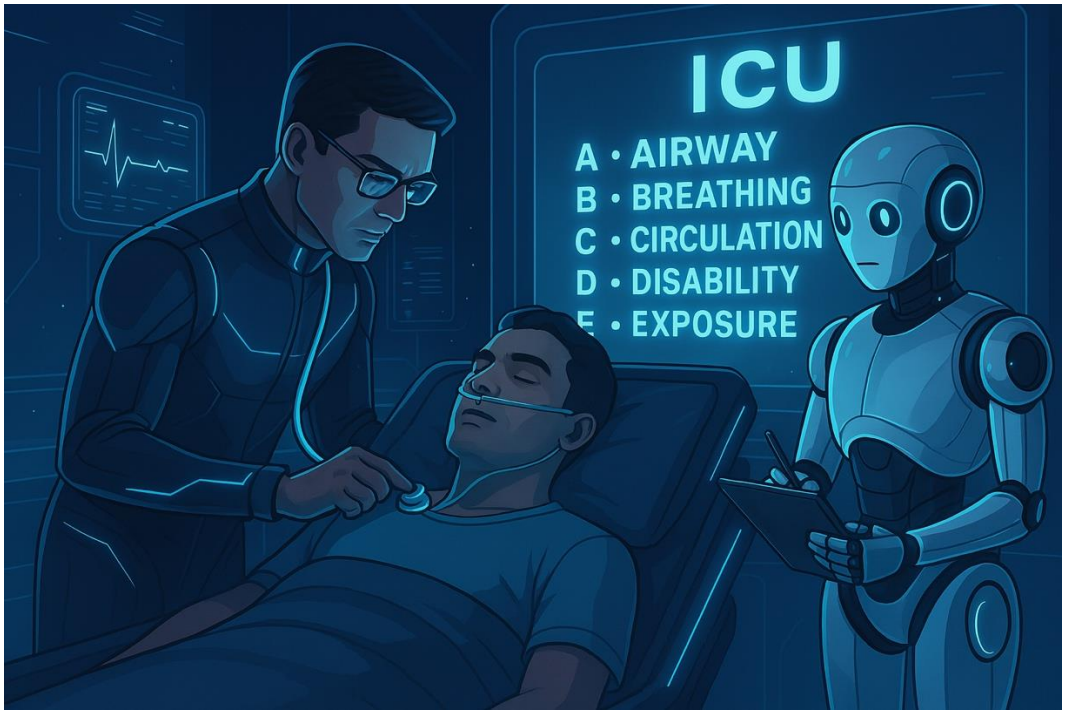
SI-Listener: Real-time intelligence at the edge

SI-Listener:

a Real-Time Voice-to-Knowledge Engine on the Edge



NOTE: most of the images in this presentation have been generated with Microsoft's Image Creator

Maurizio Martignano
Spazio IT – Soluzioni Informatiche s.a.s
Via Manzoni 40
46051 San Giorgio Bigarello, Mantova
https://spazioit.com

November 2025

Vision: Transforming Speech into Actionable Intelligence

- Real-time understanding of live audio
- Intelligent segmentation and context-aware analysis
- Structured knowledge extraction — locally, securely, efficiently

What is SI-Listener?

- An integrated pipeline that:
  - **Ingests live audio** from microphones or system sources
  - **Segments speech dynamically** using voice triggers
  - **Uses Generative AI** to understand context
  - **Extracts structured data**
  - **Queries domain knowledge bases** to check compliance with guidelines, protocols
  - **All in real-time, on a single edge device**

What is SI-Listener?

- **Audio Ingestion** — Real-time, no recordings needed
- **Dynamic Segmentation** — "Start/Stop Recording" voice triggers
- **Contextual Understanding** — GenAI pre-processing
- **Structured Extraction** — domain-specific data (e.g. FHIR)
- **Knowledge Query** — Vector DB-based protocol/reference lookup
- **Output/Integration** — API or local interface

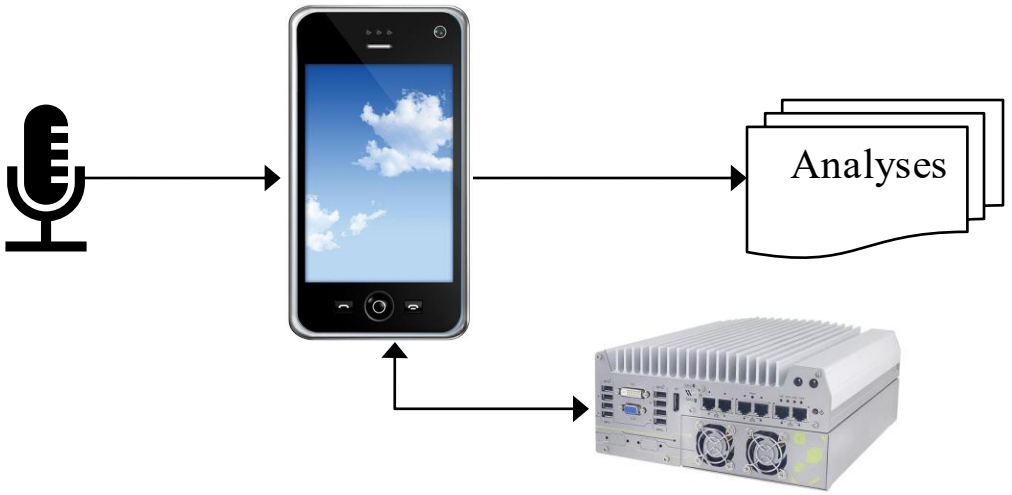
Key Innovations

- 🧠 GenAI contextualization BEFORE data extraction
- 🔊 Live dynamic segmentation — action-triggered blocks
- 🖥️ Edge-optimized pipeline on a single industrial PC
- 📚 Real-time reference to domain knowledge bases
- 🛠️ Open-source foundation for adaptability and transparency
- (( )) Bidirectional Voice Interaction

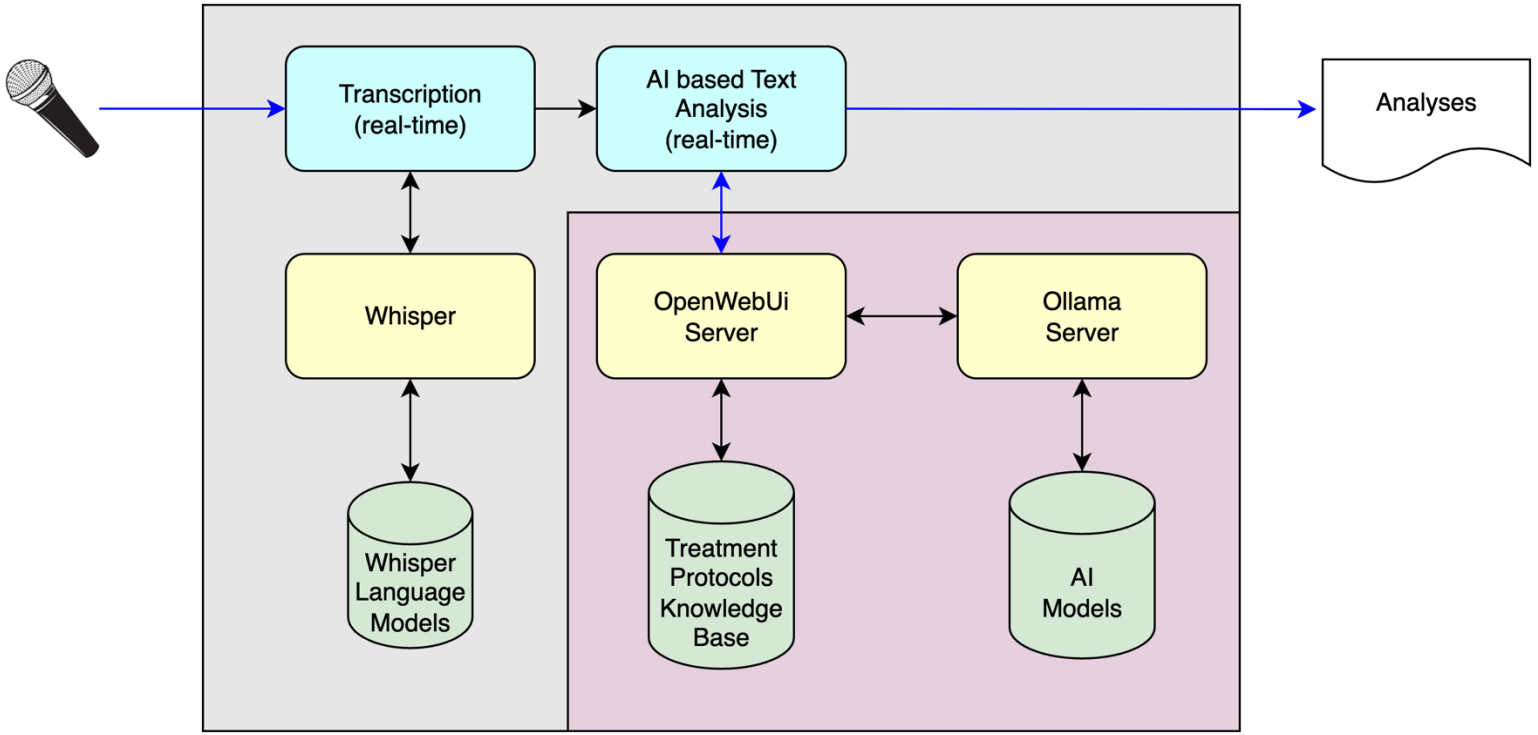
Why Edge Deployment?

- 🚫 No dependency on cloud
- 🔒 Privacy-first: sensitive data stays local
- ⚡ Ultra-low latency for real-time decisions
- 🌐 Works in offline or low-connectivity environments

Two deployments



Two deployments (Architecture)



Potential Domains of Application

- 🏭 Industrial Maintenance & Inspections
- 🚓 Law Enforcement & Emergency Services
- 🎓 Education & Training Simulations
- 🚀 Aerospace & Tactical Operations
- 🧠 Cognitive Assistants in Research Labs

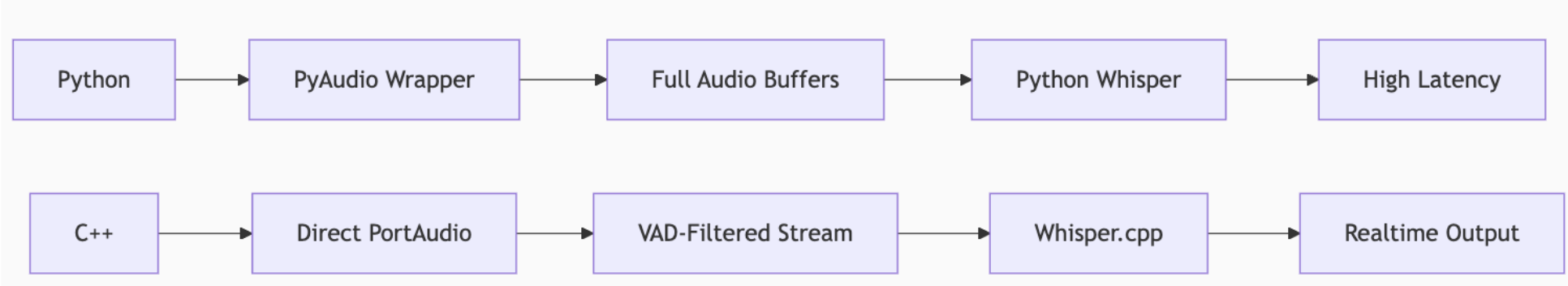
SI-Listener in Action (Use Case)
Field Technician Scenario:

- Speaks while inspecting machinery
- SI-Listener segments speech: "Start Recording", "Stop Recording”
- GenAI interprets: "bearing overheating”
- Structured output: Maintenance record with timestamp, defect code
- KB Query: Suggests SOP from vector DB

Comparison to Market

Feature	SI-Listener	Competitors
Real-time Segmentation	✅ Yes	❌ No
On-device Edge Processing	✅ Yes	❌ Mostly Cloud
GenAI Before Extraction	✅ Yes	❌ After/Not Used
Protocol KB Query	✅ Yes	❌ Rarely Integrated
Open-Source Core	✅ Yes	❌ Closed Source

C++ Version
Architectural Changes



Feature	Python	C++
Audio Pipeline	Queue-based	Zero-copy ✅
VAD	Energy threshold	RMS + Silence tracking 🔍
Resource Mgmt	GC-dependent	RAII destructors
Concurrency	GIL-limited	True parallelism 🚦

C++ Version Improvements

- **Performances**
  - 3.2× faster chunk processing
  - 60% less CPU utilization
  - Handles 2× longer audio sessions
- **Adaptive Audio Processing**
  - Processes only speech segments
  - Auto-adjusts to ambient noise
  - Phrase timeout intelligence
- **Resources Optimization**
  - Static buffer allocation
  - Reusable audio containers
  - No garbage collection pauses
  - 40% smaller memory footprint
- **Production Grade Stability**
  - Audio length padding (min 100ms)
  - Comprehensive PortAudio error checks
  - Thread-safe queue synchronization

Contacts

- GitHub: github.com/mmartign/Speech-to-Data

- **Contact**: info@spazioit.com
  Website: https://spazioit.com