

Abdul Rasheed Esa

Voice AI Infrastructure Engineer — production voice agents at ~20k calls/day | **Mastra contributor**
📍 Hyderabad, India (works US hours) | 📞 +91 8639213518 | ✉️ rasheed3086@gmail.com | 🌐 [GitHub](#) | [LinkedIn](#)

Built and run a production voice-agent platform: self-hosted LiveKit + Twilio and customer SIP trunks, ~20k calls/day for US legal and accounting brands. Hands-on from SIP and media to LLM agents, evals and infra. 4 merged fixes in Mastra's voice providers.

Experience

Technical Lead (Contract) - FlowForge @ Realbotics AI — *Los Angeles (Remote) Feb 2025 - Present*

- Built the company's voice-agent platform from zero and lead a team of 4; ~20k production calls/day, load-tested to 200 concurrent calls (402/402 completed), for **LegalZoom**, **1-800Accountant**, **1-800-THE-LAW2** and others (Python, LiveKit Agents, PostgreSQL, Temporal)
- Made mid-call handoffs between agents instant: 5.5s → 31ms (~180×), without dropping the phone line or the recording (LiveKit Agents)
- Raised capacity ~5× (6 → 31 concurrent calls per worker) by tracing four successive bottlenecks; cold starts 4.6s → 1.7s p95 (AWS ECS, Redis)
- Root-caused a SIP failure cascade that took campaign calls from 20% to 100% failing in 8 minutes; rejected calls now free their line at once (LiveKit SIP, Twilio)
- Tuned voicemail detection from a sweep over 13k+ real calls: false voicemail triggers on live callers 3.7% → 1%
- Moved production off AWS to Hetzner in under 24h when the account was frozen; zero-downtime deploys that drain live calls (Terraform, Docker)

Technical Lead - KOKO Nexus @ KOKO AI *Apr - Dec 2024*

- Took a voice AI receptionist from zero to production: answers business phone lines with RAG over each business's data at sub-500ms response latency (Twilio, Deepgram, ElevenLabs, GCP Cloud Run)

Machine Learning Engineer (Contract) - Insurtech Platform *Aug - Dec 2023*

- Raised vehicle-damage detection accuracy from 65% to 94% and shipped it as an API handling 100+ claims/day, cutting manual review time by 75% (YOLOv8, Django)

Projects

Open source: Mastra (28k-star TypeScript agent framework, ~2M weekly npm downloads) *Oct 2026*

- 4 merged fixes to Mastra's realtime voice providers (OpenAI Realtime, Inworld, AWS Nova Sonic, Gemini Live): unanswered tool calls, false barge-ins, audio streams that never closed, and timers leaked on server disconnects; each with tests, merged by Mastra's CTO

Self-Hosted GPU Voice Stack *May 2026*

- Ran a fully self-hosted voice agent on live phone calls, with Sesame CSM-1B fine-tuned (LoRA) on a custom voice: under 400ms end-to-end, first audio ~315ms, vs p50 ~580ms on hosted APIs (RunPod RTX 6000 Ada, LiveKit, PyTorch)
- Moved the end-of-turn model to GPU: p50 23ms vs 300–400ms (17×), so the agent replies as soon as the caller stops talking (CUDA)

Skills

Voice / real-time: LiveKit (Agents, SIP, self-hosted), Twilio SIP trunking, WebRTC, streaming STT/TTS

Backend / infra: Python (FastAPI), TypeScript/Node, PostgreSQL, Redis, Temporal, Docker, AWS ECS, Hetzner

AI: LLM agents (LangGraph), RAG (Qdrant), LLM-judge evaluation, OpenAI/Anthropic/Groq APIs

Education

B.E. Computer Science and Engineering (AI & ML) — *Osmania University (MJCET) 2021 - 2025*

- ICPC Asia Regionals | Smart India Hackathon 2022 Grand Finalist (15,000+ teams) | 3x hackathon winner