CSI 5180: Topics In AI: Virtual Assistants

Winter 2023

Assignment 1

Submitted To
Professor Caroline Barrière (Ph.D)

Submitted By

Shubham Kulkarni - 300276475
Snehal Sudhir Bhole - 300318105

QUESTION 1 – Comparing VAs as to Speech Recognition (10 points)

Google Assistants:

Google assistants were created by google in 2016, and are now widely used in watches, computers, smartphone speakers, electronic devices etc. Google assistants are the most accurate virtual assistants existing currently, due to vast database and machine learning algorithms. It was designed to give two-way conversational interactions. Google keeps on expanding its scope and reaches people speaking 44 languages. The google assistant can be accessed using the wake word “Hey Google” or “OK Google”.

Siri:
Siri was first made by Stanford Research Institute and later integrated with Apple's iPhone in 2011. Siri is available on a range of Apple devices, including the iPhone, iPad, Apple Watch, and Mac, and can be activated by holding down the home button on compatible devices or by saying "Hey Siri." Apple has also introduced Siri Shortcuts, which allows users to create custom voice commands for performing specific tasks or actions within supported apps. It integrates with a variety of Apple services and applications, such as Apple Music, Apple Maps, and Apple Calendar, to provide a seamless user experience.

Alexa:
Alexa is amazon’s digital assistant launched in 2014. Alexa, Amazon's virtual assistant, has become widely popular due to its integration with Amazon Echo devices. Alexa has a good accuracy rate and is able to perform a variety of tasks, such as setting reminders, playing music, and answering questions. Alexa is available in 8 languages.

Figure 1 Given below summarizes a study conducted on 800 queries using Google Assistants, Siri, Cortana and Alexa.

Query Results

	Answered Correctly		Understood Query	
	Apr-17	Jul-18	Apr-17	Jul-18
Google Assistant	74.8%	85.5%	99%	100%
Siri	66.1%	78.5%	95%	99%
Cortana	48.8%	52.4%	97%	98%
Alexa	n/a	61.4%	n/a	98%

Figure 1: Comparison of various Virtual Assistants

Bixby:
Bixby is Samsung's virtual assistant and is available on Samsung devices. Bixby is available on a range of Samsung devices, including smartphones, tablets, and home appliances, and can be activated by pressing a dedicated Bixby button on compatible devices or by saying "Hi, Bixby." Samsung has also introduced Bixby Voice, which allows users to control their Samsung devices using only voice commands.

Cortana:
Cortana is Microsoft's virtual assistant and is available on Windows devices. It also integrates with a wide range of Microsoft services, such as Bing, OneDrive, and Office 365, to provide a seamless user experience across different devices and platforms.It also integrates with a wide range of Microsoft services, such as Bing, OneDrive, and Office 365, to provide a seamless user experience across different devices and platforms.

Table 1 shows the comparison of all the virtual assistants considered for the study.

Virtual Assistant	Languages Supported	Accuracy	WER	Wake Words	Other metrics used
Google Assistant	44 Languages	95%	4.9%	Hey Google, Ok Google	Personalised speech recognition
Siri	21 Languages	78.4%	NA	Hey Siri	False accept rate and false positive rate
Alexa	7 languages	73%	NA	Alexa	Acoustic Echo Cancellation
Bixby	7 Languages	NA	NA	Hi Bixby	Absence of an open NLP-AI platform
Cortana	8 Languages	NA	5.1%	Hey Cortana	Eavesdropping Aspects

Table 1: Comparison of Virtual Assistants

Articles:

- VentureBeat - Google Assistant can now interpret 44 languages on smartphones: Source
- VentureBeat - Why Google Assistant supports so many more languages than Siri, Alexa, Bixby, and Cortana: Source
- Statista - Digital assistant performance comparison. Source
- SmartAction - Does Word Error Rate Matter? Source
- Search Engine Journal - Google Assistant is More Accurate Than Alexa, Siri, and Cortana. Source
- Makeuseof - What Is Bixby and What Can You Do With It on Your Samsung Phone? Source
- Business Insider - Alexa is getting better at answering users' questions. Source
- Becker's Hospital Review - Google's voice assistant more accurate than Alexa, Siri for medication info, study suggests. Source

QUESTION 2 – Wake Word (5 points)

If the wake words are eliminated and virtual assistants can respond without directly being called, it is undoubtedly a great advancement in the history of mankind and would significantly alter the ways of communication. Nowadays, smart devices with Intelligent Virtual Assistants are ubiquitous and have been monitoring the activities in our surroundings consistently through microphones and radar sensors. If the wake word is eliminated in the interaction, there will be some other approach used to start the conversion with Virtual Assistant. These approaches may include hand gestures, movements, and Facial Expressions. With the usability of such devices, there is an unsaid promise that they get to know you. There are various views on this more conversational way of communication with VAs.

With the ability to respond without a wake word, VAs would give a human-like feel, seamless and natural interactions with users. VAs will become more convenient and accessible. Such Intelligent systems would be very accurate and efficiently trained with past data using advanced Natural Language Processing, and Machine Learning algorithms. Ultimately, it will depend on the level of implementation and execution. Some professionals believe that, though this advancement will result in better interaction between users and VAs, there are certain concerns with privacy, the amount of data collected, and the potential for false activations.

When no specific wake word is assigned to begin communication, VA may not be unable to distinguish between intentional and unintentional commands, which could lead to frustrating situations. For instance, if virtual assistants are used in automated vehicles, the system responding unintentionally may disturb and distract the driver. Such situations may lead to fatal accidents. Thus the accuracy of the systems is an important aspect to consider in a real-time setting. Additionally, it will continually monitor its users to provide assistance which may lead to a loss of trust in technology. Thus, it will become very important for technology companies to implement robust privacy measures to address such privacy-related concerns.

Articles:

- Wired - Google Assistant's Future Is Looking Us Right in the Face- Google says its voice assistant is getting more conversational, and that face unlocking is ready to replace wake words. But don't hold your breath. Source
- VentureBeat - Don't lift a finger: AI-driven voice commands are the future of the smart home. Source
- Source Gear - Google Assistant no longer requires wake word to stop talking - Source
- TechCrunch - Google makes it easier to chat with its Assistant. Source

QUESTION 3 – Diarization (5 points)

Identifying *who talked when* is a procedure called automatic speaker diarization. In the article at hand, IBM talks about its cutting-edge speaker diarization system, which uses a variety of embedding techniques to express acoustic information in short audio clips. The system employs spectral clustering, the neural network-based embedding of uncertainty information, and acoustic embedding.

The article discusses the challenges related to speech diarization such as embedding speech segments, understanding speaker similarity and estimating the number of speakers. The process uses chopped audio input segments representing a single speaker's characteristics. Time-delay neural networks-based x-vectors and LSTM-based vectors are successfully used for generating embeddings. One of the critical challenges in this process is embedding the speech segments. Another challenge faced is to score the speaker similarity between speech segments to generate clusters. Different segments have distinct embeddings due to phonetic differences, which are significant when working with small segments. Neural networks are trained to compute speaker similarity while considering the duration-dependent within-speaker variability.

The major challenge faced while using spectral clustering is determining the number of clusters. It is a vital step to differentiate between speaker-indicative and noisy eigenvectors. The larger the eigenvalue of the similarity vector, the matrix corresponds to the speaker and corresponds to noise otherwise. However, it is hard to find a cut-off point. When more than one user is associated with

temporal response(Obtained by using eigenvectors), one of the speakers will have a larger positive value, and others will have large negative values. The temporal response will be noisy if the eigenvector is not associated with the speaker.

Overall, The article offers a general introduction to diarization and its uses in virtual assistant technology. It explores new developments that are assisting in enhancing the usefulness and accuracy of virtual assistants and shows the advantages and disadvantages of this technology.

QUESTION 4 – Safety (5 points)

Article 1: Your New Vehicle Should Have Apple CarPlay Or Android Auto—Here's Why - Source

Apple CarPlay and Android Auto are rapidly becoming commonplace platforms for navigation, entertainment and communication in a new vehicle. Traditionally used physical maps, mp3 players, and iPods are eliminated by these new advanced features added in smart vehicles. In-car infotainment systems and in-car Bluetooth capabilities have introduced hands-free calling and an unending way to stay connected with the car. These systems let users use their smartphones without distractions from actual devices. The voice-operated commands also satisfy a growing number of hands-free laws across many states.

According to the author, safety issues exist with smartphone interfaces. It is less distracting to use voice commands to operate a smartphone than to pick one up and use it while driving. Due to voice-enabled text functionality, there is less need to concentrate on the screen. The built-in in-car system will suffice if you're not connected to your smartphone.According to industry experts from AAA, voice-activated smartphone interfaces are less distracting than using a phone, but also underline the importance of drivers paying constant attention to the road.

In conclusion, the in-built infotainment systems in car and smartphone interfaces controlled by voice command are certainly an advancement in the field of speech recognition and natural language processing but the technology we can access is often limited and often expensive.

Article 2: Hands-free car systems still distract drivers, study says - Source

One of the serious concerns faced by the United States is distracted driving. Many states have laws to reduce accidents due to distracted driving and regulate the use of hands-free technology. The article here presents a study conducted by the University of Utah on the use of smartphones and in-car infotainment systems while driving and conclusions derived from the study are discussed. The study was conducted between 65 smartphone users and 251 in-car infotainment systems. For smartphones, all three AI-based assistants Siri, Cortana and Google assistants were highly disturbing when asked to perform any actions like playing music, etc while driving. On the other hand, in-car infotainment systems were not that disturbing except for Mazda6. The results obtained from these studies were more serious.

The study showed that the participants were distracted for around 27 seconds after using virtual agents.These VAs should be designed with minimal distractions and fully focus on the primary job "Driving". On the contrary, young drivers are primarily distracted due to the use of internet-based applications in cars. To attract young minds, car companies have started incorporating social

media apps, news, and weather apps in their cars, which has led to the major causes of distracted driving. Moreover, listening to music throughout the journey, facilitated by wired or wireless chargers in cars is a common practice but significantly contributes to increasing cases of distracted driving.

Thus to conclude, undoubtedly Virtual Assistants have used a large population from disabled people to visually impaired people but they are not to be used by everyone and for every purpose. The technology accessible now is at a fundamental level and extensive research is required to use them for all day-to-day activities.

From both articles, it is clear that there is a need for more accurate, accessible, well-trained and well-tested systems on the road. The in-built infotainment systems in cars and voice-controlled smartphones definitely facilitate our mundane activities and help us to increase our productivity to some extent but they can also be a cause of distraction in the mentioned scenarios. Unfortunately, there is no specific answer and way to control the distraction due to voice-controlled systems as long as we don't put in the effort to make them more accurate.

QUESTION 5 – Scaling issues in ASR (5 points)

In the presentation Adam Coates discusses 2 pieces of scale which are Data and Computation Power. He dives deep into them individually and presents challenges associated with scaling both of them.

Data Scaling: Transcribing data is an expensive process and it roughly costs 1$ per minute to transcribe new data. However, the real challenge is understanding the application for which the data will be used. Defining the purpose of data transcribing as the model will tend to perform in a direction in which it is trained. Obtaining a 'Read' style of speech is easier as there are massive samples available online and even for an annotator (speech generator), it is easy to produce in a quiet environment in a smooth tone. However, we need to take into consideration the actual input can involve nuances, such as dysfluency, and stuttering in conversational audio, and environmental factors such as reverb and echo. Along with this the Lombard effect, speaker accent, microphone quality and presence of noise cancellation can also influence the quality of audio. He also mentions, engineering a robust data pipeline is easier than engineering a robust speech engine hence if there are significant performance drops, the approach of data engineering should be applied over model fine-tuning.

There are a few approaches of audio augmentation, such as additive noise and tempo change that can be applied to the existing clean 'Reading' audio to exponentially increase the training samples providing the model more data to train on the same content in multiple contexts.

Computation Power: It is generally observed that higher computation power results in faster and better performance, but there can be multiple approaches implemented to better optimize the higher computation power. Even parallelism has its own issues such as managing network traffic and data needs to be moved with appropriate strategies. There are multiple instances where while training the data if an edge case of a library is observed, it can significantly impact the further training and can impact the training speed by a factor of 2 or 3 which can increase the computation time from one week to three weeks.

Different strategies can be used to counter the problems encountered in parallel processing. The code can be optimized to better deal with edge cases so that even if something of the sort is encountered, it is handled efficiently by the model. Also, a concept of batch processing can be

introduced which keeps similar length utterances together while training so that computation power is not lost due to different lengths of utterances and all the batches can finish processing in a similar amount of time and there are fewer wait times involved for libraries to get next inputs.

What this scaling issue represents for the research field of speech recognition

Overfitting, the choice of architecture and optimization algorithms and the distribution of the data are all important factors that can impact the performance of speech recognition systems.

University research labs and private labs typically have distinct resources and capacities in terms of data accessibility and computer capacity. University research labs frequently have access to big datasets and computing clusters that are available to the public, but they might not have the money to buy and maintain the necessary infrastructure. On the other hand, private labs could not have access to the same degree of publicly available resources as university research labs even if they usually have more financial resources to purchase and operate powerful computing systems and large datasets.

Articles:

- Speech Recognition and Deep Learning - Adam Coates, Vinay Rao. Source

QUESTION 6 – Evaluation (5 points)

An automatic speech recognition system's performance can be measured by using the metric word error rate (WER).

Word error rate can be computed as **WER** = (S + D + I) / N = (S + D + I) / (S + D + C)
S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference (N=S+D+C).

System 1: The recognized text is "the cat is crazy". Substitutions are 2, since “is” is substituted by and “lazy” is substituted by “crazy”. The WER for System 1 would be 2/4 = 0.5, or 50%.

System 2: The recognized text is "the clap is crazy". Again there are 2 substitutions for “cat” and “lazy” by “clap” and “crazy”. The WER for System 2 would be 2/4 = 0.5 or 50%.

System 3: The recognized text is "all cat is hazy". The number of errors is 2, since "all" is substituting “the” and "lazy" should be "hazy". The WER for System 2 would be 2/4 = 0.5 or 50%

The WER for all the systems is appropriate.

WER has numerous limitations. There is no distinction made between the words that are crucial to the sentence's meaning and those that are less significant. It makes no distinction between two words that differ entirely or only by one character.
There are multiple adaptations that can be applied to WER to counter its limitations.

1. **Levenshtein Distance**: It is also known as Edit Distance which calculates the smallest number of single-character alterations (insertions, deletions and substitutions) to transform one string to another. It is useful in finding matches for short strings in numerous larger

texts where only a limited number of changes are anticipated as the goal of approximate string matching. This measurement is more precise than WER and can be helpful in applications where the cost of various mistakes kinds may change.

2. **BLEU Score**: Bilingual Evaluation Understudy compares the n-gram overlap between a machine-generated translation and a set of reference translations to calculate a score that reflects the grade of the machine-generated translation. The output of BLEU is consistently an integer between 0 and 1. With values closer to 1, texts that are more comparable to the reference texts are considered to be more similar to the candidate text. As BLEU considers the fluency and sufficiency of the machine-generated translation, it has been demonstrated to be more effective than WER.

Articles:

- Wikipedia - Word Error Rate. Source
- Huggingface - Metric:wer. Source
- Wikipedia - Levenshtein distance. Source
- Wikipedia - BLEU. Source

QUESTION 7 – Datasets (5 points)

After exploring online, the following datasets can be one of the best for training speech recognition software.

Google's Speech Commands Dataset: It is a large-scale dataset of speech commands used for training machine learning models. It was developed by Google in 2017 and includes more than 105,000 brief audio samples of spoken English requests. The clips were captured in a range of settings, including silent spaces, boisterous places, and with background music. The data includes a diverse range of speakers and accents, making it a suitable dataset for training speech recognition models that can handle real-world conditions. Researchers and programmers in the fields of voice recognition and artificial intelligence have trained and evaluated their models using the Speech Commands Dataset. Additionally, it has been applied to enhance voice-activated gadgets like virtual assistants and smart home systems.

Mozilla's Common Voice Dataset: It is a crowd-sourced dataset created by the Mozilla Foundation. The datasets are available in a variety of languages such as English, French, German etc. The data is collected from volunteers and includes a wide range of accents and speaking styles. These volunteers record themselves speaking specific phrases, and the resulting speech data is used to train machine learning models for speech recognition and other applications. It has been used by researchers, programmers, and businesses to train and assess their speech recognition algorithms. It is frequently updated with fresh voice data.

Factor	Google's Speech Commands Dataset	Mozilla's Common Voice Dataset
Purpose	Focused on providing data for training speech recognition models	Producing a more inclusive and diverse dataset of speech data that can enhance the performance of speech recognition models
Size	Contains over 105,000 short audio clips of spoken commands in English	Contains thousands of hours of speech data in multiple languages.
Language	Supports only English	Supports 100+ Languages
Data Collection	Created and Collected by Google	Crowd-sourced, collected from multiple volunteers across the world
Diversity	It includes a diverse range of speakers and accents	It was designed to create a more inclusive dataset of people with various accents, speaking patterns, and racial and ethnic backgrounds

Table 2: Comparison of Datasets

Articles:

- Google Research - Launching the Speech Commands Dataset. Source
- Mozilla Foundation - Common Voice. Source

QUESTION 8 – Open Source Software (10 points)

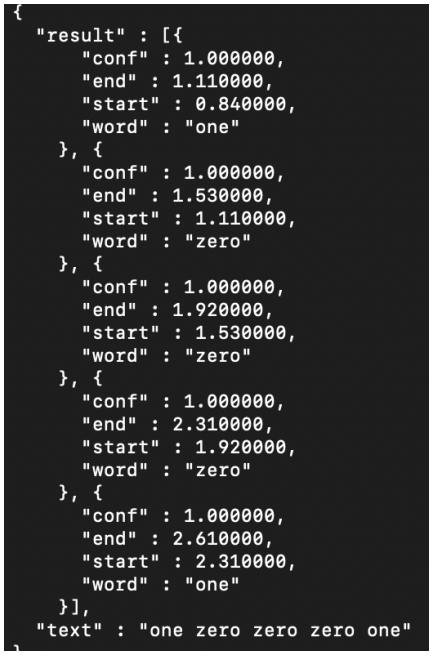
Exploring Vosk:

Vosk is a software toolkit for speech recognition and transcription. It is aimed at building and integrating speech recognition functionality into various applications. It is written in C++ and can be integrated into projects involving other programming languages like Python, Java and C#.

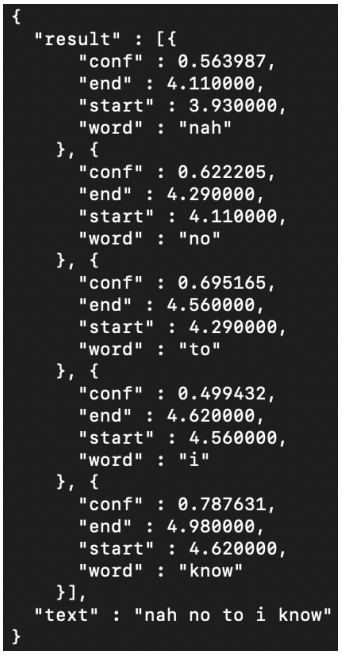
Vosk is suitable for a variety of applications, including voice-controlled smart devices, speech-to-text translation, and voice commands in video games. It supports 20+ languages and dialects and can recognize speech in real time. It also supports speaker identification along with speech recognition. It has the ability to work offline, on lightweight devices as its portable per-language models are around 50 Mb. Deep neural networks (DNNs) are used by the program to recognize speech patterns, enabling it to transcribe even in loud settings with excellent accuracy.

Vosk Tutorial: The steps to install Vosk are available on the alphacephei.com website under the "Installation" section. There are some dependencies such as homebrew, git, python3, and pip3. The installation is very straightforward and can be done within an hour. There is a sample test audio file present under the python folder and it can be easily tested out. This testing is equivalent to a "Hello World" program. After some explorations, the system gets ready and is relatively easy to use. One thing hard to contextualize is the output generated by the system which needs furthermore implementation to display the results in a user-friendly way.

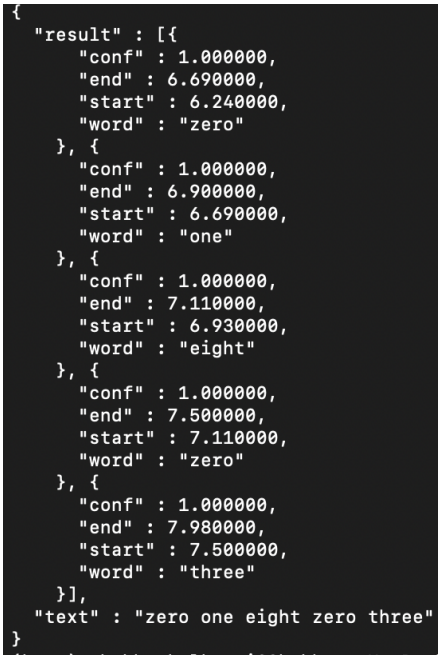
Input from audio file: “one zero zero zero one…nine o two one o….zero one eight zero three”
Sample Output: The system is able to understand most of the numbers.



```
{
  "result" : [{
      "conf" : 1.000000,
      "end" : 1.110000,
      "start" : 0.840000,
      "word" : "one"
    }, {
      "conf" : 1.000000,
      "end" : 1.530000,
      "start" : 1.110000,
      "word" : "zero"
    }, {
      "conf" : 1.000000,
      "end" : 1.920000,
      "start" : 1.530000,
      "word" : "zero"
    }, {
      "conf" : 1.000000,
      "end" : 2.310000,
      "start" : 1.920000,
      "word" : "zero"
    }, {
      "conf" : 1.000000,
      "end" : 2.610000,
      "start" : 2.310000,
      "word" : "one"
    }],
  "text" : "one zero zero zero one"
}
```



```
{
  "result" : [{
      "conf" : 0.563987,
      "end" : 4.110000,
      "start" : 3.930000,
      "word" : "nah"
    }, {
      "conf" : 0.622205,
      "end" : 4.290000,
      "start" : 4.110000,
      "word" : "no"
    }, {
      "conf" : 0.695165,
      "end" : 4.560000,
      "start" : 4.290000,
      "word" : "to"
    }, {
      "conf" : 0.499432,
      "end" : 4.620000,
      "start" : 4.560000,
      "word" : "i"
    }, {
      "conf" : 0.787631,
      "end" : 4.980000,
      "start" : 4.620000,
      "word" : "know"
    }],
  "text" : "nah no to i know"
}
```



```
{
  "result" : [{
      "conf" : 1.000000,
      "end" : 6.690000,
      "start" : 6.240000,
      "word" : "zero"
    }, {
      "conf" : 1.000000,
      "end" : 6.900000,
      "start" : 6.690000,
      "word" : "one"
    }, {
      "conf" : 1.000000,
      "end" : 7.110000,
      "start" : 6.930000,
      "word" : "eight"
    }, {
      "conf" : 1.000000,
      "end" : 7.500000,
      "start" : 7.110000,
      "word" : "zero"
    }, {
      "conf" : 1.000000,
      "end" : 7.980000,
      "start" : 7.500000,
      "word" : "three"
    }],
  "text" : "zero one eight zero three"
}
```

Advantages	**Disadvantages**
Cross-platform compatibility: Vosk can run on a variety of platforms, including desktop computers, embedded systems, and mobile devices.	**Dependence on external libraries**: It requires the installation of additional libraries to work, which can increase the complexity of setup and integration.
Real-time recognition: Vosk can recognize speech in real time, making it suitable for interactive applications such as voice-controlled devices.	**Potential for errors:** It is bound to make errors based on dialect of speech which need manual correction.
High accuracy: Vosk uses deep neural networks (DNNs) to recognize speech patterns, making it highly accurate and capable of transcribing even in noisy environments.	**High Resource Consumption**: The DNN libraries make the computation intensive and significant computing resources are needed.
Multiple language support: Vosk supports 20+ languages, making it suitable for global use.	**Dependence on Training Data**: Though there is support for multiple languages, training data for a use case may not be available and the accuracy of results can decrease.
Customizable: The source code of Vosk is available, allowing developers to customize and extend the software to meet their specific needs.	**Privacy Concerns**: It transmits data to cloud which has no specified privacy guidelines and may cause security risk for some users.

Table 3: Advantages and Disadvantages of Vosk

Exploring Kaldi:

Kaldi is an open-source toolkit for speech recognition research. Researchers and developers can simply apply new algorithms and models because of their very flexible and adaptable design. It provides a number of pre-built speech recognition models and recipes.
Kaldi is built around the concept of speech recognition pipelines, where the audio signal is processed through a series of stages, including feature extraction, acoustic modelling, and language modeling. Users can quickly replace pipeline parts because of the design's modularity, which enables system customization for particular applications or experimentation with novel concepts. It has been used to build state-of-the-art speech recognition systems that have achieved high accuracy on a variety of benchmarks and datasets. Kaldi is widely used in both academia and industry and has a sizable and active community of contributors, it is always changing and getting better.

Kaldi Tutorial: In order to do a Kaldi tutorial a lot of dependencies are involved, starting with a need for a Linux-based operating system. Along with that, multiple packages such as an atlas, autoconf, automake, git, libtool, svn, wget, zlib, awk, bash, grep, make, pearl. After this, the Kaldi repository is to be cloned and the source code should be compiled. The next steps involved are setting up environment variables, preparing the training and test data and ultimately evaluating the performance.

Overall, the setup of Kaldi is difficult, involves multiple dependencies and is time-consuming. However, there is a step-by-step tutorial available for working on this setup on Kaldi's website under 'Kaldi for dummies tutorial'. We have tried to perform installation of Kaldi on Mac and Windows operating systems and have been unsuccessful after multiple attempts as the fundamental requirement of the setup is a Linux-based machine.

Advantages	**Disadvantages**
Flexibility: Kaldi's modular architecture enables users to quickly switch out speech recognition pipeline parts, enabling them to modify the system for particular applications or to test out novel concepts.	**Steep Learning Curve**: To use Kaldi efficiently, one must possess a certain level of technical proficiency due to its sophisticated architecture.
Performance: Kaldi has been used to build state-of-the-art speech recognition systems that have achieved high accuracy on a variety of benchmarks and datasets.	**High Computing Power**: It requires a significant amount of computing resources to train and run speech recognition models
Active Community: It has a huge and active community of contributors, guaranteeing that it keeps advancing and getting better.	**Not Commercially Viable**: As an open-source toolkit, it has limited commercial support compared to proprietary speech recognition solutions.
Integration with other tools: Kaldi can be easily integrated with other open-source tools, such as Python and HTK	**Limited Use Cases**: It is designed specifically for speech recognition research and may not be suitable for all applications, such as commercial speech recognition products.

Customizable: Kaldi makes it simple to begin with voice recognition jobs by offering a number of pre-built speech recognition models and recipes.	**Documentation**: It is hard to navigate the documentation as it has been contributed by a large community, it is not very user-friendly.

Table 4 : Advantages and Disdanvatnges of Kaldi

Articles:

- Alpha Cephi - Vosk. Source
- Kaldi - About the Kaldi project. Source
- Kaldi - Kaldi for Dummies tutorial. Source