The Voice AI Pipeline

Four layers, each transforming the signal before passing it on. Voice cloning is an optional fourth layer.

ASR

Automatic Speech Recognition

Spoken audio converted to text. The first point of failure for dialect and accent.

Whisper, Deepgram

LLM

Language Model

Intelligence and reasoning. Most of the cost and latency sit here.

GPT-4, Claude

TTS

Text-to-Speech

The response converted back to audio. A generic voice unless a clone is applied.

ElevenLabs, Cartesia

CLONE

Voice Clone (optional)

Replaces generic TTS with a specific person's voice. Quality is limited by the source audio.

ElevenLabs IVC / PVC

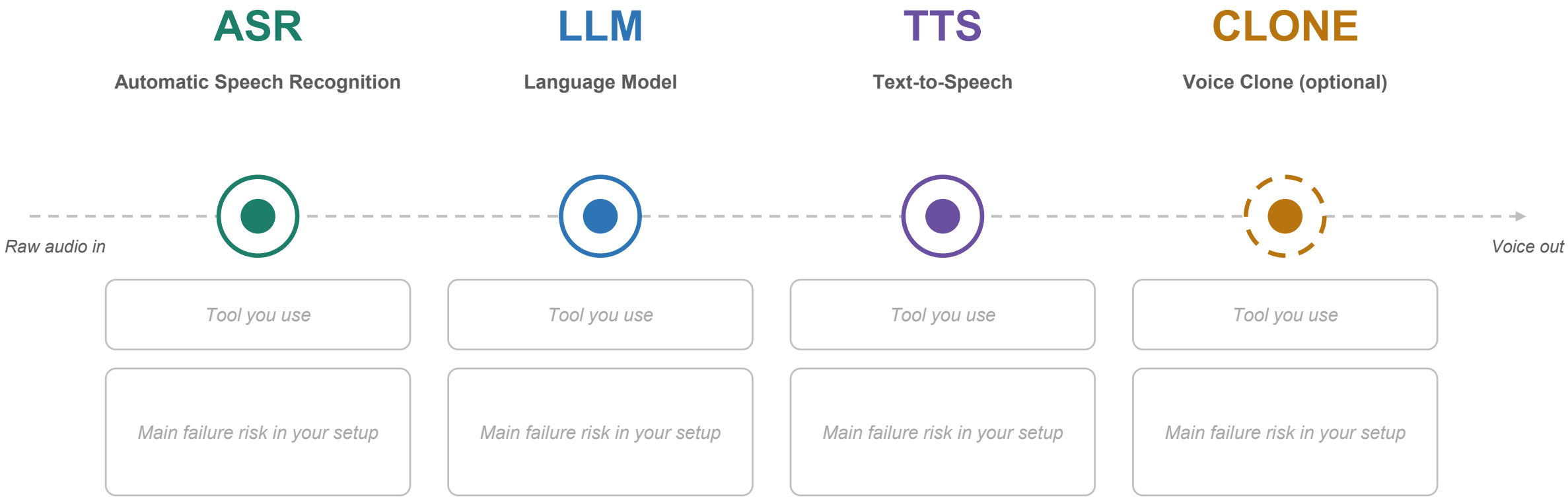


GIGU: Garbage In, Garbage Out. The quality ceiling at each layer is set by the weakest input before it. Voice agents combine all four layers into a real-time conversation that can also take actions.

Source: When AI Picks Up the Phone

Map your voice pipeline

Note the tool you use at each layer and where your setup is most likely to fail.



GIGU: Garbage In, Garbage Out. The quality ceiling at each layer is set by the weakest input before it. Voice agents combine all four layers into a real-time conversation that can also take actions.

Source: [When AI Picks Up the Phone]

About this diagram

The Voice AI Pipeline

When this fits

- Teams planning a voice agent, IVR replacement or voice-note product, who need to see where each tool fits.
- Diagnosing a voice system that underperforms: the pipeline shows which layer to look at first.
- Explaining to stakeholders why accent and dialect handling matters so much in multilingual markets.

When it doesn't fit

- Text-only chatbots. Only the language model layer applies.
- Detailed vendor selection. The tools shown are examples from the post, and this market changes quickly.

How to use this file

- Every element is a native PowerPoint shape: edit the text, recolour it, or delete what you don't need.
- Use the blank template slide (the one headed Map your or Plan your) to fill in your own version.
- Copy the diagram straight into your own deck. It is sized for a standard widescreen slide.

Source: When AI Picks Up the Phone