INNOVATION AI LAB

INO AI LAB

[AI EDUCATION • COURSE MATERIAL]

Voice AI: TTS, STT, and Voice Agents

ElevenLabs, Whisper, real-time voice agents

LEVEL	DURATION	LESSONS
Intermediate	**3.5 hours**	**7**

Innovation AI Lab — Tools, Prompts & Performance

// 2026 EDITION

[00]

Table of Contents

LESSON 01 The Voice AI Stack in 2026

LESSON 02 Voice Cloning and Ethics

LESSON 03 Real-Time Voice Agents That Work

LESSON 04 Transcription, Subtitles, and Audio Editing

[LESSON 01]

The Voice AI Stack in 2026

Three layers: speech-to-text (STT) for transcription, large language model for reasoning, text-to-speech (TTS) for response. The 2026 leaders: Whisper Large v4 or Deepgram Nova-3 for STT, GPT-5/Claude/Gemini for reasoning, ElevenLabs v3 or OpenAI TTS for output. End-to-end models like GPT-5 Voice and Gemini Live skip the middle steps entirely for lower latency but less control.

Latency is the user experience. Anything over 1 second feels broken. Use streaming everywhere — STT streams as you speak, LLM streams tokens, TTS streams audio chunks as tokens arrive. With careful engineering you can hit 600–800ms first-audio latency on phone networks in 2026. Below that range, voice conversations feel natural; above it, they feel like waiting on hold.

// KEY TAKEAWAYS

- Three layers: STT → LLM → TTS.
- Streaming is non-negotiable for natural latency.
- Target <800ms first-audio for natural feel.

[LESSON 02]

Voice Cloning and Ethics

ElevenLabs and Resemble can clone a voice from 30 seconds of audio. The technology is now indistinguishable in short utterances and very good even in long ones. Use cases that work: localizing your own podcast, accessibility (giving someone back their pre-illness voice), audiobook narration in your brand voice. Use cases that don't: anything involving someone else's voice without explicit recorded consent.

Build consent into the workflow. Require uploaded proof of identity for any cloned voice. Watermark audio output (ElevenLabs and others ship audio watermarking by default). Log every generation with the consenting party's record. The regulatory environment around voice clones tightens monthly; building consent infrastructure now is cheaper than retrofitting under a court order later.

// KEY TAKEAWAYS

- Cloning works from 30s of audio.
- Explicit recorded consent is mandatory.
- Watermark output; log every generation.

[LESSON 03]

Real-Time Voice Agents That Work

Phone agents, drive-through, customer service: real-time voice agents are now production-grade for narrow domains. The architecture: telephony provider (Twilio, Vonage) → audio stream → STT → LLM with function calling → TTS → back to caller. LiveKit, Pipecat, and Retell AI are 2026's leading orchestration platforms. They handle interruption detection, turn-taking, and graceful handoff to humans.

Design for graceful failure. Define the top 10 things callers might want that the agent can't do, and give it a clear human handoff for each. Confidence thresholds matter — if the LLM isn't sure, transfer. Customers tolerate handoffs; they hate getting stuck in a loop with a bot that won't admit it doesn't know. Measure handoff rate and reason as the primary product metric.

// KEY TAKEAWAYS

- Voice agents work in narrow domains today.
- Use LiveKit/Pipecat for interruption handling.
- Graceful human handoff is the key metric.

[LESSON 04]

Transcription, Subtitles, and Audio Editing

Whisper-based tools (Whisper itself, MacWhisper, Aiko) now transcribe at near-human accuracy in 90+ languages, free and local. Descript and CapCut layer editing on top — edit audio by editing text, remove filler words automatically, regenerate flubbed lines with voice cloning. For podcasters and creators, this collapses post-production from hours to minutes per episode.

Verify medical, legal, and technical transcriptions. Domain-specific terminology trips up even the best models — drug names, statutes, code identifiers. Build a custom vocabulary file or a per-domain post-processing pass that corrects known terms. For high-stakes use (court filings, medical records), keep a human-in-the-loop reviewer; the cost of one wrong transcription dwarfs the labor savings.

// KEY TAKEAWAYS

- Whisper is near-human in 90+ languages.
- Edit audio by editing text — game changer.
- Verify domain-specific terms; never ship raw.

[NEXT]

Keep Going

You've completed this course material. The real learning starts when you apply what you've read. Pick one idea from this PDF and run a small experiment this week. Document what worked and what didn't. Share your findings with the community.

Explore more free courses, daily AI tips, and curated tools at:

innovationailab.com

Have feedback or want to suggest a topic? We read every message.

hello@innovationailab.com

— Innovation AI Lab Team —

// part of LumiLife Tech