

Voice AI Glossary: Terms Buyers Meet in Vendor Calls

Plain definitions of the voice AI terms that vendors use: latency, barge-in, turn detection, STT and ASR, TTS, speech-to-speech, SIP trunking, warm transfer, concurrency, BAA and more.

Updated October 29, 2026

Vendors use many technical terms. This glossary explains them in plain words. Terms are in alphabetical order. Prices in examples are from public pages on 2026-10-29 and can change.

A to C

Agent. Software that talks with a caller, decides what to do and takes actions, such as booking an appointment.

ASR (automatic speech recognition). See STT.

BAA (business associate agreement). A contract that HIPAA requires between a US covered entity and a vendor that handles protected health information for it. The required terms are in 45 CFR 164.504(e).

Barge-in. The caller speaks while the agent is speaking. A good agent stops and listens. A poor agent keeps talking.

Bring your own (BYO). You use your own account for a model, voice or carrier. For example, Deepgram lists lower Voice Agent API rates for "BYO LLM" and "BYO TTS".

Cold transfer. The agent sends the call to a human without context. The human starts from zero.

Concurrency. The number of calls the agent handles at the same time. Plans cap it. Retell AI, for example, includes 20 concurrent calls on pay as you go.

Controller and processor. GDPR terms. You are usually the controller. The vendor is the processor. Article 28 sets the terms of the contract between you.

ConversationRelay. A Twilio product for connecting phone calls to an AI application. Twilio lists it at \$0.07 per minute.

D to L

Denoising. Removing background noise from the caller's audio before speech-to-text. Retell AI lists "Advanced Denoising" as a \$0.005 per minute add-on.

Function calling (tool use). The model asks your systems to do something, such as check a calendar. The agent then reads the result back to the caller.

Guardrails. Rules that stop the agent from saying or doing things it should not.

Hallucination. The model states something that is not true. Ground the agent in your own data to reduce this.

Knowledge base. Documents the agent can search during a call, for example your FAQ.

Latency. The delay between the end of the caller's speech and the start of the agent's reply. Each part of the stack adds delay: speech-to-text, the model, text-to-speech and the network. Ask vendors for measured latency on real calls.

LLM (large language model). The model that decides what the agent says.

P to S

Pass-through pricing. The vendor bills a provider's cost without a markup. Vapi, for example, bills models "pass-through at cost".

PII redaction. Removing personal data, such as card numbers, from transcripts and logs. Retell AI lists "PII Removal" at \$0.01 per minute.

Robocall. In the US, a call made with an autodialer or an artificial or prerecorded voice. In FCC 24-17 (February 2024), the FCC confirmed that AI-generated voices are an artificial voice under the TCPA.

SIP (Session Initiation Protocol). The standard protocol that sets up and ends voice calls over the internet.

SIP trunking. A connection that carries calls between your phone system and a carrier over SIP. With it you can keep your numbers and your carrier. Twilio lists its SIP interface at \$0.0040 per minute in the US.

Speech-to-speech. A model that takes audio in and gives audio out, without separate STT and TTS steps. OpenAI lists gpt-realtime-2.1 audio at \$32 per 1M input tokens and \$64 per 1M output tokens.

STT (speech-to-text). Software that turns the caller's speech into text. Deepgram lists Nova-3 streaming at \$0.0077 per minute on pay as you go.

T to W

TCPA (Telephone Consumer Protection Act). The US law that limits robocalls and autodialed calls.

Telephony. The phone network part of the call: numbers, inbound and outbound minutes. Twilio lists US local inbound calls at \$0.0085 per minute.

TTS (text-to-speech). Software that turns the agent's reply into a voice. ElevenLabs lists Flash/Turbo at \$0.04 per 1,000 characters.

Turn detection (end-of-turn detection). Deciding when the caller has finished speaking. If it fires too early, the agent interrupts. If it fires too late, the agent feels slow. LiveKit Agents uses a transformer model for this. LiveKit says it "helps to reduce interruptions."

VAD (voice activity detection). Detecting when there is speech in the audio. Pipecat lists Silero VAD among its audio processing options.

Warm transfer. The agent passes the call to a human with context, for example a summary.

The caller does not repeat the story.

WebRTC. A standard for real-time audio and video in browsers and apps. Pipecat and LiveKit use it for non-phone calls.

Zero data retention (ZDR). The vendor does not keep your call data after processing. Vapi lists ZDR on its Core and Pro plans. Ask what ZDR covers at each vendor.

Frequently asked questions

What is the difference between STT and ASR?

None in practice. Both mean software that turns speech into text. ASR stands for automatic speech recognition. STT stands for speech-to-text.

What is a speech-to-speech model?

A model that takes audio in and gives audio out, without a separate text step for speech-to-text and text-to-speech. OpenAI's Realtime models are one example.

Why does concurrency matter when I buy?

Plans cap the number of calls at the same time. If your peak is higher than the cap, callers wait or the call fails. Some vendors charge extra for each additional concurrent call.

Sources

[LiveKit Agents, GitHub repository \(fetched 2026-10-29\)](https://github.com/livekit/agents) — <https://github.com/livekit/agents>

[Pipecat, GitHub repository \(fetched 2026-10-29\)](https://github.com/pipecat-ai/pipecat) — <https://github.com/pipecat-ai/pipecat>

[LiveKit, Pricing \(fetched 2026-10-29\)](https://livekit.com/pricing) — <https://livekit.com/pricing>

[Twilio, Programmable Voice pricing, United States \(fetched 2026-10-29\)](https://www.twilio.com/en-us/voice/pricing/us) — <https://www.twilio.com/en-us/voice/pricing/us>

[Deepgram, Pricing \(fetched 2026-10-29\)](https://deepgram.com/pricing) — <https://deepgram.com/pricing>

[ElevenLabs, API pricing \(fetched 2026-10-29\)](https://elevenlabs.io/pricing/api) — <https://elevenlabs.io/pricing/api>

[OpenAI, API pricing \(fetched 2026-10-29\)](https://developers.openai.com/api/docs/pricing) — <https://developers.openai.com/api/docs/pricing>

[Vapi, Pricing \(fetched 2026-10-29\)](https://vapi.ai/pricing) — <https://vapi.ai/pricing>

[Retell AI, Pricing \(fetched 2026-10-29\)](https://www.retellai.com/pricing) — <https://www.retellai.com/pricing>

[FCC, Declaratory Ruling FCC 24-17, CG Docket No. 23-362 \(fetched 2026-10-29\)](https://docs.fcc.gov/public/attachments/FCC-24-17A1.pdf) — <https://docs.fcc.gov/public/attachments/FCC-24-17A1.pdf>

[eCFR, 45 CFR 164.504, Uses and disclosures: organizational requirements \(fetched 2026-10-29\)](https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.504) — <https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.504>

[GDPR Article 28, Processor \(gdpr-info.eu\) \(fetched 2026-10-29\)](https://gdpr-info.eu/art-28-gdpr/) — <https://gdpr-info.eu/art-28-gdpr/>