GENERATIVE AI · TAUGHT WITH ELEVENLABS

Speech and Voice AI

Transcription, synthesis, and the turn-taking that decides whether a voice agent feels alive or broken. Includes Arabic and code-switched speech, and what the models still get wrong there.

Intermediate | 5 phases | 25 sessions | 38 contact hours

Syllabus

5 phases, 25 sessions. Each phase ends in something you have built.

1 Audio That Streams

Treat audio as frames on a wire instead of a file you load once it is over.

6 sessions
$72 alone

1.1 Sample Rates and Codecs

16 kHz PCM, 8 kHz mu-law, and the high frequencies the phone network already deleted.

FREE PREVIEW

1.2 Frames, Buffers, Backpressure

20 ms frames over a WebSocket, and the queue that grows quietly until the call drops.

1.3 VAD Is Not Turn Detection

Silero VAD v6 finds speech, then a silence timer still cuts the caller off mid number.

1.4 ASR After Whisper

Whisper large-v3 is now the floor: Scribe v2 at 2.2 percent WER, open Voxtral Small at 2.8.

1.5 Streaming ASR

Partials, revision churn, and time to final measured after the speaker actually stops.

1.6 Diarization and DER

pyannote Community-1 against Precision-2, and why a 9 percent DER meant a clean meeting.

By the end: **A mic-to-transcript tool that prints partials live and a per-turn latency breakdown.**

2 WER on Your Own Audio

Replace a vendor leaderboard number with a measurement taken on your own recordings.

4 sessions
$72 alone

2.1 Computing WER Properly
jiwer, micro-averaging across the corpus, and why per-utterance means flatter a bad score.

2.2 The Normalizer Problem
whisper_normalizer can halve a reported WER, so publish the raw number right next to it.

2.3 Entities Over Averages
Ninety-five percent word accuracy with a third of account numbers wrong is a failed system.

2.4 Hallucination on Silence
Feed hold music, ringing and room noise, then count the words that nobody ever said.

By the end: **A labeled sixty-utterance test set scored raw and normalized, with entity accuracy split out.**

3 Voice Out

Pick and drive a speech model against a latency budget rather than against a demo clip.

5 sessions
$72 alone

3.1 TTS in 2026
Cartesia Sonic 3.5, ElevenLabs v3 versus Flash v2.5, and open weights like Kokoro and Fish S2.

3.2 Time to First Audio
Inference latency, server TTFB and first playable chunk are three different numbers.

3.3 Streaming Text Into TTS
Speak from a partial clause, and watch a naive chunker wreck prosody at every comma.

3.4 Voice Design and Pronunciation
Prompted voices, pronunciation dictionaries, and the Arabic name your model keeps mangling.

3.5 Cloning a Voice
Instant versus professional clones, voice CAPTCHA, and a consent record you can produce later.

By the end: **A streaming voice service that speaks from partial text, with a measured p95 time to first audio.**

4 The Loop

Build the turn-taking machinery that decides whether an agent feels alive or broken.

5 sessions
$72 alone

4.1	**Cascade or Speech-to-Speech**	Run gpt-realtime-2.1 and an STT plus LLM plus TTS pipeline on one task, then read both traces.
4.2	**Semantic Endpointing**	LiveKit's turn detector and Deepgram Flux against pauses inside addresses and phone numbers.
4.3	**Barge-In Is Not cancel()**	Stop generation, flush every buffer, and truncate history to the audio the caller really heard.
4.4	**The 800ms Budget**	Split mouth to ear across endpointing, LLM first token and TTS first audio, then instrument each.
4.5	**Tool Calls Without Dead Air**	Acknowledge inside 300 ms, run the slow lookup async, and make the retry idempotent.

By the end: **An agent a stranger can interrupt mid sentence, with barge-in stop time under 250 ms at p95.**

5 Real Calls, Real Rules

Put the agent on a phone line, in two languages, inside the rules that now actually apply.

5 sessions
$72 alone

5.1	**Arabic and Code-Switching**	One Arabic model scores 5.8 percent on Common Voice and 49.7 percent on Casablanca dialect.
5.2	**Telephony That Breaks**	Twilio Media Streams, raw 8 kHz mu-law, and the WAV header that plays back as static.
5.3	**Conversation Evals**	Simulated callers over the real SIP path, scored on task completion, not a transcript diff.
5.4	**Tracing a Call**	One trace per call, spans per turn, and the recorded audio that proves what the caller heard.
5.5	**Disclosure and Cloning Law**	EU AI Act Article 50 from 2 August 2026, TCPA consent for cloned voices, and the ELVIS Act.

By the end: **A recorded and traced bilingual phone call with disclosure, replayable in a regression suite.**

Tools you will use

ElevenLabs Scribe v2 and Flash v2.5 | OpenAI gpt-realtime-2.1 | LiveKit Agents 1.6 | Pipecat 1.6 | Deepgram Flux | pyannote.audio 4.0 | Silero VAD v6 | jiwer 4 with whisper_normalizer | Twilio Media Streams | NVIDIA Parakeet-TDT-0.6B-v3

What you will build

1. **Your Domain's WER**
   Sixty real utterances, raw and normalized, entities scored apart
2. **The Interruptible Agent**
   Barge-in under 250 ms, latency traced stage by stage
3. **The Bilingual Phone Line**
   Arabic and English over SIP, disclosed, traced, regression tested

Registration

Phase 1: Audio That Streams	$72
Phase 2: WER on Your Own Audio	$72
Phase 3: Voice Out	$72
Phase 4: The Loop	$72
Phase 5: Real Calls, Real Rules	$72
Whole track, all 5 phases	**$300**

Buying every phase separately costs $360, so the whole track saves $60.

No payment is taken online. Registering creates an order and payment instructions are sent with it. Access opens once payment is confirmed.