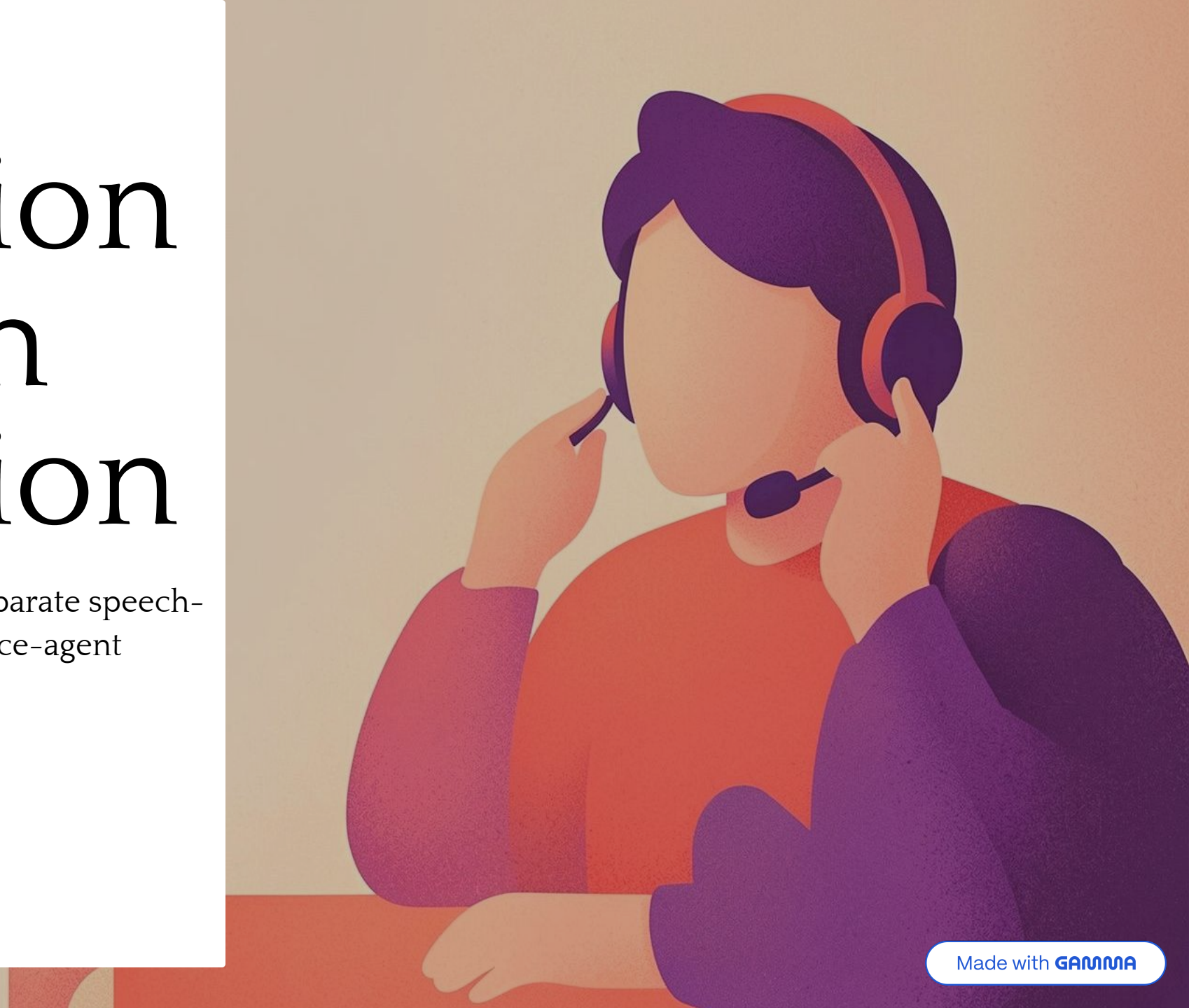
AI Voice Recognition vs Speech Recognition

Help product and technology buyers separate speech-to-text, speaker recognition, and full voice-agent systems before comparing vendors.

Back to blog



Made with GAMMA

START WITH THE DISTINCTION

“Voice recognition” usually means two different jobs

Separate what is being heard from who is speaking.

Speech recognition

Converts spoken words to text for dictation, captions, commands, and transcripts.

Speaker recognition

Identifies or verifies who is speaking.

Voice biometrics uses speaker verification for authentication.

When a vendor shows a transcript, ask whether its accuracy claim is for ASR or for identifying the speaker.

Made with GAMMA

SYSTEM VIEW

A voice assistant is a chain, not a single model

Latency, barge-in handling, and failed transcripts matter alongside word error rate.

01

Capture audio

02

Turn speech into text

Often as a live stream.

03

Extract intent and entities

04

Choose the next action or reply

05

Speak the answer

With text-to-speech.

06

Add speaker recognition

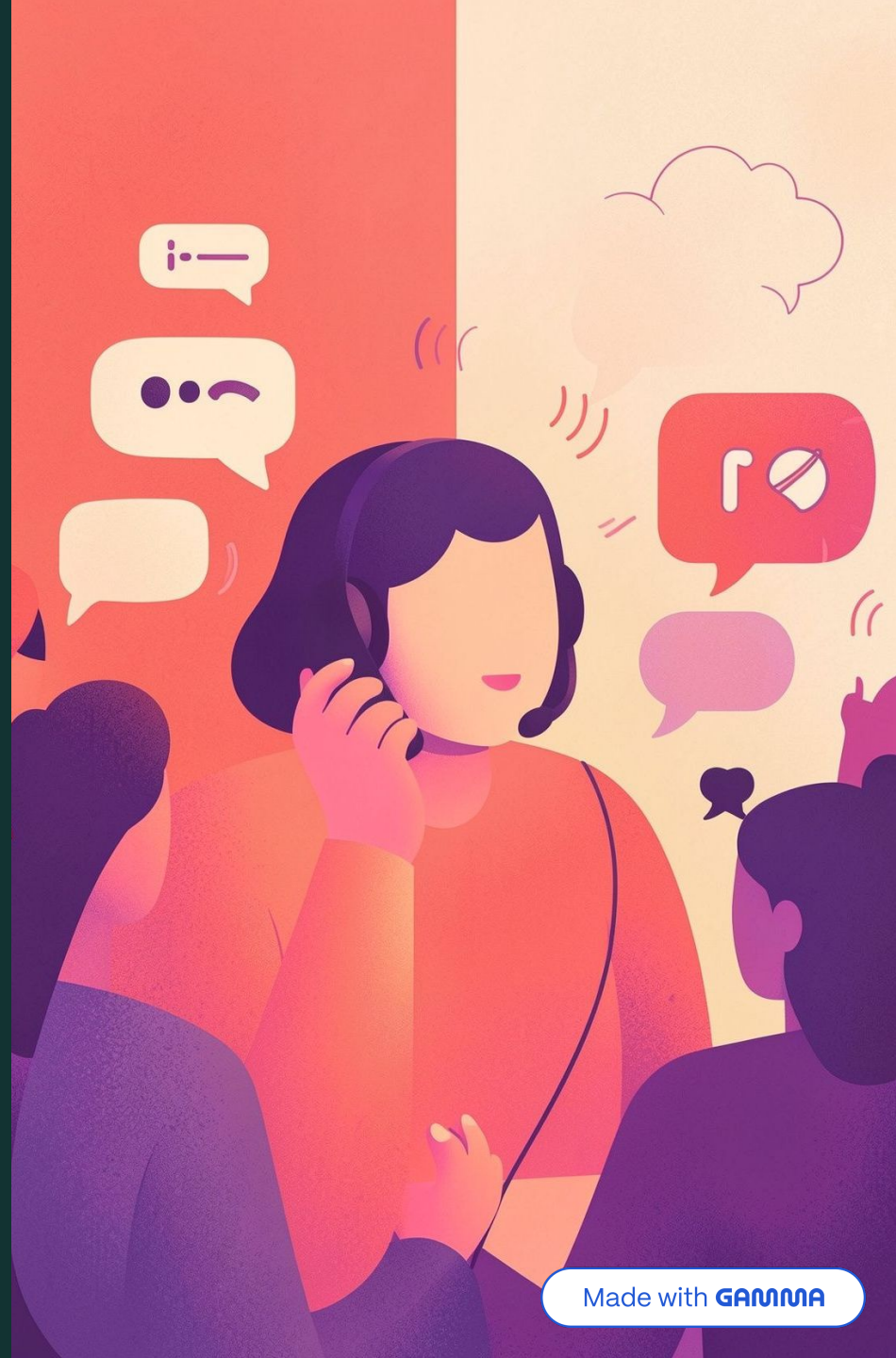
Only when identity matters.

Made with GAMMA

CONTEXT CHANGES THE RESULT

Indian English is workable until the real-world conditions change

- Clean, reasonably neutral Indian English is a sensible starting point.
- Accuracy deteriorates with phone-quality audio, noise, code-switching, and a **heavy regional accent**.
- Global models can show large gaps on Indian languages and some accents; India-focused models such as Sarvam may rank near the top across many languages and dialects.



Made with GAMMA

SHORTLIST BY FIT

Choose the engine around the language and channel

Language or channel	Start with	Keep testing
Clean Indian English	Google Cloud Speech-to-Text, Azure, Deepgram, Whisper, or AssemblyAI.	Performance on your channel.
Hindi, Hinglish, and regional languages	India-first engines such as Sarvam.	Google Chirp or local specialists.
Contact centre	India-focused platforms such as Mihup and Gnani.ai.	Google has broad Indian language coverage.

Made with GAMMA



EVIDENCE OVER BROCHURES

“High accuracy” only counts on your recordings

- A contact-centre comparison placed Mihup and Gnani in **a top tier for real-call Hindi and Hinglish**.
- Studio word error rate is not the same as performance in a moving car, on a Bluetooth headset, or across a noisy call.
- Demand a test set that matches your city, language mix, and channel.

Made with GAMMA

BUY THE OPERATING ENVIRONMENT

Call-centre buyers should evaluate the whole operating environment



Streaming and low latency.



Noise, compression, and crosstalk robustness.



Custom vocabulary for names, products, and policy numbers.



Speaker diarization for agent versus customer.

Language and code-switch handling.

Listen to recorded calls from your own queues before trusting a brochure WER.

Made with GAMMA

PRODUCT ARCHITECTURE

Mobile integration is an architecture choice

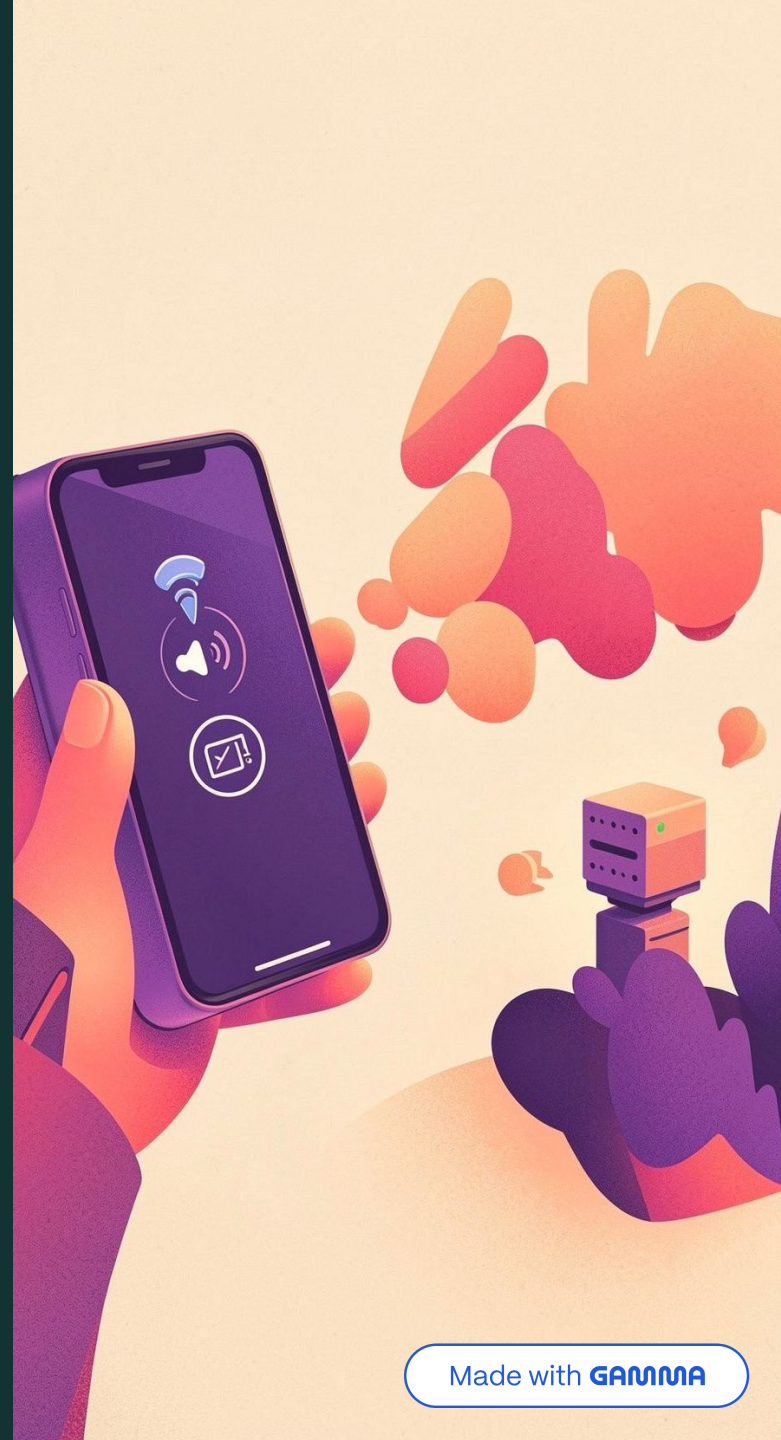
01 Choose streaming commands or batch transcription.

02 Pick an API such as Google, Azure, Deepgram, or Sarvam, or a Whisper-compatible endpoint.

03 Record and stream or upload audio through your server.

04 Handle permissions, background limits, offline fallback, language settings, partial results, and low-confidence states.

05 Keep API keys off the client; use on-device recognition when simple commands do not need to leave the device.



Made with GAMMA

EVALUATION PLAN

Build a test plan, not a vendor scorecard

Compare results on the audio, language, and channel you actually operate.

01 Measure word error rate on your audio, **not the vendor's demo**.

02 Track time to first partial and final result.

03 Check names, numbers, product terms, Hinglish, and regional-language samples.

04 Include cost per minute at your expected volume.

05 Use **your own accent** and channel when comparing results.

Made with GAMMA

THE DECISION

Make the buying decision with the right mental model

For the full calling conversation, EchoLeads.ai is positioned as a voice-agent platform rather than a raw speech-to-text API; its voice agents incorporate speech recognition for live calls.

Need	Better question
What is it?	Is this STT, speaker ID, or both?
Indian English	How does it perform on our accent and channel?
Regional languages	What does it do on our real Hindi, Tamil, Telugu, or Bengali audio?
Call centre	How do latency, noise, diarization, and vocabulary perform together?
Assistant	Where do ASR, intent, dialog, and TTS meet?

Made with GAMMA