How to Build Voice AI

The essential building blocks of a production voice agent, from the first SIP packet through speech and reasoning to the agent's configuration.

THE THESIS

Voice AI is 10% advanced speech, 20% LLM, and 70% bi-directional streams.

THE SIGNAL PATH

How a single utterance travels through the stack.

SIP Signaling › Bi-dir Stream › STT + TTS › LLM › Orchestration › Agent

Signaling is optional. It only exists when you bridge to telephony. Each block is explained on the next page.

WHERE IT BREAKS AT SCALE

Most stacks run KARMA. It's proven, but it cracks at millions of calls a day. Production AI needs patches, not plugins.

THE BUILDING BLOCKS

Six layers, one conversation

OPTIONAL

01 SIP Signaling

Sets up, controls, and tears down the call, but only when you bridge to the phone network.

IN PRACTICE

A SIP server like Kamailio or Routr is your signaling foundation.

70%

02 Bi-directional Stream

A raw, full-duplex audio stream between caller and app. The hard 70% everyone underestimates.

IN PRACTICE

Mic in the browser, or AudioSocket / External Media from a server like Asterisk.

10%

03 Speech · STT + TTS

The ears and the voice. Audio in becomes text, text out becomes audio. Streamed, never batched.

IN PRACTICE

Deepgram, Whisper, ElevenLabs, Cartesia, Rime.

20%

04 Language Model

The brain. Turns intent into a reply, and calls tools when it needs to act.

IN PRACTICE

Function-calling makes things happen and memory keeps it grounded in your systems.

CORE

05 Orchestration

The conductor. A state machine to handle turn-taking, barge-in, VAD, tool calls, and handoff.

IN PRACTICE

Where most of the real engineering actually lives.

CORE

06 Agent Mechanics

Configuration, prompts, voices, and guardrails. Everything that defines who the agent is.

IN PRACTICE

In a server like Asterisk, an agent starts in pjsip_wizard.conf.

GO DEEPER

Where to go from here

THE HARD PARTS, IN DETAIL

Latency budget

Sub-second round-trips, or it feels robotic. STT, LLM and TTS each spend part of the budget.

Turn detection & barge-in

VAD detects when the caller starts and stops talking. Barge-in lets them cut in while the agent is still speaking.

Tool calls & state

The model acts through function calls; the state machine keeps the conversation on rails.

Human handoff

Know when to transfer, and carry context across the bridge without dropping metadata.

Scaling past KARMA

Horizontal scale, clean transfers, and metadata that survives a REFER.

FROM THE SOURCE

Programmable Voice Apps with Asterisk
LinkedIn · sanders-pedro

The essentials of building Voice Applications
dev.to/psanders

Building scalable IVRs with Routr + Asterisk
pedrosanders.medium.com

Routr: a programmable SIP server
github.com/fonoster/routr

Build it once. Build it right.
Ready to wire the 70%? Let's talk.

pedrosanders.me →