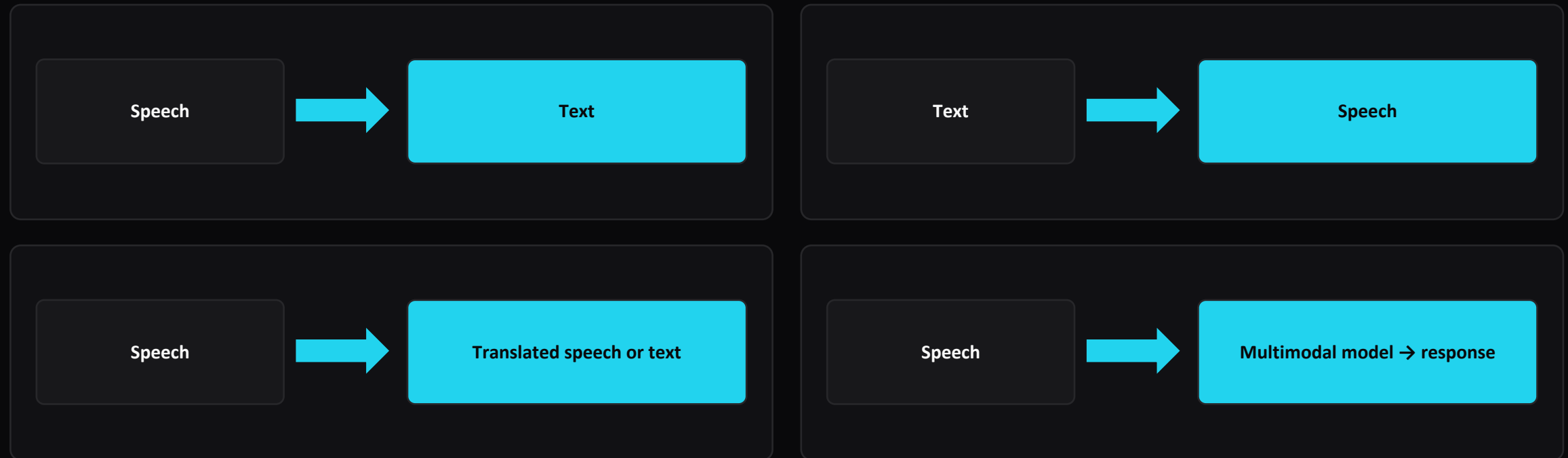


What happens to the spoken language?



Key message

Follow the direction of the transformation.

Speech recognition

Converts spoken language into text.

Also called

- Speech to text
- Transcription
- Automatic speech recognition

Example

Audio: Customer describes a damaged product
Text: "The screen cracked during delivery."

Common scenarios

- Transcribe calls and meetings
- Create captions
- Accept voice commands
- Search recorded conversations

Exam signal

Convert spoken words into written text.

Speech synthesis

Converts text into audible speech.

Also called

- Text to speech
- Voice synthesis

Common scenarios

- Read an assistant's answer aloud
- Generate accessibility narration
- Voice automated announcements
- Give a virtual assistant a voice

Recognition listens. Synthesis speaks.

Exam signal

Produce spoken output from text.

Transcription versus translation

Capability	Language changes?	Example
Transcription	No	Spoken Spanish → written Spanish
Speech translation	Yes	Spoken Spanish → written English
Speech-to-speech translation	Yes	Spoken Spanish → spoken English

Question to ask

Preserve language?

Must the output preserve the input language?

Exam trap

Audio to text is not automatically translation.

Rule

Translation requires a change between languages.

Beyond transcription

A multimodal model can respond directly to spoken input.

Example · “A customer speaks a question while showing the assistant a photograph of a damaged product.”

The model can use

- The words that were spoken
- Tone, pacing, or other audio context
- Conversation history
- Additional text or image inputs

Key distinction

- Written words only: Speech recognition
- Intelligent conversational response: Multimodal model
- An action completed: Potentially agentic AI

The model considers both

Spoken request and visual evidence.

Real-time processing

Best fit for

- Voice assistants
- Live captions
- Contact-center conversations
- Live interpretation
- Interactive voice applications

Priorities

- Low latency
- Partial or streaming results
- Rapid turn-taking

Scenario signal

A person is waiting for an immediate response.

Batch processing

Best fit for

- Recorded meetings
- Call-center archives
- Podcast transcription
- Large media collections
- Compliance analysis

Priorities

- Scale and throughput
- Processing many files
- Rich results after completion
- Immediate response not required

Scenario signal

Existing audio must be processed in volume.

Choose real-time or batch

Requirement	Best approach
Display captions during a meeting	Real time
Transcribe 50,000 recorded calls	Batch
Let a customer speak to an assistant	Real time
Analyze last month's interviews	Batch
Interpret a live conversation	Real time

Rule of thumb

Interactive now → real time. Recorded at scale → batch.

Common “which capability?” patterns

Scenario asks the system to...

Best answer

Produce captions from a lecture

Speech recognition

Read an answer aloud

Speech synthesis

Convert spoken French into English

Speech translation

Answer a spoken question about an image

Multimodal interaction

Transcribe an archive of recorded calls

Batch speech recognition

Respond during a live phone call

Real-time speech

Exam strategy

Follow the direction of the transformation, then decide whether anyone is waiting.

A company has 10,000 recorded customer-service calls. It wants searchable text versions by tomorrow morning.

Which approach is most appropriate?

A. Real-time speech synthesis

B. Batch speech recognition

C. Real-time speech translation

D. Multimodal image analysis

Answer reveal

B. Batch speech recognition

The input consists of completed audio recordings, the desired output is text, and no one is waiting for an immediate response.