Noname manuscript No.
(will be inserted by the editor)

An Energy-Efficient Spike Encoding Circuit for Speech Edge Detection

Dingkun Du · Kofi Odame

Received: date / Accepted: date

Abstract In speech processing applications, the instantaneous bandwidth of speech can be used to adaptively control the performance of an audio sensor's analog front end. Extracting the instantaneous bandwidth of speech depends on the detection of speech edges in the time-frequency plane. In this paper, we propose a spike encoding circuit for real-time and low-power speech edge detection. The circuit can directly encode the signal's envelope information by temporal spike density without additional envelope extraction. Furthermore, the spike encoding circuit automatically adapts its resolution to the amplitude of the input signal, which improves the encoding resolution for small signal without increasing the power consumption. We use the nonlinear dynamical approach to design this circuit and analyze its stability. We also develop a linearized model for this circuit to provide the design intuition and to explain its adaptive resolution. Fabricated in 0.5-μm CMOS process, the spike encoding circuit consumes 0.3-μW power and the experimental results are presented.

Keywords Spike encoding · Biologically inspired circuits · Nonlinear circuits · Asynchronous circuits · Speech processing

1 Introduction

In the mammalian auditory system, incoming sound is encoded as a series of auditory nerve firing patterns or spikes. This spike encoding has evolved to specifically enhance certain features [1]. For instance, changes in the auditory environment – like the onset of a new sound – are quickly identified and emphasized by the auditory nerve firing patterns. This ability is important for

D. Du and K. Odame
Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA
Tel.: +1-603-646-2230
Fax: +1-603-646-3856
E-mail: dingkun.du@dartmouth.com; odame@dartmouth.edu

priming a response in the mammal [2]. The representation of sounds of interest (e.g. speech) is particularly invariant, even in the presence of loud, interfering noise [3].

The robustness, efficiency and low latency of the biological auditory system has inspired several audio processing algorithms that are based on spike encoding [2], [3], [4], [5], [6]. Likewise, our recently-developed real-time *speech edge detection* algorithm [7] relies on spike encoding. Our algorithm's target applications are low-power, real-time systems. So, it is crucial that our spike encoding implementation be energy efficient and real time.

According to [2], [8], the energy (or envelope) change is an effective feature to identify speech edges. Therefore, an essential function in speech edge detection algorithm is to capture the onsets and offsets of the speech signal according to its envelope change information.

In this paper, we propose a novel spike encoding circuit to represent the signal's envelope change information with the temporal spike density. To achieve a high efficiency, we apply the nonlinear dynamical approach to this design. The resulting circuit uses one comparator and a variable threshold to generate spikes and does not need additional envelope extraction. Also, the encoding resolution is adaptive. It is finer for smaller envelope change, which makes small signals easier to capture without wasting power consumption to improve resolution for the full range of the envelope change. Our prototype integrated circuit implementation only consumes 0.3-μW power and occupies 0.03-mm^2 die size in 0.5-μm CMOS process.

The rest of this paper is organized as follows. Section II introduces the background of the spike encoding circuits, including the implications of speech edge detection and previous works related to the spike encoding circuit. Section III describes the design and analysis method of our novel circuit based on the nonlinear dynamical approach, including its nonlinear state-space model, circuit macromodel and stability issues. Section IV provides the encoding scheme, including the specifications and the linearized model to explain the adaptive encoding resolution nature of the circuit. Section V presents the CMOS implementation and measurement results of the circuit. Section VI summarizes the design and performance of the circuit.

2 Background

2.1 Implications of Speech Edge Detection

We have developed a real-time algorithm [7] for detecting the edges of speech in the time-frequency (T-F) plane, where we define the *edge* as the maximum frequency of the speech events along with the time. In this algorithm, like other related algorithms in T-F plane [2], [3], [8], the input speech is at first split into a series of frequency channels by a bank of bandpass filters. Next, the subband signal in each frequency channel is encoded into asynchronous spikes by the spike encoding circuit. Then, the spikes from each channel are

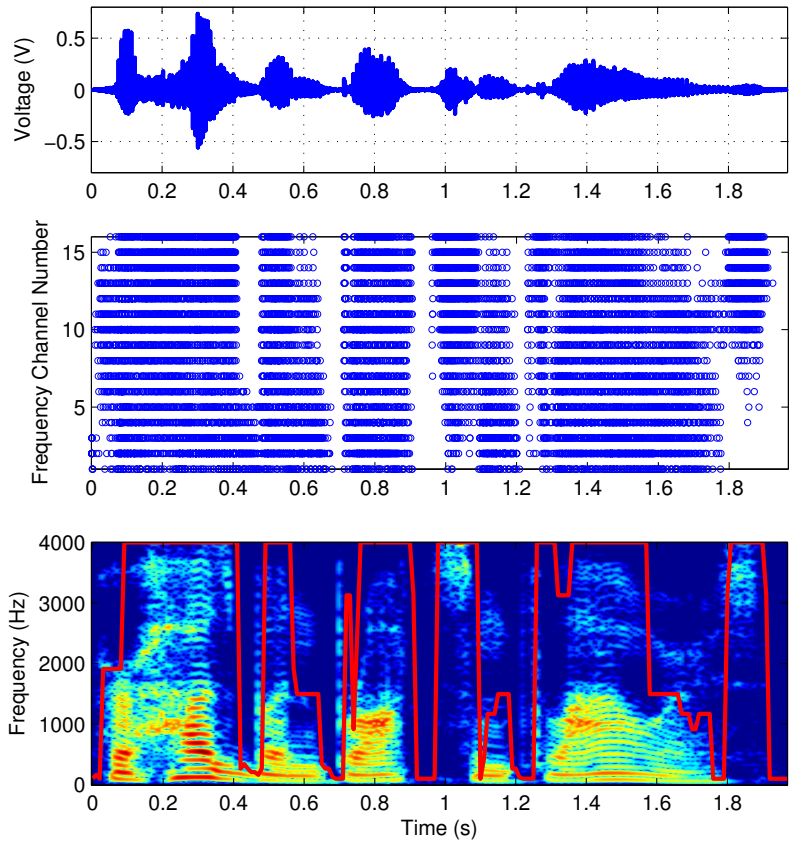

Fig. 1 A speech sample (top), the generated spikes in each frequency channel (middle) and its spectrogram (bottom), with the extracted speech edge (piecewise line) shown on the spectrogram. The spikes in each channel and the speech edge comes from the measurement results of the proposed spike encoding circuit. In this example, 16 frequency channels are logarithmically spaced to cover 100 Hz to 4 kHz. We can extract the speech edge from the temporal spike density in each frequency channel. The measurement details are illustrated in Section 5.3.

counted in a fixed time window with a ripple counter to determine the spike density. As we will explain in the next section, the spike density indicates the edges of speech in the T-F plane..

Fig. 1 depicts an example of our algorithm output for a particular speech sample. The edges were detected based on measurement results of the proposed spike encoding circuit. We can observe the speech edge adapts with the speech's energy distribution on T-F plane. The application of this algorithm is in smart audio sensors that only process the speech portions of the spectrum, while discarding any non-speech audio that may be simultaneously present. So, the smart sensors does not only save power with the bandwidth adaptation but also has decreased output noise level.

Given its application in a speech edge detection algorithm in T-F plane, the spike encoding will be applied to each of several frequency channels, so that the power budget for each encoding circuit is very limited and a highly efficient circuit implementation is required.

2.2 Related Works

Envelope change rate can be used to identify speech edge, so that a plausible way seems to be performing the edge detection by extracting the envelope first and then differentiating it to get its time derivative. After that, we could use the time derivative of the envelope to compare with a threshold to determine the onset/offset. For example, Hu and Wang proposes an speech onset/offset detection algorithm according to the time derivative of the speech's envelope [8].

Comparing to the envelope time derivative based method, the spike encoding scheme has some advantages. First of all, the analog time derivative is more sensitive to noise and interference comparing to the encoded spikes, so that the noise or interference may trigger false decision when the time derivative is near the threshold. In [8], the algorithm is not real time so that the envelope can be effectively smoothed, but in the real-time applications, the aggressive smoothing process will bring significant latency. Secondly, in the spike encoding scheme, the onset/offset decision can be implemented by the digital circuits (counters), which provide more flexibility to set the circuit parameters like the decision threshold. In the time-derivative based method, we may need additional device array, *e.g.*, the capacitor or resistor bank, to change such parameters, resulting in additional chip size and design complexity. Finally, an inherent property of our spike encoding circuit is that it has adaptive resolution. If we tried to implement such a feature in the envelope time derivative method, we would require a variable gain control loop, which demands more hardware and power resources.

Previous spike encoding circuits are usually used to encode the signal's complete waveform instead of the envelope [12], [13]. Although we can use the encoded waveform to extract the envelope information, the signal's waveform usually changes faster than the envelope does, so that the power consumption would be wasted if we just need the envelope information. We can also use envelope detectors to extract the envelope first, and then send the envelope to the spike encoding circuits [14]. However, the envelope detector brings additional power consumption, hardware complexity and latency.

The spike encoding strategy proposed in [2] can directly encode the envelope. It uses a series of comparators with different threshold voltages to generate spikes without the explicit envelope detector. However, [2] does not provide the hardware implementation and power consumption of the comparators, which could be considerable when using 8 or 16 such comparators together.

Therefore, the spike encoding circuit proposed in this paper is specifically designed to represent the signal's envelope information. It uses only one comparator with an adaptively changing threshold to generate spikes and no additional envelope detector is needed. So, we can expect high energy and hardware efficiency of this circuit. Moreover, this spike encoding circuit can adapt its resolution with different envelope amplitude to save power, which will be illustrated in Section 4.2. In addition, such spike encoding scheme is asynchronous, which falls into the event-driven signal processing strategy [9], [10], [11].

3 Model

3.1 Basic Considerations

Generally speaking, we will design a spike encoding circuit that can output a train of spikes with varying density to represent the signal's envelope information. Specifically, the spike train density should increase both as a function of signal amplitude and as a function of signal amplitude change. For the speech edge detection on T-F plane, in the time dimension, there will be a high density of spikes during speech onsets and a low density of spikes during speech offsets (due to the dependence on amplitude change). In the frequency dimension, there will be a higher density of spikes in the frequency bands where there is speech, compared to those bands where there is no speech (due to the dependence on amplitude).

Our spike encoding circuit is based on a comparator, whose input is the speech signal for a given frequency band. Whenever the comparator detects an input that exceeds a threshold level, it outputs a spike. In order to vary the density of the resulting spike train, the threshold of the comparator must adapt appropriately. The comparator gives a positive output whenever the input signal exceeds the threshold. This positive output is short lived (it is in fact a spike), because the threshold responds by increasing past the input signal level. The comparator gives a negative output whenever the threshold exceeds the input signal. The threshold then responds to the negative output by slowly decaying towards the input signal.

The speech phoneme can be modeled as amplitude-modulated (AM) signal [14], [15], so we can use an AM signal as the input to the spike encoding circuit to illustrate the encoding process. As Fig. 2 shows, the asymmetric response of the threshold will produce a high spike density during increases in signal amplitude and a low spike density during falls in signal amplitude. Also, large amplitude signals will have more opportunities for spikes to be generated than will small amplitude signals.

3.2 State-Space Model

The key characteristics of the variable threshold are as follows: It tracks the input signal, so the input signal represents the equilibrium point of the sys-

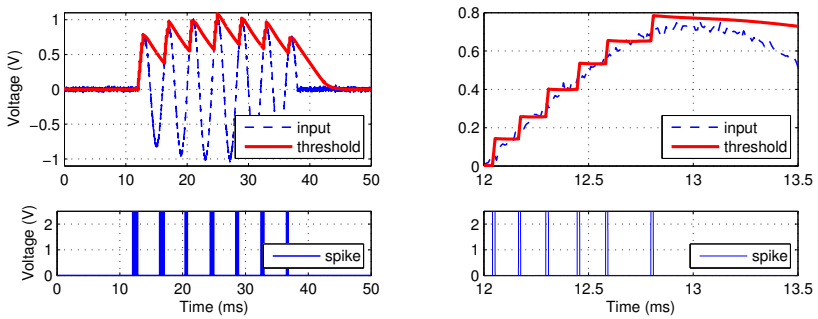

Fig. 2 Amplitude-modulated (AM) signal, the variable threshold and the corresponding generated spikes (left), with the zoomed view around 13 ms (right). On the left panel, the spike thickness indicates the temporal density due to the limited resolution of the figure.

tem. When the variable threshold is below the equilibrium point, it increases rapidly, overshoots the equilibrium, and then decays slowly back towards the equilibrium.

Representing the input as u, the variable threshold as x and the output spike as y, we can write the state model for the nonlinear system as:

$$\begin{aligned} \dot{x} &= f_{xr}(x,y,u)H(y) + f_{xf}(x,y,u)H(-y) \\ \dot{y} &= f_y(x,y,u) \end{aligned} \tag{1}$$

where $f_{xr}(x,y,u)$ is the function corresponding to the jumping of the threshold and $f_{xf}(x,y,u)$ is the function corresponding to the decaying process. The function $H(\cdot)$ is the Heaviside function (unit step function). When $y > 0$, x jumps as $f_{xr}(x,y,u)$, and when $y < 0$, x decays as $f_{xf}(x,y,u)$. All of u, x and y are normalized dimensionless variables in the interval of $[-1,1]$, where "-1" represents the minimum voltage level and "1" represents the maximum.

We set the jump of x as a constant. This jump occurs when $y > 0$, so we have:

$$f_{xr}(x,y,u) = \alpha H(y) \tag{2}$$

where α is a positive coefficient.

And, the decaying of x means x has a tendency to reach back to u, so a simple way to describe it is a linear equation as:

$$f_{xr}(x,y,u) = -\beta(x-u)H(-y) \tag{3}$$

where β is positive. Now, $\dot{x}$ can be expressed as:

$$\dot{x} = \alpha H(y) + \beta(u-x)H(-y) \tag{4}$$

Because the decaying of the threshold is much slower than the jumping process, we have that $|f_{xr}(x,y,u)| \gg |f_{xf}(x,y,u)|$, or $|\alpha| \gg |\beta(u-x)|$, and the following approximation holds:

$$\alpha \approx \alpha + \beta(u-x) \tag{5}$$

Therefore, using the approximation above, we modify the expression for $\dot{x}$ as below:

$$\begin{aligned}\dot{x} &= [\alpha + \beta(u-x)]H(y) + \beta(u-x)H(-y) \\ &= \alpha H(y) + \beta(u-x)(H(y) + H(-y)) \\ &= \alpha H(y) + \beta(u-x)\end{aligned} \tag{6}$$

Next, we can pick $f_y(x, y, u)$ such that y is positive when x is below the input, and y becomes negative only when x has overshot the input by some specific amount. A very simple way to do this is to implement a delayed comparison to the equilibrium. That is, set:

$$\dot{y} = \gamma[\mathrm{sgn}(u-x) - y] \tag{7}$$

where γ is a positive coefficient. And, $\mathrm{sgn}(\cdot)$ is the signum function, which is usually implemented as a saturation function in real-world applications:

$$\mathrm{sat}(v) = \begin{cases} v & \text{if } |v| \leq 1 \\ \mathrm{sgn}(v) & \text{if } |v| > 1 \end{cases} \tag{8}$$

To mimic the steep step of the sigmoid function, we express $\mathrm{sgn}(x)$ as $\mathrm{sat}(Ax)$, where $A \gg 1$.

Now, as we know that y is the delayed version of signum function, we have that $y = \mathrm{sgn}(y)$. $H(y)$ can be simply expressed as:

$$H(y) = \frac{\mathrm{sgn}(y) + 1}{2} = \frac{y+1}{2} \tag{9}$$

So, we can reach the state-space model for the system as:

$$\begin{aligned}\dot{x} &= \alpha\frac{y+1}{2} + \beta(u-x) \\ \dot{y} &= \gamma[\mathrm{sat}(A(u-x)) - y]\end{aligned} \tag{10}$$

3.3 Circuit Macromodel

The nonlinearity in the state model is sigmoid, which can be implemented by a comparator. The differential parts of the equations can be easily implemented by time delay. A possible circuit macromodel is show in Fig. 3, which consists of a comparator, a RC delay and a switched current source. The state-space model for this system is:

$$\begin{aligned}\dot{x} &= \frac{(y+1)I_c}{2V_aC} + \frac{u-x}{RC} \\ \dot{y} &= \frac{\mathrm{sat}(A(u-x)) - y}{\tau}\end{aligned} \tag{11}$$

where u, x and y denote the input signal, the threshold and the output spikes respectively. Also, I_c is the value of the current source, R is the resistance, C

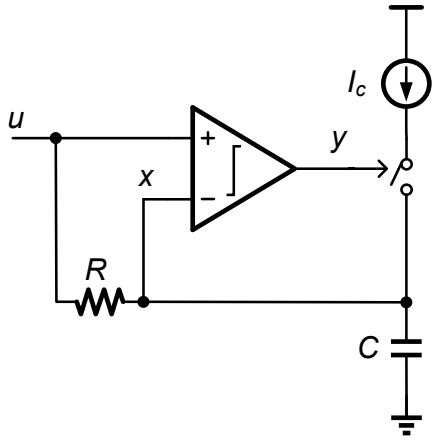

Fig. 3 One possible spike encoding circuit macromodel based on the state-space system model. The input is u and the variable threshold is x. The RC filter causes the variable threshold to adapt towards the input. The switched current source forms a charge pump that is activated whenever a spike, y, is generated.

is the capacitance, τ is the delay of the comparator, A is the gain of the comparator and V_a is highest input amplitude in order to normalize the expression for $\dot{x}$.

The coefficients of the state-space model can be expressed by circuit parameters as:

$$\alpha = \frac{I_c}{V_a C}, \quad \beta = \frac{1}{RC}, \quad \gamma = \frac{1}{\tau} \tag{12}$$

In this circuit, once u is higher than x, a high output of the comparator will open the switch of the current source to quickly charge the capacitor and make x jump to a higher level. Next, when x jumps higher than u, the low output of the comparator will shut off the charge pump, while x has a decay with time constant $\tau_d = RC$. If u rises past x, the process above will repeat. Therefore, the nonlinear state-space model has given us an efficient way to synthesize a non-traditional nonlinear system, while the resulting circuit implementation is naturally understood by traditional circuit intuition.

3.4 Stability

The spike encoding circuit should not generate spikes for direct-current (DC) input, otherwise the DC signal or small noise may generate spurious spikes to corrupt the encoding performance and cause incorrect edge detection. A traditional delta modulator has granular noise which appears on the output for direct-current (DC) or low-amplitude input [16]. As described above, the spike encoding circuit's encoding process for the signal's rising amplitude is like an asynchronous delta modulator [13], so that it is important to investigate the stability of the spike encoding circuit.

We can calculate the equilibrium point of the system for $u = u_0$ (DC input) from (10) as below:

$$\begin{aligned} x_0 &= u_0 + \frac{\alpha}{\alpha A + 2\beta} \approx u_0 + \frac{1}{A} \\ y_0 &= -\frac{\alpha A}{\alpha A + 2\beta} \approx -1 \end{aligned} \tag{13}$$

One important interpretation is that x is very close to but slightly higher than u at the equilibrium point, and the corresponding output y is at ground.

To simply our analysis below, we shift the equilibrium point to origin by the linear transformation:

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \tag{14}$$

The state-space model for the linear-shifted states with $u = u_0$ is:

$$\begin{aligned} \dot{z}_1 &= \alpha \frac{z_2 + y_0 + 1}{2} + \beta(u_0 - z_1 - x_0) \\ \dot{z}_2 &= \gamma[\mathrm{sat}(A(u_0 - z_1 - x_0)) - z_2 - y_0] \end{aligned} \tag{15}$$

which can be simplified as:

$$\begin{aligned} \dot{z}_1 &= \alpha \frac{z_2}{2} - \beta z_1 \\ \dot{z}_2 &= \gamma[\mathrm{sat}(A(q - z_1)) - z_2 - \mathrm{sat}(Aq)] \end{aligned} \tag{16}$$

where $q = u_0 - z_0$. For this system, the equilibrium point is at origin $\mathbf{z_0} = (0, 0)^T$.

We propose a Lyapunov function candidate for (16) as:

$$V(\mathbf{z}) = \frac{\gamma}{A} \int_0^{z_1} [\mathrm{sat}(A(\xi - q)) + \mathrm{sat}(Aq)] d\xi + \alpha \frac{z_2^2}{4} \tag{17}$$

We can verify that $V(\mathbf{z}) \geq 0$ with equality only when $\mathbf{z} = (0, 0)^T$. The time derivative of $V(\mathbf{z})$ is:

$$\dot{V}(\mathbf{z}) = -\beta\gamma z_1 [\mathrm{sat}(A(z_1 - q)) + \mathrm{sat}(Aq)] - \alpha\gamma \frac{z_2^2}{2} \tag{18}$$

We can also verify that $V(\mathbf{z}) \leq 0$ with equality only when $\mathbf{z} = (0, 0)^T$. Therefore, the equilibrium point $\mathbf{z_0} = (0, 0)^T$ is asymptotically stable according to Lyapunov stability theory. Because $\mathbf{z}$ is linearly shifted through $(x, y)^T$ in (14), so that the original equilibrium point $(x_0, y_0)^T$ is also asymptotically stable. This means that the spike encoding circuit has a stable equilibrium point so that there is no granular noise for DC input. Moreover, in the practical comparator circuit, the hysteresis could further stabilize the equilibrium point.

4 Encoding Scheme

4.1 Specifications

We can use the spikes generated by the circuit to determine the speech events by spike density. The density of the spike train is determined by the number of spikes that occur during a time window T_0. A low value of T_0 makes the decision latency shorter, while a large value makes the decision more robust to sudden interference and decreases the required spike encoding resolution. Through empirical testing, we chose $T_0 = 10$ ms. The onset/offset decision process is shown in Fig. 4. Since the observation time interval is fixed, the spike density is equivalent to the number of spikes N_d that occur within the T_0 time window.

For the spike density to carry useful information, the number of spikes that occurs within the T_0 time window should be able to vary over an appropriately large range. In particular, the number of spikes should be able to fall low enough to indicate a reduction in signal amplitude (an offset), and it should be able to rise high enough to indicate a rise in signal amplitude (an onset). For example, a reasonable choice for the offset threshold N_{off} is 1. The onset threshold N_{on} should be reasonably larger than N_{off}, so we could choose it as 4. N_{on} is strongly correlated to the threshold of the envelope change that can be considered as an onset, which will be discussed below.

In addition, we need also to ensure the spike density for the amplitude dropping higher than N_{off}. The decay time constant τ_d should make sure the circuit can generate at least $N_{\text{off}} + 1$ spikes in T_0 when the signal's amplitude falls to avoid wrong offset detection. Because the falling time when the signal drops to 5% of the initial value is about $3\tau_d$, τ_d needs to satisfy:

$$\tau_d < \frac{T_0}{3(N_{\text{off}} + 1)} \tag{19}$$

to make $N_{\text{off}} + 1$ spikes generated in T_0 even when the signal drops. For example, if $N_{\text{off}} = 1$ and $T_0 = 10$ ms, $\tau_d < 1.7$ ms. In addition, the decay should also be slower than 1/4 cycle of the slowest input speech signal, otherwise the threshold would decay fast, making it not able to track the signal's envelope. So, we have:

$$\tau_d > \frac{1}{3(4f_{\text{min}})} = \frac{1}{12f_{\text{min}}} \tag{20}$$

For the minimum input frequency is 100 Hz, we have $\tau_d > 0.8$ ms. In sum, we could choose $\tau_d = 1$ ms for $N_{\text{off}} = 1$ and $T_0 = 10$ ms.

The spike train density increases both as a function of signal envelope and as a function of signal envelope change. However, we are particularly interested in the relationship between spike density and envelope voltage rising, so that we can select a suitable N_{on} to decide onset. The spike density only needs to be higher than N_{off} between onset and offset, so that it is more flexible and is ensured by selecting a suitable τ_d. Therefore, we focus on the encoding transfer function between the input envelope voltage rising and output

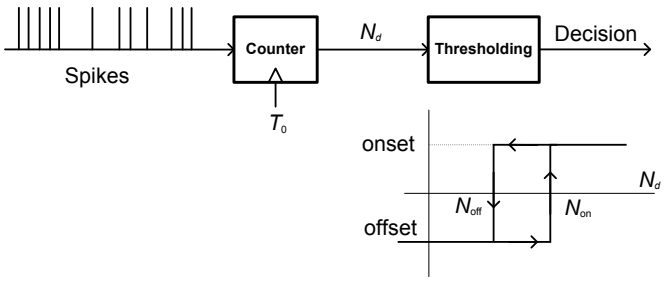

Fig. 4 Onset/offset decision process based on spike density in a time window T_0. The spikes are at first sent to the counter with reset clock T_0. In each time window T_0, the counted spike number is thresholded with dual thresholds N_{on} and N_{off} to determine onset and offset.

spike number, both observed in time window T_0. We express the input-output transfer function of the spike encoding circuit as:

$$\frac{N_d}{T_0} = f\left(\frac{V_e}{T_0}\right) \approx f(\dot{e}(t)) \tag{21}$$

where $e(t)$ is the envelope of the amplitude, which can be approximately expressed as $\Delta e/\Delta t$ if Δt is small enough. Here, we set $\Delta e = V_e$ and $\Delta t = T_0$. As both N_d and V_e are observed in the time window of T_0, we just need to find the following expression:

$$N_d = f_s\left(V_e\right) \tag{22}$$

Based on (22), if we already choose the envelope change threshold as pV_a, we could set proper circuit parameters to make $N_{\text{on}} = f_s\left(pV_a\right)$.

4.2 Adaptive Encoding Resolution

Although we have the circuit model (3) and schematic in Fig. 3, the nonlinearity of the circuit makes it difficult to gain intuition about the encoding performance, especially the feature of adaptive encoding resolution. The desired encoding transfer function (22) is difficult to explicitly express as well.

Therefore, we develop a simplified linear model for the spike encoding circuit. We assume that, within T_0, the input can be considered a smooth linear function:

$$e(t) = kt = \frac{V_e}{T_0}t \tag{23}$$

where V_e is the envelope voltage in T_0. According to Fig. 5, after threshold jumps by V_0, the decaying function $x(t)$ can be described by the differential equation:

$$\tau_d \dot{x}(t) + x(t) = e(t) \tag{24}$$

with initial condition $x(0) = V_0$. Solving this equation, we get:

$$x(t) = (V_0 + k\tau_d)e^{-t/\tau_d} + kt - k\tau_d \tag{25}$$

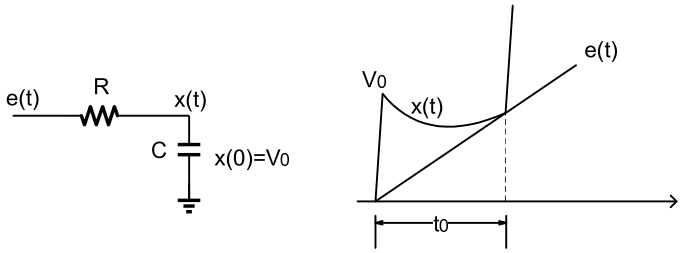

Fig. 5 The simplified circuit model for threshold decaying (left) and the threshold variation for smooth envelope (right), which neglects the fluctuations of the envelope and the hysteresis of the comparator. Here, $x(t)$ is the threshold and $e(t)$ is the linearized smooth envelope. Every time when the threshold reaches the envelope, it will jump again and also generate a spike, so that the time that the threshold uses to reach the envelope back is the spiking period, t_0.

The time t_0 the threshold takes to reach back the signal is the solution to the equation $x(t) = e(t)$, and it can be expressed as:

$$t_0 = \tau_d \ln \frac{V_0 + k\tau_d}{k\tau_d} = \tau_d \ln \left(1 + \frac{V_0 T_0}{V_e \tau_d}\right) \tag{26}$$

From this solution, we can get the spike number generated in T_0 as:

$$N_d = \left\lfloor \frac{T_0}{t_0} \right\rfloor \tag{27}$$

where $\lfloor \cdot \rfloor$ represents the rounded integer towards minus infinity. It can be further expressed as the function of V_e as:

$$N_d = f_s(V_e) = \left\lfloor \frac{T_0}{\tau_d \ln\left[1 + (V_0 T_0)/(V_e \tau_d)\right]} \right\rfloor \tag{28}$$

The encoding performance is related to T_0, τ_d and V_0. We choose T_0 and τ_d by the reasons described in Section 4.1. So, the critical parameter to determine the encoding transfer function is V_0.

In Fig. 3, the rate of change of the voltage on the capacitor is approximately I_c/C. As the comparator needs τ to shut off the switch and discontinue the charging, the jump step size of the threshold can be expressed as:

$$V_0 = \frac{I_c \tau}{C} \tag{29}$$

The rate of voltage change, I_c/C, should be fast enough to make the jump exceed the input signal quickly. In a short time Δt, we can approximately express the input signal's change as:

$$\Delta v = V_a \sin(2\pi f \Delta t) \approx V_a 2\pi f \Delta t \tag{30}$$

and we need to make:

$$\frac{I_c \Delta t}{C} > V_a 2\pi f \Delta t \tag{31}$$

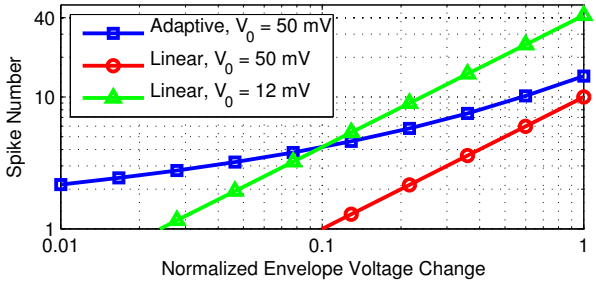

Fig. 6 Input envelope voltage change and unrounded output spike number in T_0 for adaptive encoding and linear encoding with different threshold step sizes. The input envelope voltage change is normalized with maximum input voltage 500 mV. The other parameters used for all the plots are $T_0 = 10$ ms and $\tau_d = 1$ ms.

Assuming a maximum voice band frequency of 3.4 kHz, so that we can calculate the rate of voltage change $I_c/C > 10.7$ kV/s. Because I_c/C is selected by this constraint, V_0 is then proportional to τ according to (29).

This transfer function (28) is plotted in Fig. 6 with $V_0 = 50$ mV. We find that the input-output relationship is nonlinear. The encoding is more accurate for small envelope changes while it is coarser for larger envelope changes. This means we can use smaller envelope change to generate enough spikes to trigger onset, so that the detection sensitivity is increased.

For comparison, we investigate the linear encoding at first. Assuming that τ_d is infinite so that it can hold V_0 until the threshold reaches the signal, the time t_0' used for the fast jump to reach back the signal can be estimated by the solution of following function:

$$V_0 = kt \tag{32}$$

so that $t_0' = V_0/k$. Using a similar approach as above,, we can get the encoding input-output transfer function as:

$$N_d' = f_s'(V_e) = \left\lfloor \frac{V_e}{V_0} \right\rfloor \tag{33}$$

which is a linear function if we do not take $\lfloor \cdot \rfloor$ into account. The transfer functions for the linear encoding are also shown in Fig. 6. We find that the linear encoding scheme has a lower resolution with the same V_0. If we set the onset threshold to $0.1V_a$, then the adaptive encoding scheme makes corresponding spike number $N_{\text{on}} = 4$, while linear encoding makes N_{on} less than 1. We need to make $V_0 = 15$ mV to have $N_{\text{on}} = 4$. The adaptive encoding allows that we use a larger V_0 for the same N_{on}, so that we can use a smaller τ, which reduces the comparator's speed requirement and therefore power consumption.

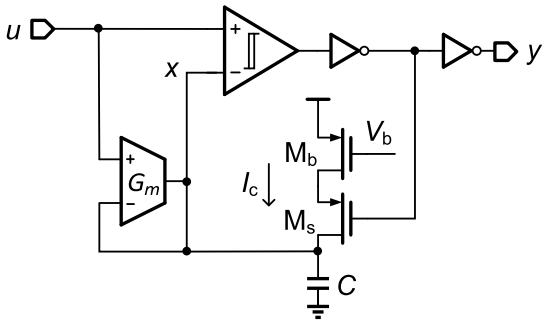

Fig. 7 Circuit implementation of the spike encoding circuit corresponding to the circuit model in Fig. 3, where the resistor is implemented by a negative feedback transconductor G_m and the current source and switch are implemented by PMOS transistors. As the switch is PMOS, we use the inverted output of the comparator to control the switch, but the state-space model and related conclusions in Section 3 will not change.

5 Prototype and Results

5.1 Prototype

We fabricated the spike encoding circuit in a 0.5-μm CMOS process. The circuit schematic is shown in Fig. 7. We use C/G_m to implement the decaying RC time constant in Fig. 3. The decaying time constant $\tau_d = 1$ ms, and we choose $C = 8$ pF and $G_m = 8$ nS.

The comparator is a three-stage amplifier with rail-to-rail input, whose schematic is shown in Fig. 8. The rail-to-rail input gives the comparator a relatively constant delay τ during the whole input amplitude range. Due to the logarithmic function in encoding transfer function (28), a small fluctuation in τ may induce considerable change of the encoding resolution, so that we need to make τ stable. The comparator has 12-mV hysteresis range by positive feedback transistors in input stage. The hysteresis makes the circuit immune to small noise signal and interference, so that few spikes are generated when speech events are absent. The delay of the comparator (and the inverter) is 12 μs.

The current source implemented by the PMOS provides 90-nA current and the transistor length L is 20 μm for high output resistance. We can calculate the rate of voltage change $I_c/C = 11.3$ kV/s and the threshold step voltage $V_0 = I_c\tau/C = 135$ mV in this design.

The micrography of the spike encoding circuit is shown in Fig. 9. Its die size is 265 μm $\times$ 105 μm.

5.2 Encoding Performance

First, we investigate the encoding performance for an ideal AM signal. Fig. 10 shows the spike encoding circuit's response to an ideal AM burst signal. By

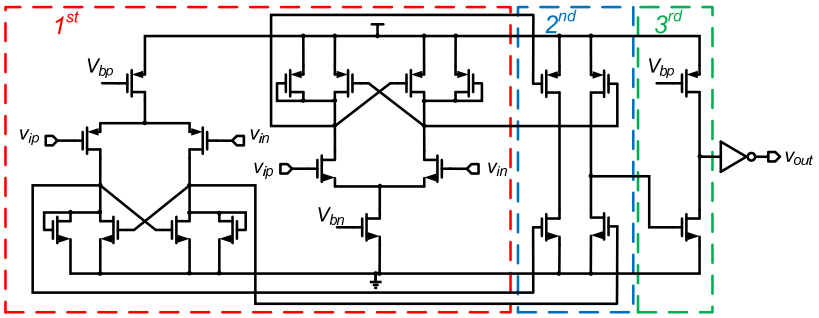

Fig. 8 Comparator schematic. The input stage is rail-to-rail and provides hysteresis, the second stage is a push-pull amplifier to interface the first stage and the common-source third stage. The output is driven by an inverter.

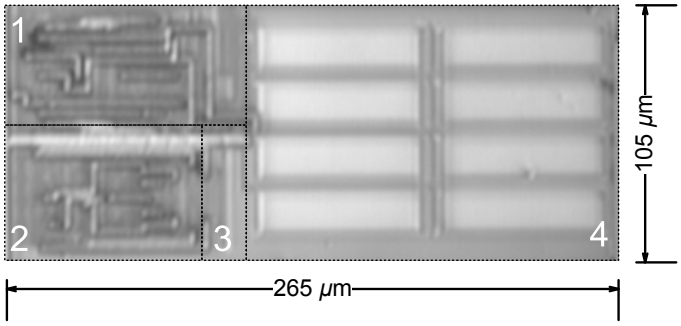

Fig. 9 Micrograph of the spike encoding circuit, where 1 is the transconductor, 2 is the comparator, 3 is the switched current source, and 4 is the poly-to-poly capacitor.

setting $T_0 = 10$ ms, $N_{\text{on}} = 4$ and $N_{\text{off}} = 1$, we can detect the onset and offset at 20 ms and 110 ms respectively. The modulation signal is a sinusoid, as is the envelope of the signal. So, envelope changes fastest when the signal's amplitude begins to rise, so the spike density should be highest around the onset (20 ms). However, the spike density of the measurement results become highest bwteen 20-30 ms, which is slightly later than the onset point. This is because the hysteresis in the comparator suppresses the circuit's sensitivity to small noise and interference. It also decrease the spike density for the initial onset process. So, the hysteresis range can be used to control the sensitivity and the noise robustness of the circuit. In this design, the spike density can still become highest right after the amplitude exceeds the hysteresis range.

Fig. 11 shows the input-output transfer function of the encoding circuit for both the envelope rising and falling. The effective encoding range is defined as the range that the output spike number monotonously increases with the increase of envelope voltage change, so the input dynamic range is 34 dB, covering the telephony quality speech dynamic range.

As predicted in Section 4.2, the spike encoding circuit shows higher resolution for smaller envelope rising, so that the encoding scheme is adaptive.

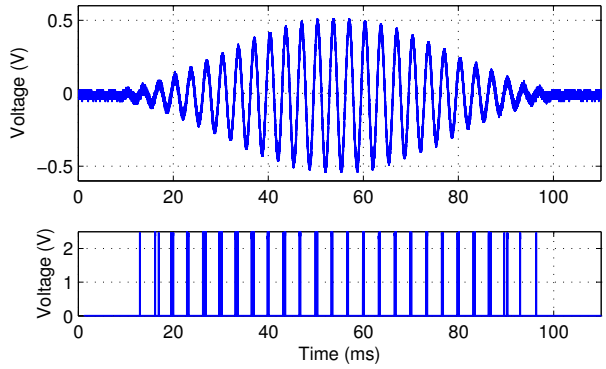

Fig. 10 An AM burst signal (top) and the corresponding spikes (bottom). If we set $T_0 = 10$ ms, $N_{\text{on}} = 4$, the onset is detected at 20 ms; if we set $N_{\text{off}} = 1$, the offset is detected at 110 ms. The thickness of the spikes indicate the spike density.

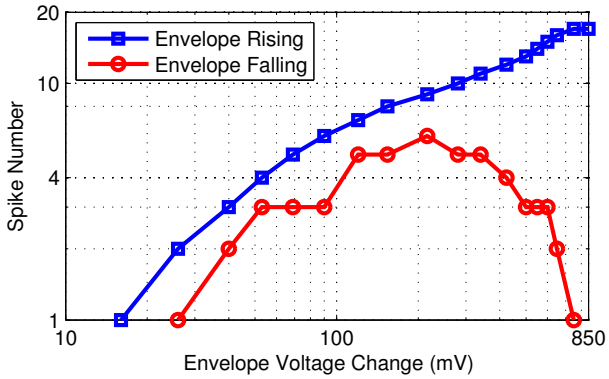

Fig. 11 Spike number and envelope voltage change both in 10-ms time window. Envelope rising and falling cases are both recorded. The carrier frequency for testing is 4 kHz. If we choose $N_{\text{on}} = 4$ and $N_{\text{off}} = 1$, then the envelope rising of more than 50 mV can trigger an onset. Also, spurious offsets are not detected for amplitude changes in the range of 50-600 mV, since the spike number for the falling envelope is greater than N_{off}, as discussed in Section 4.1.

The output spike number range is 17, or 25 dB, which is compressed by the adaptive encoding resolution. Because we just need high resolution for small envelope change to detect the onset easily, we do not need to waste power consumption to get the same resolution for high envelope change.

As discussed in [15], the speech phoneme modulation band is 2-12 Hz, which is useful for differentiating speech from non-speech signals. This is achieved by a bandpass filter in [14], but the low-cutoff-frequency filter will induce considerable latency of the decision, which is not desirable in real-time applications. The encoding spike number versus modulation frequency of the proposed spike encoding circuit is shown in Fig. 12. The modulation frequency range that can

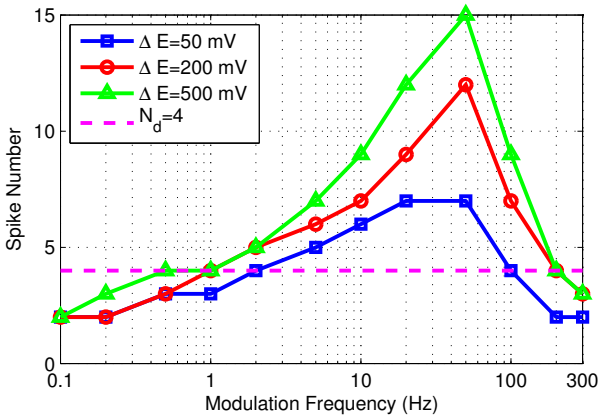

Fig. 12 Spike number in 10-ms time window versus the frequency of the AM signals at different amplitude level (ΔE), based on the speech phoneme model in [15]. If $N_{\text{on}} = 4$, the modulation frequency that can produce an onset is about 1-200 Hz.

Table 1 Encoding Performance Summary

Power Supply (V)	2.5
Input Dynamic Range (dB)	34
Output Dynamic Range (dB)	25
Smallest Detectable Rate of Envelope Change (V/s)	1.5
Input Frequency Range (Hz)	100-3400
Time Constant (ms)	1
Power Consumption (μW)	0.3
Die Size (μm^2)	265 × 105
Process	0.5-μm 2P3M CMOS

produce an onset is about 1-200 Hz, which is wider than that of the practical speech. We could use a larger decision time window T_0 to decrease the detected modulation frequency range, but the decision latency would be longer and the performance adaptation may fail. Even with this wider detected modulation frequency range, the circuit is still robust to very slow or fast background noise.

The power consumption of the circuit is 300 nW under 2.5-V power supply. The performance summary is shown in Table. 1.

5.3 Experiment with Speech Sample

Along with the spike encoding circuit, we also fabricated a bandpass filter with a tunable center frequency and bandwidth on the same chip. As shown in Fig. 13, we tuned the filter's transfer function to match each channel in a 16-channel filter bank with the center frequencies logarithmically spaced to cover the frequency range of 100 Hz to 4 kHz. We separated the speech sample in Fig. 1 into 16 constituent frequency components. Next, each signal

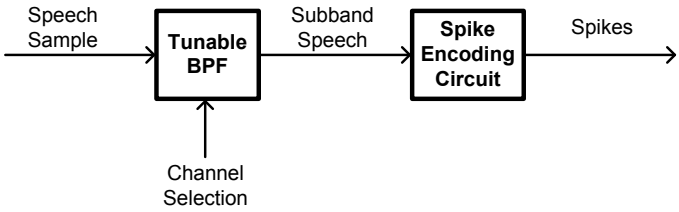

Fig. 13 Block diagram of the multi-channel speech edge detection system based on a tunable bandpass filter and a spike encoding circuit. The speech sample is sent to the bandwidth tunable bandpass filter to obtain the subband speech signal, and then the spikes are generated for each band by spike encoding circuit. The subband signals and corresponding spikes are stored for analysis.

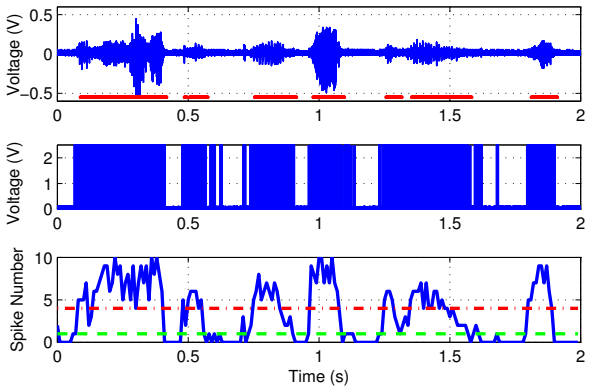

Fig. 14 Measurement results of a bandpass speech signal centered at 3.2 kHz (the highest frequency channel). The top is the input signal with the identified speech event by spike density shown as the straight lines below the waveform, the middle is the corresponding spike train and the bottom is the spike number counted in 10-ms time window with the two dashed lines indicating the onset and offset thresholds.

was processed by the spike encoding circuit. This way, we obtained the 16-channel bandpass speech signals and their corresponding spikes. Finally, we counted the spike number over a moving time window and thresholded it using $N_{\text{on}} = 4$ and $N_{\text{off}} = 1$ to get the speech event edges (onset/offset points). For example, the measurement results for the 16th channel are shown in Fig. 14.

For all 16 channels, we selected the maximum frequency channel with detected speech events at each time point, so that we obtained the speech edge detection results plotted in Fig. 1. We find the extracted speech can track the speech edge change accurately.

Finally, we added white noise to the speech sample to test the performance of spike encoding circuit. With an average SNR of 20 dB, the measurement results of the speech edge using the similar methods described above are plotted in Fig. 15, along with the results for the clean speech shown in Fig. 1. We find that the accuracy of the detected speech edge is degraded due to the presence

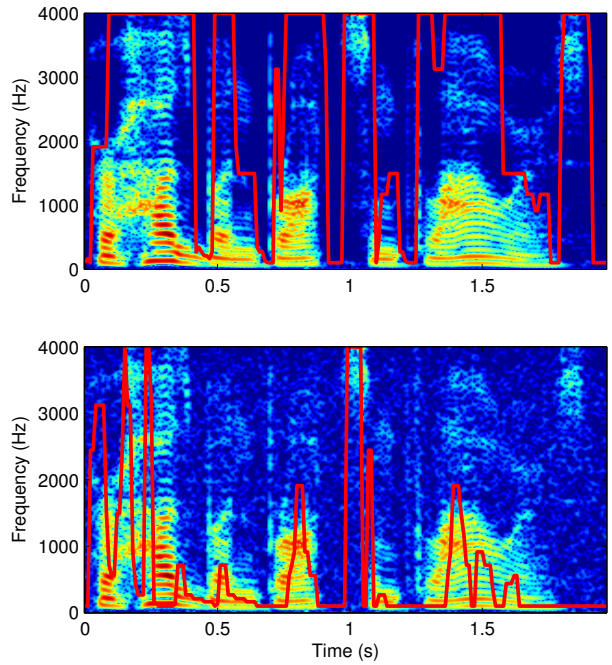

Fig. 15 The edge detection result for clean speech (top) and noisy speech with 20-dB SNR (bottom) for the speech sample in Fig. 1.

of white noise, but the speech edge still follows the trend of the ideal speech edge.

6 Conclusion

The proposed spike encoding circuit can identify the edges of speech events with self-adaptive resolution. It has small die size and power consumption, so it is promising to as an embedded component in smart audio sensors to identify speech event edge and eventually save the power consumption of the system.

In order to maintain a low latency, the spike encoding scheme has a wider detection band than the actual modulation band of speech phonemes, which means the robustness to background noise is low. As shown in Fig. 15, the speech edge detection accuracy for the speech with 20-dB SNR is degraded compared to that of the clean speech. Our future work will include investigating the quantitative relationship between latency and noise robustness, as well as the relationship between latency and speech edge detection's accuracy [7]. By understanding these relationships at the algorithm level, we can set the circuit parameters to achieve a good balance between latency and noise robustness.

The circuit is designed using a nonlinear dynamical approach, which provides us the freedom to synthesize the circuit directly from the function we desire. This has resulted in a the synthesized circuit that is highly efficient. This design methodology would be useful for other energy- and hardware-efficient circuits.

Acknowledgment

The authors would like to thank the Neukom Institute at Dartmouth College for research funding and MOSIS for chip fabrication.

References

1. E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, Feb. 2006.
2. L. S. Smith and D. S. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," *IEEE Trans. Neural Networks*, vol. 15, pp. 396–405, Sept. 2004.
3. O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109130, 1986.
4. Y. Ren and J. M. Johnson, "Auditory coding based speech enhancement," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4685–4688.
5. I. Uysal, H. Sathyendra and J. G. Harris, "Spike-based feature extraction for noise robust speech recognition using phase synchrony coding," *Proc. IEEE Int. Conf. Circuits and Systems (ISCAS)*, 2007, pp. 1529–1532.
6. H. Mino, "Encoding of information into neural spike trains in an auditory nerve fiber model with electric stimuli in the presence of a pseudospontaneous activity," *IEEE Trans. Biomed. Eng.*, vol. 54, pp. 360–369, Mar. 2007.
7. D. Du and K. Odame, "Efficient speech edge detection for intelligent audio sensors," *in preparation.*
8. G. Hu and D. Wang, "Auditory Segmentation Based on Onset and Offset Analysis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 396–405, Feb. 2007.
9. Y. Tsividis, "Event-driven data acquisition and digital signal processinga tutorial", *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 57, pp. 577–581, Aug. 2010.
10. Y. Tsividis, "Event-driven data acquisition and continuous-time digital signal processing", *IEEE Custom Integrated Circuits Conference (CICC)*, 2010, pp.1–8.
11. Y. Tsividis, "Event-driven, continuous-time ADCs and DSPs for adapting power dissipation to signal activity", *Proc. IEEE Int. Conf. Circuits and Systems (ISCAS)*, Aug. 2010, pp. 3581–3584.
12. G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. Neural Netw.*, vol. 17, pp. 211221, Jan. 2006.
13. L. C. Gouveia, T. J. Koickal and A. Hamilton, "An asynchronous spike event coding scheme for programmable analog arrays," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, pp. 791–799, Apr. 2011.
14. T. Delbruck, T. Koch, R. Berner and H. Hermansky, "Fully integrated 500μW speech detection wake-up circuit," *Proc. IEEE Int. Conf. Circuits and Systems (ISCAS)*, 2010, pp. 2015–2018.
15. M. C. Buchler, "Algorithms for sound classification in hearing instruments," Ph.D Dissertation, ETH Zurich, Zurich, Switzerland, 2002.
16. D. Goodman, "Delta modulation granular quantizing noise," *Bell Syst. Tech. J.*, vol. 48, pp. 1197-1218, May-Jun., 1969.