Arun Baby

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

Profile

Technical leader with 14+ years building conversational AI and speech systems from research to production. Currently CTO at stealth startup developing agentic AI. Previously led on-device ASR for Samsung Galaxy AI, shipping to hundreds of millions of devices. Deep expertise bridging research innovation with production systems at scale. Published researcher with 2 patents and 15+ papers in top-tier conferences.

I am passionate about research and technology, have a broad engineering background, and love developing myself, the people around me, and leading teams.

Expertise: Conversational AI | Speech Technologies (ASR/TTS) | ML Research & Engineering | Team Leadership | LLM-based Agents

Professional Experience - 14 years

Stealth Startup | Chief Technology Officer

Agentic Conversational AI Platform

Founding Technical Leadership *Jul 2024 - Present*

- **Platform Architecture:** Architected and built agentic conversational AI platform from ground up as founding CTO. Designed end-to-end voice agent architecture integrating LLMs (GPT-4, Claude, Llama), real-time speech (Deepgram, ElevenLabs, Cartesia), and agent frameworks (LiveKit, Pipecat, Ultravox).
- **Technical Implementation:** Hands-on development of core platform components including agent orchestration, voice pipeline, and real-time processing infrastructure. Built rapid prototyping framework enabling fast iteration on customer feedback.
- **Team Leadership:** Built and leading engineering team of 4, establishing technical standards, development processes, and engineering culture for shipping quickly without compromising quality.
- **Strategic Execution:** Driving technical roadmap aligned with market needs, balancing research exploration with product delivery, and ensuring architectural decisions support scale.
- **Tech Stack:** PyTorch, Python, LLM APIs, real-time voice infrastructure, cloud-native deployment.

Samsung Research (SRI-B) | Staff Engineer / Senior Chief Engineer

On-Device ASR for Galaxy AI **Acqui-hired from Zapr Media Labs**

Speech Research & Engineering *Apr 2022 - Jul 2024*

- **Galaxy AI Live Translation:** Led team of 9 engineers to develop, optimize, and deploy on-device ASR/TTS models for 6 languages in Galaxy AI Live Translation feature. Shipped to hundreds of millions of devices at S24 launch across all Galaxy flagship form factors.
- **Business Impact:** Migrated speech models from server to on-device deployment across multiple languages, **saving $6M+ in server costs annually** while **improving response time by 30%** through local processing.
- **Personalized ASR Innovation:** Pioneered on-device speaker personalization using LoRA adaptation, achieving **15% average accuracy improvement** with fully private training and deployment. Published 4 papers (ICASSP, INTERSPEECH, NCC) and filed 1 patent.
- **Technical Leadership:** Led model compression, optimization, and deployment pipeline for resource-constrained mobile devices. Developed streaming/hybrid ASR architectures and multimodal systems for production.
- **Cross-functional Collaboration:** Worked closely with product, hardware, UX, and international teams across Samsung ecosystem to deliver integrated speech experiences.
- **Environment:** TensorFlow, PyTorch, Python, C++, mobile ML optimization, on-device inference.

Zapr Media Labs | Research Scientist (Speech)

Conversational AI Platform — **Acquired by Samsung**

Speech Research Team Lead — *Jan 2020 - Apr 2022*

- **Production VoiceBot Platform:** Designed and built end-to-end conversational AI platform achieving **80% call automation rate**, handling thousands of daily conversations. Architected speech pipeline integrating custom ASR, NLU, dialogue management, and TTS.
- **ML Research & Innovation:** Achieved **10-30% relative WER improvement** over commercial ASR systems through context-aware models and domain adaptation. **20% MOS improvement** for Indian English TTS using transfer learning and data augmentation.
- **Technical Contributions:** Developed bilingual ASR/TTS models for Indian languages, controllable expressive TTS with prosody control, and voice analytics platform. Built streaming ASR system with <200ms latency.
- **ML Infrastructure:** Architected scalable training and deployment infrastructure on Kubernetes, enabling parallel training of 100+ models. Built MLOps pipeline using Kubeflow, Docker, FastAPI for rapid experimentation and production deployment on AWS/GCP.
- **Team Leadership:** Led team of 2 engineers, mentored junior researchers, and established research publication culture. 3 conference submissions, 1 patent granted.
- **Environment:** ESPnet, Kaldi, PyTorch, Huggingface, FastAPI, Kubeflow, Kubernetes, gRPC, AWS, GCP.

Cisco Systems | Software Engineer II

Data Center Debug Analytics

Enterprise Networking — *Aug 2018 - Jan 2020*

- Developed lossless logging system achieving **80% memory footprint reduction** for enterprise routing products, improving debuggability at scale.
- **10% code execution performance improvements** through profiling and optimization.
- IPv4/IPv6 protocol implementation and maintenance for data center networking products.
- **Environment:** C, Linux, Python.

MeitY (Govt of India) | Project Officer (Research)

Text-to-Speech for Indian Languages

Speech Technology for Indian Languages Consortium — *Jan 2015 - Jun 2018*

- **Program Leadership:** Led TTS development initiative across **13 Indian languages**, managing collaboration with **11 research organizations nationwide** including IIT Madras, IIT Kharagpur, and IIIT Hyderabad.
- **Technical Innovation:** Developed language-independent parser enabling unified TTS architecture, reducing development time per language by 60%.
- **Technology Transfer:** Successfully transferred technology to Samsung Research Institute, IndusOS, and ShinanoKenshi, enabling commercial deployment.
- **Research Contributions:** Deep learning-based phonetic segmentation, hybrid signal processing + ML approaches, Android platform integration.
- **Environment:** C, Python, Perl, HTK, Kaldi, HTS, Merlin, Android.

HCL Technologies | Software Engineer

Healthcare & Financial Applications

Enterprise Software Development — *Oct 2011 - Jul 2014*

- Led PoC and backend development for healthcare SaaS platform (KnowledgePoint360), managing team of 4 engineers.
- Developed multi-platform physician management system and API integrations with third-party verification services.
- Database query optimization for Thomson Reuters financial data analytics, improving job execution efficiency.
- **Environment:** C#, ASP.NET MVC, SQL Server, Delphi, .NET.

Education

M.S. (By Research) in Computer Science
Indian Institute of Technology Madras

- Supervisor: Prof. Hema A. Murthy
- Thesis: A Unified Approach to Building Speech Synthesis Systems in Indian Languages
- **Key Courses:** Pattern Recognition, Kernel Methods, Speech Technology, Advanced Algorithms

B.Tech in Computer Science
Rajagiri School of Engineering

Patents & Publications

Patents

- **Arun Baby**, George Joseph; **Speaker personalization using LoRA for on-device speech models**, 202341053222, October 2024.
- Sharath Adavanne, Nagaraj Adiga, Srikanth Konjeti, Sumukh S Badam, **Arun Baby**, et al.; **Method and system for controlling speech characteristics in speech synthesis systems**, 202141030614, July 2021.

Selected Publications (15 total) - Google Scholar

- George Joseph, **Arun Baby**; **Speaker Personalization for ASR using Weight-Decomposed Low-Rank Adaptation**, INTERSPEECH 2024.
- Nikhil Jakhar, Sudhanshu Srivastava, **Arun Baby**; **Unified Approach to Multilingual ASR with Improved Language ID for Indic Languages**, INTERSPEECH 2024.
- **Arun Baby**, George Joseph, Shatrughan Singh; **Robust Speaker Personalisation Using Generalized Low-Rank Adaptation for ASR**, ICASSP 2024.
- **Arun Baby**, et al.; **Non-native English lexicon creation for bilingual speech synthesis**, Speech Synthesis Workshop (SSW) 2021.
- **Arun Baby**, et al.; **Significance of Spectral Cues in Automatic Speech Segmentation**, Speech Communication Journal, 2020.
- **Arun Baby**, et al.; **Deep Learning Techniques with Signal Processing for Phonetic Segmentation**, INTERSPEECH 2017.

Technical Expertise

- **Leadership:** Team Building, Technical Strategy, Research to Production, Cross-functional Collaboration, Hiring
- **Conversational AI:** LLM-based Agents, Voice Assistants, Dialogue Systems, Real-time Voice Infrastructure
- **Speech Technologies:** ASR, TTS, Speaker Recognition, Speech Enhancement, Model Personalization
- **ML/AI:** Deep Learning (PyTorch, TensorFlow), Model Optimization, On-device ML, Transfer Learning, LoRA
- **ML Infrastructure:** Kubernetes, Kubeflow, Docker, FastAPI, gRPC, Cloud Deployment (AWS, GCP)
- **Frameworks & Tools:** ESPnet, Kaldi, Huggingface, LiveKit, Pipecat, Deepgram, ElevenLabs
- **Languages:** Python, C++, C, Bash