Email resume and "portfolio" to pmoeckel@frontieraudio.com.

Founding Senior Voice AI Systems Engineer

Company Overview |

Frontier is on a mission to make hundreds of millions of critical frontline workers superhuman through a hardware and software "wearable". The company headquartered in Chicago and in stealth mode. Frontier has multimillion dollar contracts with the U.S. Department of Defense (including Special Operations) and has raised funding from tier 1 investors, including the first investor in Oculus (AR/VR company sold to Meta). The company has more than 10 of the most critical U.S. companies as customers, including the largest U.S. steel company and two of the top five U.S. airlines. It has an extremely talented and ambitious team.

Job Description |

Frontier is looking for a Founding Voice AI Systems Engineer (Senior or Staff level depending on experience) who wants to do the most important work of their life. The selected individual will own huge parts of Frontier's voice platform (that lives on a "wearable") and their work will have extreme impact on the productivity, safety, and comfort of the critical frontline workers and warfighters powering and protecting the world. Being scrappy, resourceful, and flexible is critical to this role and it is 100% in-person in Chicago (Frontier will provide relocation assistance (along with significant equity)).

This role is ideal for someone who wants to (1) be a cornerstone member of a rapidly growing team and (2) iterate quickly on the cutting edge of AI voice interfaces (work deployed in the field in less than 1 year).

Responsibilities |

- Build and iterate conversational voice pipeline, from raw audio capture through STT, retrieval-augmented prompting, cloud LLM calls, low-latency TTS streaming, and back to playback — meeting the latency and power budgets defined by product.
- Design prompts, guardrails, and retrieval logic that minimize hallucinations and errors while handling voice-specific challenges (disfluencies, partial utterances, background noise)
- Select and compose best-in-class STT, embedding, large-model, and verification components and build automated benchmarks for latency, accuracy, hallucination rate, and cost.
- Implement dialog-control logic that detects barge-ins, cancels or resumes TTS on the fly, tracks short-term turn state, and keeps multi-turn exchanges fluid.
- Engineer contextual grounding & memory (persist recent intents, run similarity search against a vector DB (RAG), and attach top chunks as system context) to keep answers accurate and personalized.
- Optimize on-device vs cloud processing splits to balance latency, accuracy, and power consumption for wearable constraints
- Collaborate with backend engineers to define requirements for voice-related services (auth, device registry, telemetry) and build lightweight prototypes for testing voice pipeline integration

- Integrate voice pipeline with existing cloud services and vector databases, implementing gRPC/REST interfaces for audio/text data flow between edge devices and backend systems
- Collaborate across Android, web, firmware, and hardware teams to define protobuf/OpenAPI schemas and ensure seamless hand-offs between edge devices and cloud services; document architecture and mentor peers as the platform scales from pilot to mass deployment.

Qualifications |

Required

- Hands-on experience (≥ 5 yrs overall, ≥ 1 yrs with modern LLM or end-to-end neural speech stacks) delivering production-grade conversational systems—e.g., GPT-4/Claude-powered assistants, RAG chat-with-docs platforms, streaming Whisper/Nemo ASR fused with TTS, or modern voice assistant platforms (Alexa, Google Assistant, Siri)
- Proficiency in Python *and* one systems language (Go, Rust, or C++); hands-on with gRPC/protobuf, REST, and event-driven messaging (Kafka, SNS/SQS, or similar)
- Experience handling diverse accents, languages, and voice interface challenges
- Experience optimizing LLM prompts specifically for voice interfaces (handling disfluencies, partial utterances, conversational context)
- Depth in prompt-engineering patterns, retrieval-augmented generation, and vector databases (FAISS, pgvector, Milvus, Pinecone, etc.)
- Ability to prototype voice pipeline integrations and work effectively with backend engineers to define API requirements and data schemas
- Must have ability to be qualified to work in the U.S. in less than 3 months and have no restrictions for international travel
- Ability to commute to Frontier's Chicago office daily (relocation support in offer)

Preferred

- Experience shipping real-time conversational voice products (next-gen assistants, in-car AI copilots, field-service wearables, etc.)
- Experience with voice activity detection (VAD), endpointing, and streaming protocols for real-time audio processing
- Experience measuring and optimizing end-to-end latency, power, and cloud cost for real-time voice or LLM pipelines
- Familiarity with multi-modal context integration (using device state/location/sensors to improve voice understanding)
- Experience with on-device vs cloud processing trade-offs for voice systems, including edge AI deployment and optimization strategies
- Knowledge of ML observability and evaluation frameworks (Weights & Biases, PromptLayer, Evidently AI) and data-centric improvement loops
- Contributions to open-source LLM or speech orchestration stacks (LangChain, LlamaIndex, Semantic Kernel, ESPnet, Vosk)

Frontier Audio Labs evaluates qualified applicants without regard to race, color, religion, sex, sexual orientation, gender identity, genetic information, national origin, age, disability, veteran status, or any other legally protected characteristics.

All inquiries should be directed to pmoeckel@frontieraudio.com with your portfolio and resume.