

JAI SHANKAR

AI Engineer, Voice Agents

Bengaluru, India | +91 76440 53913 | jaimauryatech@gmail.com
linkedin.com/in/jai-shankar-b7b388234 | github.com/jai0651 | jaishankar.dev

EDUCATION

Indian Institute of Technology Delhi
B.Tech, Engineering and Computational Mechanics

New Delhi
2020 – 2024

EXPERIENCE

Plivo (voice and SMS APIs, AI voice agents)
Software Engineer I | Go, C, Python

Bengaluru
Sep 2025 – Present

- Optimized the voice media path (RTP handling, kernel and socket tuning, codec/transcoding hot paths, GC and memory-pressure reduction) and shipped neural noise cancellation as a FreeSWITCH module in C, used on **7.5M calls**, raising concurrent call capacity **2x** per machine and cutting fleet infrastructure cost by **50%**.
- Built an agentic AI support platform using RAG over production logs, internal docs and microservice codebases, with Slack and CLI workflows that automate support triage and cut support workload by **~50%**.
- Built a low-latency SMS classification and routing system with **sub-ms inference** for spam filtering, preference-aware routing and fraud shielding across production messaging workflows.
- Developed internal engineering automation integrating Kubernetes, Jenkins, GitHub, Slack and deploy pipelines, cutting staging validation and deployment time and effort by **70%**.

Hyperbots (AI for finance and accounting)
Software Engineer I | Java (Spring Boot), Python (FastAPI)

Bengaluru
Jun 2024 – Aug 2025

- Built and ran 5 microservices (notifications, RBAC, scheduling, config) handling **100K+ events a day**.
- Moved read-heavy screens to CQRS, with writes in MongoDB and Postgres and reads served from Elasticsearch views kept in sync by events, raising query throughput by **60%**.
- Cut API response times by **40%** with missing indexes, rewrites of the slowest queries and schema fixes.

2Sigma School, *SDE Intern*. Built a safe, fast remote and in-browser code execution engine (Pyodide, Cloud Run), cutting cost 70%.

Mar – May 2024

Swachh.io, *SDE Intern*. Built a Flutter affiliate app on Shopify with referrals, rewards, payments.

May – Jul 2023

PROJECTS

Voice agent from scratch | *Python, PyTorch, WebSockets*

[repo](#)

- Built every stage of a voice agent in PyTorch with no pretrained models: noise suppression, VAD, streaming Conformer ASR, Tacotron-style TTS, an ISTFT neural vocoder and ECAPA-style speaker ID.
- On the same 4.1 h of LibriSpeech, the Conformer **halved CER** against a BiGRU + CTC baseline (0.66 to 0.28) and ran 2.7x faster, while streaming in 640 ms chunks cost only 3% relative.
- Wrote a streaming OM-LSA noise suppressor that removes 8.1 dB of noise in pauses at **16 ms** latency and damages speech less than the trained GTCRN model (0.9 vs 2.0 dB).
- Served it through a real-time WebSocket server with endpointing, barge-in, an OpenAI or Claude responder and p50/p95 latency per stage, where any stage can be swapped for Silero, faster-whisper or Kokoro.

Tonecall: agent-to-agent call protocol | *TypeScript, Next.js, WebSockets*

[repo](#)

- Designed a protocol for two AI voice agents on the same phone call to detect each other with an in-band DTMF nonce handshake and switch to JSON, cutting negotiation on a live test call from about 75s to **13s** at half the cost. When the other side is a human, the agent stays in voice.
- Built the telephony runtime on Plivo media streams with mu-law/PCM transcoding, a streaming STT, LLM and TTS loop. Instances pair through a broker that signs every message with HMAC-SHA256 and rejects replays.

Patronus: AI career companion | *Next.js, TypeScript, Postgres, pgvector, OpenAI*

patronus.cv | [repo](#)

- Built and launched an LLM career companion that keeps a record of a user's work, checks jobs against it and tailors resumes and recruiter notes with RAG, rejecting any number not in the source (evals gate on **zero fabrications**).

SKILLS

Languages: Python, Go, C/C++, Java, TypeScript, Rust

Speech / AI: PyTorch, ASR (Conformer, CTC), TTS, VAD, noise suppression, speaker ID, LLM agents, RAG

Voice infra: FreeSWITCH, RTP, WebSockets, mu-law/PCM, real-time streaming, latency profiling

Backend / Infra: FastAPI, gRPC, Kafka, Redis, PostgreSQL, AWS, Kubernetes, Docker, Prometheus, Grafana