Binarized Weight Neural-network Inspired Ultra-low Power Speech Recognition Processor with Time-domain based Digital-analog Mixed Approximate Computing

Bo Liu, Hao Cai, Yu Gong, Wentao Zhu, Yan Li, Wei Ge
National ASIC System Engineering Research Center
Southeast University, Nanjing 210096, China
(liubo_cnasic, hao.cai, gongyu, zwt, yanli7, duiker)@seu.edu.cn

Zhen Wang
Nanjing Prochip Electronic Technology Co. Ltd.
Nanjing 210001, China
fanstics@prochip.com.cn

***Abstract*—In this paper, an ultra-low power speech recognition processor is implemented based on an optimized binarized weight neural-network (BWN). To accelerate the BWN and make it energy efficient, we proposed an approximate computing architecture for the quantized BWN based on time-domain digital-analog mixed addition unit and precision optimization with fault-tolerant training method. Experimental results show that the proposed digital-analog mixed approximate computing architecture can significantly reduce the power consumption while maintaining the recognition accuracy. Implemented under TSMC 28nm, the proposed processor can support 10 keywords real time recognition under different noise types and SNRs, while the power consumption is 56μW.**

***Index Terms*—Speech Recognition, Binarized Weight Neural-network, Approximate Computing, Fault-tolerant Training**

I. Introduction

The ultra-low power speech recognition processor is very widely used in those battery-powered devices with human-machine interaction, such as wearable devices, mobile devices, the Internet of Things, and so on. In the past decades, deep neural networks (DNNs) have demonstrated a more prominent advantage in speech recognition than traditional models (i.e., Hidden Markov models and Gaussian mixture models). However, the massive parameters and computations of DNNs produce too much power consumption [1] [2]. To overcome the challenges, many DNN accelerators for ultra-low power speech recognition have been proposed in recent years [4-9]. Price M., et al. presented an ultra-low power speech processor with the power consumption of 7.78 mW @40MHz and WER of 8.78% under TSMC 65nm low-power logic process, where the adopted DNN consists of three fully-connected (FC) layers and the bit width of both data and weight are 16 bits [6]. In M. Shah's work [7], the data/weight bit width of the DNN is optimized as 16/6 bits to reduce the power consumption. The DNN accelerator proposed in M. Shah's work can support 10 keywords real-time recognition with the power consumption of 11.2 mW. Bang S., et al. proposed a DNN accelerator which can support voice wake-up function (one keyword recognition) with power consumption of 321 μW [8]. In Yin's work [9], they first proposed an optimized binary neural nerwork (BNN) with 4 convolution (CONV) layers and 2 FC layers. In this BNN, the bit width of data and weight are both 1 bit, 99% of the BNN operations are additions and the multiplication operations are almost eliminated. To further reduce the power consumption, an ultra-low power DNN accelerator with approximate addition units is proposed to process the calculation of each layer in the BNN. The DNN accelerator in Yin's work can support one keyword recognition with power consumption of 141 μW.

In this paper, we proposed an ultra-low power speech recognition processor based on an optimized binarized weight neural-network (BWN). The proposed speech recognition processor can support 10 keywords recognition under various background noise. To accelerate the BWN and make it energy efficient, we proposed an approximate computing architecture for the quantized BWN based on time-domain digital-analog mixed addition unit and precision optimization with fault-tolerant training method. Implementation results show that the proposed accelerator can support 10 keywords real time recognition under different noise types and SNRs with high accuracy, while the power consumption can be significantly reduced to 56μW. Compared to the SoA architectures, this work can achieve up to 2.5× better in energy efficiency.

II. Top Architecture of Proposed Speech Recognition System

The prototype of the speech recognition system adopted in this work mainly consists of two parts: the input speech feature extraction based on MFCC and the keywords classification based on BWN. The feature extraction module is used for extracting the features of the input speech. The output of feature extraction module is 26 Mel-scale Frequency Cepstral Coefficients (MFCC) [10]. The speech classification module processes the feature classification based on a deep neural network and determines which keyword it is (or an

This work was supported by the National Science and Technology Major Project under Grant 2018ZX01031101-005.

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

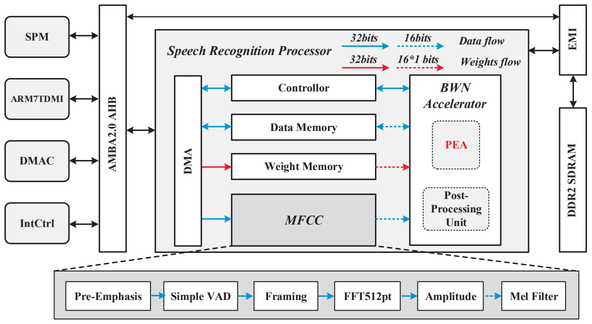

Fig. 1. Top Architecture of the Prototype Speech Recognition System

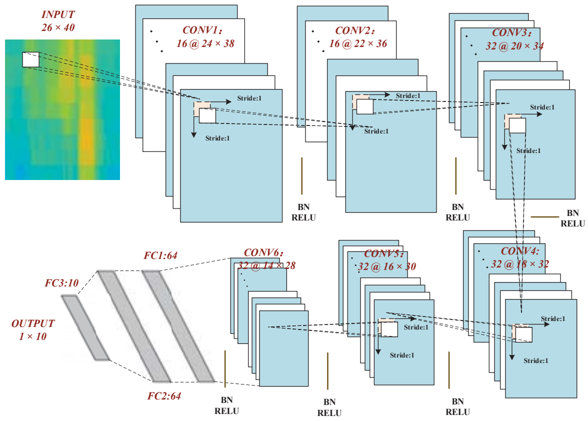

Fig. 2. BWN Topology for 10 Keywords Recognition

unknown word). The feature extraction module consists of a pre-emphasis unit, an energy-based simple voice activity detection (VAD) unit [11] [12], a framing unit, a 512-point FFT unit, a 16-stage pipeline coordinate rotation digital computer (CORDIC) based amplitude unit and a Mel Filtering unit. The top architecture of the prototype system and the speech recognition processor is shown in Fig.1. The top-level architecture consists of a system controller implemented with ARM7TDMI, a speech recognition processor, an 8Kbytes SRAM as system memory and several assistant modules for system scheduling. All modules are AMBA2.0-AHB-compatible and connected to a 32-bit AHB bus module, used as the system bus. The speech recognition processor consists of a MFCC module, a BWN accelerator, the controller and the data/weight/configure memory, each is an 8Kbytes SRAM. The accelerator can be reconfigured to process different layers in the BWN. The input speech signal is sampled at 16KHz, and all modules in Fig.1 operate on the frames of 40ms with 20ms time step size.

This paper trains a BWN for the 10 keywords recognition, based on the principle of XNOR-Net quantization framework where the bit width of both data and weight are 1 bit [13] and the quantization training method discussed in [14]. As shown in Fig.2, the BWN is composed by six convolutional (CONV) layers, three fully-connected (FC) layers, several activation (ACT) layers and batch normalization (BN) layers. The convolution kernel sizes of each CONV layer are all 3×3, the numbers of convolution kernels are 16, 16, 32, 32, 32, 32, respectively. Each CONV layer is followed by an ACT layer and a BN Layer. We use Relu as the activation function of output neurons in each BN layer. During the training process, the weights of gradient calculations during the entire forward transfer and back propagation are binarized to 1 bit (that is +1 or -1), and the feature data are quantized to 16 bits.

III. Energy-efficient Digital-analog Mixed Approximate Computing for BWN

A. BWN Processing with Time-domain Digital-analog Mixed Approximate Computing

Each of the processing element (PE) in BWN accelerator is composed of many time-domain approximate addition units (TAAUs). Fig.3 shows the structure of the TAAU. In TAAU, there are two digital input data signals of each data: X_2X_1. The rising edge signal A is the initial input starting signal of the first TAAU, Y is the output of the current TAAU and is simultaneously connected with the input starting signal line of the next TAAU. The TAAU can output four analog signal values with different time delay width. All TAAUs will only work when the input starting signal is a rising edge. When the control signals are “00”, the signal selector output $D_1D_2D_3$ is “111”, and the MOS transistors M8, M9 and M10 are all turned on. The pull-down network consists of M8, M9 and M10 in parallel. In this case, the pull-down resistor is the smallest and the delay time is t. When the control signals are “01”, the signal selector output $D_1D_2D_3$ is 011, the MOS transistor M8 is turned off, while the MOS transistors M9 and M10 are turned on. The pull-down network consists of three parallel circuits. Firstly, the MOS transistors M6, M7 and M11 form the first-stage parallel circuit; then, the first-stage parallel circuit and the MOS transistors M5 and M10 form the second-stage parallel circuit; finally, the second-stage parallel circuit and the MOS transistors M3, M4 and M9 form the whole parallel circuit. In this case, the pull-down resistor is much smaller, and the delay time is $2t$. When the control signals are “10”, the signal selector output $D_1D_2D_3$ is “001”, and the MOS transistors M8 and M9 are turned off, while the MOS transistor M10 is turned on. The pull-down network is connected by the MOS transistors M10 and M5, M6, M7, M11 in parallel, and the MOS transistors M3 and M4 are connected in series. In this case, the resistance value of the pull-down network is reduced, and the delay time is $3t$. When the two control signals are “11”, the signal selector output $D_1D_2D_3$ is “000”. In this case, the MOS transistors M8, M9 and M10 are turned off, and the pull-down network has only one path formed by the MOS transistors M3, M4, M5, M6, M7 and M11. The pull-down resistor is much larger, and the delay time is $4t$.

Fig.4 shows the PE architecture with proposed TAAUs. The PE consists of the XOR multiplication units and the proposed

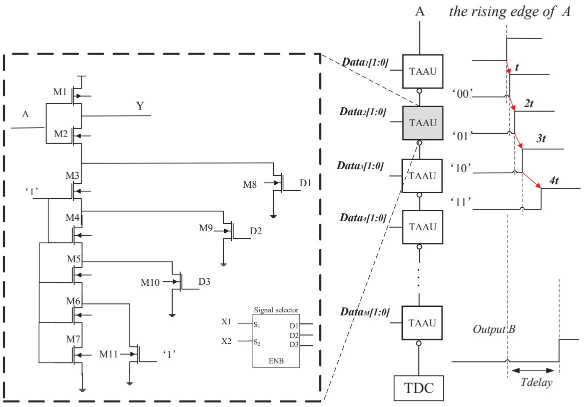

Fig. 3. Architecture of Proposed Time-domain Approximate Addition Unit

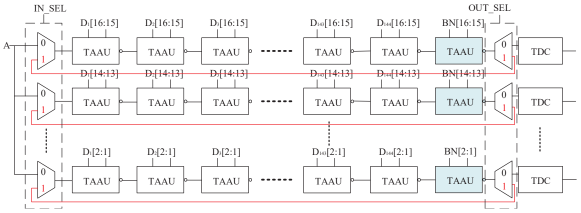

Fig. 4. PE Architecture with Time-domain Approximate Addition Units

TAAUs. The data of each CONV layer in the proposed BWN topology is a multiple of 144. Therefore, the BWN accelerator can process n×144 data at one time, controlled by the iteration controller. When the output delay signal generated by the PE is finally converted into a digital value by the Time-to-Digital Converter (TDC), it will be then loaded to the Post Process unit for future processing. For the BN layer, it is mainly composed of addition and bit-wise XOR operations. The additions of BN can be processed with the TAAUs in PEs. In this work, each TAAU can process 2-bits addition calculations. Compared to work [15], where the analog computing unit can only process 1-bit addition calculation, our work with the proposed TAAU can further improve the energy efficiency of the PE by over 20%. As shown in Fig.5, we added a set of TAAUs after the CONV layers computing to further process the additions of the BN layer. This method will further reduce the temporary data load/store operations and the power consumption.

As shown in Fig.5, the input data is firstly accumulated by the PEs to obtain the results in TAAUs for both CONV layer and BN layer. The four PEs operate independently at the same time, and then the four output results of the PEs are converted into corresponding digital results by the TDCs. If the CONV/BN layer contains a pooling layer, the output results of the four PEs are compared by bit size and the largest part of the sum are then dynamically selected. The results are finally shifted to complete the multiplication operation in BN layers. This scheduling method can reduce up to 10% of the power consumption of the BWN accelerator.

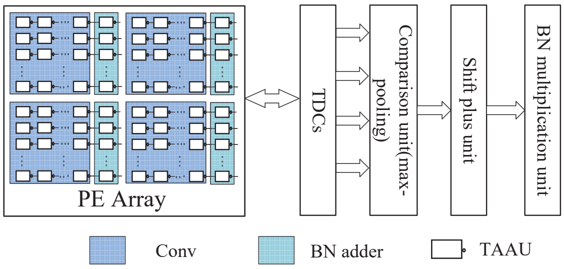

Fig. 5. BWN Workflow Based on Digital-analog Mixed Computing

In this work, we use TDCs to process the delay signals and convert them to digital values. There are many non-ideal factors in the circuit implementation, such as the process corner, temperature, etc., which will cause different degrees of circuit mismatch. To reduce the PVT impacts, the delay units of the TDCs are designed with the same delay units as the TAAU. However, the problem of network accuracy loss caused by digital-analog mixed approximate computing cannot be completely avoided. The simulation results show that the mismatch of the quantized results of the proposed time-domain approximate computing can reach up to 8% under the TSMC 28nm process. In neural networks, regularization terms are often added during the training phase in order to reduce the number of weights and avoid over-fitting. The training process is shown in Fig.6. In order to improve the robustness of neural networks and reduce the impacts of the circuit mismatch, a regularization term is added to the cost function. The optimized cost function formula is as following:

$$E_{tot} = E + \gamma * S(W) \tag{1}$$

In the equation, E is the initial cost function, $S(W)$ represents the sensitivity of the network, γ is the influence factor parameter, and E_{tot} is the optimized cost function. In this work, γ is set to 0.01. Lower sensitivity means higher robustness and higher resilience to circuit mismatch. The total cost function includes the cost function and mismatch sensitivity of the proposed BWN. It can be minimized to get a fault-tolerant network. $S(W)$ reflects the deviation of the output from the mismatch weights, the definition sensitivity $S(W)$ is as following:

$$S(W) = \sum_{k}(\sum_{\forall l,i,j} \left|W_{i,j}^{l}\right| \left|\frac{\partial O_k}{\partial W_{i,j}^{l}}\right|) \tag{2}$$

$\frac{\partial O_k}{\partial W_{i,j}^{l}}$ is the deviation of the K outputs from the W_{ij} in the layer l.

In order to further reduce the mismatch caused by the proposed approximate computing, we added some mismatch errors to the input of each layer in the BWN to optimize the mismatch error model. The training approach with mismatch errors to improve the robustness of the neural networks is called fault-tolerant training (FTT) method. In this work, the mismatch errors are sampled at a random values of a normal

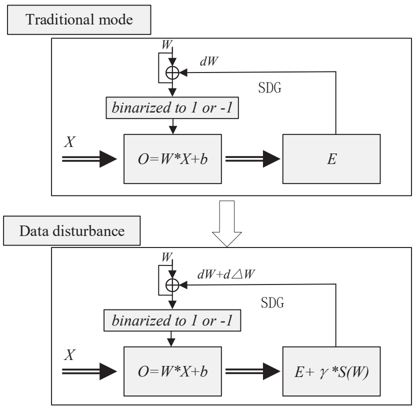

Fig. 6. BWN Fault Tolerance Training Method with Circuit Mismatch

TABLE I
RECOGNITION ACCURACY COMPARISON WITH FFT

SNR(dB)	Software accuracy	This work (without FTT)	This work (with FTT)
20dB	94.17%	91.28%	93.51%
10dB	92.53%	89.05%	91.30%
5dB	90.56%	87.33%	88.92%

TABLE II
COMPARISONS WITH OTHER SPEECH RECOGNITION ARCHITECTURES

	JSPS'18 [7]	VLSI'18 [9]	This work
Technology(nm)	40	28	28
Frequency(MHz)	50	2.5	2.5
Latency(ms)	10	25	20
Voltage(V)	0.6	0.57	0.6
DNN Structure	FC	CONV+FC	CONV+FC
Bit Width(Weight)	6	1	1
Bit Width(Data)	16	1	16
Computing Circuits	Standard Computing	Approximate Computing (digital)	Approximate Computing (digital-analog mixed)
Number of Keywords	10	1	10
Background Noise	NA	$\geq$ 5dB	$\geq$ 5dB
Power	11.2mW	141μW	56μW

distribution with a standard deviation of 0.8. The deviation value is based on the Monte Carlo simulation results of the proposed TAAU. With this approach, we can tune the BWN with circuit mismatch. The the cost functions will be updated by the proposed FTT method, and then the weights of the BWN with certain fault-tolerance are obtained. Table I shows that the recognition accuracy can be greatly improved with the proposed FTT method.

IV. IMPLEMENTATION RESULTS

The prototype system as shown in Fig.1 is implemented and evaluated on TSMC 28nm HPC+ process technology. The PE array of BWN accelerator is customized with Cadence Virtuoso Tool, and the other digital modules are described with Verilog HDL language and synthesized by Synopsys Design Compiler (DC). The SRAM blocks are functional with 0.72V. The BWN accelerator and other modules are functional with the logic supply voltage of 0.60V, and the working frequency is 2.5MHz. The timing and power consumption are evaluated with Synopsys HSIM at 25°C TT corner. The power consumption of proposed speech recognition processor is 56μW. We use the Google's Speech Commands database as our training and evaluating databases [16]. The chosen keywords are "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", along with "silence" and "unknown".

Comparisons with the state-of-the-art speech recognition processors are as shown in Table II. Compared to work [7] where the weight bit width of the adopted DNN is 6 bits, the weight bit width of the DNN used in work [9] and our work are both 1 bit. In work [9] and our work, the multiplication operations can be almost eliminated and over 99% of the operations are additions or bit-wise operations. Therefore, compared to work [7], the power consumption of the DNN accelerator can be significantly reduced. In work [9] and our work, the DNNs adopted for speech recognition contain both FC and CONV layers. The CONV layers can effectively improve the recognition accuracy under low weight bit width. In work [9], the proposed architecture is customized for a BNN where the bit width of both data and weights are 1 bit. To further reduce the energy consumption of the addition units, a digital approximate addition architecture is also proposed. Benefiting from the BNN and the approximate addition architecture, the power consumption of work [9] can be reduced to 141 μW. However, this work can only support one keyword recognition. In this paper, we use the BWN for speech recognition with data/weights bit width quantized as 16/1 bits. Compared to work [9], our work can support 10 keywords recognition under different background noise, while the power consumption can be significantly reduced to 56μW and the energy efficiency can be improved by 2.5×.

V. CONCLUSIONS

This paper proposed an ultra-low power speech recognition processor based on a binarized weight neural-network (BWN). To accelerate the BWN and make it energy efficient, we designed an approximate computing architecture for the quantized BWN based on the time-domain digital-analog mixed addition unit and precision optimization with fault-tolerant training method. Implementation results show that this accelerator can support 10 keywords real time recognition under different noise types and SNRs, while the power consumption is 56μW. Compared to the state-of-the-art architectures, this work can achieve up to 2.5× better in energy efficiency.

REFERENCES

[1] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *IEEE Int. Conf. Acoust., Speech Signal Process*, Vancouver, BC, Canada, pp. 7398–7402, 2013.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[3] S. Yin, P. Ouyang, S. Tang, F. Tu, X. Li, S. Zheng, T. Lu, J. Gu, L. Liu, and S.Wei, "A high energy efficient reconfigurable hybrid neural network processor for deep learning applications," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 53, no. 4, pp. 968-982, 2018.

[4] B. Liu, W. Dong, T. Xu, Y. Gong, W. Ge, J. Yang, and L. Shi, "E-ERA: An energy-efficient reconfigurable architecture for RNNs using dynamically adaptive approximate computing," *IEICE Electron. Express*, vol. 14, no. 15, pp. 1–11, 2017.

[5] M. Yang, C. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar and M. Seok, "A 1W voice activity detector using analog feature extraction and digital deep neural network," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 346-348, 2018.

[6] M. Price, J. Glass, and A. P. Chandrakasan, "A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 244–245, 2017.

[7] M. Shah, S. Arunachalam, J. Wang, D. Blaauw, D. Sylvester, H.-S. Kim, J.-S. Seo, and C. Chakrabarti, "A fixed-point neural network architecture for speech applications on resource constrained hardware," *Journal of Signal Processing Systems*, vol. 90, pp. 727–741, 2018.

[8] S. Bang, J. Wang, Z. Li, C. Gao, Y. Kim, Q. Dong, et al., "A 288 uW programmable deep-learning processor with 270kb on-chip weight storage using non-uniform memory hierarchy for mobile intelligence," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 250–251, 2017.

[9] S. Yin, P. Ouyang, S. Zheng, D. Song, X. Li, L. Liu, and S. Wei, "A 141μW, 2.46 pJ/neuron binarized convolutional neural network based self-learning speech recognition processor in 28nm CMOS," *IEEE Symposium on VLSI Circuits*, pp. 139–140, 2018.

[10] Z. K. Veton and A. E. Hussien, "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions," *J. Comput. Chem. Commun.*, vol. 3, no. 6, pp. 1–9, 2015.

[11] H. S. Wu, Z. Y. Zhang and M. C. Papaefthymiou, "A 13.8μW binaural dual-microphone digital ANSI S1.11 filter bank for hearing aids with zero-short-circuit-current logic in 65nm CMOS," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 348–349, 2017.

[12] K. M. H. Badami, S. Lauwereins, W. Meert and M. Verhelst, "A 90 nm CMOS, 6μW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 51, no. 1, pp. 291–302, 2016.

[13] R. Mohammad, O. Vicente, R. Joseph , et al, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," *arXiv preprint*, [Online], arXiv:1603.05279, 2016.

[14] P. Yin, S. Zhang, Y. Qi, et al,"Quantization and Training of Low Bit-Width Convolutional Neural Networks for Object Detection," *Journal of Computational Mathematics*, vol. 37, pp. 349–360, 2018.

[15] B. Liu, Z. Wang, H. Fan, J. Yang, W. Zhu, L. Huang, Y. Gong, W. Ge and L. Shi,"EERA-KWS: A 163 TOPS/W Always-on Keyword Spotting Accelerator in 28nm CMOS Using Binary Weight Network and Precision Self-Adaptive Approximate Computing," *IEEE Access*, vol. 7, pp. 82453–82465, 2019.

[16] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint*, [Online], arXiv:1804.03209, 2018.