TinyML Acoustic Classification using RAMAN Accelerator and Neuromorphic Cochlea

Adithya Krishna[1,2], Shankaranarayanan H[1], Hitesh Pavan Oleti[1], Anand Chauhan[1], André van Schaik[2], Mahesh Mehendale[1] and Chetan Singh Thakur[1]

[1]Department of Electronic Systems Engineering, Indian Institute of Science, Bangalore, India
[2]International Centre for Neuromorphic Systems, The MARCS Institute, Western Sydney University, Australia

***Abstract*—The growing use of acoustic classification systems in edge computing and Internet of Things (IoT) devices has created a demand for innovative technologies and methods that can deliver high performance and energy efficiency. In this work, we introduce a novel audio inference system that combines RAMAN, a *R*e-configurable and sp*A*rse tiny*ML* *A*ccelerator for infere*N*ce, with a hardware-efficient Neuromorphic cochlea for pre-processing. The neuromorphic cochlea mimics human hearing, specifically by employing the 'Cascade of Asymmetric Resonators (CAR)' model to replicate the basilar membrane filter in the human cochlea. In this study, we utilize a 30 cascaded-filter cochlear section to process real-time audio data and a RAMAN classifier for audio classification. RAMAN leverages activation and weight sparsity within the neural network to reduce storage, latency, and power consumption. The proposed audio inference system has been implemented on a Microchip MPFS250T SoC field-programmable gate array (FPGA) with 52.57k LUTs, all while operating with a power consumption of 237.3 mW at 40 MHz clock frequency. The proposed audio inference system is designed for low-power auditory edge applications such as speaker verification, speech detection, and keyword spotting.**

I. Introduction

Edge computing has transformed computational tasks by moving data processing and computation closer to the data source, enabling real-time analysis and rapid responses [1]. This shift away from centralized cloud servers to network edge devices is particularly significant in the context of deep neural networks (DNNs), which are widely used in various cognition and learning domains [2]–[4]. The trend of deploying DNNs directly on the edge has gained momentum due to its inherent benefits, including enhanced privacy, reduced latency, and optimized bandwidth utilization. However, challenges arise when performing computations on edge devices, such as power constraints, limited memory, and resource limitations, preventing the direct deployment of power-intensive GPUs and CPUs commonly found in cloud platforms. Specialized accelerators designed for neural computations at the edge are being developed to address these challenges. These accelerators enable improved data flow, efficient memory access, and the exploitation of network sparsity. One exciting application of this advancement is in audio classification on the edge, enabling real-time identification and categorization of audio signals with faster response times and reduced reliance on cloud resources.

In this paper, we introduce a novel approach for audio inference at the edge. Our method employs a neuromorphic cochlea as a pre-processing stage to emulate the auditory function of the human ear. This cochlea model incorporates a Cascade of Asymmetric Resonators (CAR) model [5] and a low-pass filter to replicate the behavior of in Basilar membrane and Inner hair cells found within the human inner ear. Additionally, we utilize the RAMAN tiny ML accelerator [6] for audio inference at the edge.

The neuromorphic cochlear pre-processing stage is highly compact, resource and energy-efficient compared to the existing mel-frequency cepstral coefficients (MFCCs) based pre-processing and it is shown that the cochlea outperforms MFCCs across various experimental conditions in terms of classification accuracy [7], [8].

II. Methodology

Figure 1 presents the audio inference pipeline using Neuromorphic cochlear pre-processing and RAMAN tinyML accelerator for classification. The setup involves capturing the input audio from the speaker through a microphone and subsequently, feeding it to the I^2S (Inter-IC Sound) module to obtain a 16-bit representation of the audio sample. The I^2S module assumes a crucial role in this process by facilitating the conversion and generating the necessary control, enable, and clock signals required for the proper functioning of the microphone module.

The resulting 16-bit audio sample sampled at 16 KHz serves as the input for the CAR-IHC cochlear module, which acts as a pre-processing stage. In this work, we have employed the CAR-IHC model proposed by Xu et al. [5] utilizing 30 filter banks, each specifically tuned to a distinct cut-off frequency. Consequently, the output of each filter bank is tapped for 1s duration generating a cochleagram of size 30×16000. The generated cochleagram is then downsampled to size 30×32 by employing a max pooling operation. The down-sampled cochleagram is subsequently fed as an input to the RAMAN tinyML accelerator for audio classification.

The RAMAN accelerator [6] is a compact, low-power, and energy-efficient deep neural network accelerator comprising memory, compute, and control subsystems. It employs a 3×4 processing element (PE) array for multiply-accumulate (MAC) operations, with a three-level memory hierarchy comprising global memory to store activations and parameters, a cache, and a reg-file. The salient features of the RAMAN architecture

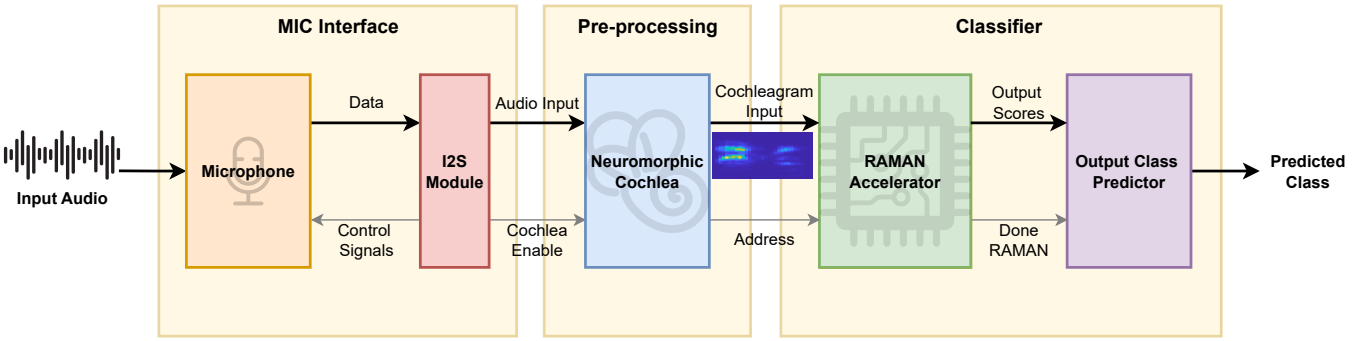

Fig. 1: The audio inference pipeline using neuromorphic cochlear pre-processing for audio feature extraction and RAMAN tinyML accelerator for audio classification.

involve employing both activation and weight sparsity to reduce latency and storage, optimized dataflow to reduce memory access, and memory optimizations like compression and peak activation memory storage reduction by cache prefetching and intelligent operation scheduling.

III. Results

We implemented a sophisticated audio inference system, leveraging a 30-CAR-section cochlear model for efficient preprocessing and the tinyML accelerator RAMAN, running in real-time on an FPGA platform. The design is extensively optimized and subsequently deployed on Microchip MPFS250T SoC FPGA, utilizing 52.57k 4-input look-up tables (LUTs) and 18.9k flip flops (FFs). The memory utilization of the system is 564 uSRAM blocks and 50 LSRAM blocks. The total power consumption of the design is 237.288 mW (123.596 mW static power and 113.692 mW dynamic power) at 40 MHz clock frequency. The resource utilization of the implemented acoustic classifier system is presented in Table I.

TABLE I: Resource utilization on MPFS250T SoC FPGA.

Resource	Used	Total	Utilization (%)
LUTs (4-input)	52.57k	254.2k	20.7
Flip Flops	18.89k	254.2k	7.43
uSRAM(64x12)	564	2352	23.98
LSRAM(1Kx20)	50	812	6.16
Math Blocks (DSPs)	82	784	10.46

IV. Conclusion

Real-time audio classification has garnered significant attention in recent years due to its broad range of applications. In this study, we have developed the audio inference system that utilizes neuromorphic cochlea as a pre-processing stage and RAMAN tinyML accelerator. The neuromorphic cochlea serves as a hardware-efficient pre-processing module that replicates the functionality of human hearing. We have implemented the "Cascade of Asymmetric Resonators (CAR) - Inner Hair Cells (IHC)" model of the cochlea on an FPGA. The CAR component mimics the basilar membrane filter, while the IHC component emulates the inner hair cells of the cochlea. The overall system has been implemented on Microchip MPFS250T SoC FPGA with 52.57k LUTs, 564 uSRAM blocks, and 50 LSRAM blocks while consuming 237.288 mW of power.

V. Acknowledgements

This work is funded by the SPARC Funds under Grant SP/MHRD-18-0006, in part by INAE under Grant SP/INAE-22-2106, and in part by the Pratiksha Trust and BCD Funds under Grant FG/SMCH-22-2106. We acknowledge the technical support provided by the APCCAS Design Committee and Microchip Technology India Pvt. Ltd.

References

[1] G. Carvalho, B. Cabral, V. Pereira, and J. Bernardino, "Edge computing: current trends, research challenges and future directions," *Computing*, vol. 103, pp. 993 – 1023, 2021.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1097–1105.

[4] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[5] Y. Xu, C. S. Thakur, R. K. Singh, T. J. Hamilton, R. M. Wang, and A. van Schaik, "A FPGA Implementation of the CAR-FAC Cochlear Model," *Frontiers in Neuroscience*, vol. 12, p. 198, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2018.00198

[6] A. Krishna, S. R. Nudurupati, D. G. Chandana, P. Dwivedi, A. van Schaik, M. Mehendale, and C. S. Thakur, "RAMAN: A Re-configurable and Sparse tinyML Accelerator for Inference on Edge," *arXiv*, 2023. [Online]. Available: http://arxiv.org/abs/2306.06493

[7] M. Buermann and T. van Meer, "Speech recognition using very deep neural networks: Spectrograms vs cochleagrams," 02 2020.

[8] Y. Muthusamy, R. Cole, and M. Slaney, "Speaker-independent vowel recognition: spectrograms versus cochleagrams," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 533–536.