A Bio-Inspired Ultra-Low-Power Spike Encoding Circuit for Speech Edge Detection

Dingkun Du and Kofi Odame
Thayer School of Engineering, Dartmouth College
Hanover, NH 03755, USA
Email: {dingkun.du, odame}@dartmouth.edu

***Abstract*—Speech edge detection can be used to adaptively control the performance and improve the energy efficiency of smart audio sensors. In this paper, we propose a spike encoding circuit inspired by mammalian auditory system for real-time and low-power speech edge detection. The circuit can directly encode the signal's envelope information by asynchronous spikes' temporal density without additional envelope extraction. Furthermore, the spike encoding automatically adapts its encoding resolution to the amplitude of the input signal, which improves encoding resolution for small signal without directly increasing power consumption. Fabricated in 0.5-μm CMOS process, the spike encoding circuit consumes 0.3-μW power, and the experimental results are presented.**

I. Introduction

In the mammalian auditory system, incoming sound is converted into a series of auditory nerve firing patterns, or spikes. This spike encoding has evolved to specifically enhance certain features. 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 priming a response in the mammal [1]. Also, the representation of sounds of interest (e.g. speech) is particularly invariant, even in the presence of loud, interfering noise [2].

The low latency and robust performance of this biological system has inspired several speech processing algorithms that are based on spike encoding [1] – [4]. Likewise, our recently-developed *speech edge detection* algorithm [5] relies on spike encoding. Current implementations of spike encoding are usually power consumptive. For example, in [1], the spikes are generated by a series of comparators with different threshold voltages; in [6], we need an additional envelope detector for the input signal; in [7], the circuit encodes the full signal waveform instead of the envelope, so power would be wasted in the specific application focusing on the envelope.

In this paper, we introduce a novel spike encoding circuit to represent the signal's envelope information by temporal spike density. It uses only one comparator with variable threshold to generate spikes and does not need additional envelope extraction. Moreover, its encoding resolution is adaptive (finer for smaller envelope change), which is equivalent to the variable-gain capability to make small signal easier to capture. Integrated all the functions together, this circuit only consumes 0.3 μW and occupies 0.028 mm^2 in 0.5-μm CMOS process.

II. Background

We have developed a real-time algorithm [5] for detecting the edges of speech in the time-frequency plane (Fig. 1). 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.

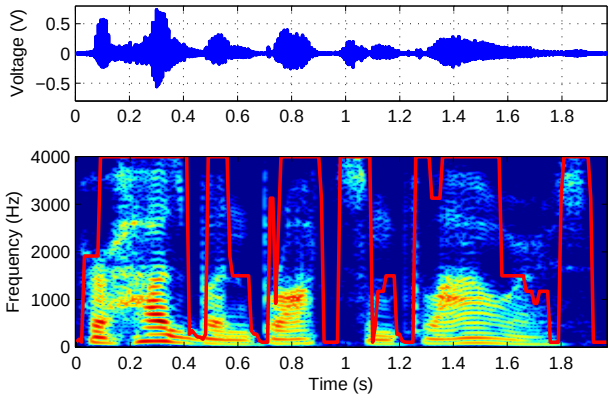

Fig. 1. A speech sample (top) and its spectrogram (bottom), with the extracted speech edge (piecewise line) shown on the spectrogram. The speech edge comes from the measurement results of the proposed spike encoding circuit. The measurement details are illustrated in Section IV-C.

Given its application in a speech edge detection algorithm, the spike encoding will be applied to each of several frequency channels, each encoding circuit must be area and power efficient. In addition, the spike encoding must robustly indicate the edges of each speech component. That is, the spiking pattern must highlight the temporal onset and offset of each speech component. Also, the spiking pattern must highlight the highest and lowest spectral components of the speech at any given time. We need to design a spike encoding circuit that can output a train of spikes with varying density. Specifically, the spike train density should increase both as a function of signal amplitude and as a function of signal amplitude change.

In the time dimension, there will be a high/low density of spikes during speech onsets/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).

III. Encoding Scheme

A. Encoding Circuit

Our spike encoding circuit is based on a comparator, whose input is the audio signal for a given frequency band. Whenever the comparator detects an input that exceeds a threshold

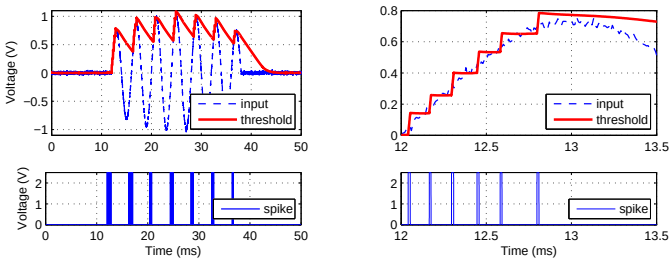

Fig. 2. Amplitude-modulated (AM) signal (the speech phoneme model in [6]), the variable threshold and the corresponding generated spikes (left). The zoomed view around 13 ms (right). On the left panel, the spike thickness indicates the temporal density due to the limited resolution of the figure.

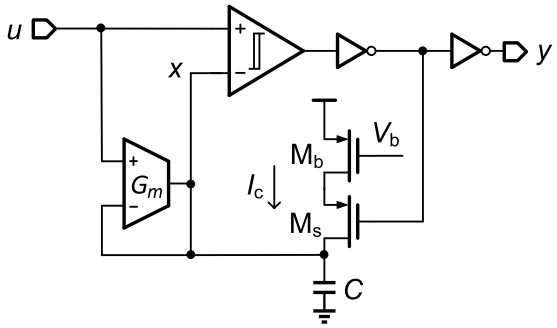

Fig. 3. Spike encoding circuit schematic. The input is u and the variable threshold is x. The G_m-C filter causes the variable threshold to adapt towards the input. The PMOS transistors form a charge pump that is activated whenever a spike, y, is generated.

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.

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.

The schematic for our spike encoding circuit is depicted in Fig. 3. The dimensionless normalized state-space model for this circuit is:

$$\begin{aligned} \dot{x} &= \frac{(y+1)I_c}{2V_aC} + \frac{G_m(u-x)}{C} \\ \dot{y} &= \frac{\mathrm{sgn}(u-x)-y}{\tau} \end{aligned} \tag{1}$$

where u, x and y denote the input signal, the threshold and the output spikes respectively, and all of them are normalized in the interval $[-1, 1]$. Also, $\mathrm{sgn}(\cdot)$ is the sigmoidal function, I_c is the value of the current source, G_m is the transconductance, C is the capacitance, τ is the delay due to the comparator and the inverter after it, and V_a is highest input amplitude in order to normalize the expression for $\dot{x}$.

In this circuit, if u is higher than x, the output of the comparator will go high, which will activate the switch of the charge pump. The charge pump will then quickly charge the capacitor and make x jump to a higher level. Typically, x will jump higher than u, forcing the output of the comparator low, which will in turn shut off the charge pump. Then, x will start to decay towards u with a time constant of $\tau_d = C/G_m$. The entire process will repeat itself when u gets higher than x again.

B. Encoding 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.

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 able to rise high enough to indicate a rise in signal amplitude (an onset).

For $T_0 = 10$ ms, we have found that a reasonable number of spikes to represent low spike density is $N_{\text{off}} = 1$, while a reasonable number of spikes to represent high spike density is $N_{\text{on}} = 4$. We also need to ensure that the spike density remains higher than N_{off} when there is no offset, but the signal is experiencing a drop in amplitude. The decay time constant τ_d must be chosen such that the circuit can generate at least $N_{\text{off}} + 1$ spikes, when there is no offset. For the parameters we have picked, a value of $\tau_d = 1$ ms works well.

C. Adaptive Encoding

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 spike number, both observed in time window T_0.

Although we have the circuit model (1) and schematic in Fig. 3, the nonlinearity of the circuit makes us difficult to gain intuition of the encoding performance, especially the feature of adaptive encoding resolution, directly. 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

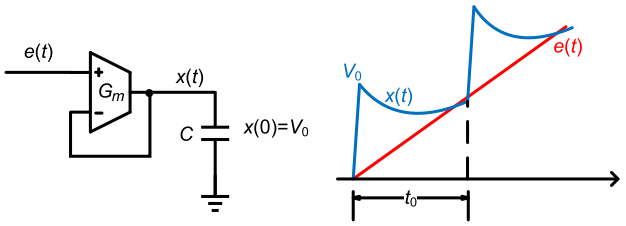

Fig. 4. The simplified circuit model for threshold decaying (left) and the threshold variation for smooth envelope (right), which neglects the fluctuations of the envelope and the hysteresis of the comparator.

linear function as $e(t) = kt = V_e t/T_0$ mimicking an ideal envelope, where V_e is the envelope voltage in T_0. According to Fig. 4, after threshold jumps by V_0, the decaying function $x(t)$ can be described by:

$$\tau_d \dot{x}(t) + x(t) = e(t) \tag{2}$$

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

$$x(t) = (V_0 + k\tau_d)e^{-t/\tau_d} + kt - k\tau_d \tag{3}$$

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

$$t_0 = \tau_d \ln \frac{V_0 + k\tau_d}{k\tau_d} = \tau_d \ln\left(1 + \frac{V_0 T_0}{V_e \tau_d}\right) \tag{4}$$

So, we can get the spike number generated in T_0 as:

$$N_d = f_s(V_e) = \left\lfloor \frac{T_0}{\tau_d \ln\left[1 + (V_0 T_0)/(V_e \tau_d)\right]} \right\rfloor \tag{5}$$

where $\lfloor \cdot \rfloor$ represents the rounded integer towards minus infinity. 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 III-B. 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 $V_0 = I_c\tau/C$. The rate of voltage change, I_c/C, should be fast enough to make the jump exceed the input signal quickly, so that V_0 is proportional to τ when I_c/C is selected by this constraint.

This transfer function (5) is plotted in Fig. 5 with $V_0 = 50$ mV. We can 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, the transfer functions for the linear encoding are also shown in Fig. 5. 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 $V_0 = 12$ mV to have $N_{\text{on}} = 4$. The adaptive encoding allows that we use a larger V_0 for same N_{on}, so that we can use a smaller τ, which reduces the comparator's speed requirement and therefore power consumption.

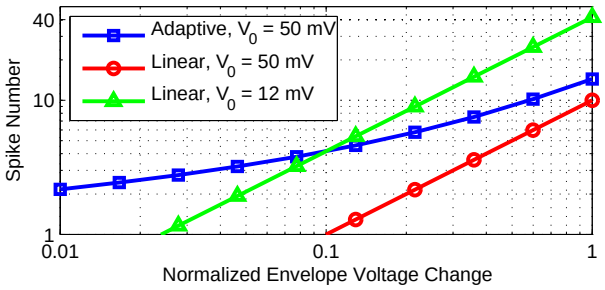

Fig. 5. Input envelope voltage change and unrounded output spike number in T_0 for adaptive encoding and linear encoding with different threshold step sizes. The input envelope voltage change is normalized with $V_a = 500$ mV. The other parameters used for all the plots are $T_0 = 10$ ms and $\tau_d = 1$ ms.

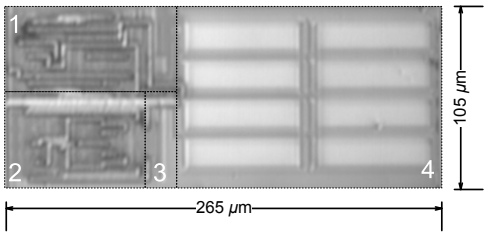

Fig. 6. Micrograph of the spike encoding circuit, where 1 is the transconductor, 2 is the comparator, 3 is the charge pump, and 4 is the poly-to-poly capacitor.

IV. Prototype and Results

A. Implementation

We use 0.5-μm CMOS process to fabricate a spike encoding circuit. The micrography of the spike encoding circuit is shown in Fig. 6. Its die size is $265\ \mu\text{m} \times 105\ \mu\text{m}$.

The comparator is a three-stage openloop amplifier. The schematic is shown in Fig. 7. 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 (5), a small fluctuation in τ may induce considerable change of the encoding resolution, so that we need to make τ stable. Also, it 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 there would be few spikes generated when speech events are absent. The delay of the comparator (and the inverter) is $12\ \mu$s.

The decaying time constant $\tau_d = C/G_m = 1$ ms, and we choose $C = 8$ pF and $G_m = 8$ nS. The current source implemented by the PMOS provides 90-nA current and the transistor length L is $20\ \mu$m for high output resistance. We can calculate $V_0 = I_c\tau/C = 135$ mV in this design.

B. Encoding Performance

Fig. 8 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.

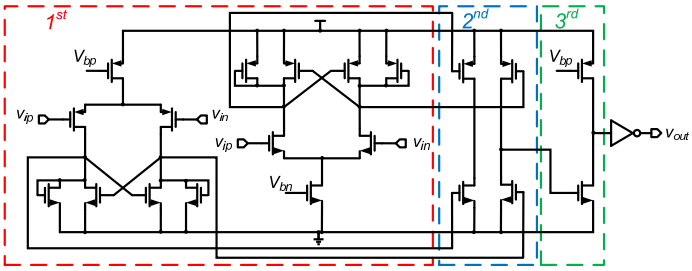

Fig. 7. Comparator schematic. The input stage is rail-to-rail and provides hysteresis, the second stage is a push-pull amplifier to interface the first stage and the common-source third stage. The output is driven by an inverter.

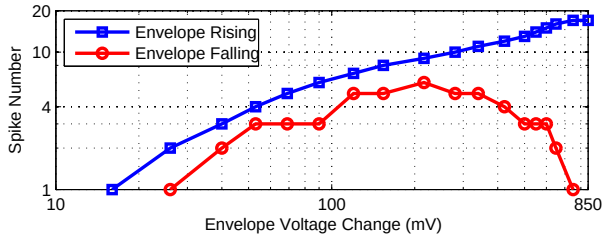

Fig. 8. Spike number and envelope voltage change both in 10-ms time window. Envelope rising and falling cases are both recorded. The carrier frequency for testing is 4 kHz. If we choose $N_{\rm on} = 4$ and $N_{\rm off} = 1$, then the envelope rising of more than 50 mV can trigger an onset. Also, spurious 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 $N_{\rm off}$, as discussed in Section III-B.

As predicted in Section III-C, the spike encoding circuit shows higher resolution for smaller envelope rising, so that the encoding scheme is adaptive. 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 the full range of envelope change.

The power consumption of the circuit is 0.3 μW under 2.5-V power supply. The performance summary is shown in Table. I.

C. Experiment for Practical 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. We tuned the filter's transfer function to match each channel in a 16-channel filter bank. We separated the speech sample in Fig. 1 into 16 constituent frequency components. Next, each signal 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 and thresholded it using $N_{\rm on} = 4$ and $N_{\rm off} = 1$ to get the speech event edges (onset/offset points). For example, the measurement results for the 16th channel are shown in Fig. 9. The results for all 16 channels are plotted in Fig. 1.

TABLE I
ENCODING PERFORMANCE SUMMARY

Power Supply (V)	2.5
Input Dynamic Range (dB)	34
Output Dynamic Range (dB)	25
Decision Time Window (ms)	10
Input Frequency Range (Hz)	100-3400
Power Consumption (μW)	0.3
Die Size (μm^2)	265×105
Process	0.5-μm 2P3M CMOS

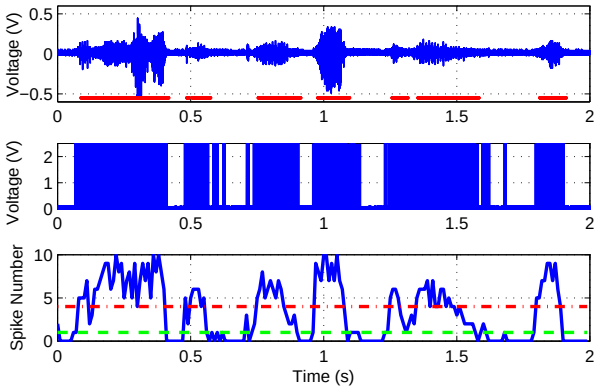

Fig. 9. Measurement results of a bandpass speech signal centered at 3.2 kHz (the highest frequency channel). The top is the input signal with the identified speech event by spike density shown as the straight lines below the waveform, the middle is the corresponding spike train and the bottom is the spike number counted in 10-ms time window with the two dashed lines indicating the onset and offset thresholds.

V. 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 be embedded in smart audio sensors to identify speech event edge and eventually save the power consumption of the system.

ACKNOWLEDGMENT

The authors would like to thank the Neukom Instuitute at Dartmouth College for research support.

REFERENCES

[1] 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.

[2] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109–130, 1986.

[3] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, Feb. 2006.

[4] 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.

[5] D. Du and K. Odame, "Efficient speech edge detection for mobile health applications," *Proc. IEEE Int. Conf. Biomedical Circuits and Systems (BioCAS)*, 2011, *to appear*.

[6] 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.

[7] 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.