A 108-nW 0.8-mm^2 Analog Voice Activity Detector (VAD) Featuring a Time-Domain CNN with Sparsity-Aware Computation and Sparsified Quantization in 28-nm CMOS

Feifei Chen, Ka-Fai Un, *Member*, *IEEE*, Wei-Han Yu, *Member*, *IEEE*, Pui-In Mak, *Fellow*, *IEEE*, and Rui P. Martins, *Fellow*, *IEEE*

***Abstract*—This paper reports a passive analog feature extractor for realizing an area-and-power-efficient voice activity detector (VAD) for voice-control edge devices. It features a switched-capacitor circuit as the time-domain convolutional neural network (TD-CNN) that extracts the 1-bit features for the subsequent binarized neural network (BNN) classifier. TD-CNN also allows area savings and low latency by evaluating the features temporally. The applied sparsity-aware computation (SAC) and sparsified quantization (SQ) aid in enlarging the output swing and reducing the model size without sacrificing the classification accuracy. With these techniques, the diversified output also aids in desensitizing the 1-bit quantizer from the offset and noise. The TD-CNN and BNN are trained as a single network to improve the VAD reconfigurability. Benchmarking with the prior art, our VAD in 28-nm CMOS scores a 90% (94%) speech (non-speech) hit rate on the TIMIT dataset with small power (108 nW) and area (0.8 mm^2). We can configure the TD-CNN as a feature extractor for keyword spotting (KWS). It achieves a 93.5% KWS accuracy with the Google speech command dataset (2 keywords). With two TD-CNNs operating simultaneously to extract more features, the KWS accuracy is 94.3%.**

***Index Terms*—Approximate computing, convolutional neural network (CNN), feature extraction, keyword spotting (KWS), reconfigurable, switched-capacitor circuits, quantization, sparsity, voice activity detection (VAD).**

I. INTRODUCTION

VOICE-control devices provide a convenient communication interface between the humans and Internet-of-Things (IoT) nodes, but handling the complicated language model can require extensive computation and memory. To cope with the limited battery size at the edge devices, a voice activity detection (VAD) [1-10] can aid the speech-recognition system to save power since the activity rate of the voice-control IoT node [11] is typically low. The VAD activates the speech-recognition system only when a human voice is detected. Thus, it is crucial to develop a power-efficient always-on VAD by reducing the number of operations per classification or the energy consumed by each operation without degrading the classification accuracy. Also, a compact VAD is desirable for integration into the system chip.

The VAD typically consists of a feature extractor [12-14] and a classifier [15-17]. Their implementation can be entirely in the digital domain, but the always-on full-bandwidth analog-to-digital converter (ADC) and digital feature extractor can consume a significant power of >20 μW [1]. An analog feature extractor (AFE), adopted in [2-5], avoids the full-bandwidth ADC and the digital feature extractor. The AFE approximates the functionality of the mel-frequency cepstral coefficients [18] with the analog band-pass filters. These filters regrettably occupy a large chip area and require channel-wise quantizers for further classification in the digital domain. [2] utilized the rectifier, integrator, and comparator as the quantizer of the AFE, where the AFE consumed a consequential power of 3.0 μW, followed by a digital decision-tree classifier. [3] used the integrate-and-fire as the quantizer to reduce the power budget of the AFE to 380 nW. Yet, it requires 8-bit ripple-counters to convert the number of pulses into digital inputs for the classifier that also can consume significant power and area. A mixer-based AFE [4] proposed to extract one feature at a time, squeezing the AFE power to 100 nW. Yet, the latency increases drastically to 512 ms, and the incomplete output features limit the classification accuracy.

Inspired by recent advances in the raw-data convolutional neural network [19], we propose a switched-capacitor time-domain convolutional neural network (TD-CNN) AFE (Fig. 1) as presented in [20]. The passive switched-capacitor circuit can extract the charge domain more area-and-power-efficient than

This work was in part supported by the Macao Science and Technology Development Fund (FDCT), Macao SAR (File no. 0110/2019/A2) and SKL-AMSV(UM)-2020-2022), and in part by the University of Macau under Grant MYRG2019-00124-AMSV and MYRG2020-00191-IME. *(Corresponding author: Ka-Fai Un.)*

F. Chen, K.-F. Un, W.-H. Yu, P.-I. Mak, R. P. Martins are with the State-Key Laboratory of Analog and Mixed-Signal VLSI/ Institute of Microelectronics, and Faculty of Science and Technology - ECE, University of Macau, Macao, China (e-mail: kafaiun@umac.mo).

R. P. Martins is also with the Instituto Superior Técnico, Universidade de Lisboa, Portugal (e-mail: rmartins@umac.mo).

Color versions of one or more figures in this article are available at https://doi.org/XXXXXXXXXXXXXXX

Digital Objective Identifier XXXXXXXXX.

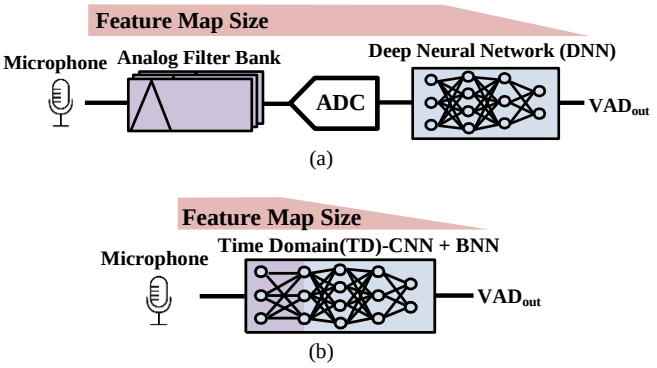


Fig. 1. (a) Conventional VAD with analog filter bank. (b) Proposed VAD with TD-CNN embedded classifier.

the analog filter bank [2-3]. The TD-CNN samples the input signal directly from the low-noise amplifier (LNA) window-by-window and performs in-memory computation (IMC) similar to the mixed-signal domain multiply-and-accumulate (MAC) operation [21-23]. The mixed-signal TD-CNN compresses the features before the quantizer to reduce the power consumption and chip area of the VAD. Unlike in the conventional AFE, the TD-CNN is the first layer of the classifier, trained with the whole neural network. Thus, TD-CNN can adapt its features to suit different applications rather than extract fixed features. Also, the mixed-signal MAC features are sensitive to analog non-idealities [19], including circuit noise, comparator offset, and capacitor mismatch. In this work, we propose sparsity-aware computation (SAC) and sparsified quantization (SQ) in the TD-CNN. The SAC handles the sign-bit of the weight and enlarges the TD-CNN output range by opening the capacitor, making the quantized weight able to take a large number of values. The SQ uses the SAC to nullify the weight with a small absolute value, making the quantized weights has a dual peak distribution, and diversifying the TD-CNN output leading to a feature quantization that is more robust to circuit non-idealities. It also improves the quantization precision where a 3-bit SQ can achieve higher accuracy than a 7-bit binary quantized model, reducing the memory size and the complexity of the TD-CNN cell.

The paper organization is as follows. Section II introduces the proposed VAD architecture and its circuit implementation. Section III details the proposed SAC and SQ techniques. Section IV provides the measurement results, and Section V draws the conclusions.

II. Proposed VAD Architecture

A. TD-CNN

Fig. 2 depicts the system architecture of the proposed VAD, which windowed the input voice at a sampling rate of f_s = 8 kS/s for feature extraction. The window length $t_{window} = N/f_s$ relates to the lowest recognizable frequency of the feature extractor. Fig. 3 shows the simulated hit rate (HR) over the channel number N of the TD-CNN layer. The simulated HR is >92% when N>80, while it drops significantly when N≤70. Herein we choose the preliminary channel number as 80, where we discard

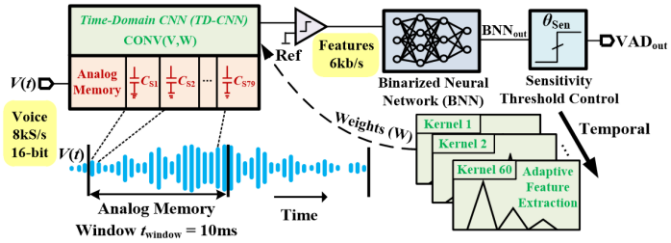


Fig. 2. System architecture of the proposed VAD.

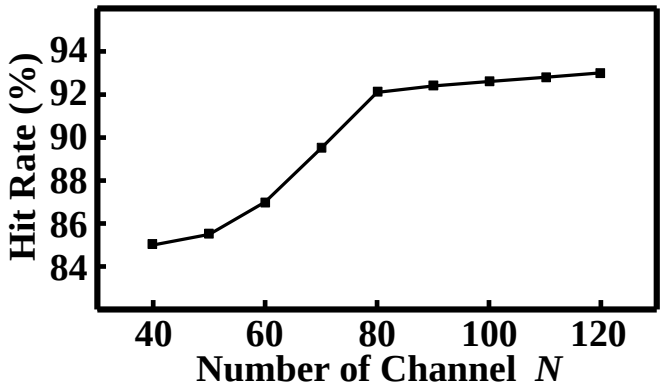


Fig. 3. Simulated test hit rate versus the channel number of the TD-CNN.

the last sample of each window for computation. Selecting a larger channel number will benefit the HR but leads to a longer window length. The leakage of the analog memory increases (to be detailed later), penalizing the computational precision. The circuit stores the sampled input voltage in the sampling-capacitor array acting as the analog memory. The switched-capacitor TD-CNN evaluates its 60 kernels temporally to extract sufficient features for the VAD. The extracted features are then 1-bit-quantized by a comparator to evaluate the binarized neural network (BNN) [16]. Thus, the conversion rate of the extracted features is only 6 kb/s. When comparing it with the A/D conversion of the raw voice signal data, which has a sampling rate of 8 kS/s with 16-bit resolution, the data conversion of the extracted 1-bit features reduces by 21.3×. The traditional BNN classifier has binary inputs, outputs, and weights. The sensitivity threshold control smooths out the BNN outputs, allowing the reconfiguration for different applications favoring the sensitivity or specificity.

Fig. 4 depicts the circuit diagram of the VAD feature extractor. The differential input voltages, buffered by an on-chip LNA with a 3.5-kHz bandwidth, charge the sampling-capacitor array. The LNA is an open-loop fully differential single-stage amplifier with a common-mode feedback circuitry. The typical input peak-to-peak voltage is 60 mV. The HR of the VAD is insensitive to the input swing of 40 to 140 mV, verified by simulations. The bandwidth of the LNA is adequate as the frequency range of the speech lies mainly between 0.1 to 2 kHz [10] while avoiding aliasing. Following the LNA, the TD-CNN comprises the analog memory and the SAC. The analog-memory array stores the windowed inputs. The SAC, performing an analog convolution, evaluates the 1-dimensional

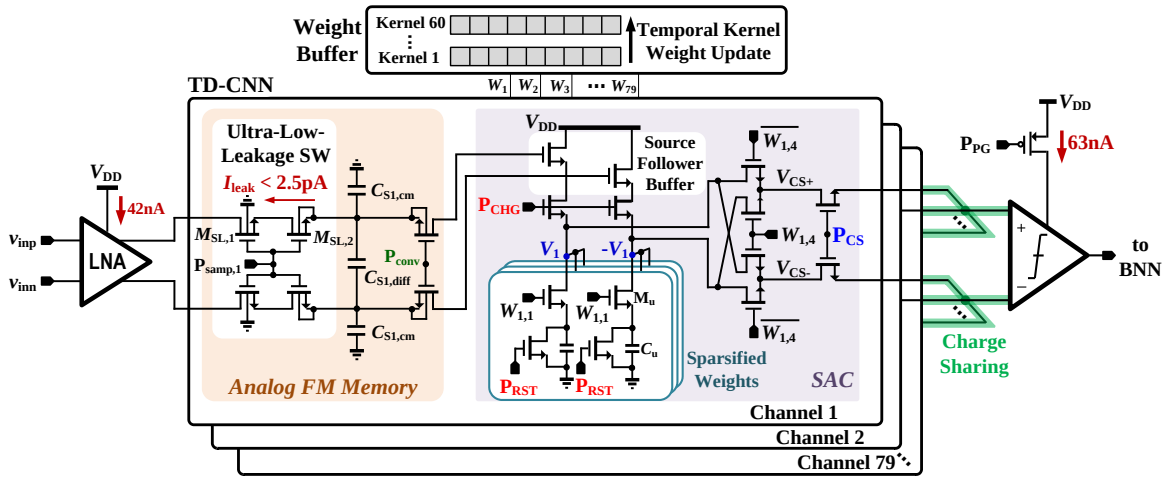


Fig. 4. Circuit details of the TD-CNN.

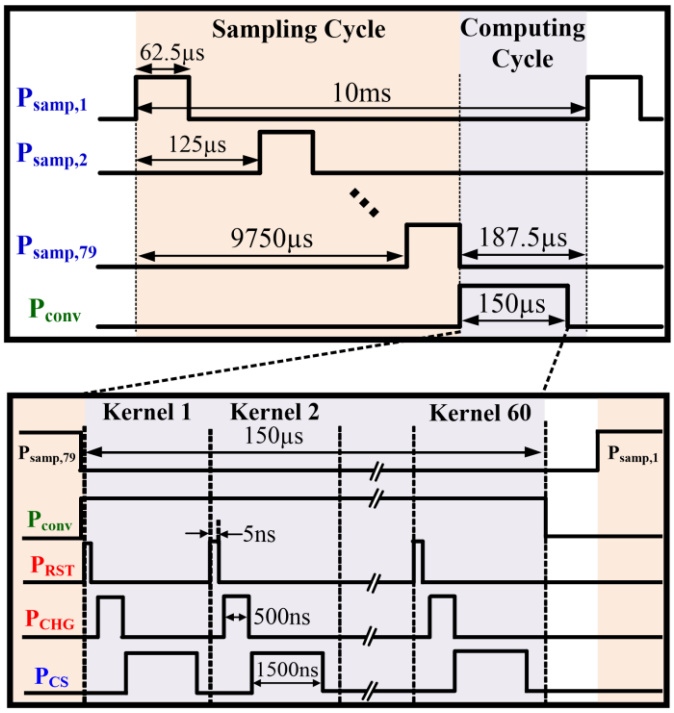


Fig. 5. Timing diagram of the TD-CNN.

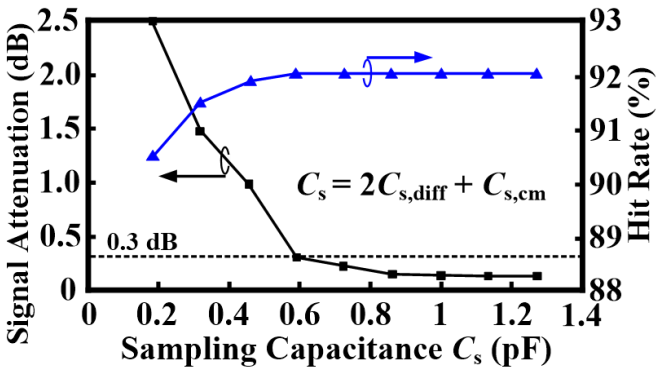


Fig. 6. The signal attenuation of the 1st channel and the HR versus C_s.

CNN operation. The output of the TD-CNN is at the charge-sharing node and is 1-bit quantized by a comparator. The comparator (a differential pair with a latch load) consumes 2 nW with its power gated with a duty cycle of 1/80 for each TD-CNN window.

Fig. 5 exhibits the timing diagram of the TD-CNN. We discard the last sample of each TD-CNN window, and the channel number of the TD-CNN N becomes 79. We also reserve the 80th sample of each window for the computing cycle. Otherwise, the computation of the current window will be delayed to the next window. It implies extra analog memory for storing data in the forthcoming window before finishing the computation of the current one. The simulation verifies that discarding the last sample of each window will not cause substantial HR degradation. The sampling time of 62.5 µs is sufficient for the signal sampling, while the off-cycle between samples reserves 187.5 µs for the computation cycle, allowing a longer settling time for the kernel computation (2.5 µs).

P_{samp1} to P_{samp79} sample the input signal for the TD-CNN evaluation. The sampling capacitance (C_S) value should compromise power and signal precision. A smaller capacitance can save the power budget of the input buffer but suffers from a larger signal attenuation due to the leakage current. Other than canceling part of the leakage error by the differential circuit, we size the sampling capacitor with a bearable leakage error. Fig. 6 plots the simulated signal attenuation of the 1st channel and the corresponding HR versus the sampling capacitance. The sampling capacitance (C_S) is >0.6 pF to sustain a sub-0.3 dB attenuation, causing no HR degradation to the VAD. We choose C_S to be 0.6 pF to balance signal attenuation and area. If we reduce C_s to 0.45 pF, it starts degrading the HR by 0.25%. We can further reduce C_s if the HR degradation is tolerable by different applications. The attenuation mismatch between different channels is within 0.3 dB. This nonlinearity is small enough to cause no HR degradation. Part of the sampling capacitor is in a differential ($C_{S,diff}$) connection to reduce the area, while the single-ended capacitor ($C_{S,cm}$) reduces the common-mode leakage by 60%. We design $C_{S,diff}$ and $C_{S,cm}$ as 0.14 and 0.32 pF, respectively. The switch $M_{SL,2}$ in the deep N-well is in series with the sampling switch $M_{SL,1}$ where the body

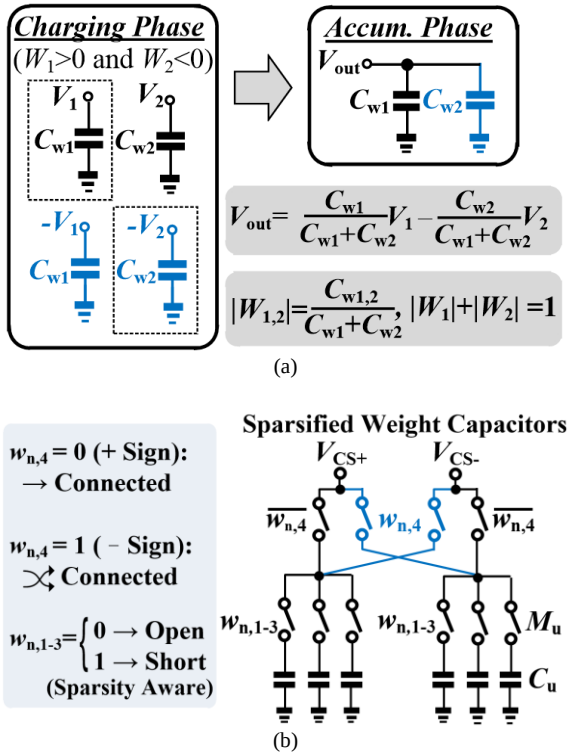


Fig. 7. (a) Demonstration of the analog convolution with a window size of 2. (b) Schematic of the SAC unit.

of $M_{SL,2}$ is connected to its source to reduce the leakage current from the sampling capacitor to the substrate from 3.6 to 2.5 pA. The switch $M_{SL,2}$ also offers better isolation, simulated to be 47 dB, between the different channels of the analog memory. For the computing cycle, P_{conv} accesses the analog memory for CNN operation. It also reduces the leakage current of V_{DD} from 83 to 17 pA. The circuit charges the sampled voltage into the computation unit capacitor array (C_u) through the source-follower buffer. The reset of the unit capacitor array during P_{RST} prevents excess charge over the computation, which otherwise degrades the computational precision. Subsequently, it will charge during P_{CHG}, where the weights of the TD-CNN model (read from the on-chip register array) determine the number of unit capacitors joining the computation. The pulse width of P_{RST} is 5 ns, generated by the on-chip deskew circuit, while P_{CHG} with a 500-ns pulse width defines the operating frequency (2 MHz) of the control circuit. To implement the MAC operation, we realize the multiplication through the capacitor characteristic $q_n = C_nV_n$. The voltages and the capacitances represent the TD-CNN's inputs and weights, respectively. The charge sharing accumulates the products (details in Section III-A).

B. BNN Classifier

The TD-CNN output is 1-bit quantized for further classification by the BNN. The size of the four-layer BNN is 60-36-12-2. It follows a sensitivity threshold control block serving as the output stage to handle the intra-sentence pause by smoothing the BNN output. It outputs '1' if the number of '1' from the recent BNN outputs is larger than θ_{sen}. Otherwise, it outputs '0'. We can adjust θ_{sen} to bias the classifier towards higher sensitivity or specificity for different applications, which

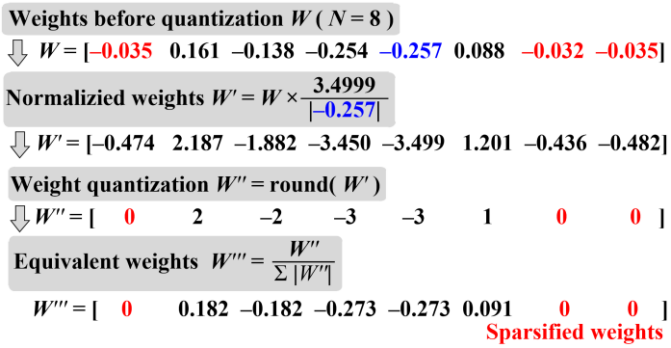


Fig. 8. A numerical example of the SQ when N=8.

typically requires retraining and updating the parameters of the entire DNN. The latency of the sensitivity threshold control is $\theta_{sen}t_{window}$.

III. Proposed Techniques for TD-CNN

The training of the TD-CNN and the classifier happens simultaneously with a floating-point precision in Python. After training the network for several epochs, we quantize the weights of the TD-CNN and the classifier to 3 bits and 1 bit, respectively, where the precision of the TD-CNN's weight is critical to the HR of the VAD. Consequently, we release the weights to floating-point precision for several iterations of this "training-quantization" procedure to uphold the HR of the VAD classifier after quantization. It generally takes 2 to 3 iterations to obtain a model. This section will present the SAC and SQ techniques for implementing the TD-CNN with low-resolution weight quantization without significant HR degradation.

A. Sparsity-Aware Computation (SAC)

We use the charge equation and charge sharing to implement the multiplication and accumulation, respectively. Fig. 7(a) presents a SAC unit with two channels demonstrating its operation. The source-follower buffers charge the differential voltages, representing signed inputs in the TD-CNN, into the unit capacitor arrays. Yet, the capacitance can only represent a positive-valued weight. We should swap the differential input for charge sharing if the corresponding weight is negative to implement the signed weight. The number of the unit capacitor joining the computation is proportional to the absolute value of the weight.

Fig. 7(b) provides the schematic of the SAC unit. The sign-bit of a weight $w_{n,4}$ determines whether to swap the polarity of the input voltage for the charge sharing. The remaining 2 bits of a weight are converted to thermometer-code ($w_{n,1-3}$) to determine whether the corresponding unit capacitor should join the charge sharing. The open capacitors do not contribute to the total capacitance of the charge-sharing node, improving the gain and output swing of the SAC. We used the minimum-sized NMOS switches (M_u) to reduce the buffer power, charge injection, and clock feedthrough. Hence, the average power of the weight buffer for each weight update is only 3.3 pJ/channel.

B. Sparsified Quantization (SQ)

Fig. 8 demonstrates the SQ with a numerical example when N =

N=2, C_1=mC_u, C_2=nC_u (m, n = 0, ±1, ±2, ±3)

(m, n)	w_1
(0, 0) , (0,±1) , (0,±2) , (0,±3)	0 , 0 , 0 , 0
(±1, 0), (±1,±1), (±1,±2), (±1,±3)	±1, ±1/2, ±1/3, ±1/4
(±2, 0), (±2,±1), (±2,±2), (±2,±3)	±1, ±2/3, ±1/2, ±2/5
(±3, 0), (±3,±1), (±3,±2), (±3,±3)	±1, ±3/4, ±3/5, ±1/2

W_1 = 0, ±1/4, ±1/3, ±2/5, ±1/2, ±3/5, ±2/3, ±3/4, ±1

W_1 can represent totally 17 possible values

(a)

N=3, C_1=mC_u, C_2=nC_u, C_3=pC_u (m, n, p = 0, ±1, ±2, ±3)

m	w_1 ($\|n\|+\|p\|$ = 0, 1, 2, 3, 4, 5, 6)
0	0 , 0 , 0 , 0 , 0 , 0 , 0
±1	±1, ±1/2, ±1/3, ±1/4, ±1/5, ±1/6, ±1/7
±2	±1, ±2/3, ±1/2, ±2/5, ±1/3, ±2/7, ±1/4
±3	±1, ±3/4, ±3/5, ±1/2, ±3/7, ±3/8, ±1/3

W_1 = 0, ±1/7, ±1/6, ±1/5, ±1/4, ±2/7, ±1/3, ±3/8, ±2/5, ±3/7, ±1/2, ±3/5, ±2/3, ±3/4, ±1

W_1 can represent totally 29 possible values

(b)

Fig. 9. Demonstration of 3-bit SQ to the weight for (a) N = 2, and (b) N = 3.

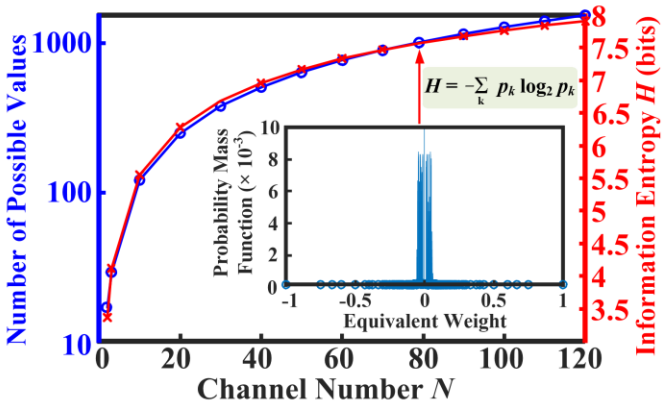


Fig. 10. Number of possible values for the equivalent weight, and the corresponding information entropy versus the channel number.

8. To sparsify and quantify the TD-CNN's weights, we normalize the weights in each kernel with their largest absolute weight and multiply them by 3.4999. Then, we can round them to the nearest integers, representing the unit capacitor numbers for the corresponding weights. The multiplier is slightly less than 3.5 to avoid rounding up any numbers to 4.

Fig. 9 demonstrates how the SQ scheme with a 3-bit precision improves the resolution of the weight quantization when N is 2 and 3. The capacitance for its corresponding weight can exhibit 7 values (-3C_u, -2C_u, …, 3C_u). The equivalent weight can take 17 and 29 possible values when N is 2 and 3, respectively. Herein the equivalent weights after 3-bit SQ, with N to be 79, can assume 1,017 different values, much more than the eight values from the conventional 3-bit quantization scheme due to the kernel dependent normalized factor. Fig. 10 plots the possible values of

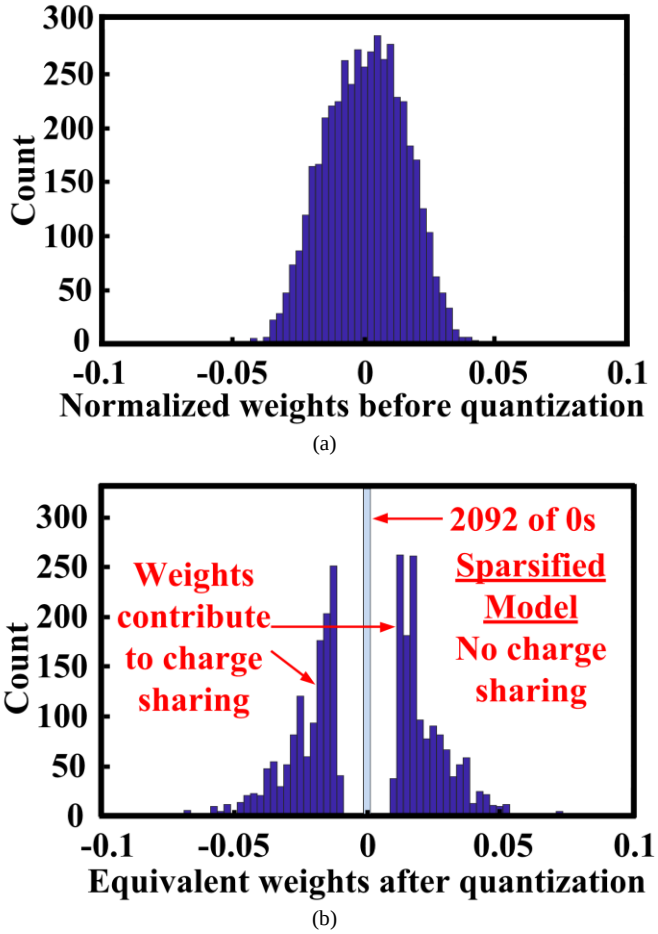


Fig. 11. Distribution of the weight: (a) before quantization and (b) after quantization.

the equivalent weights and their corresponding information entropy over the channel number. We calculate the information entropy of the equivalent weight with the assumption that before quantization, the weights follow a normal distribution. The entropy of the equivalent weights is 7.57 bits when N is 79. We observe non-uniform spacing of the equivalent weights with a concentration near zero. This benefits the quantization precision since the concentration of the weights before quantization is also around zero.

For a passive switched-capacitor circuit, the summation of the absolute equivalent weight W_n should be 1. The common normalized factor $\sum_{n=1}^{N} C_n$ which appears in the denominators of the equivalent weight, relates to the quantization of the kernel. These kernel-dependent normalized factors increase the resolution of the weight quantization. The weight before quantization exhibits a normal distribution as usual [Fig. 11(a)]. Out of the 4,800 weights in the TD-CNN, the quantization of 2,092 is zero, thus sparsifying the TD-CNN [Fig. 11(b)]. Therefore, the SAC lets the corresponding unit capacitors open from the charge-sharing node, which will not attenuate the signal swing of the TD-CNN output. Here, we can consider the distribution of the weight as dual-peak by ignoring the bar representing zero-weight. The energy efficiency of other DNN accelerators [24, 25] usually improves by skipping the sparsity of a network without causing significant

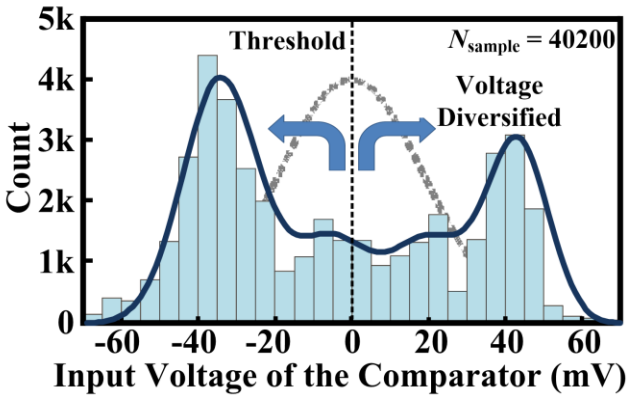


Fig. 12. Distribution of the TD-CNN output.

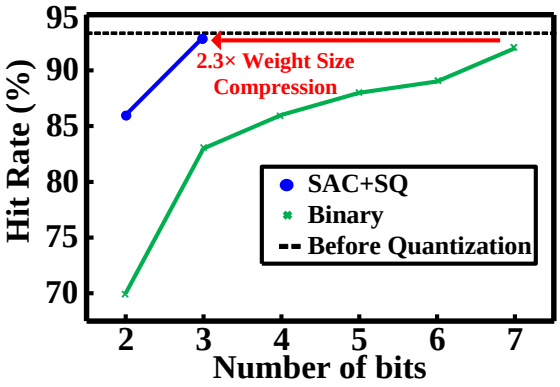


Fig. 13. Simulated HR over different resolution of the weight quantization.

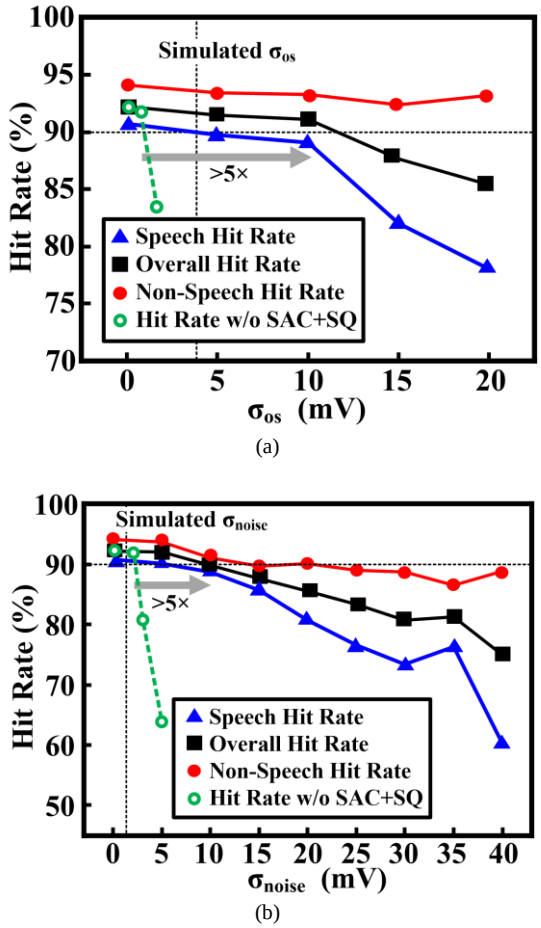


Fig. 14. Simulated speech and non-speech hit rates versus the comparator's (a) offset, and (b) noise.

accuracy degradation. The proposed quantization scheme increases the sparsity of the TD-CNN during andling by the SAC.

Fig. 12 plots the dual-peak distribution of the TD-CNN output while merging in the middle. Conventionally, the output distribution of a MAC is normal according to the central limit theorem. Yet, we can consider that the TD-CNN is the weighted sum of the weights. The output distribution is the convolution of the weight distributions with different *x*-axis scaling. Also, we obtain a dual-peak distribution, and the channel number is not sufficiently large to fulfill the central-limit theorem. Thus, the diversification of the output distribution improves the robustness of the 1-bit quantizer to the circuit non-idealities. To quantize a normal distributed output, the offset and noise of the comparator will cause a larger error as the threshold voltage of the comparator is close to the mode of the distribution. Here, the diversified output brings the threshold voltage to a low-density region.

Fig. 13 shows the simulated HR of the weight quantization with and without SAC+SQ. The HR of the 3-bit SAC+SQ TD-CNN layer is higher than the one with a 7-bit binary-quantized model, which matches the calculated entropy. It reduces the memory size for storing the weights of the TD-CNN by 2.3× and the number of unit capacitors for computation by 21×, thus significantly reducing the area and computational power. Fig. 14(a) and (b) plot the Monte Carlo simulation of the HRs with the comparator's offset and noise. The HR of the model with binary quantization starts to drop when the standard deviation of the offset and noise are both >2 mV. The HR of the TD-CNN with SAC+SQ is >90% when the value of the comparator's offset and noise is up to 10 mV, meaning an improvement of 5×. The simulated standard deviations of the comparator's offset is 3.9 mV. The simulated input-referred noise from the LNA, TD-CNN and the comparator are 0.5, 1.5 and 1.4 mV, respectively, guaranteeing the proposed TD-CNN's robustness.

C. Mismatch Analysis of the SAC

To quantify HR degradation due to the mismatch of the unit capacitance array, we construct a mismatch model by generalizing the equation in Fig. 7(a) as,

$$V_{out} = \frac{\sum_{n=1}^{N}(C_{wn}+\Delta C_{wn})V_n}{C_p+\sum_{n=1}^{N}(C_{wn}+\Delta C_{wn})} = \alpha\frac{\sum_{n=1}^{N}C_{wn}V_n}{\sum_{n=1}^{N}C_{wn}} + \beta, \quad (1)$$

where

$$\alpha = \frac{\sum_{n=1}^{N}C_{wn}}{C_p+\sum_{n=1}^{N}(C_{wn}+\Delta C_{wn})}, \quad (2)$$

and

$$\beta = \frac{\sum_{n=1}^{N}\Delta C_{wn}V_n}{C_p+\sum_{n=1}^{N}(C_{wn}+\Delta C_{wn})}. \quad (3)$$

V_{out} is the output voltage of the SAC; C_{wn}, ΔC_{wn} and V_n are the capacitance, its mismatch and input voltage of the n^{th} channel,

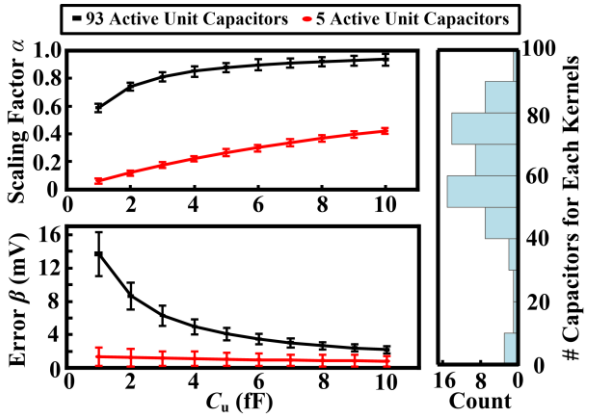


Fig. 15. α and β versus C_u with the histogram of the number of active unit capacitors of the 60 kernels.

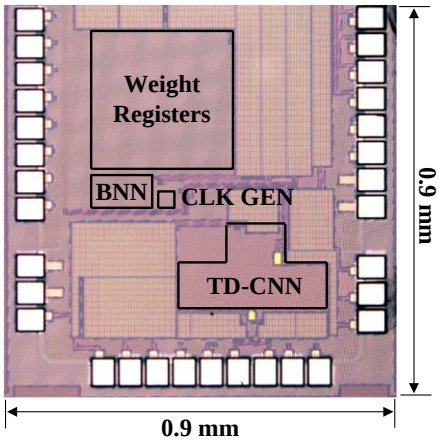
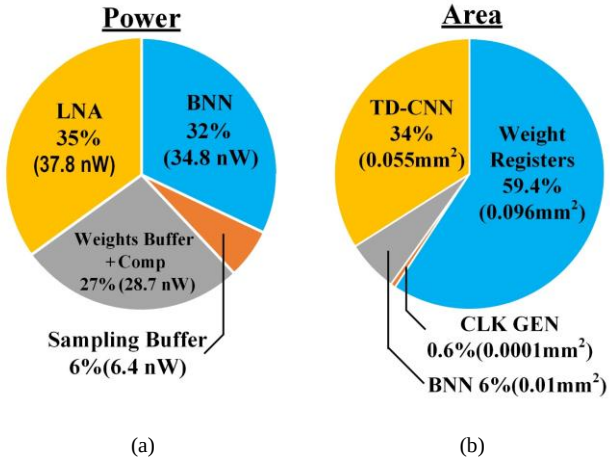


Fig. 16. Chip micrograph of the VAD.

respectively. C_p is the parasitic capacitance on the charge-sharing node. α and β are the scaling factor and the error due to capacitor mismatch and parasitic capacitance. The extracted value of C_p is 65 fF including grounded and coupling capacitances. Fig. 15 shows the simulated scaling factor α and offset β with their 95% confidence intervals over the unit capacitance C_u. It also shows the histogram of the number of active unit capacitors of the 60 kernels. The number of active unit capacitors ranges from 5 to 93 within the 60 kernels. Thus, the possible values of the absolute nonzero equivalent weight range from 1/93 to 3/5, verifying the quantized weight distribution in Fig. 11(b). We choose C_u to be 4.3 fF to balance the power consumption and computation precision. The simulation verifies that no significant attenuation and offset due to the parasitic capacitance on the HR degradation occurs when the standard deviation of the mismatch is up to 30% of C_u, where Monte Carlo simulations show the standard deviation of the mismatch is less than 10%.

IV. Experimental Results

Fabricated in 28-nm CMOS, the proposed analog VAD occupies a chip area of 0.9×0.9 mm^2, including the pads (Fig. 16). With dual supplies of 0.9 V (analog) and 0.6 V (digital), the VAD consumes 108 nW, with 73 nW associated with the AFE. Fig. 17 summarizes the power and area breakdown.



(a) (b)

Fig. 17. Measured (a) power and (b) area breakdowns of the VAD.

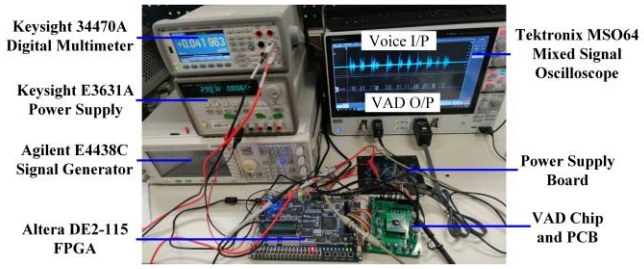


Fig. 18. The experimental setup.

Limited by our process access, we used 2.7 kB weight registers to store the model, including the TD-CNN and BNN. The register density is 4.2 μm^2/bit, while a typical static random-access memory (SRAM) can achieve a memory density of 0.3 to 0.5 μm^2/bit. There will be a further reduction of the corresponding power and area if an SRAM is available. Fig. 18 illustrates the measurement setup with the input signal generated by the Agilent Signal Generator E4438C. We use an Altera DE2-115 field-programmable gate array to write the model parameters into the registers through an on-chip serial-to-parallel interface. A Tektronix MSO64 Oscilloscope stores the output of the VAD classifier for HR analysis.

A. Voice Activity Detector (VAD)

To validate the proposed VAD performance, we deploy the speech from the TIMIT dataset mixed with the factory noise from the NOISEX-92 dataset of a 10-dB signal-to-noise ratio (SNR) unless specified otherwise. The speech activity rate is 50% by inserting a non-speech segment into each dataset file to minimize the prior knowledge. The measured testing set consists of 200 wav-files in the dataset with a total length of 20 minutes. Fig. 19(a) displays the time domain input signal and its label, while Fig. 19(b) and (c) show the simulated and measured extracted features for the corresponding signal. We can observe similar patterns from the 40 selected features with a higher correlation verifying the HR of the VAD. Fig. 20 illustrates the measured speech versus non-speech HR at SNR =10 dB. The speech and non-speech HRs are 90% and 94%, respectively, when θ_{sen} is 5, and the corresponding latency is 50 ms. We can adjust θ_{sen} to balance the performance of the speech and non- speech HR for different applications. Fig. 21 exhibits

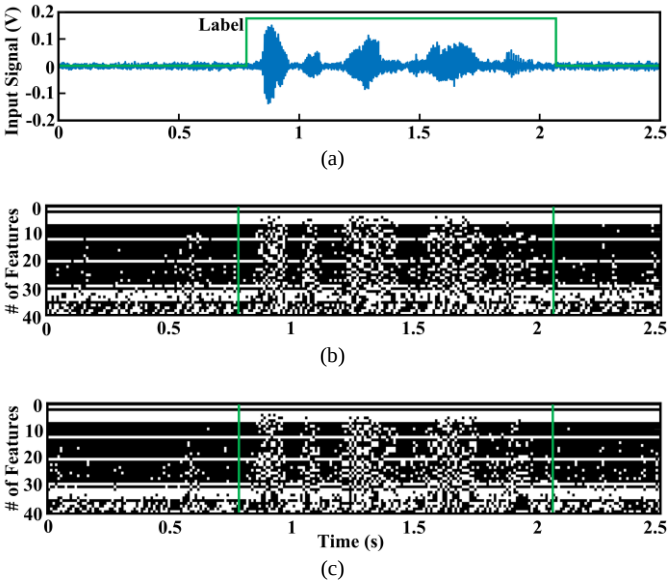


Fig. 19. (a) Time domain input signal with the label, (b) simulated features, and (c) measured features.

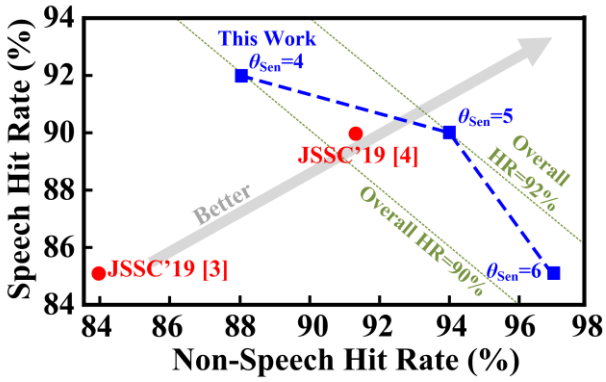


Fig. 20. Measured speech versus non-speech hit rates of the VAD at SNR = 10 dB.

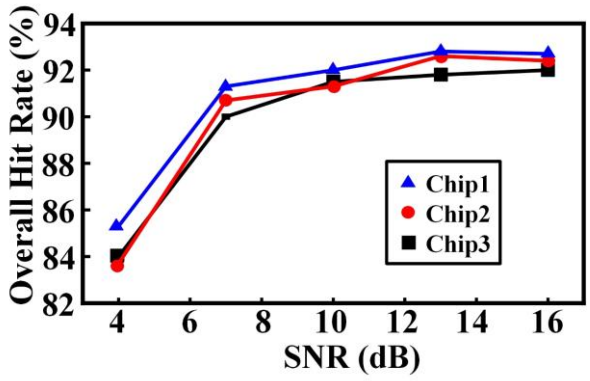


Fig. 21. Measured overall hit rate of the proposed VAD for 3 chip samples.

the measured overall HR versus different SNRs for multiple chips without chip-to-chip calibration, with the network trained with SNR = 10 dB. The overall HR is >91% when SNR is ≥10-dB for 3 test chips.

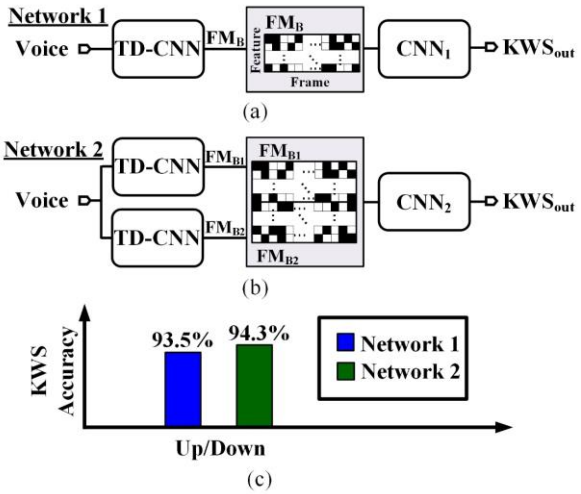


Fig. 22. Feature extractor of KWS with (a) one TD-CNN, (b) two TD-CNNs in parallel, and (c) the 2 keywords KWS accuracy.

TABLE I
PERFORMANCE COMPARISON OF THE FEATURE EXTRACTOR

Feature Extractor	This Work	[5]	[4]	[3]
Feature Extraction Topology	TD-CNN Feature Extraction	Time-Interleaved-Mixer-based Frequency Ext.	Analog-to-Event Filter Bank	Analog Filter Bank
Programmable Feature Extractor	**Yes (TD-CNN + SAC + SQ)**	Yes (b-DCT Sequence)	No	No
Channel Number	60	16-48	16	16
Frequency Range (Hz)	100 to 4k	75 to 4k	100 to 5k	75 to 5k
Feature Type	**Binary Adaptive**	Digital Frequency	Event-based Frequency	Analog Frequency
Power Consumption (nW)	73*	60	380	6000
Area (mm²)	**0.055**	0.56	1.6	2.56
Technology	**28 nm**	180 nm	180 nm	90 nm

** power consumed by the clock buffers for the sampling, charge-sharing buffers and LNA*

B. Keyword Spotting (KWS)

We can also reconfigure the proposed TD-CNN as the feature extractor of a keyword spotting (KWS) system. We concatenate the extracted features of the TD-CNN window-by-window to form a 2-dimensional feature map [Fig. 22(a)]. An off-chip CNN further processes the features. To evaluate the performance of this system, we use two keywords 'Up' and 'Down' from the Google speech command dataset (GSCD). Network 2 achieves a 93.5% accuracy. To improve the accuracy, we can deploy two TD-CNNs in parallel to extract 2× more features [Fig. 22(b)]. The accuracy of Network 2 improves up to 94.3%. It is necessary to further increase the number of features and their resolution to cover more keywords.

C. Performance Benchmarks

We compare the performance of our feature extractor with the state-of-the-art in Table I. The TD-CNN with SAC+SQ extracts the binary features for further classification. It consumes 73 nW power and occupies an active area of 0.055 mm^2. Compared with [3], the power consumption is reduced by 5×. Compared with [4], we reduce by 10× the area of the TD-CNN while its power consumption is low due to a low extraction rate which causes a low classification rate. Table II compares the

TABLE II
PERFORMANCE COMPARISON OF THE VAD

Voice Activity Detector	This Work	[5]	[4]	[2]	[7]
Classifier	**TD-CNN + BNN**	Neural Network	BNN	Digital Fixed-Point Deep Neural Network	Analog SNR Decision Rule
Classification Rate (Hz)	100	2	100	100	31.25
Dataset	TIMIT + NOISEX-92	LibiSpeech + NOISEX-92	AURORA4 + DEMAND	AURORA2	Custom
Accuracy (SP HR/Non-SP HR)	90.1%/94% @ 10dB SNR	91.5%/90% @ 10dB SNR	84%/85% @ 10 dB SNR	90% @ 7 dB SNR	Not Comparable
Power Consumption (nW)	**108**	142	1000	22300	760
Energy/classification (nJ)	**1.08**	73	10	223	24.32
Chip Area (mm^2)	0.8	17.6**	2.5	2.1	0.14 (Active)
Technology	28 nm	180 nm	180 nm	65 nm	180 nm

** *Area including an audio compressor and a processor*

performance of the VADs. The achieved energy per classification is 1.08 nJ which is 9× less than [3]. The chip area is also 2.6× less than [1].

V. CONCLUSIONS

This paper presented a VAD using a switched-capacitor TD-CNN as a passive analog feature extractor and a BNN classifier. The TD-CNN temporally extracted the features reducing the area and power of the VAD while sustaining high HR and low latency. The required 1-bit features significantly reduced the number of data conversions. The SAC increases the output swing of the TD-CNN, and the SQ improves the resolution to the weight quantization of the TD-CNN. These two techniques diversified the TD-CNN output, which desensitized the 1-bit quantizer from the offset and noise. We can also configure the TD-CNN as the feature extractor of the KWS.

REFERENCES

[1] M. Price, J. Glass, and A. Chandrakasan, "A low-power speech recognizer and voice activity detector using deep neural networks," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 66–75, Jan. 2018.

[2] K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6μW power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan. 2016.

[3] M. Yang, C. Yeh, Y. Zhou, J. Cerqueira, A. Lazar, and M. Seok, "Design of an always-on deep neural network-based 1-μW voice activity detector aided with a customized software model for analog feature extraction," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1764–1777, June 2019.

[4] S. Oh et al., "An acoustic signal processing chip with 142-nW voice activity detection using mixer-based sequential frequency scanning and neural network classification," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 3005–3016, Nov. 2019.

[5] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, pp. 1534–1538, Sept. 2014.

[6] S. Mun, S. Shon, W. Kim, D. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 796–800, Mar. 2017.

[7] S. Thomas, G. Saon, M. Van Segbroeck, and S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 4500–4504, Apr. 2015.

[8] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. Tschanz, and V. De, "A 2.3 nJ/frame voice activity detector-based audio front-end for context-aware system-on-chip applications in 32-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, Aug. 2013.

[9] M. Croce, B. Friend, F. Nesta, L. Crespi, P. Malcovati and A. Baschirotto, "A 760-nW, 180-nm CMOS fully analog voice activity detection system for domestic environment," *IEEE J. Solid-State Circuits*, vol. 56, no. 3, pp. 778-787, Mar. 2021.

[10] E. Shi, X. Tang, and K. Pun, "A 270 nW switched-capacitor acoustic feature extractor for always-on voice activity detection," *IEEE Trans. Circuits and Syst. I: Reg. Papers*, vol. 68, no. 3, pp. 1045-1054, Mar. 2021.

[11] M. Cho et al.,"A 6×5×4 mm^3 general purpose audio sensor node with a 4.7μW audio processing IC," in *Proc. Symp. VLSI Circuits*, pp. C312-C313, June 2017.

[12] N. C. Bui, J. J. Monbaron, and J. G. Michel, "An integrated voice recognition system," *IEEE J. Solid-State Circuits*, vol. 18, no. 1, pp. 75-81, Feb. 1983.

[13] Y. Kuraishi, K. Nakayama, K. Miyadera, and T. Okamura, "A single-chip 20-channel speech spectrum analyzer using a multiplexed switched-capacitor filter bank," *IEEE J. Solid-State Circuits*, vol. 19, no. 6, pp. 964-970, Dec. 1984.

[14] E. Fragniere, "A 100-channel analog CMOS auditory filter bank for speech recognition," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 140–589, Feb. 2005.

[15] B. Moons, D. Bankman, L. Yang, B. Murmann, and M. Verhelst, "BinarEye: An always-on energy-accuracy-scalable binary CNN processor with all memory on chip in 28nm CMOS", in *Proc. Custom Integr. Circuits Conf. (CICC)*, pp. 1-4, Apr. 2018.

[16] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks", in *Proc. Neural Information Processing Systems (NIPS)*, pp. 4114-4122, Dec. 2016.

[17] J. Li, K. -F. Un, W. -H. Yu, P. -I. Mak and R. P. Martins, "An FPGA-based energy-efficient reconfigurable convolutional neural network accelerator for object recognition applications," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 68, no. 9, pp. 3143-3147, Sept. 2021.

[18] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. on Speech Audio Process.*, vol. 7, no. 5, pp. 525-532, Sept. 1999.

[19] R. Zazo, T. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. Interspeech*, pp. 3668-3672, Sept. 2016.

[20] F. Chen, K. -F. Un, W. -H. Yu, P. -I. Mak and R. P. Martins, "A 108nW 0.8mm^2 analog voice activity detector (VAD) featuring a time-domain CNN as a programmable feature extractor and a sparsity-aware computational scheme in 28nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 368-369, Feb. 2022.

[21] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8 μJ/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, Jan. 2019.

[22] W. -H. Yu et al., "A 4-bit mixed-signal MAC array with swing enhancement and local kernel memory," in *Proc. IEEE Int. Midwest Symp. Circuits and Syst. (MWSCAS)*, pp. 326-329, Aug. 2021.

[23] Z. Jiang, S. Yin, J. Seo, and M. Seok, "C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism," *IEEE J. Solid-State Circuits*, vol. 55, no. 7, pp. 1888-1897, July 2020.

[24] D. Park et al., "A 7.3 M output non-zeros/J, 11.7 M output non-zeros/GB reconfigurable sparse matrix–matrix multiplication accelerator," *IEEE J. Solid-State Circuits*, vol. 55, no. 4, pp. 933-944, Apr. 2020.

[25] J. Kim, J. Lee, J. Lee, J. Heo, and J. Kim, "Z-PIM: A sparsity-aware processing-in-memory architecture with fully variable weight bit-precision for energy-efficient deep neural networks," *IEEE J. Solid-State Circuits*, vol. 56, no. 4, pp. 1093-1104, Apr. 2021.

Feifei Chen received his B.Sc. degree in Electronic Science and Technology from Chongqing University, Chongqing, China in 2011, and his M.Sc. degree in electrical engineering and computer science from National Taiwan University (NTU), Taipei, Taiwan, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Science and Technology and the State Key Laboratory of Analog and Mixed-Signal VLSI, Department of Electrical and Computer Engineering, University of Macau, Macao, China.

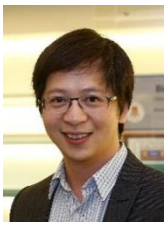
From 2011 to 2015, Mr. Chen was an analog IC design engineer in Shenzhen, developing high-speed interface ICs. His current research interests are ultra-low-power (ULP) voice activity detector and keyword spotting system.

Ka-Fai Un (S'09-M'14) received his Ph.D. degree from the University of Macau (UM), Macau, in 2014, respectively. He was a Post-Doctoral Researcher and a Lecturer (Macao Fellow) in 2014 and 2015, respectively. He is an assistant professor since 2018 with the State Key Laboratory of Analog and Mixed-Signal VLSI, UM. He was on leave from UM and was a visiting Post-Doctoral Researcher with the School of Electrical and Electronic Engineering, University College Dublin, Dublin, Ireland, from 2017 to 2018.

His research interests include analog artificial intelligence (AI) circuits, digital AI accelerator design, and analog and RF circuit design. In 2003, He won the Macau Mathematics Olympics and represented Macau in the Chinese Mathematics Olympiad (CMO) and the International Mathematics Olympiad (IMO), in Changsha and Tokyo, respectively. He is the recipient of the 2008 APCCAS Merit Student Paper Certificate.

Wei-Han Yu (S'09-M'18) received the Ph.D. degree from the University of Macau (UM), Macau, in 2018. From 2019 to 2021, He was a Visiting Scholar at Muramnn Mixed-Signal Group, Stanford University. He has been an Assistant Professor with the State-Key Laboratory of Analog and Mixed-Signal VLSI (AMSV), University of Macau (UM), Macao, China, since 2021.

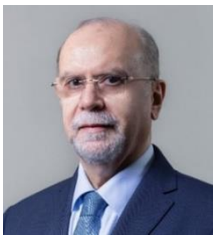
His research interests include edge AI, in-memory computing, switched capacitor circuits, energy-harvested RF transceivers, and RF/mm-wave power amplifiers. He received the IEEE ISSCC Student Travel Grant Award, the FDCT Science and Technology Postgraduate Student Award in 2016, and the IEEE SSCS Predoctoral Achievement Award in 2018.



Pui-In Mak (S'00-M'08-SM'11-F'19) received the Ph.D. degree from University of Macau (UM), Macao, China, in 2006. He is currently Full Professor at UM Faculty of Science and Technology, ECE Department; Interim Director at the State Key Laboratory of Analog and Mixed-Signal VLSI and Deputy Director (Research) at the Institute of Microelectronics. His research interests are on analog and radio-frequency (RF) circuits and systems for wireless and multidisciplinary innovations.

His involvements with IEEE are: Editorial Board Member of IEEE Press ('14-'16); Member of Board-of-Governors of IEEE Circuits and Systems Society ('09-'11); Senior Editor of IEEE Journal on Emerging and Selected Topics in Circuits and Systems ('14-'15); Associate Editor of IEEE Journal of Solid-State Circuits ('18-), IEEE Solid-State Circuits Letters ('17-), IEEE Transactions on Circuits and Systems I ('10-'11, '14-'15) and II ('10-'13). He is/was the TPC Vice Co-Chair of ASP-DAC ('16), TPC Member of A-SSCC ('13-'16,'19), ESSCIRC ('16-'17) and ISSCC ('17-'19). He is/was Distinguished Lecturer of IEEE Circuits and Systems Society ('14-'15) and IEEE Solid-State Circuits Society ('17-'18). He was the chairman of the Distinguished Lecturer Program of IEEE Circuits and Systems Society ('18-'19).

Prof. Mak (co)-received the DAC/ISSCC Student Paper Award'05, CASS Outstanding Young Author Award'10; National Scientific and Technological Progress Award'11; Best Associate Editor of IEEE Transactions on Circuits and Systems II'12-13, A-SSCC Distinguished Design Award'15 and ISSCC Silkroad Award'16. In 2005, Prof. Mak was decorated with the Honorary Title of Value for scientific merits by the Macau Government. Prof. Mak is inducted as Overseas Expert of the Chinese Academy of Sciences since 2018; Fellow of the UK Institution of Engineering and Technology (IET) for contributions to engineering research, education and services since 2018; and Fellow of the IEEE for contributions to radio-frequency and analog circuits since 2019, and Fellow of the UK Royal Society of Chemistry since 2020.



Rui P. Martins (**IEEE** Member'88 – Senior Member'99 – Fellow'08), born in April 30, 1957, received the Bachelor, Masters, and Ph.D. degrees, as well as the Habilitation for Full-Professor in Electrical Engineering and Computers from the Department of Electrical and Computer Engineering (DECE), Instituto Superior Técnico (IST), U. of Lisbon, Portugal, in 1980, 1985, 1992 and 2001, respectively. He has been with the DECE / IST, U. of Lisbon since October 1980.

Since Oct. 1992, has been on leave from U. of Lisbon and with the DECE, Faculty of Science and Technology (FST), University of Macau (UM), Macao, China, where he is a Chair-Professor since Aug. 2013. In FST, he was Dean (1994-1997), and has been UM's Vice-Rector since Sep. 1997. From Sep. 2008 to Aug. 2018, Vice-Rector (Research) and from Sep. 2018 to Aug. 2023, Vice-Rector (Global Affairs). Within the scope of his teaching and research activities he has taught 21 bachelor and master courses and, in UM, has supervised (or co-supervised) 47 theses, Ph.D. (26) and Masters (21). Authored or Co-authored: 9 books and 12 book chapters; 49 Patents, USA (39), Taiwan (3) & China (7); 675 papers, in scientific journals (289) and in conference proceedings (386); as well as other 70 academic works, in a total of 815 publications. He created in 2003 the Analog and Mixed-Signal VLSI Research Laboratory of UM, elevated in January 2011 to State Key Laboratory (SKLAB) of China (the 1st in Engineering in Macao), being its Founding Director. He was the Founding Chair of UMTEC (UM company) from January 2009 to March 2019, supporting the incubation and creation in 2018 of Digifluidic, the first UM Spin-Off, whose CEO is a SKLAB PhD graduate. He was also a co-founder of Chipidea Microelectronics (Macao) [later Synopsys-Macao, and now Akrostar, where the CEO is one of his Ph.D graduates] in 2001/2002.

Prof. Rui Martins is an IEEE Fellow, was Founding Chair of IEEE Macau Section (2003-2005) and IEEE Macau Joint-Chapter on Circuits And Systems (CAS) / Communications (COM) (2005-2008) [2009 World Chapter of the Year of IEEE CAS Society (CASS)], General Chair IEEE Asia-Pacific Conference on CAS – APCCAS'2008, Vice-President (VP) Region 10 (Asia, Australia and Pacific) (2009-2011) and VP-World Regional Activities and Membership of IEEE CASS (2012-2013), Associate-Editor of IEEE Transactions on CAS II: Express Briefs (2010-2013), nominated Best Associate Editor (2012-2013). He is a member of the Advisory Board of the Journal of Semiconductors of the Chinese Institute of Electronics (CIE), Institute of Semiconductors, Chinese Academy of Sciences, since January 2021, and a Fellow of the Asia-Pacific Artificial Intelligent Association since October 2021. He was also member of: IEEE CASS Fellow Evaluation Committee (2013, 2014, 2018 – Chair, 2019, 2021 & 2022 – Vice-Chair); IEEE Nominating Committee of Division I Director (CASS/EDS/SSCS) (2014); and IEEE CASS Nominations Committee (2016-2017). In addition, he was General Chair of ACM/IEEE Asia South Pacific Design Automation Conference – ASP-DAC'2016, receiving the IEEE Council on Electronic Design Automation (CEDA) Outstanding Service Award in 2016, and also General Chair of the IEEE Asian Solid-State Circuits Conference – A-SSCC 2019. He was Vice-President (2005-2014), President (2014-2017) and now again Vice-President (2021-2024) of the Association of Portuguese Speaking Universities (AULP), and received 3 Macao Government decorations: the Medal of Professional Merit (Portuguese-1999); the Honorary Title of Value (Chinese-2001) and the Medal of Merit in Education (Chinese-2021). Since July 2010 was elected, unanimously, to the Lisbon Academy of Sciences, as Corresponding (2010-2022) and Effective Member (from 2022), being the only Portuguese Academician working and living in Asia.