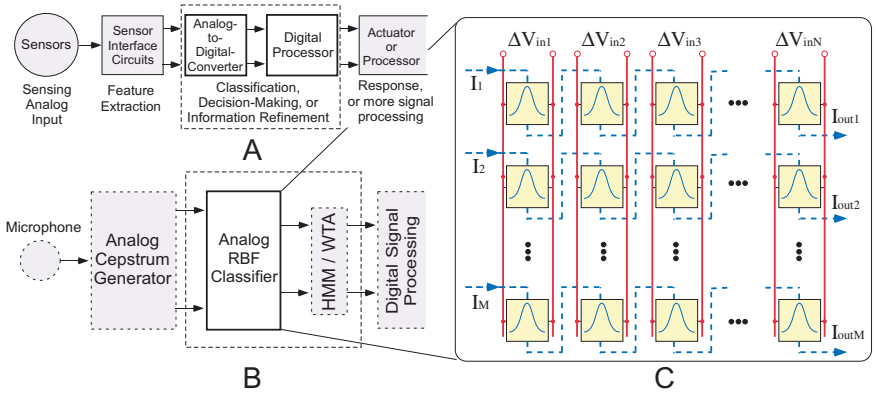

Fig. 1. A: The block diagram of a typical microsystem. **B:** An analog RBF-based classifier in an analog front-end for speech recognition includes a band-pass-filter bank based analog Cepstrum generator, an analog RBF-based classifier, and a continuous-time hidden Markov model. **C:** The block diagram of an analog RBF-based classifier which is composed of an array of the proposed floating-gate bump cells. Followed by a winner-take-all circuit, it results in a highly compact and power-efficient analog vector quantizer.

be avoided altogether. In such systems, multivariate Gaussian response functions are critical building blocks for a variety of applications, such as radial-basis-function(RBF)-based classifiers, Gaussian mixture modeling of data, and vector quantizers. This chapter discusses the development of an analog Gaussian response function having a diagonal covariance matrix and demonstrates its application to vector quantization.

Fig. 1**B** illustrates one possible application of this work as part of an analog speech recognizer [1] that includes a band-pass-filter bank based analog Cepstrum generator, an analog RBF-based classifier, and a winner-take-all (WTA) stage, or 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 power efficient.

In this chapter, a highly compact and power-efficient, programmable analog RBF-based classifier is demonstrated. It is at least two orders of magnitude more power efficient than the digital counterparts. As illustrated in Fig. 1**C**, the analog RBF-based classifier is composed of an array of proposed floating-gate bump cells having bell-shaped transfer characteristics that can realize the Gaussian functions. The height, the width, and the center of a bump circuit transfer curve, which represent the maximum likelihood, the variance, and the

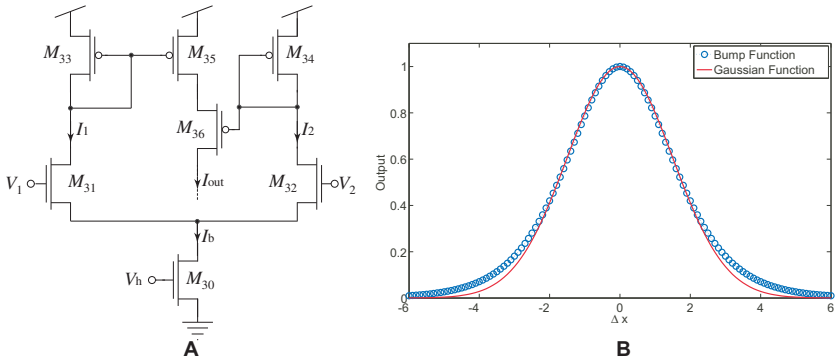

Fig. 2. A: Schematic of a conventional bump circuit introduced in [7]. **B:** Comparison between the normalized Gaussian function and the normalized Bump function.

mean of a template distribution respectively, can be independently programmed. The ability to program these three parameters empowers the classifiers to fit into different scenarios with the full use of statistical information up to the second moment. When followed by a winner-take-all stage, an RBF-based classifier forms a multi-dimensional 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 technique used in pattern recognition and data compression. Crucial issues of the vector quantizer implementation concern the storage efficiency and the computational cost for 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 because of their simplicity.

2 Bump circuits

The schematic of a conventional bump circuit in [7] is shown in Fig. 2**A**. If all transistors operate in the subthreshold region, the branch currents in the differential pair can be expressed as

$$I_1 = \frac{I_\mathrm{b}}{1 + e^{-\kappa \Delta V_\mathrm{in}/U_\mathrm{T}}},\ I_2 = \frac{I_\mathrm{b}}{1 + e^{\kappa \Delta V_\mathrm{in}/U_\mathrm{T}}}, \tag{1}$$

where κ is the subthreshold slope factor, U_{T} is the thermal voltage, and $\Delta V_{\mathrm{in}} = V_{\mathrm{in1}} - V_{\mathrm{in2}}$. The output current is the harmonic mean of I_1 and I_2 and can be described as

$$I_{\mathrm{out}} = \frac{I_1 I_2}{I_1 + I_2} = \frac{I_{\mathrm{b}}}{2 + e^{\kappa \Delta V_{\mathrm{in}}/U_{\mathrm{T}}} + e^{-\kappa \Delta V_{\mathrm{in}}/U_{\mathrm{T}}}} = \frac{I_{\mathrm{b}}}{2} \mathrm{sech}^2 \left(\frac{\kappa \Delta V_{\mathrm{in}}}{2U_{\mathrm{T}}} \right). \quad (2)$$

The normalized bump function is compared with the normalized Gaussian function as shown in Fig. 2**B**. This simple circuit can implement the exponential decay behavior of a Gaussian function. It is noticeable that, from (2), the width of the transfer characteristic is fixed by the ratio of κ/U_{T}.

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, 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, 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. As will be shown later, the floating-gate bump circuit introduced in this chapter has the potential to adapt to both the mean and the variance of the distribution.

3 Programmable Floating-gate Bump circuit

In the proposed analog classifier, the Gaussian response function is approximated by the bell-shaped transfer characteristics of a 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. 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. 3. All floating-gate transistors have two input capacitances and all input capacitances are of the same size. The proposed floating-gate bump circuit is composed of three parts: an inverse generation block, a conventional bump circuit, and in between a fully differential variable gain amplifier (VGA).

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

$$V_{\mathrm{in1}} + V_{\mathrm{1c}} = V_{\mathrm{in2}} + V_{\mathrm{2c}} = V_{\mathrm{const}}, \quad (3)$$

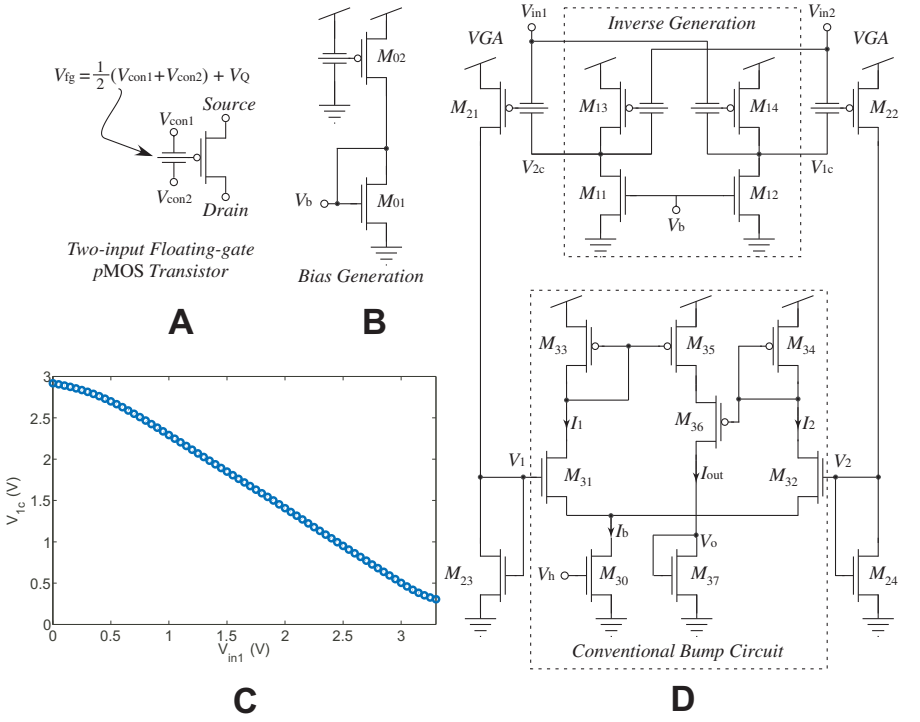

Fig. 3. A: The symbol of a two-input floating-gate pMOS transistor. **B:** The schematic of the bias generation circuit for the proposed floating-gate bump circuit. **C:** The transfer characteristic of the inverse generation block. **D:** The schematic of the proposed bump circuit that is composed of an inverse generation block, a fully differential variable gain amplifier (VGA), and a conventional bump circuit.

where V_{const} only depends on the bias voltage, V_{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_{b}, provides the summing amplifiers an operating range that is one V_{DSsat} away from the supply rails, as shown in Fig. 3**C**.

The floating-gate voltages on M_{21} and M_{22} can be expressed as

$$\text{V}_{\text{fg},21} = \frac{1}{2}(V_{\text{in1}} + V_{\text{const}} - V_{\text{in2}}) + \frac{Q_{21}}{C_{\text{T}}} = \frac{1}{2}\Delta V_{\text{in}} + V_{\text{Q,cm}} + \frac{1}{2}V_{\text{Q,dm}} \tag{4}$$

$$\text{V}_{\text{fg},22} = \frac{1}{2}(V_{\text{in2}} + V_{\text{const}} - V_{\text{in1}}) + \frac{Q_{22}}{C_{\text{T}}} = -\frac{1}{2}\Delta V_{\text{in}} + V_{\text{Q,cm}} - \frac{1}{2}V_{\text{Q,dm}}, \tag{5}$$

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

$$V_{\text{Q,cm}} = \frac{1}{2}\left(\frac{Q_{21} + Q_{22}}{C_{\text{T}}} + V_{\text{const}}\right), V_{\text{Q,dm}} = \frac{Q_{21} - Q_{22}}{C_{\text{T}}}. \tag{6}$$

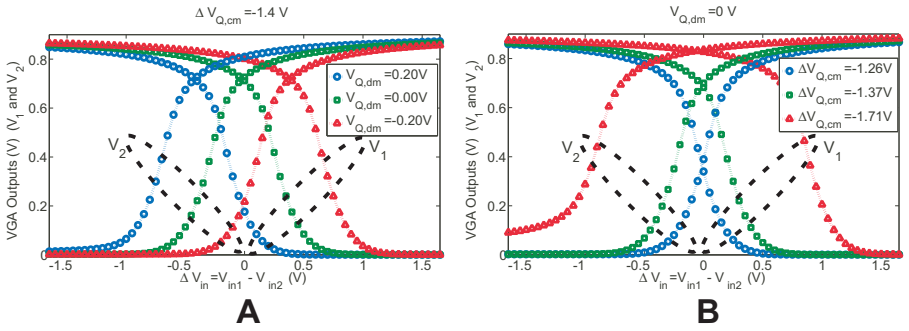

Fig. 4. Measured variable gain amplifier transfer characteristics. $V_{\rm in2}$ is fixed at $V_{\rm DD}/2$ and $V_{\rm in1}$ is swept from 0V to $V_{\rm DD}$, where $V_{\rm DD}$ is 3.3V. In the programming mode, the control gate voltages are set to be $-\Delta V_{\rm Q,cm} \mp V_{\rm Q,dm}/2$ and the floating-gate transistors are programmed to have 1 μA of current. **A:** The differential charge on M_{21} and M_{22} are programmed to several different levels and the amount of the common-mode charge is fixed. **B:** The common-mode charge on M_{21} and M_{22} are programmed to several different levels and the amount of the differential charge is fixed.

From (4) and (5), 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 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 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. On the other hand, as shown in Fig. 4**B**, 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. 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

$$\frac{\Delta V_{\rm out}}{\Delta V_{\rm in}} \approx -\gamma \left(1 + e^{-\frac{\gamma\kappa_{\rm p}}{2U_{\rm T}}(V_{\rm DD} - V_{\rm Q,cm} - V_{\rm T0,p})}\right) = \eta, \tag{7}$$

where $\gamma = \frac{\kappa_{\rm p}}{\kappa_{\rm n}}\sqrt{\frac{I_{0,\rm p}W_{\rm p}L_{\rm n}}{I_{0,\rm n}L_{\rm p}W_{\rm n}}}$, the subscripts "p" and "n" refer to the pMOS and nMOS transistors respectively, I_0 is the subthreshold pre-exponential current

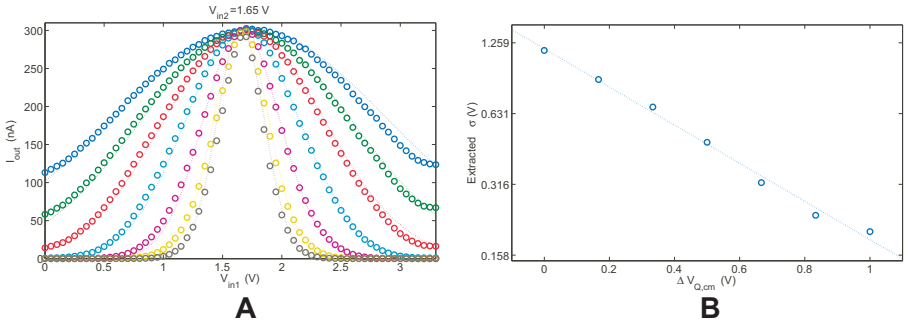

Fig. 5. Gaussian fits of the transfer curves and the width dependance. **A:** Comparison of the measured 1D bumps (circles) and the corresponding Gaussian fits (dashed lines). One of the bump input voltages is fixed at $V_{\mathrm{DD}}/2$, where V_{DD} is 3.3V through the measurement. The extracted standard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

factor, W and L are the dimensions of a transistor, κ is the subthreshold slope factor, V_{T0} is the threshold voltage, and U_{T} is the thermal voltage. From (2), the transfer function of the complete bump circuit can be expressed as

$$I_{\mathrm{out}} = \frac{2I_{\mathrm{b}}}{2 + e^{\kappa\eta\Delta V_{\mathrm{in}}/U_{\mathrm{T}}} + e^{-\kappa\eta\Delta V_{\mathrm{in}}/U_{\mathrm{T}}}}, \tag{8}$$

which is used to approximate a Gaussian function. By adjusting $V_{\mathrm{Q,cm}}$, the magnitude of the VGA gain increases exponentially and the extracted standard deviation decreases exponentially.

In Fig. 5**A**, the common-mode charge is programmed to several different levels and the transfer curves with different widths are measured. 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 (7). The minimum achievable extracted standard deviation from the measurements is 0.199V, 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

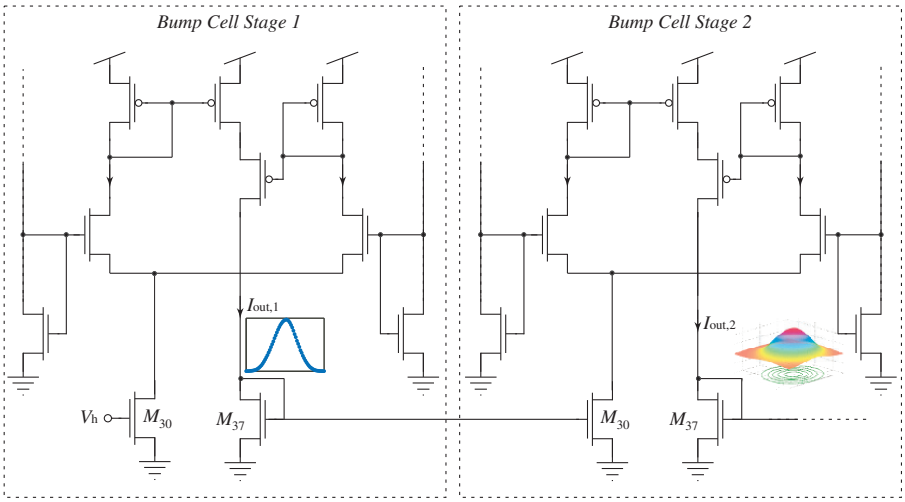

Fig. 6. By connecting the diode-connected output transistor to the tail transistor of the next stage bump cell, the resulting output current can approximate a multivariate Gaussian function with a diagonal covariance matrix.

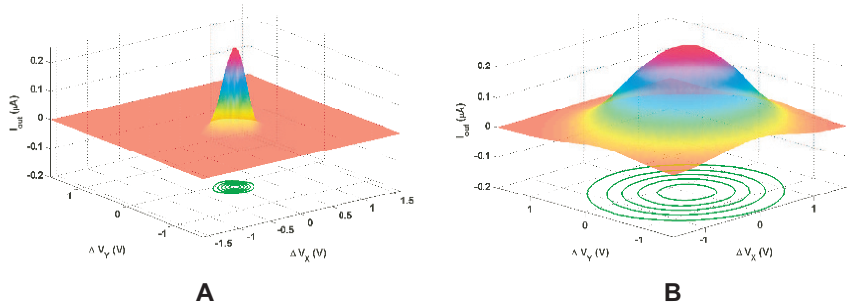

Fig. 7. Measurement results from two cascading floating-gate bump circuits. ΔV_X is the input voltage difference $\Delta V_{in} = V_{in1} - V_{in2}$ of the first stage floating-gate bump circuit and ΔV_Y is the input voltage difference of the second stage. In both stages, $V_{in2} = V_{DD}/2$. The common-mode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

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, two 2-D "bumps" with different widths approximating bivariate Gaussian functions with different standard deviations are shown. The output currents of an array of these floating-gate bump circuits can easily be summed up to implement GMMs.

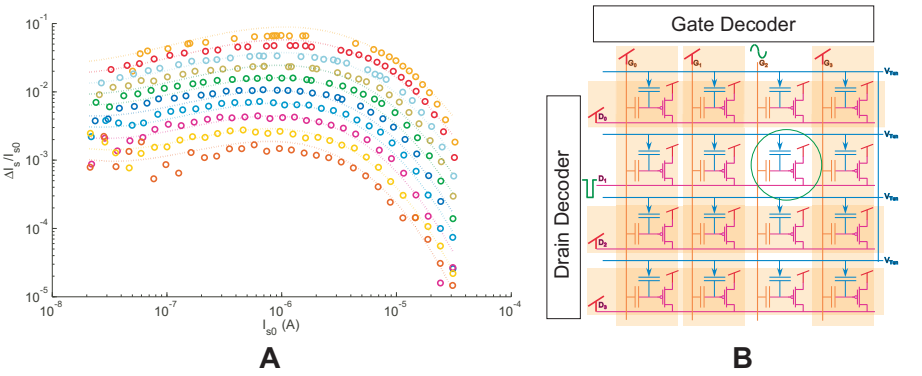

Fig. 8. A: Measured injection characterization points (circles) and the corresponding curve fits (dashed lines). The pulse width is fixed at 200μsec. 10 different values of $V_{\rm ds}$ ranging from 5.6V to 6.5V and 30 channel current levels ranging from 20nA to 20μA are used to obtain the curve fits for each curve. Cubic functions are used to regress the nonlinear functions $g(\cdot)$ and $f(\cdot)$ in (10). **B:** The block diagram of 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.

4 Programming Floating-gate Transistor Array

How to accurately programming an array of floating-gate transistors is a critical technique in the development of the proposed analog classifier. Fowler-Nordheim tunneling and channel hot electron 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]. The floating-gate programming method and the way to program an array of floating-gate transistors will be briefly reviewed in this section.

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, channel hot electron injection are employed. As detailed in [16], the injection current can be modeled as

$$I_{\rm inj} = I_{\rm inj0} \left(\frac{I_{\rm s}}{I_{\rm s0}} \right)^{\alpha} e^{-\Delta V_{\rm ds}/V_{\rm inj}}, \tag{9}$$

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 (9), logarithmic of this ratio should be a linear function of V_{ds} and a nonlinear function of $\log(I_{\mathrm{s0}}/I_{\mathrm{u}})$, where I_{u} is an arbitrary unity current. It can be expressed as

$$\log\left(\frac{\Delta I_{\mathrm{s}}}{I_{\mathrm{s0}}}\right) = g\left(\log\left(\frac{I_{\mathrm{s0}}}{I_{\mathrm{u}}}\right)\right) V_{\mathrm{ds}} + f\left(\log\left(\frac{I_{\mathrm{s0}}}{I_{\mathrm{u}}}\right)\right), \tag{10}$$

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_{ds} and I_{s0} are given and ΔI_{s} can be measured. Thus, $g(\log(I_{\mathrm{s0}}/I_{\mathrm{u}}))$ and $f(\log(I_{\mathrm{s0}}/I_{\mathrm{u}}))$ can be regressed by high order polynomial functions. After the characterization process, we obtain the resulting polynomial regressive functions, $\hat{f}(\log(I_{\mathrm{s0}}/I_{\mathrm{u}}))$ and $\hat{g}(\log(I_{\mathrm{s0}}/I_{\mathrm{u}}))$. In the programming process, with the regressive functions, the appropriate V_{ds} value for injection can be predicted by

$$V_{\mathrm{ds}} = \frac{\log\left(\frac{\Delta I s}{I_{\mathrm{s0}}}\right) - \hat{f}\left(\log\left(\frac{I_{\mathrm{s0}}}{I_{\mathrm{u}}}\right)\right)}{\hat{g}\left(\log\left(\frac{I_{\mathrm{s0}}}{I_{\mathrm{u}}}\right)\right)}, \tag{11}$$

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

The measured and the regressive results for the 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_{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 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 (7).

To program an array of the floating-gate bump circuits, floating-gate transistors are arranged as in Fig. 8**B** in the programming mode. There are two conditions required for 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.

5 A Programmable Analog Vector Quantizer

A *"FG-pFET & Mirror"* block shown in Fig. 9**A** is added in front of the first bump cell to program its tail current, which sets the height of the "bump."

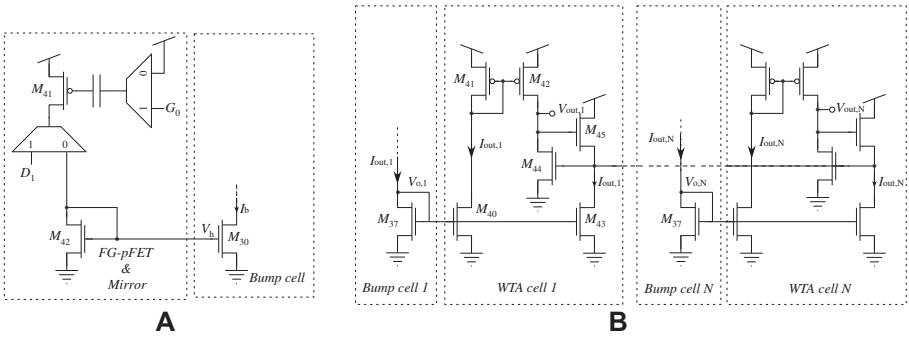

Fig. 9. A: The schematic of the *"FG-pFET & Mirror"* block. The charge on the *p*MOS transistor can be programmed to set the height of the bell-shaped transfer curve. **B:** The schematic of a current mode winner-take-all circuit. Only the output voltage of the winning cell will be high to indicate the best-matching template.

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 winner-take-all circuit, the schematic of which is shown in Fig. 9**B**. Only the output voltage of the winning cell 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. 10. Most of the multiplexers are in the inverse generation and bias generation blocks. Since only one bias generation block is needed for the whole system, when the system is scaled up, the bias generation block does not cost extra complexity. In the analog RBF-based classifier and the 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. 11.

Two examples are used to demonstrate the reconfigurability of the classifiers as shown in Fig. 12. Four templates are used and their outputs are superposed in a 3-D plot. 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.

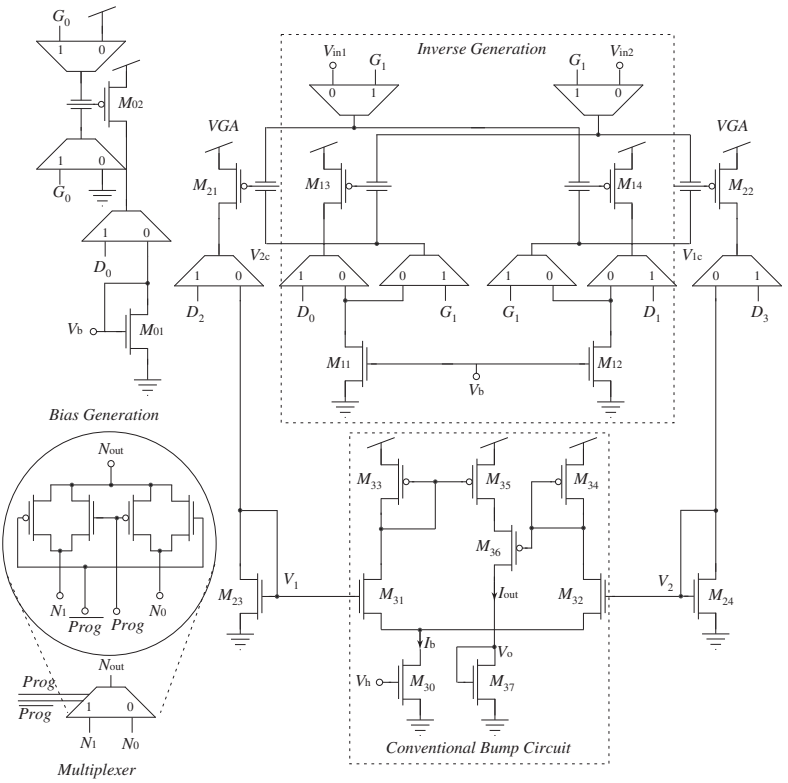

Fig. 10. The complete schematics of the floating-gate bump circuit. Multiplexers for floating-gate programming are inserted into the original circuits. The "1" on the multiplexer indicates the connection in the programming mode and the "0" indicates the connection in the operating mode. The tunneling junction capacitors are not shown for simplicity. Most of the multiplexers are in the bias generation and inverse generation blocks. Only two multiplexers are added in the bump cell that includes the VGA and the conventional bump circuit.

6 Performance of The Analog Vector Quantizer

We have fabricated a prototyped analog vector quantizer in a $0.5\,\mu$m CMOS process. We also fabricated a 16×16 highly compact low-power version of an analog vector quantizer in the $0.5\,\mu$m CMOS process occupying less than $1.5 \times 1.5mm^2$. Some important parameters and measured results are listed in the TABLE 1.

To measure the power consumption, several "bumps" are programmed with identical width while other "bumps" are deactivated 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. 13**A** indicates the average power consump-

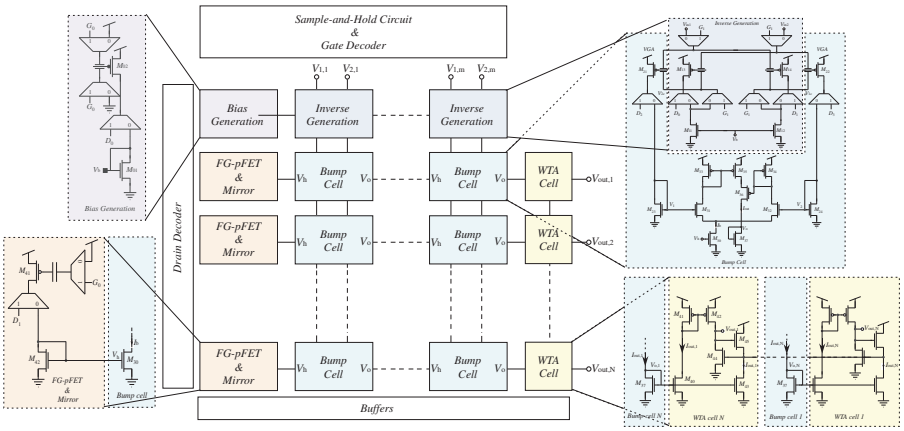

Fig. 11. Architecture of an analog vector quantizer. The core is the bump cell array followed by a WTA circuit. The main complexity from programming are at the peripheries and the system can be scaled up easily.

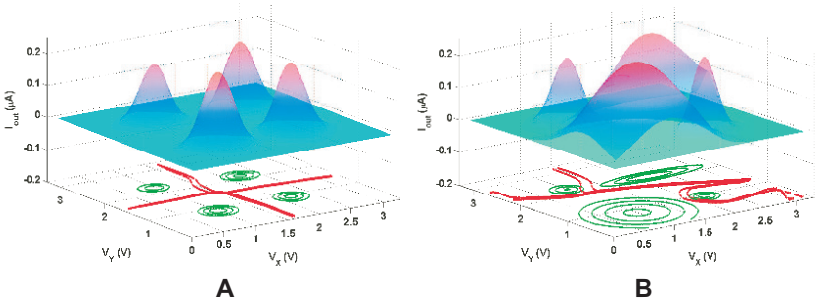

Fig. 12. Configurable classification results. The measured bump output currents (circle contours) and the WTA voltages (thick solid lines at the bottom) of four templates are superposed in a single plot. V_{X} and V_{Y} are the V_{in1} in the first stage and the second stage floating-gate bump circuits respectively. Both of their V_{in2} terminals are fixed at $V_{\mathrm{DD}}/2$. **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.

tion per bump cell with a specific value of width. The relation between the power consumption and the extracted standard deviation is shown in Fig. 13**B**.

The VGA is the main source of the power consumption. The gain is tunable when the *n*MOS 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 *p*MOS transistors when they are in transition between saturation and ohmic region. From simulation, to save the power consumed in

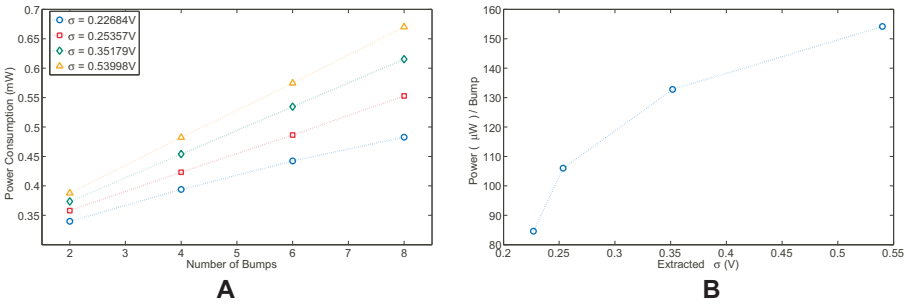

Fig. 13. Relation between the power consumption and the extracted variance. **A:** Measured power consumption of the analog vector quantizer with different number of floating-gate bump cells being activated with a fixed width. The slope of the curves indicate the average power consumption per bump cell. **B:** The 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.

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 1KHz at that current level, we can not measure the speed of our floating-gate bump circuit directly, which is expected to be around mega-Hz range. We can only measure the response time from the input to the WTA outputs. The measured transient response of the analog vector quantizer is shown in Fig. 14**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 with the cost of more power consumption as shown in Fig. 14**B**. The

Table 1. Analog Vector Quantizer Parameters

Size of VQ	7(templates)×2(components)
Area/Bump Cell	$42 \times 82\,\mu\text{m}^2$
Area/WTA Cell	$20 \times 35\,\mu\text{m}^2$
Power Supply Rail	$V_{\text{DD}} = 3.3V$
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

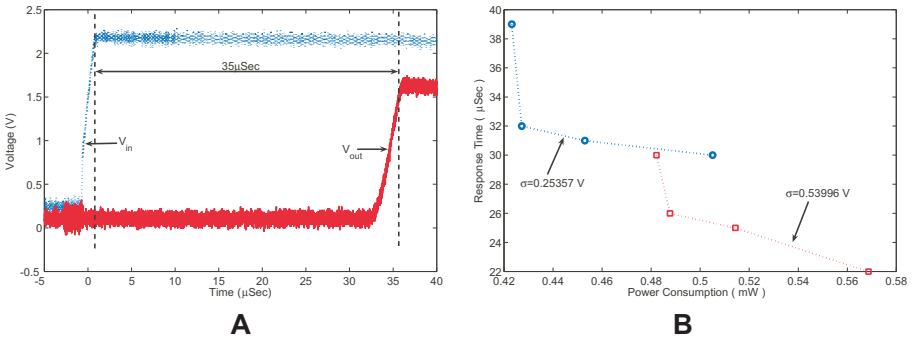

Fig. 14. Response time and speed-power trade-off of an analog vector quantizer. **A:** The response time between the input voltage and the WTA output. **B:** The relation between the response time and the power consumption for a given bump width. The inverse generation block dominates the response time 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 trade-off.

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. 14**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 trade-off between the speed and the power consumption in the inverse generation block.

Finally, we 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, we choose to evaluate the impact of using the analog RBFs on system performance. To this end receiver operating characteristic (ROC) curves and equal error rate (EER) are adopted. Two separate 2D bumps are programmed to have the same variance with a fixed separation as shown in Fig. 15. The corresponding Gaussian fits are used as the actual probability density functions (pdf) of two classes. Comparing these two pdf's using different thresholds renders a ROC curve of these two Gaussian distributed classes that is used as the evaluation reference. With the knowledge of the class distributions, comparing the output currents using different thresholds generates a ROC curve for the 2D 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. The EER, which is the intersection of the ROC curve and the $-45°$ line as shown in Fig. 16**A**, is the usual operating point of classifiers. In

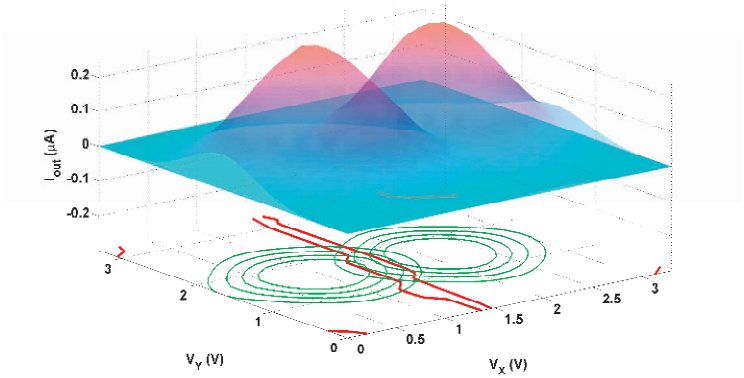

Fig. 15. Distributions of two "bumps" used to evaluate the classifier performance. In the measurements for performance evaluation, the separation of the center is kept constant but the widths of these two "bumps" varies. The measured bump output currents (circle contours) and the WTA voltages (thick 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 $V_{\mathrm{DD}}/2$.

Fig. 16**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 the analog 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 chapter.

7 Power Efficiency Comparison

To compare the efficiency of our analog system with the DSP hardware, we estimate the metric of millions of multiply accumulates per second per milli-watt (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 10μ sec (which is still an overestimate) with the power consumption of $120\mu W$ 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

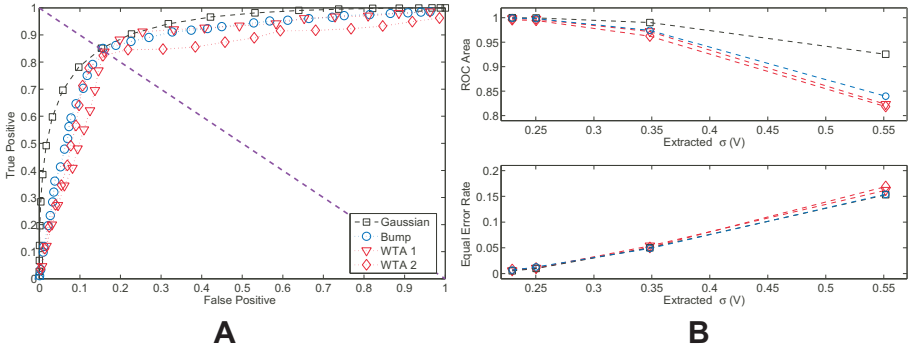

Fig. 16. ROC and EER performance of the classifiers. **A:** The ROC curves of the Gaussian fits (squares), output currents of the 2D bumps (circles) and WTA output voltages (triangles and diamonds) with the extracted $\sigma = 0.55$V. The Gaussian fits are used as the actual pdf's of the two classes 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. **B:** The effects of different bump widths on the receiver operating characteristic (ROC) area and the equal error rate (EER) performance. The separation of the means of two classes is 1.2V. The results show that the analog VQ is comparable to an ideal maximum-likelihood (ML) classifier.

is expanded to include the WTA function, the efficiency of the proposed 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], the efficiency of which ranges from 37 to 175 MMAC/s/μW. 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, the power consumption can be easily reduce 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.

8 Conclusion

In this chapter, a new programmable floating-gate bump circuit is demonstrated. The height, the center and the width of its bell-shaped transfer characteristics can be programmed individually. A multivariate radial basis function with a diagonal matrix can be realized by cascading these bump cells. Based on the new bump circuit, a novel compact RBF-based soft classifier is built. By adding a simple current mode winner-take-all 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 *n*MOS transistors in the VGA are assumed in the transition between the above-threshold and the subthreshold regions. The *p*MOS 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 *p*MOS and *n*MOS transistors, we can have

$$\begin{aligned} & I_{0,\mathrm{p}}\left(\frac{W_{\mathrm{p}}}{L_{\mathrm{p}}}\right)\frac{1}{4U_{\mathrm{T}}^2}\left[\kappa_{\mathrm{p}}(V_{\mathrm{DD}}-V_{\mathrm{fg},21}-V_{\mathrm{T0,p}})\right]^2 \\ & \quad = I_{0,\mathrm{n}}\left(\frac{W_{\mathrm{n}}}{L_{\mathrm{n}}}\right)\ln^2\left(1+e^{\frac{\kappa_{\mathrm{n}}}{2U_{\mathrm{T}}}(V_1-V_{\mathrm{T0,n}})}\right) \end{aligned} \tag{12}$$

where the subscripts of "p" and "n" refer to *p*MOS and *n*MOS 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, $V_{\mathrm{Q,dm}} = 0$ and

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

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 (12) with respect to $V_{\mathrm{fg},21}$ and have

$$\begin{aligned} \frac{\Delta V_{\mathrm{out}}}{\Delta V_{\mathrm{in}}} &= \frac{dV_1}{dV_{\mathrm{fg},21}} = -\gamma\left(1+e^{-\frac{\kappa_{\mathrm{n}}}{2U_{\mathrm{T}}}(V_1-V_{\mathrm{T0,n}})}\right) \\ &= \frac{-\gamma}{1-e^{-\frac{\gamma\kappa_{\mathrm{p}}}{2U_{\mathrm{T}}}(V_{\mathrm{DD}}-V_{\mathrm{fg},21}-V_{\mathrm{T0,p}})}} \\ &\approx -\gamma\left(1+e^{-\frac{\gamma\kappa_{\mathrm{p}}}{2U_{\mathrm{T}}}(V_{\mathrm{DD}}-V_{\mathrm{Q,cm}}-V_{\mathrm{T0,p}})}\right), \end{aligned} \tag{13}$$

where $\gamma = \frac{\kappa_{\mathrm{p}}}{\kappa_{\mathrm{n}}}\sqrt{\frac{I_{0,\mathrm{p}}W_{\mathrm{p}}L_{\mathrm{n}}}{I_{0,\mathrm{n}}L_{\mathrm{p}}W_{\mathrm{n}}}}$. 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," in *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 *Symposium on VLSI Circuits,* pp.244-247, 13-15 June 2002.
3. M. Bracco, S. Ridella, and R. Zunino, "Digital Implementation of Hierarchical Vector Quantization," in *IEEE Trans. Neural Networks,* vol. 14, no. 5, pp.1072-1084, Sep. 2003.
4. G. Cauwenberghs and V. Pedron,"A low-power CMOS analog vector quantizer," in *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, D. Anderson, "A floating-gate vector-quantizer," in *Midwest Symposium on Circuits and Systems*, Vol.1,4-7, Aug. 2002, pp. I-196-9.
6. T. Yamasaki and T. Shibata,"Analog soft-pattern-matching classifier using floating-gate MOS technology," in *IEEE Trans. Neural Networks*, 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 *Int. Joint Conf. Neural Networks*, 1992, vol. 2, pp. 607V612.
9. J. Choi, B. J. Sheu, and J. C.-F. Chang, "A Gaussian synapse circuit for analog neural networks," in *IEEE Trans. VLSI Syst.,* vol. 2, pp. 129V133, Mar. 1994.
10. S.-Y. Lin, R.-J. Huang, and T.-D. Chiueh, "A Tunable Gaussian/Square Function Computation Circuit for Analog Neural Networks" in *IEEE Transactions on Circuits and System II*, vol. 45, no. 3, 1998, pp. 441-446.
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 *Intelligent Control, 2002. Proceedings of the 2002 IEEE International Symposium on*, Vancouver, Cananda, Oct. 2002, pp. 454-459.
12. D. Hsu, M. Figueroa, and C. Diorio, "A silicon primitive for competitive learning," in *Conference on Neural Information Processing Systems*, Dec. 2000.
13. P. Hasler, "Continuous-Time Feedback in Floating-Gate MOS Circuits," in *IEEE Trans. Circuit and system II*, 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," in *IEEE Trans. Circuit and system II*, 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," in *IEEE J. Solid-State Circuits,* vol. 41, no. 9, pp.2107-2114, Sept. 2006.
16. P. Hasler and J. Dugger,"Correlation Learning Rule in Floating-Gate pFET Synapses," in *IEEE Trans. Circuit and system II*, 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. CICC05, Sept. 2005, pp. 734737.
18. J. Glossner, K. Chirca, M. Schulte, H. Wang, N. Nasimzada, D. Har, S. Wang; A. J. Hoane, G. Nacer, M. Moudgill, M., S. Vassiliadis, "Sandblaster low power DSP," in *IEEE Prec. Custom Integrated Circuits Conference*, pp.575-581, oct. 2004.
19. R. Chawla, A. Bandyopadhyay, V. Srinivasan, and P. Hasler, "A 531nW/MHz, 128x32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity," in *IEEE Prec. Custom Integrated Circuits Conference*, pp.651-654, oct. 2004.

20. R. Karakiewicz, R. Genov, A. Abbas, and G. Cauwenberghs, "175 GMACS/mW Charge-Mode Adiabatic Mixed-Signal Array Processor," in *Symposium on VLSI Circuits*, June, 2006.