API Studying

1. Google Speech-To-Text

speech-api-leadConsidering that Google is essentially the nervous system of the Internet at this point, it’s no surprise their Speech-To-Text API is among the most popular – and most powerful – APIs available to developers.

Google Speech-To-Text was unveiled in 2018, just one week after their text-to-speech update. Google’s Speech-To-Text API makes some audacious claims, reducing word errors by 54% in test after test. In certain areas, the results are even more encouraging.

One of the reasons for the APIs impressive accuracy is the ability to select between different machine learning models, depending on what your application’s being used for. This also makes Google Speech-To-Text a suitable solution for applications other than short web searches. It can also be configured for audio from phone calls or videos. There’s a fourth setting, as well, which Google recommends using as default.

The Speech-To-Text API also features an impressive update for extended punctuation options. This is designed to make more useful transcriptions, with fewer run-on sentences or punctuation errors.

The newest update also allows developers to tag their transcribed audio or video with basic metadata. This is more for the company’s benefit than for the developers, however, as it will allow Google to decide which features are most useful for programmers.

The Google Speech-To-Text API isn’t free, however. It is free for speech recognition for audio less than 60 minutes. For audio transcriptions longer than that, it costs $0.006 per 15 seconds.

For video transcriptions, it costs $0.006 per 15 seconds for videos up to 60 minutes in length. For video longer than one hour, it costs $0.012 for every 15 seconds. Make sure you factor that into your pricing models when developing applications and web services.

Pros

Recognizes over 120 languages

Multiple machine learning models for increased accuracy

Automatic language recognition

Text transcription

Proper noun recognition

Data privacy

Noise cancellation for audio from phone calls and video

Cons

Costs money

Limited custom vocabulary builder

2. Microsoft Cognitive Services

Microsoft is also a major player in the world of voice recognition APIs. Microsoft Cognitive Services is more than just another speech recognition API, however. It's also a part of the Microsoft Trust Services which offer unparalleled security options for developers looking for the most secure data for their applications.

The main thing that separates Microsoft Cognitive Services' Speech to Text API is the Speaker Recognition function. This is the auditory version of security software like face recognition. Think of it as a retina scan for the sound of the user's voice. It makes it incredibly easy for different levels of users.

This same voice recognition capability allows software to adapt to specific user's speech styles and patterns. It also offers more custom vocabulary options than Google, as an additional benefit.

Beyond that, Microsoft Cognitive Service's speech recognition API has many of the same benefits of other voice APIs. It can perform real-time transcription, as well as converting text-into-speech. Thus, Microsoft Cognitive Services can cover most of your text and speech-based needs. It can also be used for call center log analysis, if you've got large amounts of audio that needs to be analyzed.

Considering the widespread popularity of Microsoft products and services, Microsoft Cognitive Services is growing faster than many of the other APIs on our list. If you're looking to join in with a vibrant, active community of developers, Microsoft Cognitive Services could be a good fit.

Pros

Enhanced data security via voice-recognition algorithms

Real-time transcription

Real-time translation

Customizable vocabulary

Text-to-speech capabilities for natural speech patterns

Cons

Built-in constraints due to the API being created for general purposes

Uses microservices, which can be useful for solving individual problems but falls short for larger problems

3. Dialogflow (Formerly API.AI, Speaktoit)

Dialogflow is also owned by Google. The main advantage over other voice APIs is Dialogflow's ability to take context into consideration when analyzing speech, which makes for more accurate transcriptions. It also allows developers to customize their voice-based commands for different devices, such as smart devices, phones, wearables, cars, and smart speakers.

Dialogflow's earlier incarnation, Api.ai, was used to power the Assistant app, one of the earliest virtual voice-based assistants, way back in 2014. It's since been discontinued but demonstrates that Dialogflow has been in the AI/machine learning/voice recognition game for longer than most.

The Dialogflow voice recognition API also has a number of analytics built into the platform. You can measure user engagement or session metrics, as well as usage patterns or latency issues. This is bound to be helpful when getting investors, sales and marketing teams, and developers on the same page.

Dialogflow currently only supports 14 languages, however. This makes it less useful for multilingual software than Google Speech-To-Text or Microsoft Cognitive Services.

Pros

Free

Easy to use

Easy to set up

Integrates with a wide variety of software

Easily integrated with other web services

Can integrate with non-Google devices like Amazon's Alexa

Cons

Cannot handle math functions

Cannot match intent with common phrases

Cannot create clickable links in the text box

Cannot search across intents

Can only provide one webhook

Voice Recognition APIs for Longform and Offline Processing

4. IBM Watson

It's no secret we're generating, processing, and analyzing larger quantities of data than any other time in history. Not all of that data is going to be clean and well-organized, especially if you're designing or developing an API. As API developers, it's our job to make sure that the data is organized and usable.

IBM Watson is perhaps one of the purest expressions of AI as a virtual assistant. IBM Watson is very adept at processing natural language patterns, which is one of the holy grails of AI and machine learning developers.

The IBM Watson Speech to Text API is particularly robust in understanding context, relying on hypothesis generation and evaluation in its response formulation. It's also able to differentiate between multiple speakers, which makes it suitable for most transcription tasks. You can even set a number of filters, eliminating profanities, adding word confidence, and formatting options for speech-to-text applications.

IBM Watson offers three different interfaces for developers. There's a WebSocket interface, an HTTP REST interface, and an asynchronous HTTP interface.

IBM Watson is simple to set up and implement, which makes it a wonderful option for those looking for a Speech-To-Text API but aren't completely technically proficient. IBM provides extensive documentation and one of the most thorough API reference manuals on the market. If you're looking for a speech-to-text API that's simple to set up and start using immediately, IBM Watson might be a good fit.

Of course, IBM Watson is more than just a speech-to-text API. It's one of the most fully-developed machine learning libraries in existence. It continues to learn and evolve, the more you use it. This makes it suitable for preventing outages and disruptions as well as accelerating research and data. Most applications that would benefit from structuring unstructured data will benefit from using the IBM Watson API.

As one of the best-developed machine learning APIs out there, IBM Watson isn't cheap. It is quick to get up and running, however, meaning you won't waste money on downtime or having to hire multiple developers just to get started. The peace of mind of a nearly plug-and-play Speech-To-Text API may be worth the cost of admission alone.

Pros

Processes unstructured data

Assists humans instead of replacing them

Helps overcome human limitations

Improves productivity be delivering relevant data

Improves user experience

Can process large quantities of data

Easy to set up and get started with

Cons

Doesn't directly support structured data

Expensive to switch to

Requires maintenance

Only supports a limited number of languages

Takes time to implement fully

Requires education and training to make full use of its resources

5. Speechmatics

Speechmatics offers an easy-to-use cloud-based API for automatic transcription services. Its main claim to fame is that it supports a wide range of file formats, meaning it can be used for offline file processing.

speechmaticsIt also supports a truly impressive array of languages, so you won't be limited to English. It's also been found to be more accurate than most of the other speech recognition APIs out there, so you won't have to proofread your transcriptions quite as extensively, so you can focus on other things.

The Speechmatics API is also highly adept at speaker recognition. It processes an impressive array of different variables, from confidence values to timing and speaker indications. This makes Speechmatics useful for machine learning applications, as it gets to know a speaker more thoroughly with each iteration.

Speechmatics has been found to be one of the fastest and most reliable automatic transcription APIs available for developers. It also supports nine languages, including different variants on English, including British and Australian English.

There are a couple of drawbacks to the Speechmatics API, however, although none of them are major enough to be a dealbreaker. First and most notably, there’s no app interface. If you’ll be using the transcription services, you’ll need to upload the audio to the website.

Secondly, each query does cost money. It costs .06 GBP per 1 minute of processed audio. If you’re going to be using the Speechmatics API for any sort of commercial app or web service, make sure to consider that when setting your processing. They do offer a discount for over 1000 minutes of processed audio. Perhaps you can work out some sort of bulk rate if you’re going to be using the Speechmatics API extensively.

Pros

Fast

Easy to use

Accurate

Supports multiple languages

Supports multiple English variants

Multi-speaker support

Multiple file formats supported

Does well with noisy audio

Easily integrated via REST API

Speaker recognition

Can be used for cloud-based transcription services and private usage, using the same API

Cons

No app interface

Costs money for each query

Final Thoughts

Python Coding Voice Recognition Reference

```
def audio_int(num_samples=50):
    """ Gets average audio intensity of your mic sound. You can use it to get
        average intensities while you're talking and/or silent. The average
        is the avg of the 20% largest intensities recorded.
    """


    p = pyaudio.PyAudio()


    stream = p.open(format=FORMAT,
                    channels=CHANNELS,
                    rate=RATE,
                    input=True,
                    frames_per_buffer=CHUNK)


    values = [math.sqrt(abs(audioop.avg(stream.read(CHUNK), 4)))
              for x in range(num_samples)]
    values = sorted(values, reverse=True)
    r = sum(values[:int(num_samples * 0.2)]) / int(num_samples * 0.2)
    print(" Average audio intensity is ", r)
    stream.close()
    p.terminate()


    if r > THRESHOLD:
        listen(0)


    threading.Timer(SILENCE_LIMIT, audio_int).start()


def listen(x):
    r=rs.Recognizer()
    if x == 0:
        system('say Hi. How can I help?')
    with rs.Microphone() as source:
        audio=r.listen(source)
    try:
        text = r.recognize_google(audio)
        y = process(text.lower())
        return(y)
    except:
        if x == 1:
            system('say Good Bye!')
        else:
            system('say I did not get that. Please say again.')
            listen(1)
```

Reference:

https://nordicapis.com/5-best-speech-to-text-apis/

https://cloud.google.com/speech-to-text