Accuracy Comparison of Open-Source Speech-to-Text Models

Hina Joshua, Nihal Sundarrajan, Anisha Patel, Mohamed Farag

Contents

Introduction

Speech Recognition is a technology useful in processes that benefit from the automation of transcript generation from spoken words. Historically, this technology has been supported by proprietary software which made access for all difficult. Currently, open source speech recognition libraries and systems has removed access limitations and expanded the community utilizing speech recognition tools.

Speech-to-text (STT) system is the method of converting spoken language in an audio file into text files. These systems are used by developers to create programs that can be used by the end user. The STT systems themselves are not meant to be used by end users, but rather are the mechanisms that can function 'under the hood' of speech recognition programs.

Open Source software is free of charge and can also allow (based on license preconditions) for the freedom to :

- Run it for a purpose you need or desire
- Study how it works and change it per your need
- Redistribute it in your community
- Redistribute your modified version in your community

The idea behind open source software is that the source code be available to the whole community, which can then benefit from the work of others and the work remains in progress and constantly evolving.

Overview

For the task of converting speech to text there are two main solution options available:

- **API** (e.g. Google Speech-to-text, AssemblyAI, AWS transcribe, Speechmatics, Azure Speech-to-Text)
- **open source library** (e.g. Kaldi , VOSK, Deep Speech, Speech Recognition, Silero, Flashlight)

Speech-to-text APIs tend to be easy to use as they do not require an understanding of the underlying model; Since they apply a state-of-the-art model they have high accuracy; they also have multiple additional features such as entity detection and sentiment analysis. The biggest disadvantage of using an API for a larger scale project is its high cost (except free tier for small projects) as well as the consistent need of an internet connection.
Open source speech-to-text libraries are completely free of charge, allow the freedom to use for any purpose. You can see what is going on 'under the hood' and therefore learn a great deal in the process. They have an added plus of data security, since the data doesn't have to be shared on the cloud or to a third party. The draw backs of using an open source libraries are that they are difficult to set up and require a lot of prerequisites including a good GPU, programming skills, deep learning knowledge. The main drawback however is that open source libraries they tend to be less accurate than the APIs.

For the purpose of our project, we were looking to find the most accurate speech-to-text system to carry out transcription of podcast audio files (English language only) of varying lengths and sources.
Currently, there are several options of open source speech-to-text systems available for download and use, making this choice a challenge. We chose to include open source speech-to-text libraries based on the following criteria:

- simple installation (using pip command)
- minimal dependencies
- open source

We chose to test the accuracy for the following 4 options for our Speech-to-text needs:

- Deep Speech
- Speech Recognition
- VOSK (small model)
- VOSK (large model)

Models

VOSK development started in 2020 and is one of the newest open source speech-to-text models. It allows speech recognition in over 20 languages. There are language models available to download for each of the supported language. Getting started with VOSK is easy with a simple pip command and a download of a copy of the language model available on https://alphacephei.com/vosk/models. There are 50 MB sized portable small models as well as larger over 1 GB models available based on the need. Small models are meant for small projects on mobile applications such as smartphones and Raspberry Pi's as well as desktop applications. Large models are for highly accurate transcription needs on servers. VOSK is licenses under the Apache License 2.0 which is a permissive license allowing commercial use, modification, distribution, patent and private use with main conditions requiring preservation of copyright and license notices [1].

Deep Speech is an open source end-to-end speech-to-text model. It is based on Baidu's original research paper about speech recognition technologies. The project is available to download on Mozilla's GitHub. Since it is an end-to-end model, it takes an audio input and outputs text in words from the audio. It is easy to get started with and download requiring a simple pip command. Implementation of DeepSpeech uses Goggle's TensorFlow. Since the computer power required is large, running DeepSpeech requires a large GPU. DeepSpeech is licensed under the Mozilla Public license 2.0 which permits commercial use, modification, distribution, patent and private use conditioned on making source code of files under the same license [2].

Speech Recognition is a Python library for speech recognition tasks which supports several APIs. It acts as a wrapper around multiple speech to text solutions and integrates them all into one package so you can use several services with only one library. It is easy to get started with this library with a simple pip command. The Recognizer class in the library recognizes language in an audio file. The Source code for the library is available on Uberi's GitHub and the library is available for use under the 3-clause BSD license [3].

Method

Data

We obtained 40 audio files of publicly available podcast episodes of varying duration (5 minutes - 1 hour), varying style of presentation and topics from 5 different podcast creators. Episodes from the following publicly available podcasts are represented in the data:

- A Slight Change of Plans | Pushkin
- Happiness Lab | Pushkin
- Stuff You Should Know | iHeart
- Conan O'Brien Needs a Friend | Team Coco
- Ted Talk Daily | TED

The podcast audio data was further manipulated using the python package *inaspeechsegmenter* in order to create variables of interest. These variables included:

- Duration of the podcast
- Percentage of music

These audio files were used to study the accuracy of the 4 selected Speech- to text models:

1. DeepSpeech
2. SpeechRecognition
3. VOSK (small model)
4. VOSK (large mdoel)

Each of the four Speech-to-text models being studied were used to transcribe the audio files. This resulted in the creation of 4 different transcripts for each episode. Each text file was then analyzed for accuracy by calculating the word error rate (WER) using a reference transcript generated from a Speech-to-text API (otter.ai) .

WER for Accuracy

In order to quantify the speech-to-text accuracy of each of the speech-to-text model being studied, we calculated the word error rate (WER) for their transcripts.
WER is the number of errors (substitutions, deletions and insertions) divided by the total number of words spoken. This is a popular and metric used to quantify the performance of a speech-to-text system.
In order to calculate the WER, words in the generated transcript being tested were compared to the words in the reference transcript. Since WER is the ratio of errors in the generated transcript to the total words in the reference transcript, a low WER indicates high speech-to-text accuracy. A WER of 10% indicates that the model was 90% accurate. This means that there is a 10% chance of errors and therefore with 90% accuracy, a reference transcript of 1,500 words could have 150 errors [4].

Results

Comparing Accuracy Between Models

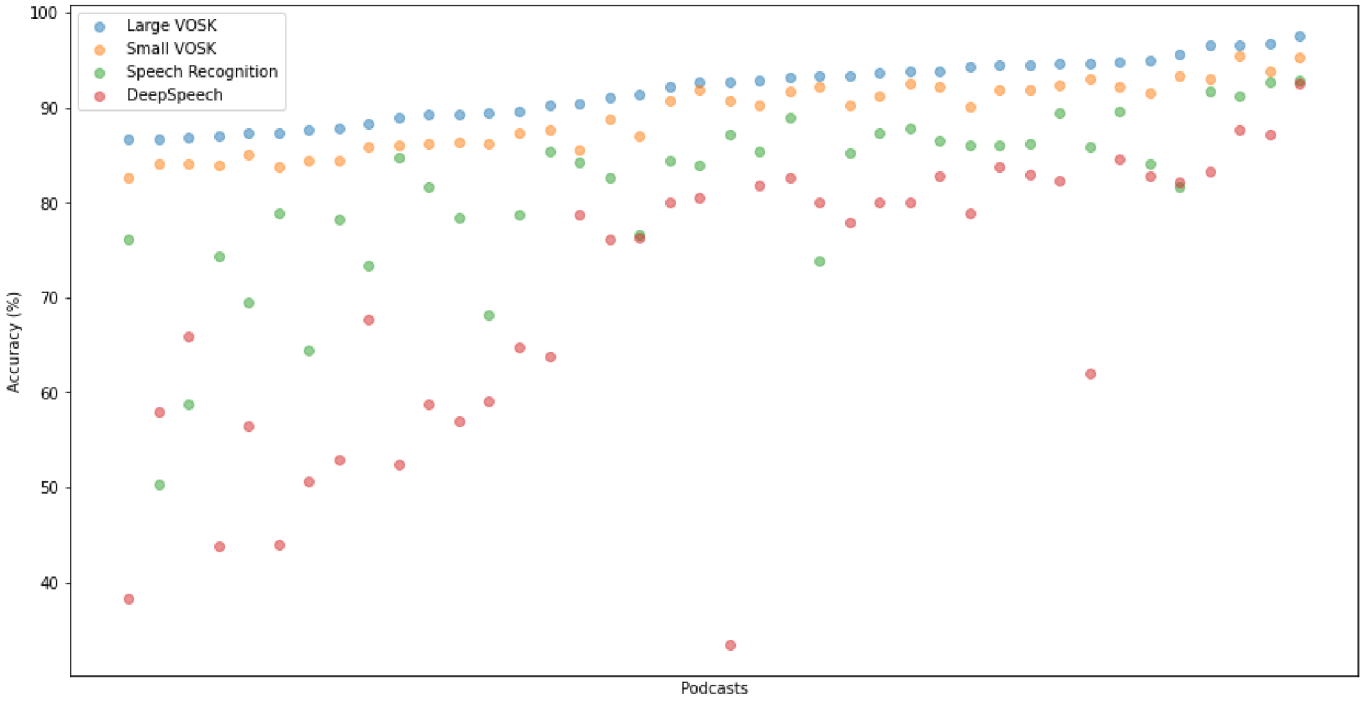
Summary statistics for the accuracies computed for each of the 4 models in all 40 podcast episodes are shown in *table 1*. The Large VOSK model has the highest mean accuracy of 91.8 % and the lowest standard deviation (3.26) indicating that it is the most accurate and consistent among all 4 models being studied.

The consistency of the large VOSK model can also be visualized in *figure 1.* For each of the 40 podcasts, the Large VOSK accuracy is consistently above the other 3 models.

Table1: Descriptive Statistics for Model Accuracies

	Large Vosk	Small Vosk	Speech Recognition	Deep Speech
mean	91.83	89.19	81.33	70.36
std	3.27	3.67	9.21	15.33
min	86.69	82.67	50.28	33.44
25%	89.25	85.96	77.99	58.61
50%	92.70	90.29	84.35	78.35
75%	94.50	92.26	86.75	82.44
max	97.63	95.43	92.89	92.54

Figure 1: Accuracy Comparison Between Transcription Models



Variability in Model Accuracy Based on Duration of Podcast

Since the data included 40 podcasts of varying duration, we were interested to know if the duration of a podcast has an effect on the accuracy of the speech-to-text models. *Figure 2* is a visual representation of the variability in accuracy of each model based on the duration of the podcast.

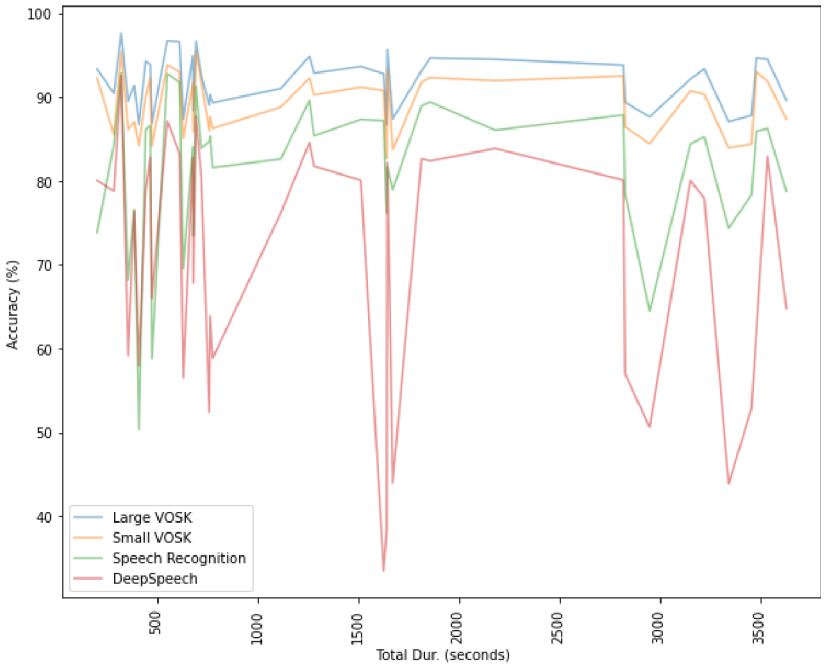
A Pearson Correlation (n = 40) examined the relationship between the duration of the podcast (mean = 25.12 minutes) to the computed accuracy for each model(means: large vosk = 91.83, small vosk = 89.19, speech recognition = 81.33, deep speech = 70.36). The correlation results for all 4 models are shown in *table 2*. All 4 models show weak correlation with low significance and the 95% CI include zero. This indicates that the duration of the podcast does not appear to have an impact on the accuracy of the model; however, since the p-values for all models are high, it would be of value to study a larger sample size in order to obtain statistically significant results.

(The effect of the percentage of music on the computed accuracy was also studied. The results are available in *table 3* and *figure 3* in the appendix.)

Table 2: Pearson Correlation for Relationship Between Duration of Podcast and Accuracy of the Models

	Large VOSK	Small VOSK	Speech Recognition	Deep Speech
n	40	40	40	40
r	-0.099	-0.055	0.073	-0.23
p-value	0.55	0.74	0.65	0.15
95% CI	-0.4 , 0.22	-0.36 , 0.26	-0.24 , 0.38	-0.51 , 0.08

Figure 2: Model Accuracy Based on Duration of Podcast



Conclusion

In order to produce text from audio with the goal of conserving meaning, it is essential to ensure that the generated transcript has low word error rate. While there are many options available to address your speech-to-text requirements, we found the VOSK large model to have a high accuracy, usability as well as scalability. It does not require a large GPU (unlike DeepSpeech), produces transcripts of over 90 percent accuracy with the highest consistency among all open source speech-to-text models we studied. The performance of our models did not seem to be affected by the duration of or the percentage of music in the audio file.
It would be of value to further study speaker attributes such as accent, gender, voice pitch etc. as a contributing factor to the accuracy of a speech-to-text model.

References

1. Vosk offline speech recognition API. (n.d.). Retrieved August 15, 2022, from https://alphacephei.com/vosk/
2. Mozilla. (n.d.). Mozilla/DeepSpeech. GitHub. Retrieved August 15, 2022, from https://github.com/mozilla/DeepSpeech
3. Speechrecognition. PyPI. (n.d.). Retrieved August 15, 2022, from https://pypi.org/project/SpeechRecognition/
4. Taylor, R. (2021, May 21). What is wer? what does word error rate mean? Rev. Retrieved August 15, 2022, from https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean

Appendix

Table 3: Pearson Correlation for Relationship Between Percentage of Music in the Podcast and Accuracy of the Models

	Large VOSK	Small VOSK	Speech Recognition	Deep Speech
n	40	40	40	40
r	0.19	0.18	-0.04	0.34
p-value	0.23	0.26	0.80	0.02
95% CI	-0.13 , 0.48	-0.14 , 0.47	-0.35 , 0.27	-0.06 , 0.61

Figure 3: Model Accuracy Based on Percentage of Music