Speech-to-Text

Speech-to-text conversion powered by machine learning.

TRY IT FREE (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/FREETRIAL)

View documentation (https://cloud.google.com/speech-to-text/docs) for this product.

Powerful speech recognition

Google Speech-to-Text enables developers to convert audio to text by applying powerful neural network models in an easy-to-use API. The API recognizes 120 languages and variants to support your global user base. You can enable voice command-and-control, transcribe audio from call centers, and more. It can process real-time streaming or prerecorded audio, using Google's machine learning technology.

Convert your speech to text right now

Language

English (United States)

Speaker diarization BETA

Off

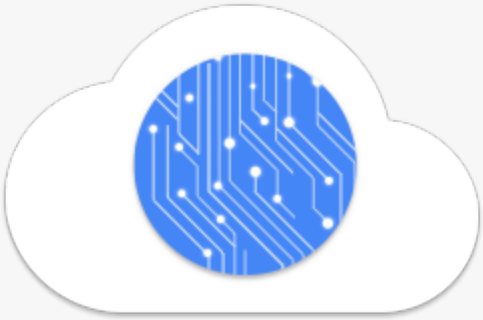
Speakers

1 speaker

Punctuation

Show JSON

CHOOSE FILE

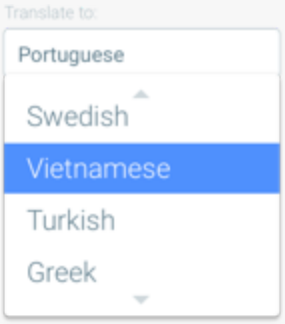


Powered by machine learning

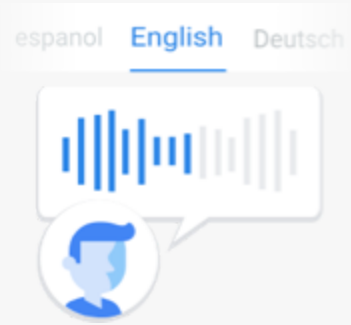
Apply the most advanced deep-learning neural network algorithms to audio for speech recognition with unparalleled accuracy. Accuracy improves over time as Google improves the internal speech recognition technology used by Google products.

Recognizes 120 languages and variants

Speech-to-Text can support your global user base, recognizing 120 languages and variants (https://cloud.google.com/speech-to-text/docs/languages)



. You can also filter inappropriate content in text results for all languages.



Automatically identifies spoken language

Using Speech-to-Text you can identify what language is spoken in the utterance (up to four languages). This can be used for voice search (such as, “What is the temperature in Paris?”) and command use cases (such as, “Turn the volume up.”)

Returns text transcription in real time for short-form or long-form audio

Speech-to-Text can stream text results, immediately returning text as it’s recognized from streaming audio or as the user is speaking. Alternatively, Speech-to-Text can return recognized text from audio stored in a file. It’s capable of analyzing short-form and long-form audio.

Automatically transcribes proper nouns and context-specific formatting

Speech-to-Text is tailored to work well with real-life speech and can accurately transcribe proper nouns (e.g., names, places) and appropriately format language (e.g., dates, phone numbers). Google supports more than 10x proper nouns compared to the number of words in the entire Oxford English Dictionary.

Offers selection of pre-built models, tailored for your use case

Speech-to-Text comes with multiple pre-built speech recognition models so you can optimize for your use case (such as voice commands). Example: Our pre-built video transcription model is ideal for indexing or subtitling video and/or multispeaker content and uses machine learning technology that is similar to YouTube captioning.

MODEL	DESCRIPTION

MODEL	DESCRIPTION
command_and_search	Best for short queries such as voice commands or voice search.
phone_call	Best for audio that originated from a phone call (typically recorded at an 8khz sampling rate).
video	Best for audio that originated from video or includes multiple speakers. Ideally the audio is recorded at a 16khz or greater sampling rate. This is a premium model that costs more than the standard rate.
default	Best for audio that is not one of the specific audio models. For example, long-form audio. Ideally the audio is high-fidelity, recorded at a 16khz or greater sampling rate.

Features

Automatic speech recognition

Automatic speech recognition (ASR) powered by deep learning neural networking to power your applications like voice search or speech transcription.

Global vocabulary

Noise robustness

Handles noisy audio from many environments without requiring additional noise cancellation.

Inappropriate content filtering

Filter inappropriate content in text results for some languages.

Recognizes 120 languages and variants with an extensive vocabulary.

Customized speech recognition

Manually customize speech recognition for your business by specifying up to 5,000 words or phrases that are likely to be spoken (such as product names). Also automatically convert spoken numbers into addresses, years, or currencies, or do other conversions, depending on context.

Real-time streaming or prerecorded audio support

Audio input can be streamed from an application's microphone or sent from a prerecorded audio file (inline or through Google Cloud Storage). Multiple audio encodings are supported, including FLAC, AMR, PCMU, and Linear-16.

Auto-Detect Language (beta)

When you need to support multilingual scenarios, you can now specify two to four language codes and Cloud Speech-to-Text will identify the correct language spoken and provide the transcript.

Automatic Punctuation (beta)

Accurately punctuates transcriptions (e.g., commas, question marks, and periods) with machine learning.

Model selection

Choose from a selection of four pre-built models: default, voice commands and search, phone calls, and video transcription.

Speaker Diarization (beta)

Know who said what you can now get automatic predictions about which of the speakers in a conversation spoke each utterance.

Multichannel recognition

In multiparticipant recordings where each participant is recorded in a separate channel (e.g., phone call with two channels or video conference with four channels), Cloud Speech-to-Text will recognize each channel separately and then annotate the transcripts so that they follow the same order as in real life.

Pricing

Speech-to-Text is priced per 15 seconds of audio processed after a 60-minute free tier. For details, please see our pricing guide. (https://cloud.google.com/speech-to-text/pricing/)

FEATURE	STANDARD MODELS (ALL MODELS EXCEPT ENHANCED PHONE AND VIDEO)		PREMIUM MODELS* (ENHANCED PHONE, VIDEO)	
	0-60 Minutes	Over 60 Mins up to 1 Million Mins	0-60 Minutes	Over 60 Mins up to 1 Million Mins
Speech Recognition (without Data Logging - default)	Free	$0.006 / 15 seconds **	Free	$0.009 / 15 seconds **
Speech Recognition (with Data Logging opt-in)	Free	$0.004 / 15 seconds **	Free	$0.006 / 15 seconds **

This pricing is for applications on personal systems (e.g., phones, tablets, laptops, desktops). Please contact us (https://services.google.com/fb/forms/speech-api-pricing-request/) for approval and pricing to use the Cloud Speech-to-Text API on embedded devices (e.g., cars, TVs, appliances, or speakers).

* Currently available for US English only

** Each request is rounded up to the nearest increment of 15 seconds. For example, if you make three separate requests (Standard model), each containing 7 seconds of audio, you are billed $0.018 USD for 45 seconds (3 × 15 seconds) of audio. Fractions of seconds are included when rounding up to the nearest increment of 15 seconds. That is, 15.14 seconds are rounded up and billed as 30 seconds.

TRY IT FREE (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/FREETRIAL)

Optimize Speec...

WATCH NEXT '19 VIDEO

(https://www.youtube.com/watch?v=71jzm19xn4U&autoplay=1)

A product or feature listed on this page is in beta. For more information on our product launch stages, see here (https://cloud.google.com/terms/launch-stages).

Cloud AI products comply with the SLA policies listed here (https://cloud.google.com/terms/sla/). They may offer different latency or availability guarantees from other Google Cloud services.