HyIMC: Analog-Digital Hybrid In-Memory Computing SoC for High-Quality Low-Latency Speech Enhancement

Wanru Mao[1,2,*],Hanjie Liu[1,2,*],Guangyao Wang[1,2,*],Tianshuo Bai[1], Jingcheng Gu[1], Han Zhang[1], Xitong Yang[2], Aifei Zhang[2], Xiaohang Wei[2], Meng Wang[2], Wang Kang[1], Senior Member, IEEE
[1]School of Integrated Circuit Science and Engineering, Beihang University, Beijing, China
[2]Zhicun Research Lab, Beijing, China
Email: wang.kang@buaa.edu.cn, *Equal Contributions

***Abstract*—In-memory computing (IMC) holds significant promise for accelerating deep learning-based speech enhancement (DL-SE). However, existing IMC architectures face challenges in simultaneously achieving high precision, energy efficiency, and the necessary parallelism for DL-SE's inherent temporal dependencies. This paper introduces HyIMC, a novel hybrid analog-digital IMC architecture designed to address these limitations. HyIMC features: 1) a hybrid analog-digital design optimized for DL-SE algorithms; 2) a schedule controller that efficiently manages recurrent dataflow within skip connections; and 3) non-key dimension shrinkage, a model compression technique that preserves accuracy. Implemented on a 40nm eFlash-based IMC SoC prototype, HyIMC achieves 160 TOPS/W energy efficiency, compresses the DL-SE model size by ∼600%, improves the feature of merit by ∼1200%, and enhances perceptual evaluation of speech quality by ∼120%.**

***Index Terms*—Speech enhancement, In-memory computing, eFlash memory, Analog-digital hybrid architecture**

I. Introduction

Speech enhancement, crucial for applications from communication to hearing aids, aims to improve the quality and intelligibility of speech signals [1]. Deep learning-based speech enhancement (DL-SE) has demonstrated superior pattern extraction capabilities [2], but its computational and memory demands pose challenges for deployment on resource-constrained hardware [3]. In-memory computing (IMC), by performing computations within memory, offers a promising solution to mitigate data movement bottlenecks. Various memory technologies, including eFlash, are being explored for IMC [4]. IMC can be implemented digitally or analogically. Digital IMC offers high precision but suffers from power and area overhead, making it more suitable for cloud computing [5], [6]. Analog IMC provides high energy efficiency, ideal for DL inference on edge devices, but is limited by SNR and basic operation support [7], [8].

Existing state-of-the-art (SOTA) DL-SE models, such as DCCRN, DTLN, DC-UNet, and DCTCRN [9]–[12], primarily consist of convolutional layers (CLs) and recurrent layers (RLs), where matrix-vector multiplication (MVM) is the dominant computation and parameter access is a key bottleneck in RLs [13]. To address these challenges, we propose a hybrid in-memory computing (HyIMC) SoC based on eFlash memory for energy-efficient DL-SE. Our main contributions are: 1) a hybrid analog-digital IMC architecture for high-quality, low-latency DL-SE; 2) a schedule controller optimizing recurrent dataflow in skip connections; and 3) non-key dimension shrinkage (NKDS), a model compression technique that preserves accuracy.

This work was supported by the Beijing Natural Science Foundation (L223004) and Beijing MSTC Program (Z231100007423019).

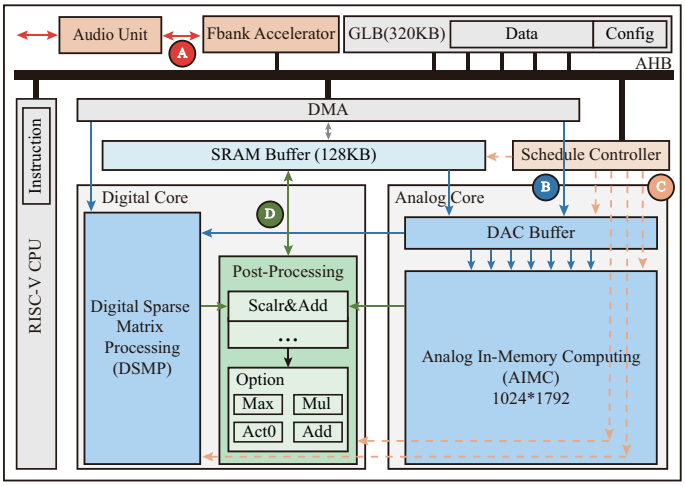

Fig. 1. The overall architecture of the HyIMC SoC and critical modules.

II. Chip Architecture

In HyIMC (Fig. 1), four modules are designed to support all operations in the speech enhancement process, including signal processing as well as CLs, RL, and other layers in DL-SE.

Fbank accelerator (A) works for signal feature extraction and inverse signal process.

The quantized weights of CLs in DL-SE approximately follow a Gaussian distribution, with over 90% of the weights being within 7 bits. The 8-bit weights are divided into high-order and low-order. We design a digital sparse matrix processing (DSMP) unit to handle the high-bit weights, while the analog in-memory computing (AIMC) unit processes the dense low-bit weights. Inputs are simultaneously sent to both

978-3-9826741-0-0/DATE25/© 2025 EDAA

AIMC and DSMP (B). The results from both parts are then processed through a post-processing module involving shifting and summation to generate the final output. This hybrid processing approach effectively enhances inference accuracy while maintaining computational power.

1) AIMC: The AIMC uses an eFlash-based cross-array structure, consisting of 1792 rows of weights and 16 rows of biases, with 1024 columns per row, totaling a capacity of 1.8MB, meeting the deployment requirements of the compressed DL-SE model. In the input module, we design a parallel sliding window mechanism to update the concatenated speech signal data across multiple frames, updating only a small portion of the data at a time, which further reduces data movement. In the array computation module, block-based activation is employed, effectively reducing the power consumption.

2) DSMP: Supports the use of multi-bit and mixed-precision modes across different layers.

a) multi-bit mode combination: The 1-bit mode decomposes 8-bit weights into high 1-bit and low 7-bit components. Additionally, 2-bit and 3-bit modes have been developed to better suit the requirements of different network layers.

b) mixed-precision modes: AIMC maintains 8-bit precision, while DSMP adjusts the weight bit-width to achieve variable precision of 16, 12, or 10 bits.

Schedule controller (C) manages time-dependent operations, such as RL jump connections and loops. It reads configuration data and updates all module settings. These configurations and scheduling optimized are generated by the compiler [14].

Post-processing module (D) handles the results of B, as well as other digital computations, including activation, pooling, and other tasks.

III. Proposed NKDS Design

We use optimized, noise-aware quantization-aware training (inspired by IMC systems [15]) to reduce DL-SE model mapping complexity. NKDS prunes redundant features based on task relevance (layer/channel-wise sensitivity) and IMC chip hardware adaptability (complex-valued convolutions). This achieved $\sim$600% size reduction vs. DCCRN with comparable inference accuracy (Table I).

IV. Evaluations and Results

A. Experiment Setup

1) Platform: To evaluate this architecture, we design a HyIMC SoC using 40nm technology and wafer-level chip scale packaging (WLCSP).

2) Dataset and models: We evaluate the NKDS compressed DCCRN on HyIMC with the ICASSP2020 DNS Challenge dataset and compare it with the original baseline model and three outstanding models in the competition, which are all tested on Intel quad-core i5 CPU.

B. Experiment Results

Table I presents a comparative analysis with prior SOTA DL-SE research. To simplify this comparison, we introduce a feature of merit (FoM) defined in (1):

$$FoM = \left(\frac{\text{PESQ}}{\text{Model Size} \times \text{Latency}} \right) \times 100\% \qquad (1)$$

TABLE I
Comparison of DL-SE

Model	Model Size	PESQ	Latency (ms)	FoM
Noisy	-	2.45	-	-
DNS Baseline [9]	1.3 MB	2.68	40	5.15%
DCCRN [9]	3.7 MB	3.27	62.5	1.41%
DCUNET [10]	3.6 MB	3.22	62.5	1.43%
DTLN [11]	987 KB	3.04	40	7.88%
Our Work	**640 KB**	**3.16**	**30**	**16.85%**

When compared to the original DCCRN implemented on an Intel quad-core i5 CPU, which was the top-performing model in the ICASSP2020 DNS Challenge [9], our work achieves approximately a 600% reduction in model size, a 210% improvement in latency, and a 1200% enhancement in FoM, while maintaining similar perceptual evaluation of speech quality (PESQ [16]) scores (ranging between -0.5 and 4.5, with higher scores indicating better quality). Additionally, our work demonstrates a 2.1$\times$ reduction in model size, a 1.2$\times$ increase in PESQ, and a 1.3$\times$ decrease in latency compared to the DNS baseline. These findings underscore the efficacy of the proposed NKDS and HyIMC in model compression and acceleration while preserving superior accuracy.

References

[1] D. Michelsanti *et al*., "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation, " TASLP, vol. 29, pp. 1368-1396, 2021.

[2] D. Hepsiba *et al*., "Role of Deep Neural Network in Speech Enhancement: A Review," SLAAI-ICAI, pp. 103-112, 2019.

[3] A. Shrestha *et al*., "Review of Deep Learning Algorithms and Architectures," Access, vol. 7, pp. 53040-53065, 2019.

[4] W. Banerjee, "Challenges and Applications of Emerging Nonvolatile Memory Devices," Electronics, vol.9, no. 6, pp. 1029, 2020.

[5] Y. -D. Chih *et al*., "16.4 An 89TOPS/W and 16.3 TOPS/mm 2 All-Digital SRAM-Based Full-Precision Compute-in Memory Macro in 22nm for Machine-Learning Edge Applications," ISSCC, pp. 252-254, 2021.

[6] Y. He *et al*., "7.3 A 28nm 38-to-102-TOPS/W 8b Multiply-Less Approximate Digital SRAM Compute-in-Memory Macro for Neural-Network Inference," ISSCC, pp. 130-132, 2023.

[7] F. Tu *et al*., "ReDCIM: Reconfigurable Digital Computing-in-Memory Processor with Unified FP/INT Pipeline for Cloud AI Acceleration," JSSC, vol. 58, no. 1, pp. 243-255, 2022.

[8] S. -E. Hsieh *et al*., "7.6 A 70.85-86.27 TOPS/W PVT-Insensitive 8b Word-Wise ACIM with Post-Processing Relaxation," ISSCC, pp. 136-138, 2023.

[9] Y. Hu *et al*., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," arXiv:2008.00264, 2020.

[10] H. S. Choi *et al*., "Phase-Aware Speech Enhancement with Deep Complex -Net," ICLR, 2018.

[11] N. Westhausen *et al*., "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," arXiv:2005.07551, 2020.

[12] Q. Li *et al*., "Real-Time Monaural Speech Enhancement With Short-time Discrete Cosine Transform," arXiv:2102.04629, 2021.

[13] Y. -H. Chen *et al*., "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, pp. 367-379, 2016.

[14] T. Bai *et al*., "An End-to-End In-Memory Computing System Based on a 40-nm eFlash-Based IMC SoC: Circuits, Toolchains, and Systems Co-Design Framework," TCAD, vol. 43, no. 6, pp. 1729-1740.

[15] G. Wang *et al*., "A 40nm 5-16Tops/W@INT8 eFlash In-Memory Computing SoC Chip with Noise Suppression and Compensation Techniques to Improve the Accuracy," ICTA, pp. 128-129, 2023.

[16] A.W. Rix *et al*., "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," ICASSP, pp. 749-752, 2001.