A Ternary Weight Mapping and Charge-mode Readout Scheme for Energy Efficient FeRAM Crossbar Compute-in-Memory System

Tiancheng Cao[1,2], Zhongyi Zhang[1,2], Wang Ling Goh[1], Chen Liu[2], Yao Zhu[2], Yuan Gao[2]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Republic of Singapore
[2]Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Republic of Singapore
Email: CAOT0006@e.ntu.edu.sg

***Abstract*—This work presents an edge-AI system built on capacitive ferroelectric random-access memory (FeRAM) crossbar array, which is compatible with CMOS backend-of-line (BEOL) fabrication process. A novel capacitive crossbar circuit and a ternary mapping technique are proposed. Compared to the conventional binary representation, the proposed ternary mapping improves the storage efficiency exponentially in weight resolution. The feasibility of neuromorphic computing system implemented on FeRAM crossbar array is explored with speech command classification task. A ResNet-32 model with 0.45M parameters is implemented on 64×64 FeRAM crossbar array with the measured FeRAM model. It achieved 97.12% inference accuracy with 2 ternary digits and 5% device variation on Google Speech Command dataset 35-command classification task.**

Keywords—Capacitive crossbar, FeRAM, Speech recognition, Edge-AI, neuromorphic system

I. INTRODUCTION

Computing-in-memory (CIM) [1] has shown its remarkable compatibility with Artificial Intelligence (AI) algorithms implementation on edge devices [2]. Resistive synapses like resistive random-access memory (RRAM) and phase change memory (PCM) for CIM have attracted much research interest recently [3] because of its tunable resistance and advantage to compute the vector-matrix multiplication (VMM) operation which is intensively used in AI in parallel. However, the high static read current and leakage current are inevitable for resistive synapse, which results in high static power especially for on-state devices with only a few thousand ohms. Although an extra transistor can be introduced to form 1T1R structure [4] to eradicate sneak path and static current of unselected cell, the increase of cell area and circuit footprint is the problem. Furthermore, the random telegraph noise (RTN) for small scale current-driven device will also challenge the performance of the resistive synapses [5].

Capacitive crossbar array with capacitive synapse device has been proposed recently as a potential candidate to overcome the aforementioned challenges [6-7]. With capacitive devices, the only power consumption reading operation of the crossbar is the transient power following the principle of charge sharing and preservation. As a result, the capacitive synapse has great advantage in energy efficiency and much smaller cell area compared to resistive synapse [8]. Capacitive crossbar arrays with different ferroelectric random-access memory (FeRAM) synapses like HfO_2 [9-11], SiN [12] and HZO [13] have been reported. However, all the above synapses have sparse coercive field distribution, which means the state of the cell is sensitive to the applied voltage [14]. Therefore, precise capacitance tuning on high density array is unfeasible due to the sensitive half-selected cells on the same row or column with the selected cell. Furthermore, the input of the CIM is restricted to a small window to avoid the capacitance fluctuation with different input voltage. Some methods have been proposed to improve the neural network mapping efficiency. For example, multi-level capacitance has been presented under different program voltage [9], but the 11 levels are concentrated within 1 pF and sensitive to the read voltage obviously. Implementing weight by multiple columns with digital shift & add is another attractive solution [13], but the extra circuit and binary cell limit the efficiency of the crossbar array. Moreover, to our knowledge, previous works evaluate small capacitive crossbar array on simple dataset due to the restriction of the reported capacitive synapse.

The ferroelectricity is recently found in the aluminum scandium nitride ($Al_xSc_{1-x}N$) thin film in 2019 [15]. Unlike HfO_2, the distribution of the coercive field of $Al_xSc_{1-x}N$ is quite tight [15], making it possible to construct the selector-free array without unexpected state changing. Derived from the experiments on a CMOS (BEOL) compatible FeRAM array, [16], the edge-AI application framework implemented on $Al_{0.7}Sc_{0.3}N$-based capacitive crossbar array are explored in this work. Readout circuit for FeRAM-based synapse is designed and a high efficiency ternary weight mapping method is proposed. To evaluate the performance of the framework, a speech recognition task is performed with ResNet [17] on the proposed framework. The rest of the paper is organized as follows. Section II introduce the readout circuit design for FeRAM array and the proposed ternary mapping. Section III presents the and evaluate the speech recognition system implemented on capacitive crossbar array. Section IV concludes the paper.

This work was supported by the Agency for Science, Technology and research (A*STAR), Singapore under the Nanosystems at the Edge programme, grant no.A18A1b0055 and the Ferroelectric Aluminum Scandium Nitride ($Al_{1-x}Sc_xN$) Thin Films and Devices for mm-Wave and Edge Computing programme, grant no. A20G9b0135.

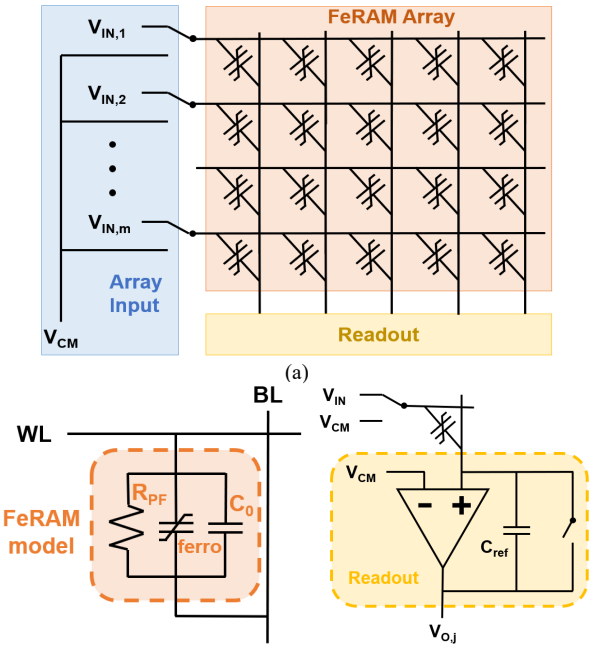


Fig. 1. (a) The structure of general FeRAM capacitive crossbar array with word line connected to array input and bit line connected to readout circuit. (b) The circuit model of an $Al_{0.7}Sc_{0.3}N$-based FeRAM cell. (c) The structure of the proposed readout circuit.

II. Capacitive FeRAM Crossbar

A. Capacitive Crossbar and Readout Circuit

The structure of a general capacitive crossbar array is shown in Fig. 1(a). The synaptic devices are assigned at every cross point between word line (WL) and bit line (BL). The voltage collected at each BL is the sum of multiplication of WL input voltage and capacitance of synaptic device. Thanks to the tight coercive field distribution of $Al_{0.7}Sc_{0.3}N$-based synapse, it is possible to build an ultra-high density selector-free FeRAM array. Furthermore, input voltage can be represented in analog way and implemented in only one cycle without concerns about the state change of FeRAM.

The readout process of capacitive crossbar includes two phases, charging phase and charge transfer phase [8]. However, conventional readout circuit in two phases is not appropriate for FeRAM -based synapse due to leakage current and the circuit model of the synapse cell is shown in Fig. 1(b). Therefore, a novel readout circuit shown in Fig. 1(c) is designed in this work.

During the charging phase, output switch is closed to avoid the influence of leakage current on reference capacitance, C_{ref}, BLs are connected to common-mode voltage, V_{CM}, and input vector is transferred to voltage, V_{IN}, applied to WLs. The synapse is then charged based on the input voltage and the FeRAM state. During the charge transfer phase, output switch is open, BLs and WLs are connected to V_{CM}. Leakage current is neglected because of zero potential difference. Thereafter, the charges on the synapses is transferred to C_{ref} at the end of each BL. The output voltage of the j^{th} column, $V_{O,j}$, is represented by following equation:

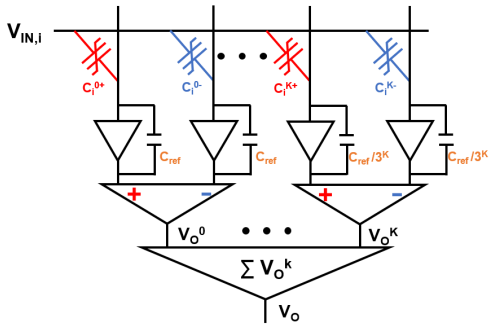
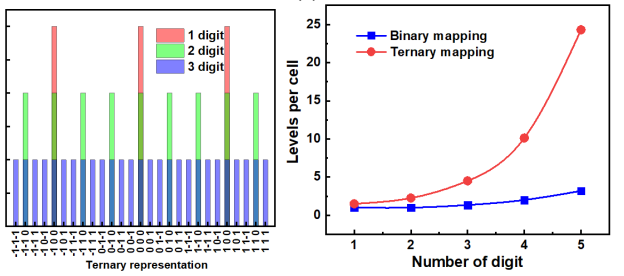


Fig. 2. The diagram of one weight representation with the proposed ternary mapping. Two column bit lines represent one digit, one for positive and one for negative. Each digit is different from reference capacitance and summed.

Positive state	Negative state	Positive weight	Negative weight	Reference	Equivalent weight
C_{ON}	C_{OFF}	1	0	$C_{ref}/3^k$	1
C_{OFF}	C_{ON}	0	1	$C_{ref}/3^k$	-1
C_{OFF}	C_{OFF}	0	0	$C_{ref}/3^k$	0

(a)



(b) (c)

Fig. 3. (a) The working principle of the proposed ternary digits on FeRAM. (b) The example of 3-ternary-digit representation. (c) The mapping efficiency with increasing number of digits.

$$V_{O,j} = \frac{\sum V_{IN,i} \times C_{i,j}}{C_{ref}} \quad (1)$$

where $V_{IN,i}$ is the input voltage of i^{th} row and $C_{i,j}$ is the conductance of the FeRAM at i^{th} row and j^{th} column.

B. Ternary Mapping Method

To represent the negative weight in FeRAM, differential pair is employed here [18]. Two adjacent FeRAM cells are used to represent one weight, one cell for positive and another for negative. The interim output voltage collected from BLs are subtracted to get the differential output as:

$$V_O = V_O^+ - V_O^- = \frac{\sum V_{IN,i} \times (C_i^+ - C_i^-)}{C_{ref}} \quad (2)$$

where C_i^+ and C_i^- are the conductance of differential pair FeRAM at i^{th} row.

Because of the small on/off ratio of FeRAM, binary synapse is chosen for large array considering the precision of weight and conventional binary mapping method implements multi-bit weight by multiple columns and shift&add. To improve the efficiency of the weight mapping, a ternary mapping method is

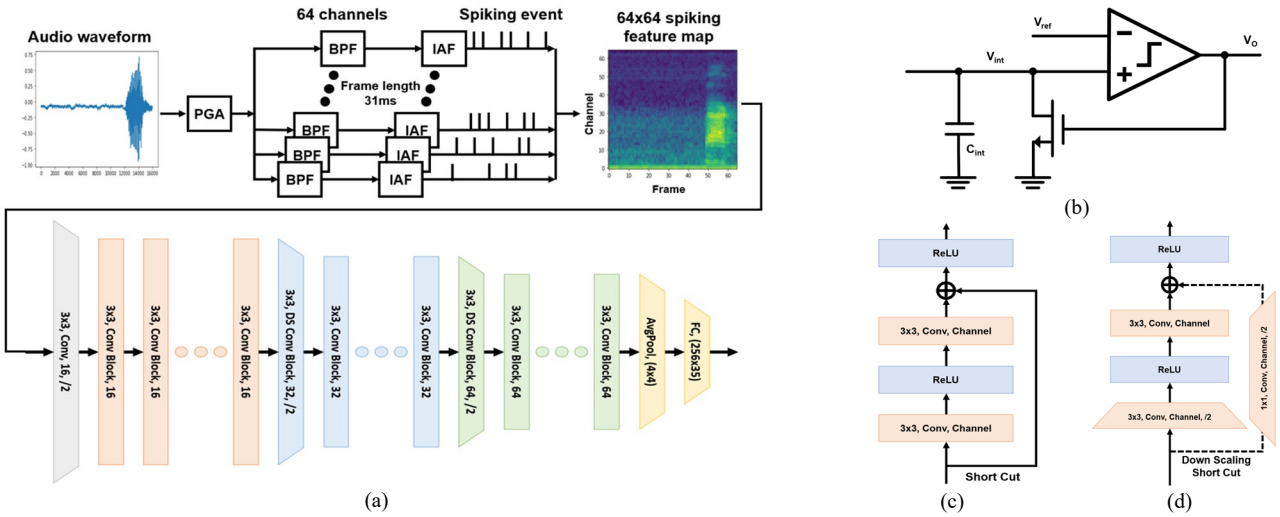


Fig. 4. (a) The diagram of the speech recognition system with feature extraction module and ResNet implemented by FeRAM crossbar array. (b) The circuit of integrate-and-fire (IAF) module. (c) Architecture of convolution block (Conv Block), consisting of convolution, ReLU and short cut operations. (d) Architecture of down scaling convolution block (DS Conv Block), consisting of convolution, ReLU and down scaling short cut operations.

proposed here. The example of K ternary digits is given in Fig. 2. To map k^{th} ternary digit, $k = 0, 1, \dots, K$, the reference capacitance is enlarged 3^k times to decrease the weight of digit. The output voltage can be described as following equation:

$$V_O = \sum {V_O}^k = \sum_k \frac{\sum_i V_{IN,i} \times ({C_i}^{k+} - {C_i}^{k-})}{3^k \times C_{ref}} \quad (3)$$

where ${V_O}^k$ is the k^{th} ternary digit of output.

The principle of the proposed ternary digits is shown in Fig. 3(a) and the 3-ternary-digit levels in Fig.3(b). Compared to conventional binary mapping, the area efficiency of the proposed ternary mapping shows great potential with the increase of digit number and the trend is described in Fig. 3(c).

III. FeRAM Crossbar-based Keyword Spotting System

To demonstrate the feasibility of the proposed FeRAM - based capacitive crossbar array and ternary mapping method on edge-AI applications, a speech recognition system of spiking feature extraction and ResNet model is built for Google Speech Commands dataset [19]. As shown in Fig. 4(a), the system transfers the waves into feature maps and applied the ResNet model implemented by capacitive crossbar array.

A. Spiking Feature Extraction

In the design of a 64-channel analog spiking feature extraction (SFE) system, shown in Fig. 4(a), the system exhibits strong performance in accurately detecting and responding to variations in the amplitude, frequency, and noise characteristics of speech signals. This ensures a high degree of robustness in the system's ability to extract relevant features from the input signal.

The first block in the analog SFE is a programmable gain amplifier (PGA) [20] that can adjust its gain from -12dB to 12dB in order to accommodate various microphone signal intensities. The PGA employs a two-stage, fully differential, Miller-compensated operational amplifier with an active-RC configuration. This amplifier exhibits an open-loop DC gain of 58 dB and has a unity gain bandwidth of 100 kHz. The speech signal is subsequently input into a 64-channel bandpass filter (BPF) [21] with a Q-factor that can be varied from 1 to 10 and a frequency range from 20 to 4000 Hz. Finally, the features generated by the BPF are transformed into asynchronous events by the integrate-and-fire (IAF) module, and then transmitted to the digital ResNet model for further processing.

B. ResNet Implemention

ResNet is a widely used deep convolutional neural network model in AI applications. VMM operation is the foundation of the key components of ResNet, such as identical dimension convolutional layer, down scaling short cut and fully-connected classifier. Consequently, FeRAM-based capacitive crossbar array is a potential candidate for low power implementation.

The mapping of convolutional layer and fully-connected layer on resistive crossbar array have been discussed in previous works [18] and is applicable for capacitive crossbar in this work. Besides, Down scaling short cut in this work is a modified 2-D convolutional operation with 1 kernel size and 2 stride size. Thus, complete implementation of ResNet on capacitive crossbar array is feasible and simulated in this work.

IV. Simulation Results

The synapse model is built with the experiments on multiscale modeling of $Al_{0.7}Sc_{0.3}N$-based FeRAM [16]. The individual device area is $10^{-6} cm^2$ with measured C_0 around $0.15 nF/cm^2$ and 64×64 array size is selected. The system is trained 200 epochs on a 16-core 3.2GHz AMD Ryzen 5800H CPU with 32GB RAM with the non-idealities-aware training [18].

To find the appropriate model for the system, three models with different sizes are evaluated in software, ResNet-20,

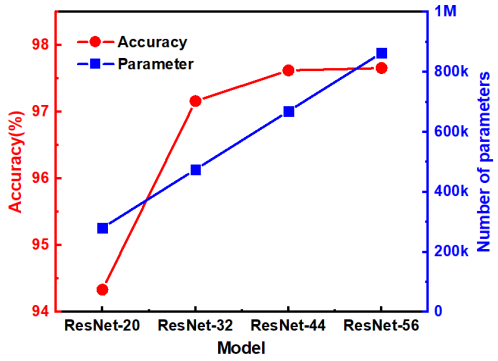
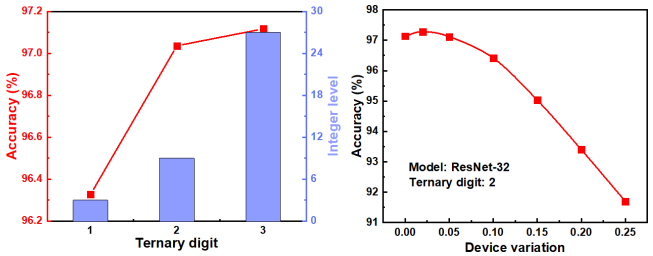


Fig. 5. The performance comparison of different ResNet models on speech recognition task.



(a) (b)

Fig. 6. (a) The effect of ternary digits on ResNet-32 model with corresponding integer levels. (b) The impact of device variation (σ/μ) on inference accuracy.

ResNet-32, ResNet-44 and ResNet-56, with 0.27M, 0.45M, 0.64M and 0.82M parameters, respectively. The performance of different models on speech recognition task is given in Fig. 5 where ResNet-32 shows high recognition accuracy with small model size. Therefore, ResNet-32 is chosen for the system implementation in following evaluation.

Thereafter, the proposed ternary mapping is applied to quantize the ResNet-32 model. Fig. 6(a) illustrates how the ternary resolution affects training accuracy. The results show high efficiency of ternary mapping. The training accuracy achieves over 97% with only 2 ternary digits, 9 quantized integer levels.

Although FeRAM device is relatively more stable during the reading procedure, device-to-device variation still affects the precision of the synapse. Fig. 6(b) shows the impact of device variation (σ/μ) to inference accuracy. It is worth noting that with the increase of the variation, the system shows its robustness and still achieves >97% accuracy with up to 5% device variation. Besides, because of the small random unpredicted variation during the training, the model reduces the overfitting effect, and the system even performs better with 2% device variation.

The performance of this work is compared with other recently reported CIM solutions for speech recognition in Tab. I. The performance of the final system built with Python on the Google Speech Commands dataset 35-command task is shown in Fig. 7. The proposed FeRAM crossbar with ternary mapping achieved outstanding 97.12% inference accuracy with only 2 ternary digits and 5% high device variation. The overall network size is reduced significantly to only 0.45M parameters.

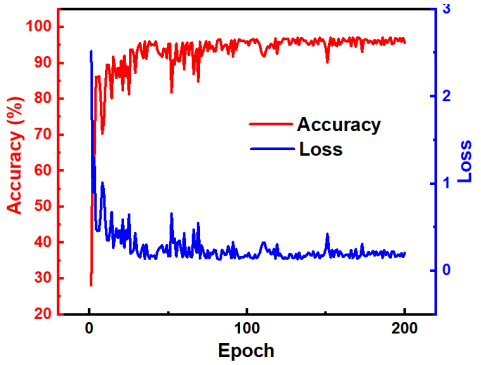


Fig. 7. The performance of the final $Al_{0.7}Sc_{0.3}N$ -based FeRAM crossbar array system on the Google Speech Commands dataset 35-command task

Table. I. Performance comparison

	[22]	[23]	[24]	[25]	This work
Dataset	Google Speech Command 12	Google Speech Command 35	Google Speech Command 10	Google Speech Command 12	Google Speech Command 35
Architecture	RNN	RNN	CNN	SNN	CNN
Model	LSTM	LSTM	NR	NR	ResNet-32
Parameter	NR	0.27M	NR	0.22M	0.45M
Device	ReRAM	ReRAM	SRAM	PCM	FeRAM
Device Level	analog	8	2	1k	2
Cell Structure	1T1R	1T1R	8T	3T1R	1C
Crossbar Size	128×128	16×16	96×288	NR	64×64
Accuracy	90%	85%	91%	78%	97%

Considering the power and area efficiency of FeRAM crossbar array, the proposed system is a promising candidate for edge-AI applications.

V. Conclusion

In this work, an edge-AI speech recognition system implemented by CMOS BEOL compatible $Al_{0.7}Sc_{0.3}N$-based FeRAM capacitive crossbar array is presented. Thereafter, a customized capacitive crossbar circuit and a novel ternary mapping method are proposed for CIM architecture. The proposed ternary mapping significantly increases the storage efficiency in weight resolution as compared to the conventional binary representation. With the modeling of FeRAM-based synapse, a feature extraction system and 0.45M ResNet-32 model are implemented on 64 × 64 capacitive crossbar array for Google Speech Command dataset 35-command classification task. The system showed its outstanding performance in AI implementation and mapping efficiency and achieved 97.12% inference accuracy with 2 ternary digits and 5% device variation (σ/μ). The proposed system provides a potential solution to ferroelectric capacitor implementation as the fundamental component of a capacitive artificial neural network for neuromorphic edge computing.

REFERENCES

[1] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260-285, Feb. 2018.

[2] C. Wang et al., "Multi-state memristors and their applications: an overview," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 12, no. 4, pp. 723-734, Dec. 2022.

[3] Y. Xi et al., "In-memory learning with analog resistive switching memory: a review and perspective," *Proc. IEEE*, vol. 109, no. 1, pp. 14-42, Jan. 2021.

[4] X. Peng, S. Huang, H. Jiang, A. Lu and S. Yu, "DNN+NeuroSim V2.0: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators for On-Chip Training," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 11, pp. 2306-2319, Nov. 2021.

[5] J. Kang et al., "Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition," *2017 IEEE International Electron Devices Meeting (IEDM)*, pp. 6.4.1-6.4.4, 2017.

[6] Q. Zheng et al., "Artificial neural network based on doped HfO2 ferroelectric capacitors with multilevel characteristics," *IEEE Electron Device Lett.*, vol. 40, no. 8, pp. 1309–131, Aug. 2019.

[7] D. Kwon and I.-Y. Chung, "Capacitive neural network using chargestored memory cells for pattern recognition applications," *IEEE Electron Device Lett.*, vol. 41, no. 3, pp. 493–496, Mar. 2020.

[8] Y. -C. Luo, A. Lu, J. Hur, S. Li and S. Yu, "Design of non-volatile capacitive crossbar array for in-memory computing," *2021 IEEE International Memory Workshop (IMW)*, pp. 1-4, 2021.

[9] Q. Zheng et al., "Artificial neural network based on doped HfO2 ferroelectric capacitors with multilevel characteristics," *IEEE Electron Device Lett.*, vol. 40, no. 8, pp. 1309-1312, Aug. 2019.

[10] Z. Wang, M. Rao, JW. Han, et al., "Capacitive neural network with neuro-transistors," *Nat. Commun.*, vol. 9, no. 3208, Aug. 2018.

[11] U. Schroeder, M.H. Park, T. Mikolajick, et al., "The fundamentals and applications of ferroelectric HfO_2," *Nat. Rev. Mater.*, vol. 7, pp. 653–669, Mar. 2022.

[12] D. Kwon and I. -Y. Chung, "Capacitive neural network using charge-stored memory cells for pattern recognition applications," *IEEE Electron Device Lett.*, vol. 41, no. 3, pp. 493-496, Mar. 2020.

[13] Y. -C. Luo, A. Lu, J. Hur, S. Li and S. Yu, "Design and optimization of non-volatile capacitive crossbar array for in-memory computing," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 69, no. 3, pp. 784-788, Mar. 2022.

[14] T. Mikolajick et al., "Next generation ferroelectric materials for semiconductor process integration and their applications," *J. Appl. Phys.*, vol. 129, no. 10, Mar. 2021.

[15] S. Fichtner, N. Wolff, F. Lofink, L. Kienle, and B. Wagner, "AlScN: A III-V semiconductor based ferroelectric," J. Appl. Phys., vol. 125, no. 11, 2019, doi: 10.1063/1.5084945.

[16] C. Liu et al., "Multiscale modeling of Al0.7Sc0.3N-based FeRAM: the steep switching, leakage and selector-free array," *2021 IEEE International Electron Devices Meeting (IEDM)*, pp. 8.1.1-8.1.4, 2021.

[17] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[18] T. Cao et al., "A non-idealities aware software–hardware co-design framework for edge-AI deep neural network implemented on memristive crossbar," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 12, no. 4, pp. 934-943, Dec. 2022.

[19] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," arXiv:1804.03209, 2018, [online] Available: https://arxiv.org/abs/1804.03209.

[20] 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.

[21] M. Yang, C. -H. Chien, T. Delbruck and S. -C. Liu, "A 0.5V 55µW 64×2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2554-2569, Nov. 2016.

[22] C. J. Schaefer, M. Horeni, P. Taheri and S. Joshi, "LSTMs for keyword spotting with ReRAM-based compute-in-memory architectures," *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, Daegu, Korea, pp. 1-5, 2021.

[23] W.Wan, R.Kubendran, C.Schaefer *et al.*, "A compute-in-memory chip based on resistive random-access memory," *Nature* 608, pp.504–512, 2022.

[24] X. Jia *et al*., "A 11.6µW computing-on-memory-boundary keyword spotting processor with joint MFCC-CNN ternary quantization," *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, Austin, TX, USA, pp. 2816-2820, 2022.

[25] V. P. K.Miriyala, M.Ishii, "Ultra-low power on-chip learning of speech commands with phase-change memories," arXiv: 2010.11741, 2020, [online] Available: https://arxiv.org/abs/2010.11741.