Arun Baby

www.arunbaby.com • www.linkedin.com/in/arunbaby0

Summary

Technical leader with 14+ years building conversational AI and speech systems. CTO at stealth startup developing agentic AI. Previously led Samsung Galaxy AI on-device ASR (shipping to 200M+ devices, $6M+ savings). Deep expertise bridging research and production. 2 patents, 15+ papers (INTERSPEECH, ICASSP).
Expertise: Conversational AI • LLM-based Agents • Speech Tech (ASR/TTS) • ML Engineering • Team Leadership

Experience

Stealth Startup | **Agentic Conversational AI**
Chief Technology Officer *Jul 2024–Present*
- Architected and built agentic conversational AI platform integrating LLMs (GPT-4, Claude, Llama), real-time speech (Deepgram, ElevenLabs), and agent frameworks (LiveKit, Pipecat, Ultravox)
- Leading team of 4 engineers; hands-on development of core agent orchestration and voice pipeline infrastructure
- Driving technical roadmap and product architecture from concept to customer deployment

Samsung Research | **Galaxy AI (Acqui-hired from Zapr)**
Staff Engineer / Senior Chief Engineer *Apr 2022–Jul 2024*
- Led team of 9 to deploy on-device ASR/TTS models for Galaxy AI Live Translation (6 languages, 200M+ devices at S24 launch)
- **Business Impact:** Migrated models from server to on-device, **saving $6M+ annually** and **improving latency by 30%**
- Pioneered on-device speaker personalization using LoRA, achieving **15% accuracy improvement**. Published 4 papers (ICASSP, INTERSPEECH), filed 1 patent
- Developed model compression pipeline and streaming ASR architectures for resource-constrained mobile devices

Zapr Media Labs | **Conversational AI (Acquired by Samsung)**
Research Scientist (Speech) *Jan 2020–Apr 2022*
- Built production VoiceBot platform achieving **80% call automation**, handling thousands of daily conversations
- Achieved **10-30% WER improvement** over commercial ASR and **20% MOS improvement** for Indian English TTS
- Architected ML training/deployment infrastructure on Kubernetes; led team of 2 engineers. Published 3 papers, 1 patent

Cisco Systems | **Enterprise Networking**
Software Engineer II *Aug 2018–Jan 2020*
- Developed lossless logging system achieving **80% memory reduction** and **10% performance improvements** for enterprise routing

MeitY (Govt of India) | **TTS for Indian Languages**
Project Officer (Research) *Jan 2015–Jun 2018*
- Led TTS development across **13 Indian languages**, managing collaboration with 11 research organizations nationwide
- Developed unified TTS architecture reducing per-language development time by 60%. Technology transferred to Samsung, IndusOS

HCL Technologies | **Healthcare & Financial Applications**
Software Engineer *Oct 2011–Jul 2014*
- Led backend development for healthcare SaaS platform, managing team of 4. Query optimization for Thomson Reuters analytics

Education

Indian Institute of Technology Madras
M.S. (Research) in Computer Science
Thesis: Unified Approach to Speech Synthesis in Indian Languages (Advisor: Prof. Hema A. Murthy)

Rajagiri School of Engineering
B.Tech in Computer Science

Patents & Publications

Patents (2): Speaker personalization using LoRA for on-device speech models (2024) • TTS prosody control (2021)
Selected Publications (15 total - Google Scholar): Speaker Personalization for ASR using LoRA (INTERSPEECH 2024) • Multilingual ASR with Language ID (INTERSPEECH 2024) • Robust Speaker Personalisation using LoRA (ICASSP 2024) • Automatic Speech Segmentation (Speech Communication Journal 2020)

Technical Skills

Conversational AI: LLM-based Agents (GPT-4, Claude, Llama) • Voice Assistants • Dialogue Systems • Real-time Voice Infrastructure
Speech Technologies: ASR • TTS • Speaker Personalization • Speech Enhancement • Model Compression
ML/AI: PyTorch • TensorFlow • Model Optimization • On-device ML • Transfer Learning • LoRA • ESPnet • Kaldi • Huggingface
Infrastructure: Kubernetes • Kubeflow • Docker • FastAPI • gRPC • AWS • GCP • LiveKit • Pipecat • Deepgram • ElevenLabs
Languages: Python • C++ • C • Bash