14.1 A 510nW 0.41V Low-Memory Low-Computation Keyword-Spotting Chip Using Serial FFT-Based MFCC and Binarized Depthwise Separable Convolutional Neural Network in 28nm CMOS

Weiwei Shan[1], Minhao Yang[2], Jiaming Xu[1], Yicheng Lu[1], Shuai Zhang[1], Tao Wang[1], Jun Yang[1], Longxing Shi[1], Mingoo Seok[3]

[1]Southeast University, Jiangsu, China
[2]EPFL, Neuchâtel, Switzerland
[3]Columbia University, New York, NY

Ultra-low power is a strong requirement for always-on speech interfaces in wearable and mobile devices, such as Voice Activity Detection (VAD) and Keyword Spotting (KWS) [1-5]. A KWS system is used to detect specific wake-up words by speakers and has to be always on. Previous ASICs for KWS lack energy-efficient implementations having power <5μW. For example, deep neural network (DNN)-based KWS [1] has a large on-chip weight memory of 270KB and consumes 288μW. A binarized convolutional neural network (CNN) used 52KB of SRAM,141μW wakeup power at 2.5MHz, 0.57V [2]. An LSTM-based SoC used 105KB of SRAM and reduced power to 16.11μW for KWS with 90.8% accuracy on the Google Speech Command Dataset (GSCD) [3]. Laika reduced power to 5μW [4], not including the Mel Frequency Cepstrum Coefficient (MFCC) circuit. High compute and memory requirements have prevented always-on KWS chips from operating in the sub-μW range.

We propose four ways to implement a low-memory, low-computation, sub-μW KWS chip with 2KB on-chip memory. It is composed of an MFCC feature extrator and a depthwise-separable convolutional NN (DSCNN). The main contributions are: 1) A serial FFT-based MFCC circuit (Fig. 14.1.1, left) with reduced data storage and power as compared to a parallel approach; 2) A binarized DSCNN (Fig. 14.1.1, top right), decomposing a standard 3D convolution to a 2D convolution [5], reducing data storage by 7× and computation by 7× as compared to a CNN; 3) Elimination of the redundant memory and computations, reducing computation by 17.4× and intermediate data storage by 3.5×, compared to traditional whole-word processing [1-5], without loss of detection accuracy; 4) Near-threshold design of the whole chip with customized register-file-based memory blocks. Finally, we reduce the frequency to 40KHz to compute one frame in 16ms with an ultra-low power of 0.51μW in 28nm CMOS, achieving 94.6% accuracy on two-word GSCD KWS.

Aiming for 1-2 keyword spotting, the NN topology is shown at the bottom of Fig. 14.1.1. Our network comprises 4 layers (Conv + DS-convolution + Pooling + FC), where the DS-convolution (DSC) layer convolves each channel in the input feature map by a separate 2D filter and then uses pointwise convolutions (DSCPW). Each depthwise layer is followed by batch-normalization and a sign activation. Input data are 10 dimensions of 8b features from the MFCC. For two-word detection, the number of total coefficients (weights) is 3,456. The number of computations is 150,496, composed of 102,400 8b multiply-accumulates (MACs) and 48,096 1b MACs. The only difference between one/two-word is the dimension of FC layer, that is, 288•2 vs. 288•3 for one/two-word.

Figure 14.1.2 shows the MFCC circuit to generate 8b features, including pre-emphasis, framing, windowing, Fast Fourier Transformation (FFT), MEL filter, log2 operation and a Discrete Cosine Transform (DCT). Input speech data from GSCD are cut to 500ms and sent to the MFCC. A 16b 256-point FFT is executed at an 8KHz sampling rate and frames of 32ms with a 16ms step. The FFT is the main part of the MFCC consuming most of the power. Instead of using a traditional parallel FFT circuit, we design a serial-pipeline FFT circuit with memory compressed by 8× and power reduced by 11×. It has four Radix-2^2 Single-path Delay Feedback (R2^2SDF) units, each composed of an independent memory and a butterfly (BF) unit. After the FFT, we simplify the 20 MACs in the Mel filters for the Mel cepstrum energy to only 2 MACs, due to the overlap in the Mel filter coefficients. To further reduce power, we replace the complicated natural logarithm operation by a log2 operation, implemented by a look-up-table (LUT). Moreover, we use hybrid precision quantization to reduce power. The MFCC and CNN circuits are operated in a pipeline, allowing the circuits to be clocked at a lower frequency to reduce power.

Figure 14.1.3 shows the hardware optimizations of the DSCNN based on eliminating the redundant memory/computations. Most KWS chips use the whole word (for example, 31 frames•13 dimensions in [1], and 11 frames•40 dimensions in [2]), which consumes significant memory power. As shown in Fig. 14.1.4 (bottom left), memory consumes 70% of power in the whole-word approach. However, since only one frame differs between two adjacent inputs, there are data overlaps, and hence repeated computations within the convolutional layers. [2] reused the convolution results to eliminate some redundant computations but had no reduction in memory. We propose to eliminate the redundant computations and also compress the memory. Specifically, we calculate and store the data related to the latest frame (Fig. 14.1.3 bottom), where the bottom parts are ours and the gray parts are the eliminated storage. For example, 10 frames of data are stored because the kernel of the 1st layer is 10×4, and only one frame of the data is updated in each 16ms step to further reduce dynamic power. Pooling results with step size equal to 2 are sent to the FC layer to reduce the size of the FC layer, which needs to be stored separately for odd and even frames since the FC layer has no overlaping computations. In total, we eliminate 71.4% of the data storage and 94.3% of computations without losing detection accuracy. The number of computations per frame after the hardware optimizations is reduced to 8,736 (5,120 8b MACs and 3,616 1b MACs).

The hardware architecture of DSCNN is shown in Fig. 14.1.4 (top left), composed of a control Finite State Machine (FSM), a Processing Element (PE) array, five memory blocks and a mapping module. Except for the 8b input layer, all activations are binarized to 0/1 to represent -1/1, reducing the data transitions (toggles) as compared to a complement code computation. We design our chip to operate at near-threshold voltage (NTV), using extra-high-threshold (EHVT) transistors for ultra-low leakage. All standard cells are pruned and re-characterized at 0.5V. Memory design is a challenge because: 1) The MFCC and NN need many small-sized memory blocks, including both single-port and dual-port memories (Fig. 14.1.4, top right); 2) Leakage power dominates since the clock frequency is reduced to 40KHz (allowable owing to the reduced amount of computation); 3) The foundry-provided SRAM is unable to work at NTV. Thus, we design a customized register-file-type memory with low leakage, flexible to form small-size memories and able to work at NTV. Overall, the power consumption of memory and logic are reduced dramatically.

Fabricated in 28nm CMOS, Fig. 14.1.7 shows the die micrograph with a core area of 0.44×0.52mm^2. The measured minimum voltages of 30 chips are in the top left of Fig. 14.1.5, with the lowest voltage ranging from 0.41V to 0.48V. The power consumption related to voltages for the best chip are shown in Fig. 14.1.5 (top right), achieving a minimum power of 0.51μW at 0.41V. Fig. 14.1.5 (bottom left) shows the accuracies of one-word (happy, dog, marvin, up, down, etc.) detection and two-word (happy + dog) detection on the GSCD with the ratio of the number of fillers to keywords set as 6:1, achieving 94.6% accuracy for two-word spotting and around 98% for one-word spotting. Complex words have higher accuracies than simple keywords. The receiver operating characteristics (ROC) curve for two-keyword spotting (Fig. 14.1.5, bottom right) based on software testing shows the false reject rate (FRR) vs. the false alarm rate (FAR) under a 1-hour-long voice, which is randomly concatenated by 1200 words composed of fillers, two keywords and white noise in GSCD. Fig. 14.1.6 shows a comparison of the presented KWS system with other prior work. Our chip offers a small memory size, a small area, a low frequency and reduces power consumption by 10-564×.

Acknowledgments:
Projects supported by National Natural Science Foundation of China (61574033 and 61774038), China Major S&T Project (2018ZX01031-101) and Columbia Research Initiatives in Science and Engineering (RISE).

References:
[1] S. Bang et al., "A 288μW Programmable Deep-Learning Processor with 270KB On-Chip Weight Storage Using Non-Uniform Memory Hierarchy for Mobile Intelligence," *ISSCC*, pp. 250-250, Feb. 2017.
[2] S. Yin et al., "A 141 UW, 2.46 PJ/Neuron Binarized Convolutional Neural Network Based Self-Learning Speech Recognition Processor in 28NM CMOS," *IEEE Symp. VLSI Circuits*, pp. 139-140, 2018.
[3] J. Giraldo et al., "18μW SoC for Near-Microphone Keyword Spotting and Speaker Verification," *IEEE Symp. VLSI Circuits*, pp. 52-53, 2019.
[4] J. Giraldo et al., "Laika: A 5μW Programmable LSTM Accelerator for Always-on Keyword Spotting in 65nm CMOS," *ESSCIRC*, pp. 166-169, 2018.
[5] Y. Zhang et al., "Hello Edge: Keyword Spotting on Microcontrollers," *arXiv:1711.07128*, 2017.

 978-1-7281-3205-1/20/$31.00

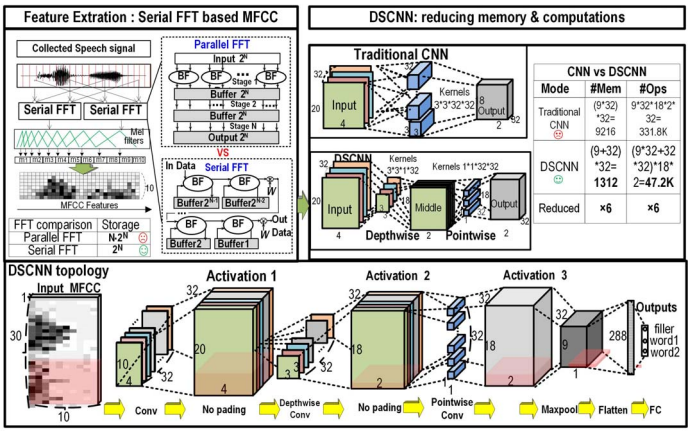


Figure 14.1.1: Concepts of serial FFT-based MFCC in reducing memory (top left), DSCNN network vs. CNN (top right) and our DSCNN topology (bottom).

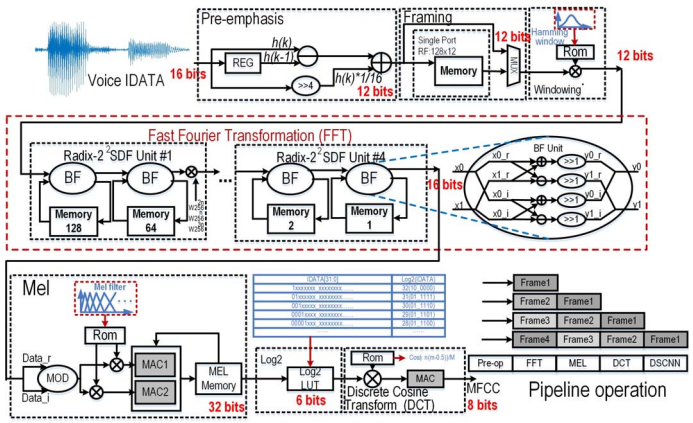
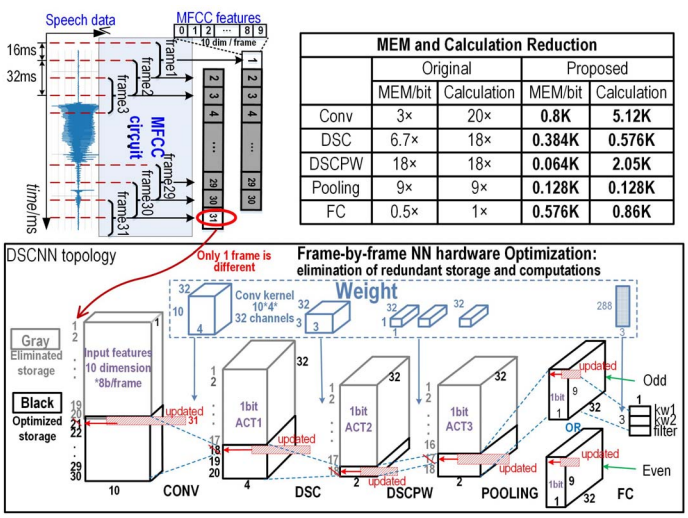
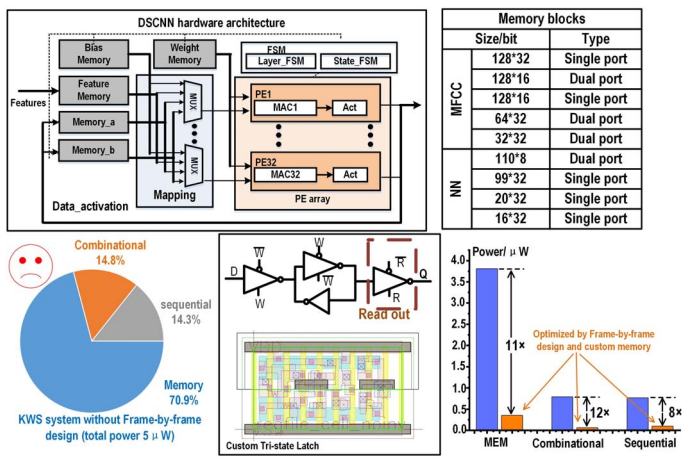


Figure 14.1.2: MFCC circuit composed of pre-emphasis, framing, windowing, FFT, MEL, $\log_2$ operation and DCT.



MEM and Calculation Reduction	Original		Proposed	
	MEM/bit	Calculation	MEM/bit	Calculation
Conv	3×	20×	0.8K	5.12K
DSC	6.7×	18×	0.384K	0.576K
DSCPW	18×	18×	0.064K	2.05K
Pooling	9×	9×	0.128K	0.128K
FC	0.5×	1×	0.576K	0.86K

Figure 14.1.3: Hardware design of DSCNN based on elimination of redundant memory/computation.



Memory blocks	Size/bit	Type
MFCC	128*32	Single port
	128*16	Dual port
	128*16	Single port
	64*32	Dual port
	32*32	Dual port
NN	110*8	Dual port
	99*32	Single port
	20*32	Single port
	16*32	Single port

Figure 14.1.4: DSCNN hardware architecture (top left), memory blocks (top right), original KWS power distribution (bottom left), custom-designed memory cell (bottom middle) and power reduction results (bottom right).

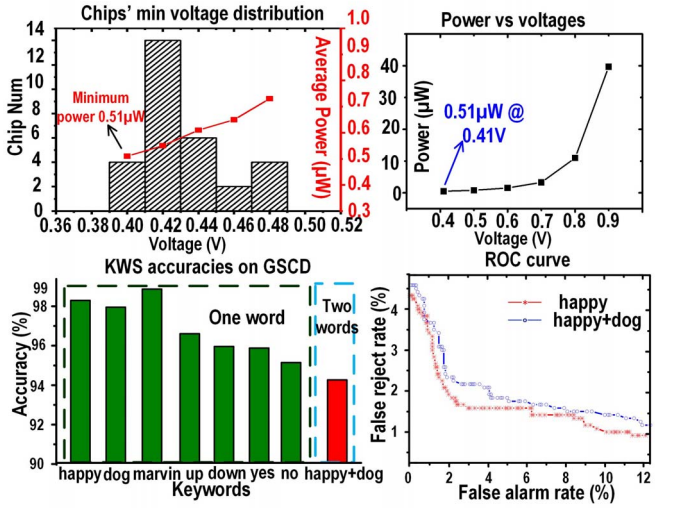
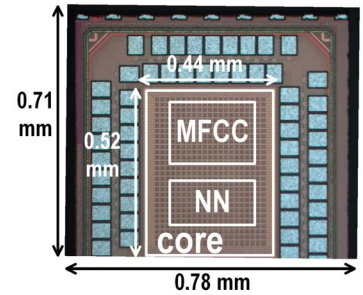


Figure 14.1.5: Measurement results: chip minimum voltage distribution, power at 0.4-0.9V, accuracy of keyword detection on three datasets (bottom left, 2-word), accuracy with 0-to-40dB noise (bottom right, 2-word).

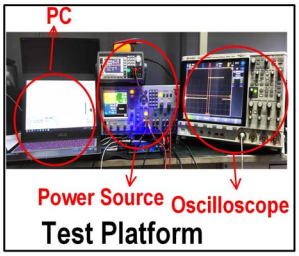
	ISSCC 2017 [1]	VLSI 2018 [2]	VLSI 2019 [3]	ESSCIRC 2018 [4]	This work
Tech.	40 nm	28 nm	65 nm	65 nm	28 nm
Algorithm	DNN	CNN	LSTM	LSTM	DSCNN
Voltage	0.63-0.9V	0.57-0.9V	0.6V	0.575V	0.41V
Memory	270KB	52KB	65KB	32KB	2KB
Core Size	7.1mm²	1.29mm²	2.56mm²	1.04mm²	0.23mm²
Frequency	1.9MHz	2.5MHz	250KHz	250KHz	40KHz
Latency	6.5ms	0.5-25ms	16ms	16ms	64ms
Keyword Num	10 words	1 word	10 words	4 words	1~2 words
Power	288 μ W	141 μ W	16.1 μ W*	5 μ W**	0.51 μ W
Dataset	NA	TIDIGIT	GSCD	NA	GSCD
Accuracy	NA	96%	90.87%	91.2%	98@1 word; 94.6%@2 words

(*16.1 μ W refers to the power of digital KWS, not including Analog Front-End
**5 μ W in [5] does not include MFCC)

Figure 14.1.6: Chip measurement results and comparison with state-of-the-art.



Process	28 nm CMOS
Supply Voltage	0.41V
Frequency	40kHz
Core Size	0.44×0.52 mm²
Die Size	0.78×0.71 mm²
Memory	2KB
MFCC power	0.34 μW
NN power	0.17 μW



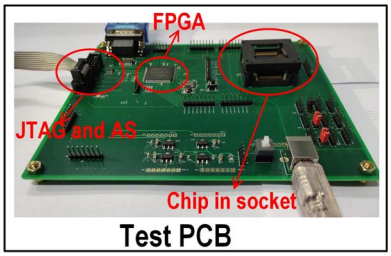
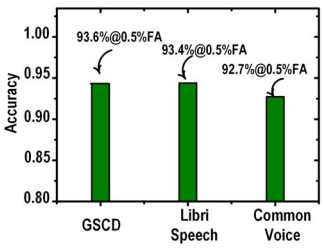


Figure 14.1.7: Chip die photo, characteristics, testing platform and PCB board.



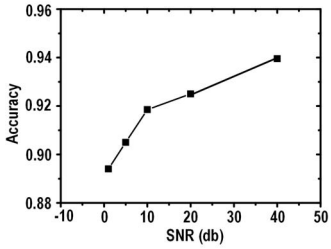


Figure 14.1.S1: Accuracy of the key words detection on three datasets (left), and accuracy with 0-40dB noise (right).

978-1-7281-3205-1/20/$31.00 ©2020 IEEE