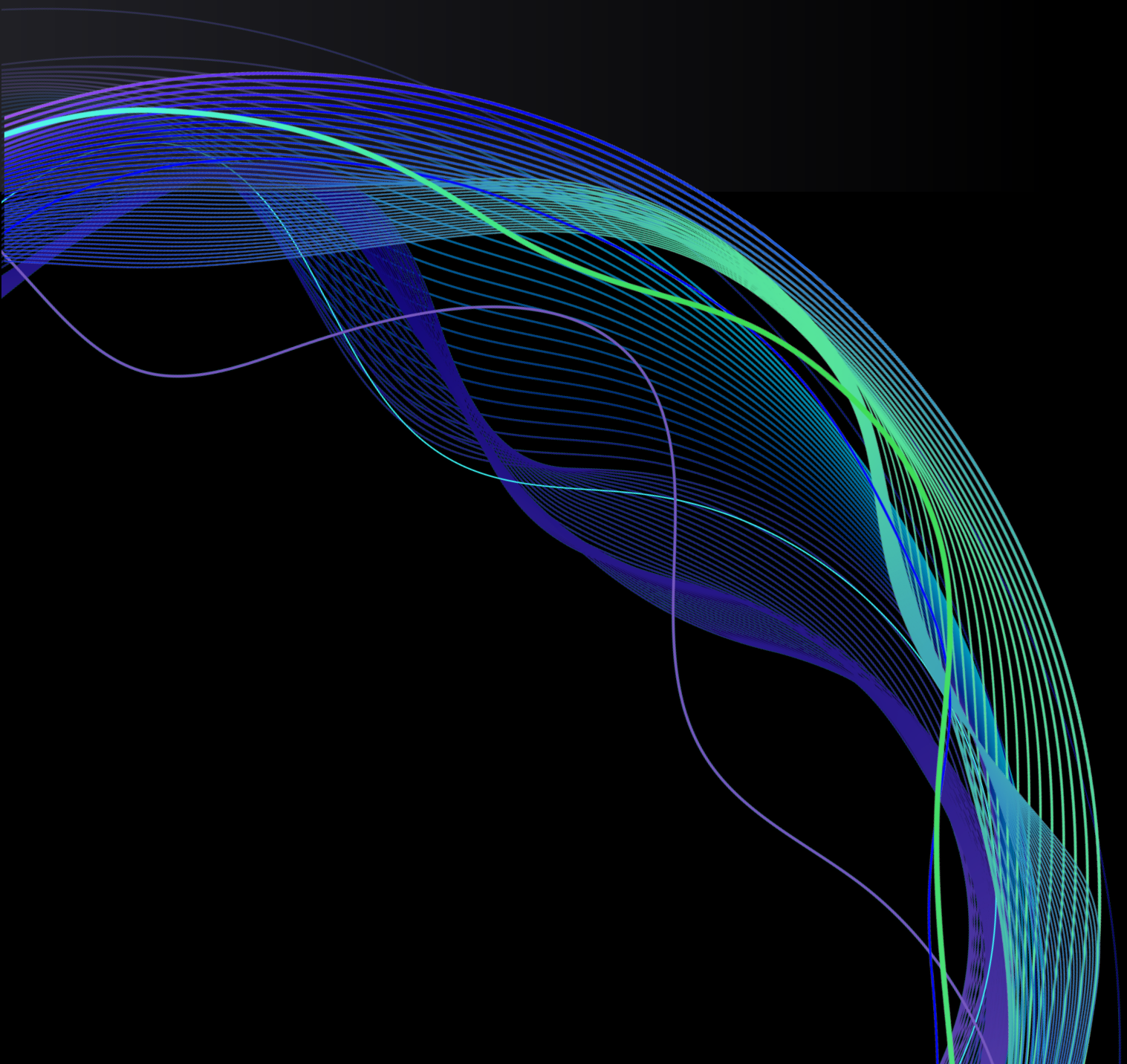
Deepgram

2024 Guide:

# Best Speech-to-text APIs



# Table of contents

# Introduction

If you've been shopping for a speech-to-text (STT) solution for your business, you're not alone. In our recent State of Voice Technology report, 82% of respondents confirmed their current utilization of voice-enabled technology, a 6% increase from last year.

The vast number of options for speech transcription can be overwhelming, especially if you're unfamiliar with the space. From Big Tech to open source options, there are many choices, each with different price points and feature sets. While this diversity is great, it can also be confusing when you're trying to compare options and pick the right solution.

This article breaks down the leading speech-to-text APIs available today, outlining their pros and cons and providing a ranking that accurately represents the current STT landscape. Before getting to the ranking, we explain exactly what an STT API is, and the core features you can expect an STT API to have, and some key use cases for speech-to-text APIs.

# What is a speech-to-text API?

At its core, a speech-to-text (also known as automatic speech recognition, or ASR) application programming interface (API) is simply the ability to call a service to transcribe audio containing speech into written text. The STT service will take the provided audio data, process it using either machine learning or legacy techniques (e.g. Hidden Markov Models), and then provide a transcript of what it has inferred was said.

# What are the most important things to consider when choosing a speech-to-text API?

What makes the best speech-to-text API? Is the fastest speech-to-text API the best? Is the most accurate speech-to-text API the best? Is the most affordable speech-to-text API the best? The answers to these questions depend on your specific project and are thus certainly different for everybody. There are a number of aspects to carefully consider in the evaluation and selection of a transcription service and the order of importance is dependent on your target use case and end user needs.

**Accuracy** - A speech-to-text API should produce highly accurate transcripts, even while dealing with varying levels of speaking conditions (e.g. background noise, dialects, accents, etc.). “Garbage in, garbage out,” as the saying goes. The vast majority of voice applications require highly accurate results from their transcription service to deliver value and a good customer experience to their users.

**Speed** - Many applications require quick turnaround times and high throughput. A responsive STT solution will deliver value with low latency and fast processing speeds.

**Cost** - Speech-to-text is a foundational capability in the application stack, and cost efficiency is essential. Solutions that fail to deliver adequate ROI and a good price-to-performance ratio will be a barrier to the overall utility of the end user application.

**Modality** - Important input modes include support for pre-recorded or real-time audio:

- Batch or pre-recorded transcription capabilities - Batch transcription won't be needed by everyone, but for many use cases, you'll want a service that you can send batches of files to to be transcribed, rather than having to do it one-by-one on your end.
- Real-time streaming - Again, not everyone will need real-time streaming. However, if you want to use STT to create, for example, truly conversational AI that can respond to customer inquiries in real time, you'll need to use a STT API that returns its results as quickly as possible.

**Features & Capabilities** - Developers and companies seeking speech processing solutions require more than a bare transcript. They also need rich features that help them build scalable products with their voice data, including sophisticated formatting and speech understanding capabilities to improve readability and utility by downstream tasks.

**Scalability and Reliability** - A good speech-to-text solution will accommodate varying throughput needs, adequately handling a range of audio data volumes from small startups to large enterprises. Similarly, ensuring reliable, operational integrity is a hard requirement for many applications where the effects from frequent or lengthy service interruption could result in revenue impacts and damage to brand reputation.

**Customization, Flexibility, and Adaptability** - One size, fits few. The ability to customize STT models for specific vocabulary or jargon as well as flexible deployment options to meet project-specific privacy, security, and compliance needs are important, often overlooked considerations in the selection process.

**Ease of Adoption and Use** - A speech-to-text API only has value if it can be integrated into an application. Flexible pricing and packaging options are critical, including usage-based pricing with volume discounts. Some vendors do a better job than others to provide a good developer experience by offering frictionless self-onboarding and even including free tiers with an adequate volume of credits to help developers test the API and prototype their applications before choosing the best subscription option to choose.

**Support and Subject Matter Expertis** - Domain experts in AI, machine learning, and spoken language understanding are an invaluable resource when issues arise. Many solution providers outsource their model development or offer STT as a value-add to their core offering. Vendors for whom speech AI is their core focus are better equipped to diagnose and resolve challenge issues in a timely fashion. They are also more inclined to make continuous improvements to their STT service and avoid issues with stagnating performance over time.

## What are the most important features of a speech-to-text API?

In this section, we'll survey some of the most common features that STT APIs offer. The key features that are offered by each API differ, and your use cases will dictate your priorities and needs in terms of which features to focus on.

Multi-language support - If you're planning to handle multiple languages or dialects, this should be a key concern. And even if you aren't planning on multilingual support now, if there's any chance that you would in the future, you're best off starting with a service that offers many languages and is always expanding to more.

Formatting - Formatting options like punctuation, numeral formatting, paragraphing, speaker labeling (or speaker diarization), word-level timestamping, profanity filtering, and more, all to improve readability and utility for data science.

- Automatic punctuation & capitalization - Depending on what you're planning to do with your transcripts, you might not care if they're formatted nicely. But if you're planning on surfacing them publicly, having this included in what the STT API provides can save you time.
- Profanity filtering or redaction - If you're using STT as part of an effort for community moderation, you're going to want a tool that can automatically detect profanity in its output and censor it or flag it for review.

Understanding - A primary motivation for employing a speech-to-text API is to gain understanding of who said what and why they said it. Many applications employ natural language and spoken language understanding tasks to accurately identify, extract, and summarize conversational audio to deliver amazing customer experiences.

- Topic detection - Automatically identify the main topics and themes in your audio to improve categorization, organization, and understanding of large volumes of spoken language content..
- **Intent detection** - Similarly, intent detection is used to determine the purpose or intention behind the interactions between speakers, enabling more efficient handling by downstream agents or tasks in a system in order to determine the next best action to take or response to provide.
- **Sentiment analysis** - Understand the interactions, attitudes, views, and emotions in conversational audio by quantitatively scoring the overall and component sections as being positive, neutral, or negative.
- **Summarization** - Deliver a concise summary of the content in your audio, retaining the most relevant and important information and overall meaning, for responsive understanding, analysis, and efficient archival.

Keywords **(a.k.a. Keyword Boosting)** - Being able to include an extended, custom vocabulary is helpful if your audio has lots of specialized terminology, uncommon proper nouns, abbreviations, and acronyms that an off-the-shelf model wouldn't have been exposed to. This allows the model to incorporate these custom terms as possible predictions.

Custom models - While keywords provide inclusion of a small set of specialized, out-of-vocabulary words, a custom model trained on representative data will always give the best performance. Vendors that allow you to tailor a model for your specific needs, fine-tuned on your own data, give you the ability to boost accuracy beyond what an out-of-the-box solution alone provides.

Accepts multiple audio formats - Another concern that won't be present for everyone is whether or not the STT API can process audio in different formats. If you have audio coming from multiple sources that aren't encoded in the same format, having a STT API that removes the need for converting to different types of audio can save you time and money.

## What are the top speech-to-text use cases?

As noted at the outset, voice technology that's built on the back of STT APIs is a critical part of the future of business. So what are some of the most common use cases for speech-to-text APIs? Let's take a look.

**Smart assistants** - Smart assistants like Siri and Alexa are perhaps the most frequently encountered use case for speech-to-text, taking spoken commands, converting them to text, and then acting on them.

Conversational AI - Voicebots let humans speak and, in real time, get answers from an AI. Converting speech to text is the first step in this process, and it has to happen quickly for the interaction to truly feel like a conversation.

**Sales and support enablement** - Sales and support digital assistants that provide tips, hints, and solutions to agents by transcribing, analyzing and pulling up information in real time. It can also be used to gauge sales pitches or sales calls with a customer.

Contact centers - Contact centers can use STT to create transcripts of their calls, providing more ways to evaluate their agents, understand what customers are asking about, and provide insight into different aspects of their business that are typically hard to assess.

Speech analytics - Broadly speaking, speech analytics is any attempt to process spoken audio to extract insights. This might be done in a call center, as above, but it could also be done in other environments, like meetings or even speeches and talks.

**Accessibility** - Providing transcriptions of spoken speech can be a huge win for accessibility, whether it's providing captions for classroom lectures or creating badges that transcribe speech on the fly.

# How do you evaluate performance of a speech-to-text API?

All speech-to-text solutions aim to produce highly accurate transcripts in a user-friendly format. We advise performing side-by-side accuracy testing using files that resemble the audio you will be processing in production to determine the best speech solution for your needs. The best evaluation regimes employ a holistic approach that includes a mix of quantitative benchmarking and qualitative human preference evaluation across the most important dimensions of quality and performance, including accuracy and speed.

The generally accepted industry metric for measuring transcription quality is Word Error Rate (WER). Consider WER in relation to the following equation:

***WER + Accuracy Rate = 100%***

Thus, an 80% accurate transcript corresponds to a WER of 20%

WER is an industry standard focusing on error rate rather than accuracy as the error rate can be subdivided into distinct error categories. These categories provide valuable insights into the nature of errors present in a transcript. Consequently, WER can also be defined using the formula:

***WER = (# of words inserted + # of words deleted + # of words substituted) / total # of words.***

We suggest a degree of skepticism towards vendor claims about accuracy. This includes the qualitative claim that OpenAI's model "approaches human level robustness on accuracy in English," and the WER statistics published in Whisper's documentation.

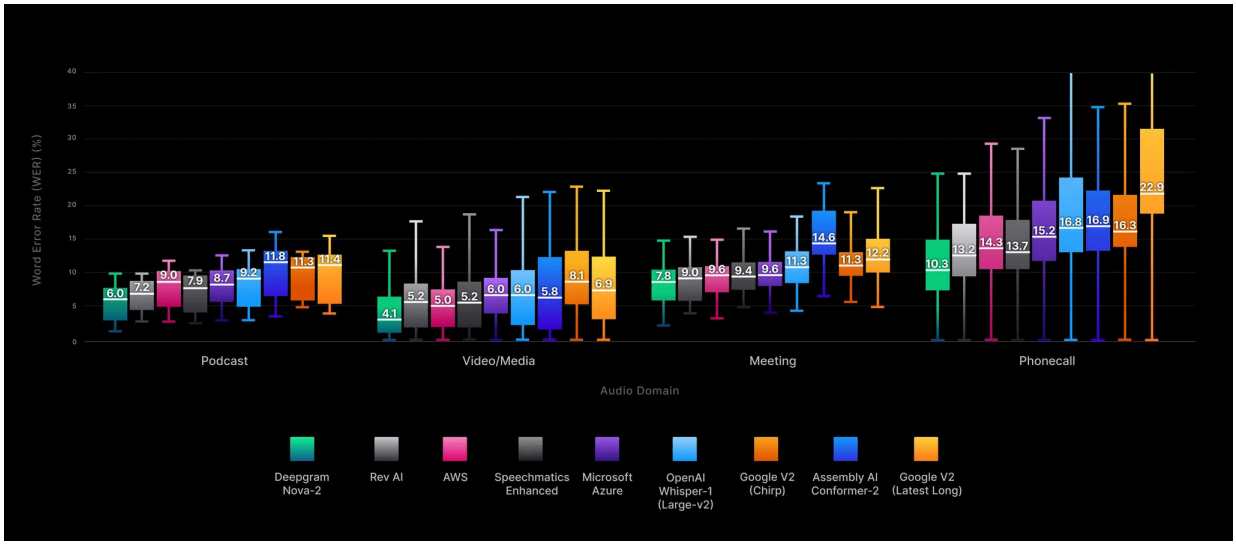


*Figure 1: The figure above compares the average Word Error Rate (WER) of Deepgram Nova-2 with other popular models across four audio domains: video/media, podcast, meeting, and phone call. It uses a boxplot chart, which is a type of chart often used to visually show the distribution of numerical data and skewness. The chart displays the five-number summary of each dataset, including the minimum value, first quartile (median of the lower half), median, third quartile (median of the upper half), and maximum value.*

One limitation of WER as a benchmarking tool is its high sensitivity to the difficulty of the audio data it measures. For example, testing our product using two different audio files—one with "easy" audio (i.e., slowly-spoken, simple vocabulary, and good diction, recorded with high-quality equipment in a quiet environment), and another with challenging real-world audio (i.e., a fast paced conversation full of industry jargon, where the speakers are far from the microphone in a noisy environment and frequently speak over each other)—can result in significant variance in WER from a single model. The self-reported WER figures from other vendors often represent easy audio. We strongly recommend thorough testing using real-world data for any STT API under consideration to validate such claims.

The best benchmarking methodology utilizes holdout data sets (i.e. not used for training) taken from real-life scenarios. These should encompass a wide range of audio lengths, diverse accents, varying environments, and subjects, and be representative of the data the target speech-to-text API will encounter in production.

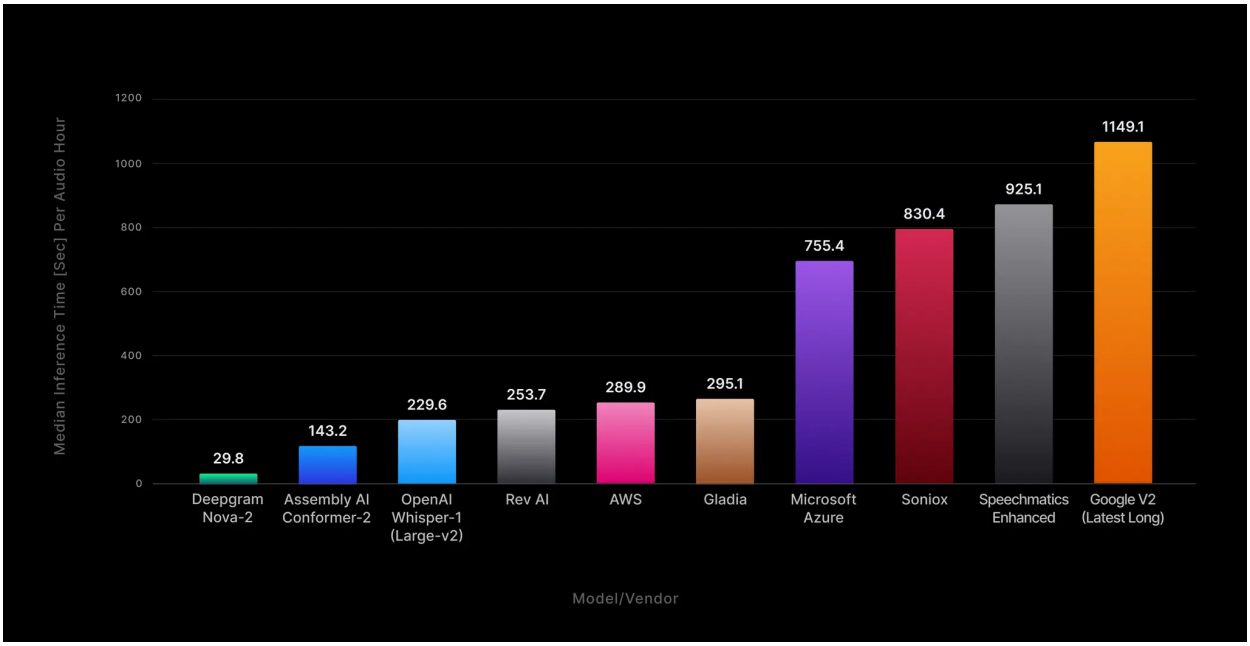



*Figure 2: The figure above compares the median inference time per audio hour for Deepgram Nova-2 and other popular models. Median inference time was determined using longer files, and diarization was included where applicable. To ensure a fair comparison, results are based on numerous trials of the same file for each vendor, giving them a fair opportunity to showcase their performance.*

# The Ranking: Top 10 speech-to-text APIs in 2024

With that background out of the way, let's dive into the rankings of the best speech-to-text APIs available today!

### 1. Deepgram Speech-to-Text API

Deepgram is the leading STT API provider in the market, offering several classes of deep learning-based transcription models—Base, Enhanced, and most recently released, Deepgram Nova-2—as well as custom model training. Deepgram is a developer-focused provider with a rich developer ecosystem, dedicated support, and a wide array of SDK options.

Deepgram's platform is designed for versatility and flexible deployment–either on-premises or public or private cloud–and handles pre-recorded audio and real-time streams from an array of sources. Offering an extensive feature set including multiple languages, smart formatting, speaker diarization, filler words, and high-value language understanding capabilities, Deepgram has processed trillions of words in production from esteemed clients such as NASA, Citibank, and Spotify.

**Pros**

- Highest accuracy
- Fastest speed
- Lowest cost
- Native real-time support with low latency
- Most flexible (deployment options, custom model training, etc.)
- Advanced feature set
- Developer-friendly and easy to start with Console or API Playground

**Cons**

- Fewer languages supported than some providers–primarily ones with lower usage–but regularly releasing new languages over time

**Price**

- Price: $0.25/audio hour

Uniquely, Deepgram eliminates the need for compromise typically found with alternative STT providers. With our models, users enjoy optimal accuracy without sacrificing speed or computational cost, underscoring our commitment to delivering a top-notch STT solution. In September 2023, Deepgram released Deepgram Nova-2, the fastest, most accurate STT model in the world. Nova-2 offers the following advantages:

- An average 30% reduction in word error rate (WER) over competitors for both pre-recorded and real-time transcription
- Lightning-fast inference times that are 5 to 40 times faster than alternative providers
- A cost-effective solution, with prices starting at a mere $0.0043/min, making it 3 to 5 times lower cost than the competition

If you'd like to give Deepgram a try, you can sign up for a free API key or contact us if you have any questions.

## 2. OpenAI Whisper API

**Pros**

- Decent transcription accuracy
- Broad language support (although accuracy plummets beyond the top dozen highest performing languages)
- Low acquisition cost
- Language and voice activity detection

**Cons**

- Tradeoffs between accuracy and speed; largest models are slow and no support for real-time transcription out-of-the-box
- No model customization
- No built-in diarization, word-level timestamps, or keyword detection
- Whisper has a number of known failure modes (e.g. hallucinations, repetition, issues with silent segments, etc.) that developers need to handle
- Not truly "free" (see below)

**Price**

- Price: Free to use*

Originally released in September 2022 and most recently updated in November 2023, OpenAI Whisper is a versatile, open-source model designed for automatic speech recognition (ASR) and translation tasks. This model was crafted by OpenAI's researchers to delve into the intricacies of speech processing systems trained through extensive weak supervision. OpenAI describes Whisper as a tool for AI researchers to explore various facets such as robustness, generalization, capabilities, biases, and limitations of the current model.

Whisper is available in five different sizes, with its models ranging from 39 million to over 1.5 billion parameters. Generally, the larger the model, the higher its accuracy, but this comes at the cost of increased processing time and computational expense. To boost processing speed in these larger models, additional computing resources are necessary.

As an open-source software package, Whisper presents an appealing option for enthusiasts, researchers, and developers. It's particularly useful for those looking to build product demonstrations or prototypes, or for conducting in-depth AI research in speech recognition and translation. Nevertheless, for developing large-scale production systems that require real-time processing of streaming voice data, Whisper may have limitations compared to some commercially available ASR solutions and requires ongoing engineering resources to stand up and operate a complete, self-hosted solution. Or alternatively, a 3rd-party vendor's Whisper implementation can be used, such as Deepgram's fully managed Whisper API that's faster, more reliable, and cheaper than OpenAI's.

Compare Whisper and Deepgram

**OpenAI Whisper requires extensive computing resources you'll need to requisition yourself. Whether that's a bank of high-end GPUs for a local deployment, or costly cloud computing credits, there are significant costs to running OpenAI Whisper at any kind of production scale. In addition to the initial capital expense to requisition equipment, you'll need to monitor and manage this infrastructure, as well as find developers to fix bugs and create workarounds for Whisper's common failure modes. These initial capital investments and operating expenses for full-time site reliability engineers and developers should be carefully considered in your Total Cost of Ownership (TCO) analysis.*

## 3. Microsoft Azure

The large public cloud providers–Amazon, Microsoft, and Google–each offer an array of AI/ML services and tools. Their primary motivation is to entice large enterprises to run their applications in the cloud and secure recurring subscription revenue from those workloads. Microsoft Azure Speech-to-Text is part of the Azure Cognitive Services suite, and currently the most capable of the Big 3's offerings, with a better overall combination of accuracy and speed.

Compare Microsoft and Deepgram

**Pros**
- Decent transcription accuracy
- Multilingual support
- Real-time streaming support
- Integration with Azure ecosystem
- Security and scalability

**Cons**
- Expensive
- Slow speeds for pre-recorded audio and latency issues in real-time transcription
- Privacy concerns
- Limited custom model support
- Cloud vendor lock-in

**Price**
- Price: $1.10/audio hour

## 4. Google Speech-to-Text

Google Speech-to-Text is offered as a part of the Google Cloud Platform. Like Microsoft's STT solution, it seamlessly integrates with other Google Cloud services (e.g. Google Drive, Google Meet, Google Docs, etc.) and offers a similar feature set, but with poor overall accuracy and among the slowest speeds for transcribing pre-recorded audio overall. Transcribing long files is cumbersome when the source does not reside in a Google Cloud Storage Bucket.

Compare Google and Deepgram

**Pros**
- Multilingual support
- Real-time streaming support
- Integration with Google Cloud ecosystem
- Security and scalability

**Cons**
- Poor overall accuracy
- Expensive
- Slow speeds for pre-recorded audio and latency issues in real-time transcription
- Privacy concerns
- Limited custom model support
- Cloud vendor lock-in (difficult to transcribe source files that don't reside in Google Cloud Storage)

**Price**
- Price: $1.44/audio hour for standard models, $2.16/audio hour for enhanced models (assumes data logging opt-out; rounded up to 15-second increments in utterances)

## 5. AssemblyAI

AssemblyAI is a privately held company employing modern deep learning models in its speech-to-text service. AssemblyAI provides faster transcription speeds than the public cloud providers, but still middle-of-the-road accuracy across pre-recorded and real-time use cases. AssemblyAI also offers a comprehensive feature set including diarization, language detection, keyword boosting, and higher-level language understanding features like summarization and topic detection.

Compare AssemblyAI and Deepgram

**Pros**
- Decent accuracy for some use cases (e.g. video/media)
- Faster speed for transcribing pre-recorded audio than public cloud providers
- Feature set

**Cons**
- Overall accuracy is lagging, especially for real-time audio
- Middle-of-the-road price per performance
- Limited customization
- Scalability (concurrent stream limits and # of speaker labels)

**Price**
- Price: $0.65/audio hour

## 6. Rev AI

Rev AI is a branch of the well-known transcription service Rev. In contrast to Rev, which offers human transcription and professional closed captioning services at a premium price point, Rev AI provides more affordable, automated speech-to-text services powered by advanced machine learning algorithms, as well as features like language detection and English-only sentiment analysis and topic detection.

**Pros**
- Decent accuracy for some use cases (e.g. podcasts, video/media, etc.)
- Faster speed for transcribing pre-recorded audio than public cloud providers
- Feature set

**Cons**
- Expensive
- Poor accuracy for non-English languages
- Poor real-time performance
- Limited customization
- Scalability (concurrent stream limits and # of speaker labels)

**Price**
- Price: $1.20/audio hour

## 7. Speechmatics

Speechmatics is a UK company primarily focused on the UK market, with accuracy that is middle of the road and very slow for transcribing pre-recorded audio. They're also one of the higher-priced ASR solutions on the market. They have limited customization support with a custom library where you need to also provide the phonetic "sounds like" words for training. Speechmatics can be deployed in the cloud or on premises.

Compare Speechmatics and Deepgram

**Pros**
- Decent accuracy for some non-English languages
- Good performance with British accents and UK spellings

**Cons**
- High cost
- Slow speed
- Poor real-time streaming support
- Limited customization

**Price**
- Price: $1.04/audio hour

## 8. Amazon Transcribe

Amazon Transcribe is offered as a part of the overall Amazon Web Services (AWS) platform. With similar features as Google and Microsoft's speech-to-text solutions, Amazon Transcribe offers good accuracy for pre-recorded audio, but poor accuracy for real-time streaming use cases. Its transcription speeds are much faster than Google and Microsoft, but still lag the top STT providers. Also similar to Google, Amazon Transcribe can only transcribe audio and video files stored in S3 buckets, part of the overarching strategy employed by the hyperscalers to deeply integrate their services within their ecosystem in an attempt to create high switching costs and vendor lock-in.

Compare Amazon and Deepgram

**Pros**
- Good accuracy for pre-recorded audio
- Easy to integrate if you are already in the AWS ecosystem
- Multilingual support
- Real-time streaming support
- Integration with Google Cloud ecosystem
- Security and scalability

**Cons**
- Expensive
- Poor accuracy for real-time audio
- Slow speeds for pre-recorded audio and latency issues in real-time transcription
- Privacy concerns
- Limited custom model support
- Cloud vendor lock-in (must transcribe from S3 storage)
- Cloud deployment only

**Price**
- Price: $1.44/audio hour general, $4.59/audio hour medical

## 9. IBM Watson

IBM Watson Speech-to-Text was an early ASR pioneer but has been outpaced by other providers and is largely considered a legacy player at this point. It has very poor accuracy compared to leading alternatives, is slow, and provides poor customization support that is expensive and takes a long time to deploy.

Compare Speechmatics and Deepgram

**Pros**
- Brand recognition

**Cons**
- Expensive
- Poor accuracy
- Slow speed
- No self-training
- Limited customization support

**Price**
- Price: $1.20/audio hour

## 10. Kaldi

Kaldi isn't technically a STT API, but it is one of the best-known open-source tools, so it's worth discussing here. Because Kaldi is not a ready-built ASR solution, the ASR solution needs to be built from Kaldi and trained with various audio corpora with Kaldi to have an actual ASR solution. The biggest issue with Kaldi is the training data that is available to use. If the training data matches your real-world audio well, the accuracy is fair, if not, then it will be very poor. There is no support but documentation and an open source community.

**Pros**
- Low acquisition cost

**Cons**
- Very poor real world accuracy (20-40% with public training data)
- Requires a lot of self training to be usable
- Speed will be very slow due to architecture
- Lots of developer work needed to integrate well with your systems.

**Price**
- Price: Free to use*

**OpenAI Whisper requires extensive computing resources you'll need to requisition yourself. Whether that's a bank of high-end GPUs for a local deployment, or costly cloud computing credits, there are significant costs to running OpenAI Whisper at any kind of production scale. In addition to the initial capital expense to requisition equipment, you'll need to monitor and manage this infrastructure, as well as find developers to fix bugs and create workarounds for Whisper's common failure modes. These initial capital investments and operating expenses for full-time site reliability engineers and developers should be carefully considered in your Total Cost of Ownership (TCO) analysis.*

See how Kaldi compares

## STT Comparison Summary Table

| Vendor | Accuracy | Speed | Cost | Customization |
|---|---|---|---|---|
| **Deepgram** | Highest | Fastest | Lowest | High |
| **OpenAI Whisper** | High | Slow | Low | Low |
| **Microsoft Azure** | High | Slow | High | Medium |
| **Google STT** | Medium | Very slow | High | Medium |
| **AssemblyAI** | Medium | Medium | Medium | Medium |
| **Rev AI** | High | Medium | High | Low |
| **Speechmatics** | High | Very slow | High | Medium |
| **Amazon Transcribe** | High | Medium | High | Medium |
| **IBM Watson** | Low | Slow | High | Medium |
| **Kaldi** | Low | Slow | Low | Medium |

# Conclusion

There you have it–the top 10 speech-to-text APIs in 2024. We hope that this helps you demystify some of the confusion around the proliferation of options that exist in this space, and gives you a better sense of which provider might be the best for your particular use case. If you'd like to give Deepgram a try for yourself, you can sign up for a free API key or contact us if you have questions about how you might use Deepgram for your transcription needs.

If you have any feedback about this post, or anything else around Deepgram, we'd love to hear from you. Please let us know in our GitHub discussions or contact us to talk to one of our product experts for more information today.

## About Deepgram

Deepgram is a foundational AI company on a mission to understand human language. We give any developer access to the most advanced speech AI transcription and understanding with just an API call. Our models deliver the fastest, most accurate transcription alongside contextual features like summarization, sentiment analysis, and topic detection. Contact us to learn more at deepgram.com/contact-us.