Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

Voice and Multimodal AI for Developers

Class Duration

14 hours of live training delivered over 2-3 days to accommodate your scheduling needs.

Student Prerequisites

- Professional software development experience in TypeScript or Python
- Familiarity with REST APIs and async programming

Target Audience

Software engineers adding voice, image, or document understanding capabilities to applications. Relevant for teams building voice assistants, accessibility features, document processing pipelines, image-aware support tools, or multimodal AI features in web and mobile applications.

Description

This course covers the practical implementation of voice and multimodal AI features in production applications. We cover real-time voice APIs (OpenAI Realtime API, speech-to-text, text-to-speech), vision capabilities (image understanding with Claude Opus/Sonnet 4.x and GPT-5.x, screenshot analysis, diagram interpretation), document understanding (PDF and image extraction, structured extraction from documents), and audio input/output pipelines. Labs build working multimodal features in TypeScript and Python.

Learning Outcomes

- Integrate real-time speech-to-text and text-to-speech APIs into a web application.
- Use the OpenAI Realtime API (or equivalent) for low-latency voice conversation features.
- Submit images to vision-capable models and extract structured information.
- Build a document understanding pipeline that extracts structured data from PDFs and scanned documents.
- Handle the unique latency, error, and UX considerations of voice and multimodal features.

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

- Apply cost and privacy considerations specific to voice and image data.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Node.js 20+ or Python 3.12+, API keys for OpenAI and/or Anthropic, and a microphone for voice labs.

Training Topics

Speech-to-Text and Text-to-Speech

- Whisper and OpenAI transcription API
- Real-time vs. batch transcription tradeoffs
- Speaker diarization and timestamping
- Text-to-speech options: ElevenLabs, OpenAI TTS, Cartesia
- Voice selection, speed, and tone controls
- Streaming TTS for low-latency UX

OpenAI Realtime API for Voice

- Realtime API architecture: WebSockets and audio streams
- Session configuration, voice activity detection, and interruption handling
- Tool calling within a realtime voice session
- Integrating realtime voice into a web app
- Latency optimization and fallback strategies
- Cost management for long-running voice sessions

Vision: Image Understanding

- Submitting images to Claude and GPT vision-capable models
- Use cases: screenshot analysis, diagram interpretation, UI review
- Structured extraction from images with JSON output
- Bounding-box and region-of-interest extraction
- Token and cost considerations for vision inputs

Document Understanding

- PDF and image-based document processing

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

- Extracting structured data from forms and tables
- Multi-page document handling and chunking
- Combining vision and text extraction pipelines
- OCR fallback strategies for scanned documents

Multimodal Application Patterns

- Combining voice + vision in a single session
- Audio + image input for accessibility features
- Multimodal RAG: images in the retrieval pipeline
- Multimodal embeddings and cross-modal search
- Caching and cost management for multimodal inputs

Real-Time Streaming and UX

- Audio buffering and partial-transcript UX
- Barge-in and interruption design
- Latency budgets for conversational features
- Mobile and browser audio capture pitfalls

Reliability and Error Handling

- Handling transient transcription errors
- Fallback paths when realtime sessions disconnect
- Recovery patterns for long-running voice agents
- Observability for voice and multimodal pipelines

Privacy and Compliance for Voice and Image Data

- Data retention policies for voice recordings
- PII in images: detection and redaction
- Consent and disclosure requirements (including dual-party-consent jurisdictions)
- Data residency for voice and image processing
- Air-gapped or self-hosted alternatives for regulated data

Workshop

- Voice conversation feature implementation lab (Realtime API)
- Document extraction pipeline lab
- Multimodal RAG mini-project
- Q&A session