Saikiran Chepa | DSP Engineer | C & Assembly | Signal Processing | AI & ML

Hyderabad, India

☎ +91 8977890154 • ✉ sai.406@gmail.com

in saikiran-chepa-81938449/ • chbsaikiran

Professional Summary

DSP and AI/ML Engineer with 5+ years of experience in deep learning and large language model development, built on a strong foundation of 11+ years in embedded DSP optimization. Skilled in designing, training, and deploying models such as CNNs, RNNs, ResNet, transformers, and LLMs, with expertise in pre-training, fine-tuning (SFT), and RLHF. Proven track record of optimizing speech and audio algorithms across ARM, QDSP, Intel, RISC-V, SHARC, and TI platforms, achieving significant performance gains through assembly coding, SIMD intrinsics, and floating-to-fixed-point conversions. Experienced in automating performance profiling with Python, Simpleperf, and ADB, as well as FFT/IFFT and matrix multiplication optimizations. Adept at bridging AI innovation with high-performance embedded systems, combining deep domain knowledge in signal processing with modern machine learning practices.

Key Highlights

- **AI/ML Expertise**: Hands-on experience with CNNs, RNNs, ResNets, transformers, and LLMs; skilled in pre-training, supervised fine-tuning (SFT), RLHF, PPO/DPO/GPRO, and deployment on AWS and Hugging Face.
- **DSP & Embedded Systems**: 11+ years of expertise in optimizing speech/audio codecs and DSP algorithms (FIR/IIR filters, FFT/IFFT, multirate processing, matrix multiplication) across ARM, Intel AVX2, RISC-V, SHARC, TI C64x+/C66x, and QDSP platforms.

- **Performance Optimization**: Proficient in floating-point to fixed-point conversions, assembly-level coding, and SIMD intrinsics (ARM NEON, Intel AVX2), delivering up to 60% cycle reduction and significant throughput gains.
- **Cross-Platform Profiling & Automation**: Built automated performance measurement pipelines using Python, batch/shell scripts, Simpleperf, and ADB; developed MCPS analysis frameworks for Android and Raspberry Pi.
- **Toolchain & Development**: Experienced with MATLAB codegen, CSIM API creation, C/Python development, version control (Git/SVN), and build systems (CMake).
- **End-to-End AI Systems**: Designed and deployed production-ready AI solutions, including a Gmail Query Assistant using Gemini LLM, OAuth, regex filtering, WebSockets, and MCP-based tool orchestration.
- **Certifications & Learning**: Completed advanced certifications, including NPTEL Domain Scholar in Signal Processing & Communications (60-week program), Coursera Deep Learning Specialization, and ERA V3 AI Engineer program.
- **Domain Knowledge**: Strong theoretical foundation in signal processing, adaptive filtering, and information theory, supported by advanced coursework (IIT Madras & NPTEL).

Education

- **MTech in Communication Systems**, IIT Madras
- **BTech in Electronics and Communication Engineering**, GRIET, JNTU Hyderabad

Experience

Engineer III, Capgemini, Client Qualcomm Hyderabad

Dec 2016 – May 2023 | Jan 2025 – Mar 2025

- Floating point to Fixed Point conversion. Implementing log10 and pow using fixed point basicops functions log2 and pow2.
- Using Matlab's codegen to convert Matlab code to C code.(Making changes to original Matlab code so as resolve codegen errors). Dynamically increasing arrays in matlab code were implemented using fixed size arrays.
- Good understanding of writing CSIM APIs for a Matlab code.
- Coded FIR filter followed by decimation by 2 in assembly, by not calculating the alternate samples. The Left and right channel of input used the same FIR coefficients so in the inner most loop of FIR filter both left and right outputs were calculated together this allowed all slot in the packets to be utilized. Also implemented up sampling followed by FIR filter.
- Batch files, Shell scripts, python scripts and cmake to automate compiling and testing of code.
- Good Understanding about Digital Signal Processing theory. FIR and IIR filters, Multirate signal processing.
- Optimized Matrix Multiplication function for AGVC(Advanced Generative Voice Coder) for Intel AVX2 architecture. AGVC uses AI/ML for speech compression.

DSP Engineer , Capgemini, Client Goodix Bangalore

Aug 2023 – Jul 2024

- Used RISCV DSP Lib's fft and ifft function and integrated them into VoiceExperience Code. The code will not be BitExact, so validated the outputs using matlab scripts which generated rmse in dB and also checked the output wav files ADOBE Audition.
- SHARC is floating point DSP, the fft and ifft library functions were available only in floating point. Integrated these float codes into CarVoice code by converting fixed point input to float and using fft and ifft lib function and then again converting there outputs back to fixed point.
- Optimized loops of function by converting input to float and doing processing in float and then converting back to fixed point. This helped in reducing cycles by 60 percent, compared to just using fixed point code.
- Optimized CarVoice Code for ARMV8A architecture, brought the factor between ARMV7A optimized code and ARMV8A optimized from factor of 2.2 to 1.4. For this coded 25 functions in arm neon intrinsics. Wrote python scripts to compare ARMV7A flat profile with ARMV8A flat profile and arrive at functions not optimized for ARMV8A.
- Obtained MCPS on android device using adb commands, used simpleperf for getting flat profile on android device.
- Worked on Raspberry Pi 4, and generated MCPS for ARMV8A and ARMV7A code. For this wrote scripts in python that would extract minimum MCPS out of 5 runs of the executable for each frame and then get average and peak of these minimums.

Senior DSP Engineer , Couth Infotech

Apr 2012 – Mar 2015

- Optimisation on C66x & C64xplus of SILK, Speex, Opus Codecs, Linear assembly code was written.
- Scratch and Channel were split into blocks of 4KB.
- Stack usage was reduced by using scratch based implementation.
- NDK setup was done for C6678 and C6472 boards to test the codec for standard & Non-Standard test vectors.
- Wrappers written for testing are
  - DataMove(Channel & Tables run time relocatable).
  - Illegal read/write.
  - Register preservation, Input buffer corruption, Input/Output buffer alignment.
  - Interrupt testing.
  - Stack and buffer calculation.
  - Code Coverage.
  - Scratch contamination

Senior Software Engineer , Aricent

Jul 2005 – Jul 2009

- eAAC plus Decoder Optimizations on C64xplus
- Interleaved and non-interleaved output support was developed.
- optimized audio and speech codec for TI C64xplus and C66x.
- Worked 32 bit float to fixed conversion of iLBC codec.

Certifications

- **Domain Scholar Certificate in Signal Processing and Communications** (NPTEL, 60-week program)
  - Courses Done :
    Discrete Time Signal Processing
    Applied Linear Algebra
    Probability Foundations for Electrical Engineers
    Multirate DSP
    Introduction to Information Theory
    Principles of Signals and Systems
    Adaptive Signal Processing
- **Deep Learning Specialization – Coursera**
  - Gained expertise in designing, training, and optimizing deep neural networks, including CNNs, RNNs, and transformer-based models using TensorFlow.
  - Applied advanced techniques such as neural style transfer, object detection, and recognition for image/video tasks.
  - Implemented vectorized deep learning architectures, optimized hyperparameters, performed bias-variance analysis.
- **ERA V3 (Extensive Reimagined AI, Version 3) – The School of AI, Bangalore**
  - Gained hands-on experience in Python programming, version control, and web development basics for AI workflows. Developed practical expertise in PyTorch, including implementing neural networks, CNNs, ResNets, and optimization techniques. Trained transformer models from scratch, covering fundamentals to advanced architectures and efficient training strategies. Learned and applied LLM fine-tuning, optimization, and scaling strategies, including RLHF, PPO, DPO, and GPRO. Gained exposure to multi-modal AI, AI agents, Model Context Protocol (MCP), and Retrieval-Augmented Generation (RAG) for real-world applications.

- **Capstone Project: Intelligent Email Query Assistant**
  - Designed and implemented a **client-server application** enabling users to query their Gmail accounts using **natural language**.
  - Integrated **Gmail API OAuth authentication** on the client side, ensuring user login and data access remained secure on local machines.
  - Deployed a **Gemini LLM-powered backend on AWS** to interpret user queries, generate structured Gmail search queries (including date filters), and process mail data.
  - Built client-side logic to apply **regex-based filtering** of promotional emails and limit results to the top 20 most relevant mails before sending them to the server.
  - Implemented server-side **information extraction pipelines** to identify transaction amounts (e.g., "How much did I spend on Zomato?") and aggregated them using an **external calculator service via Model Context Protocol (MCP)**.
  - Delivered results back to the client through a **real-time chat interface** built on **WebSockets**, enabling an interactive, conversational experience for continuous queries.
  - Achieved seamless integration of **LLM reasoning, API-based data retrieval, regex filtering, and external tool orchestration**, demonstrating end-to-end design and deployment of a production-ready AI system.

Technical Skills

- **Programming Languages:** C, Python, MATLAB
- **Platforms:** ARM, TI, Intel, QDSP, SHARC, RISCV
- **IDEs:** Visual Studio, CCS, Eclipse, DS-5, Spyder, ARM Workbench, Cursor AI, VS Code, AWS
- **Environment:** Windows, Linux
- **Version Control:** GIT, SVN