21.2 A 1μW Voice Activity Detector Using Analog Feature Extraction and Digital Deep Neural Network

Minhao Yang, Chung-Heng Yeh, Yiyin Zhou, Joao P. Cerqueira, Aurel A. Lazar, Mingoo Seok

Columbia University, New York, NY

Voice user interfaces (UIs) are highly compelling for wearable and mobile devices. They have the advantage of using compact and ultra-low-power (ULP) input devices (e.g. passive microphones). Together with ULP signal acquisition and processing, voice UIs can give energy-harvesting acoustic sensor nodes and battery-operating devices the sought-after capability of natural interaction with humans. Voice activity detection (VAD), separating speech from background noise, is a key building block in such voice UIs, e.g. it can enable power gating of higher-level speech tasks such as speaker identification and speech recognition [1]. As an *always-on* block, the power consumption of VAD must be minimized and meanwhile maintain high classification accuracy. Motivated by high power efficiency of analog signal processing, a VAD system using analog feature extraction (AFE) and mixed-signal decision tree (DT) classifier was demonstrated in [2]. While it achieved a record of 6μW, the system requires machine-learning based calibration of the DT thresholds on a chip-to-chip basis due to ill-controlled AFE variation. Moreover, the 7-node DT may deliver inferior classification accuracy especially under low input SNR and difficult noise scenario, compared to more advanced classifiers like deep neural networks (DNNs) [1,3]. Although heavy computational load in conventional floating-point DNNs prevents their adoption in embedded systems, the binarized neural networks (BNNs) with binary weights and activations proposed in [4] may pave the way to ULP implementations. In this paper, we present a 1μW VAD system utilizing AFE and a digital BNN classifier with an event-encoding A/D interface. The whole AFE is 9.4× more power-efficient than the prior art [5] and 7.9× than the state-of-the-art digital filter bank [6], and the BNN consumes only 0.63μW. To avoid costly chip-wise training, a variation-aware python model of the AFE was created and the generated features were used for offline BNN training. Measurements show 84.4%/85.4% mean speech/non-speech hit rate with 1.88%/4.65% 1-σ standard deviation among 10 dies using the same weights for 10dB SNR speech with restaurant noise.

Figure 21.2.1 shows the system architecture. Audio signal from a microphone is amplified by a low noise amplifier (LNA), and then sent to 16 parallel channels. Each channel is composed of a bandpass filter (BPF), a full-wave rectifier (FWR), and an integrate-and-fire (IAF) event encoder. The central frequencies of the BPFs are geometrically scaled from about 100Hz to 5kHz. The IAF in each channel produces asynchronous events whose rate is roughly proportional to the signal energy of the respective band. The BNN input layer has 48 neurons, derived from the 16-channel AFE output. Three hidden layers respectively have 60, 24 and 11 neurons. The output layer consists of 2 neurons, with one's activation larger than the other's indicating voice, and smaller noise.

Figure 21.2.2 shows the capacitive LNA. The gain is programmable from 24dB to 42dB with a 6dB step via the input capacitor C_{in}. Current-reuse inverter-based input is used in the main amplifier to enhance noise efficiency. Existing designs employ two tail transistors, one supplying the bias current and the other giving CMFB. However four stacking transistors between supply and ground makes it challenging if not impossible to maintain saturation of all transistors over PVT under a low supply, which in turn may largely degrade the LNA closed-loop gain. This design eliminates one tail transistor, and sets the input DC voltage and bias current via a scaled replica of the input inverter in diode connection. C_c and R_c form the pseudo-cascode compensation for stability. With the load of 16 BPFs' input capacitance, sufficient phase margin (PM) requires large bias current in the second stage of the main amplifier, which is in conflict with microwatt system power budget. To solve this dilemma, positive feedback via C_f and R_f is used to boost 3dB bandwidth and therefore C_c can be increased for PM while keeping bias low. The input DC of the main amplifier is established by the DC-servo-loop (DSL) amplifier. The DSL amplifier is biased with pA current to give a high-pass corner frequency of the LNA smaller than 100Hz.

Figure 21.2.3 (upper) shows the super-source-follower-based 2nd-order BPF. Its central frequency f_0, quality factor Q, and peak gain A_0 are derived as:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}} \quad Q = \sqrt{\frac{g_{m2}C_2}{g_{m1}C_1}} \quad A_0 = \frac{C_2}{C_1}$$

where g_{m1} and g_{m2} are the transconductance of pFETs and nFETs, respectively. To mitigate the signal swing constraints aggravated by PVT variation, a diode-connected pFET replica with its source fixed at V_{refup} is used to provide the DC gate voltage of the input pFETs. Figure 21.2.3 (lower) shows the FWR and IAF. Buffered by source followers, the BPF output voltage is converted to current by a differential OTA, and then rectified by the cross-coupled precision current rectifier. The cross-coupled topology halves the output swing of the OTA compared to the single-ended version, essential for low supply voltage operation. To alleviate the dead-zone problem of the rectifier exacerbated by PVT variation, the gate voltages of the transistors are set by a single-ended replica biased at I_{leak}. The OTA is in DC closed-loop to avoid quiescent output offset current, and its output common-mode DC is set to V_{mid}, the same as the source voltages of the transistors in the rectifier replica. The rectified current is integrated on C_{int}, and whenever the integrated voltage crosses above V_{refdn}, an event is generated at the comparator output, and the integration starts over from ground potential.

Figure 21.2.4 shows the BNN implementation. The parallel event streams from the AFE are collected by ripple counters. Events are counted every 25ms frame with a 10ms frame shift, and each frame is stored in the DMEM, replacing values in one 16×9 block. Three DMEM blocks, including the current frame n, previous frames (n-3) and (n-6), compose the 48 input neurons to classify the frame (n-3). This technique of incorporating neighboring contextual information improves classification performance [3]. To compute the pre-activations of hidden neurons, the input operand of each accumulation is either the data directly from DMEM or the 84-bit register file (RF) that temporarily stores computed activations of the previous layer, or their negated values, depending on the 1-bit weight from WMEM. The activation function hard sigmoid HS(·) [4] is a simple negation of the sign bit of the pre-activations. The classification output of each frame is obtained by comparing the activations of the 2 output neurons without applying HS(·).

The chip was fabricated in 0.18μm CMOS with a core area of 1.66×1.52mm^2. Figure 21.2.7 shows the chip micrograph. Figure 21.2.5 (upper left) shows the LNA transfer function, together with the input-referred noise spectrum density at a gain of 42dB and 24dB. The respectively calculated noise efficiency factor (NEF) and power efficiency factor (PEF) at 0.6V are 1.73 and 1.80, and 4.39 and 11.6. Figure 21.2.5 (upper right) shows the event number counted every 25ms (a frame) at the output of IAFs of all 16 channels as the function of input frequency. For classification evaluation, 300 randomly selected clean utterances from AURORA4 dataset are concatenated [4] with duration of 37 minutes and mixed with DEMAND noise dataset for training, and another 300 with the non-speech period balanced lasting 1 hour in total for testing. Figure 21.2.5 (lower left and lower right) shows the speech/non-speech hit-rate testing points of 10dB SNR speech with restaurant noise and 5dB SNR speech with metro noise using the same weights respectively over 10 dies without any AFE calibration. Figure 21.2.6 shows the comparison of the AFE and the VAD system with prior works.

Acknowledgements:
This work was supported by Swiss National Science Foundation (SNF) Early Postdoc Mobility Fellowship and Columbia University Research Initiatives in Science and Engineering (RISE). The authors thank N. Mesgarani, Y. Tsividis, M. Verhelst, X.-L. Zhang for discussions and help.

References:
[1] M. Price, et al., "A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating," *ISSCC Dig. Tech. Papers*, pp. 244-245, Feb. 2017.
[2] K. Badami, et al., "Context-Aware Hierarchical Information-Sensing in a 6μW 90nm CMOS Voice Activity Detector," *ISSCC Dig. Tech. Papers*, pp. 430-431, Feb. 2015.
[3] X. -L. Zhang, et al., "Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 252-264, 2016.
[4] I. Hubara, et al., "Binarized Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1-9, 2016.
[5] M. Yang, et al., "A 0.5V 55μW 64×2-Channel Binaural Silicon Cochlea for Event-Driven Stereo-Audio Sensing," *ISSCC Dig. Tech. Papers*, pp. 388-389, Feb. 2016.
[6] H. -S. Wu, et al., "A 13.8μW Binaural Dual-Microphone Digital ANSI S1.11 Filter Bank for Hearing Aids with Zero-Short-Circuit-Current Logic in 65nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 348-349, Feb. 2017.

 978-1-5090-4940-0/18/$31.00

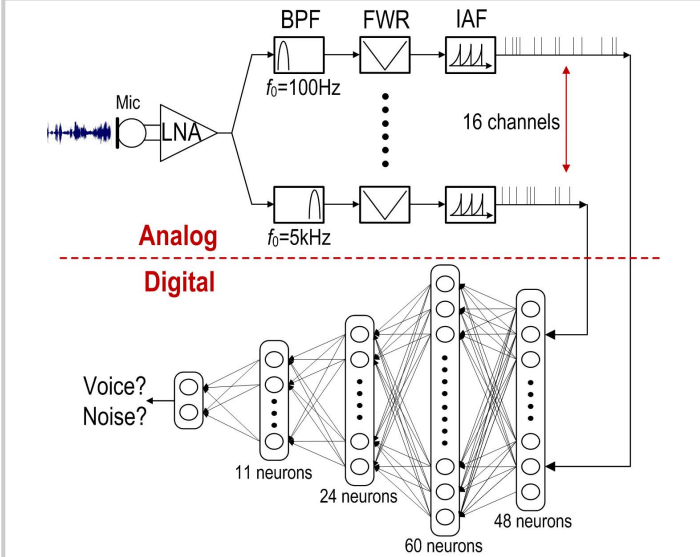

Figure 21.2.1: System architecture of the VAD.

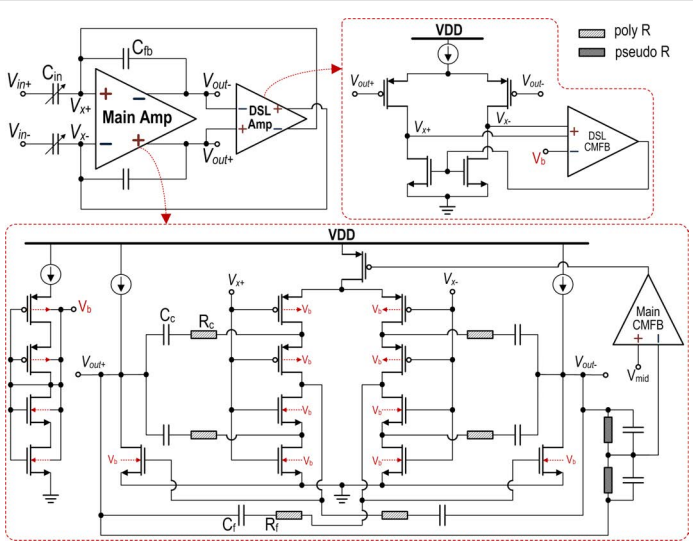

Figure 21.2.2: Circuit schematic of the LNA.

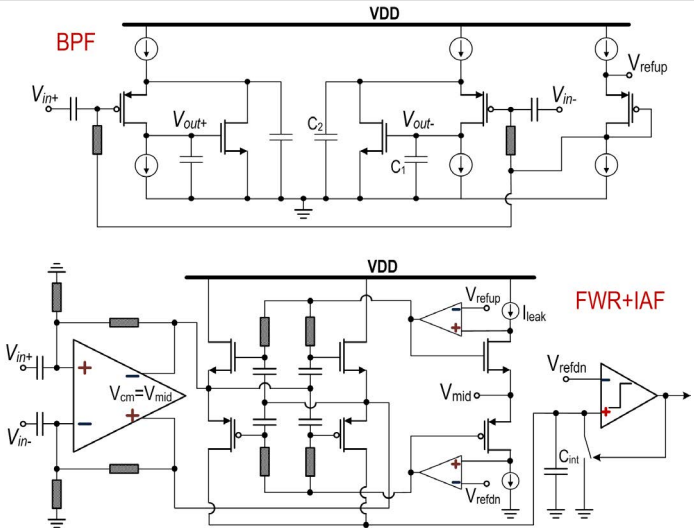

Figure 21.2.3: Circuit schematics of the BPF (top), and the FWR and the IAF (bottom).

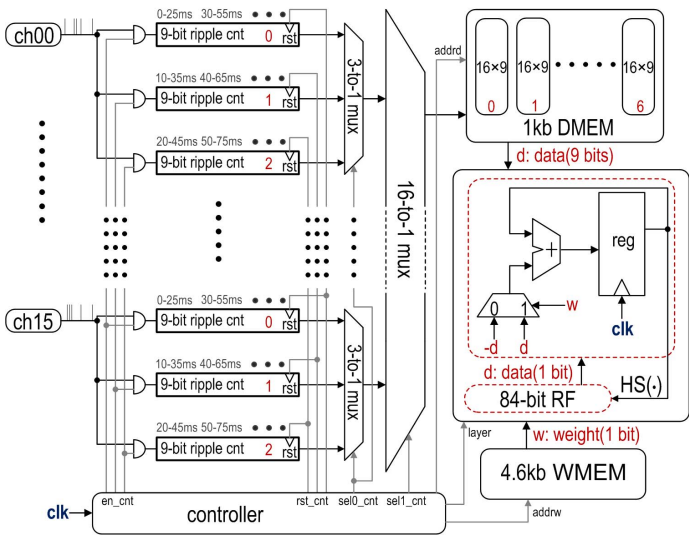

Figure 21.2.4: Architecture of the digital BNN.

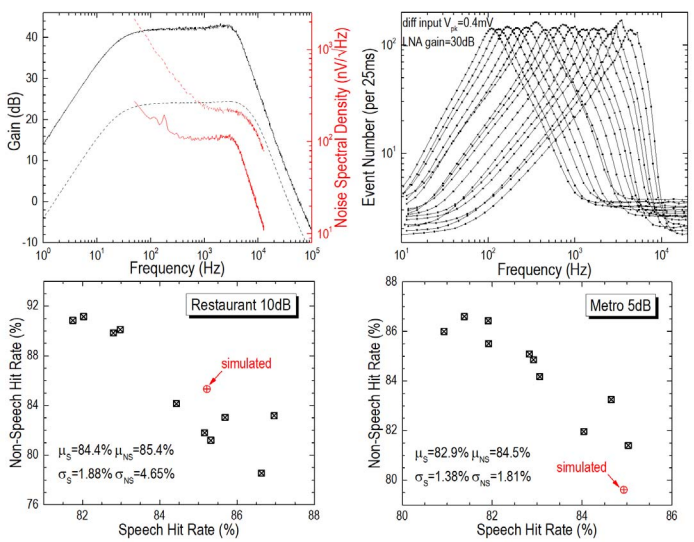

Figure 21.2.5: Chip Measurements.

Feature Extractor	This Work	Wu, ISSCC 2017	Yang, ISSCC 2016	Badami, ISSCC 2015
Technology (nm)	180	65	180	90
Feature Type	Analog to events	Digital	Analog to events	Analog
Channel Number	16	18×4	64×2	16
Frequency Range (Hz)	100 – 5k	160 – 8k	8 – 20k	75 – 5k
Power (μW)	0.38	13.8	55	6
Power/Channel (nW)	24	190	430	380
Normalized Power[a] (nW)	85	N/A	800	1480
Area/Channel (mm²)	0.1	0.00934	0.26	0.13
Dynamic Range (dB)	~40 @IAF[b]	N/A	55 @BPF Q=1.3,THD=1%	45 @LNA,THD=5%
Building Blocks	LNA, BPF, FWR, IAF	BPF	BPF, ADM	LNA, BPF, FWR, LPF

Voice Activity Detector	This Work	Price, ISSCC 2017	Esser, PNAS 2016	Badami, ISSCC 2015	Raychowdhury, JSSC 2013
Technology (nm)	180	65	28	90	32
System Input	Passive mic	Digitized sound	Digital feature	Passive mic	Digitized sound
Feature Type	Analog to events	Digital	Off-chip software	Analog	Digital
Classifier	Digital Binarized deep neural network	Digital Fixed-point deep neural network	Digital Spiking neural network	Mixed-signal Decision tree	Digital Energy-based decision rule
Power (μW)	1.0[c]	22.3	26100	6	~300
Classification Rate (/s)	100	100	1539	N/A	32600
Classification Dataset	AURORA4 mixed w/ DEMAND	AURORA2	TIMIT mixed w/ NOISEX	NOISEUS	N/A
Classification Accuracy	Speech/non-speech hit rate 84%/85% @10dB SNR, restaurant noise[c,d]	10% EER @7dB SNR, unspecified context	95.42% accuracy @unspecified SNR/context	Speech/non-speech hit rate 89%/85% @12dB SNR, babble noise	97% accuracy @unspecified SNR/context

a. calculated according to the equation in [5]; b. the ratio of max event rate at input V_{pk}=1mV and LNA gain of 30dB to min event rate at V_{pk}=0mV; c. measured at analog V_{dd}=0.6V and digital V_{dd}=0.55V; d. averaged over 10 dies.

Figure 21.2.6: Comparison table.

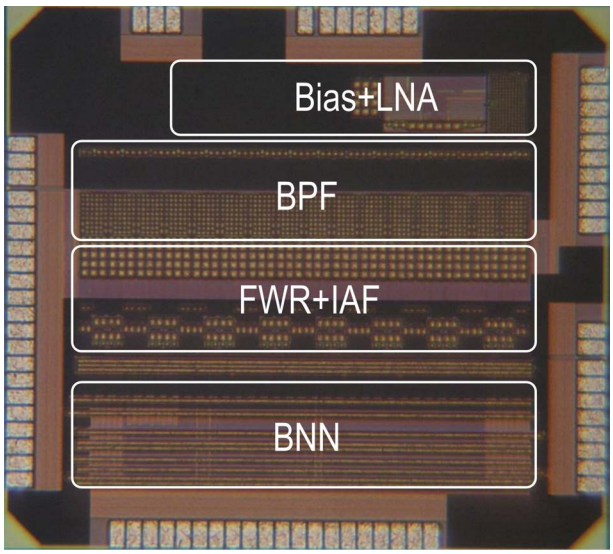

Figure 21.2.7: Chip micrograph.

978-1-5090-4940-0/18/$31.00