CSI5180: Topics in AI Virtual Assistants

Assignment 1
Speech Recognition

Submitted To:
Prof. Caroline Barrière

Team Member:
Abhisht Makarand Joshi: 300311048
Nidhi Kumari Chauhan: 300279256

Question 1 – Comparing VAs as to Speech Recognition

With respect to the market view of Speech Recognition in Virtual Assistants, the following article

- Globalme language and technology (2021), A Complete Guide to Speech Recognition Technology provides information about Apple's Siri, Amazon Alexa, and Google Assistant.

For Google Assistant, the article mentions the accuracy rate and the WER. Unfortunately, this measure is not provided for Siri or Alexa. And what about other assistants (e.g Samsung Bixby, Microsoft Cortana)?

TO DO:

- Find some blogs/articles/company's posts which discuss the speech recognition systems of the 5 companies. Try to find information about their accuracy rate or WER.
- If the companies don't mention their WER, what do they mention to try to sell their product (in relation to Speech recognition)? What about number of languages recognized, or variety of commands?
- Do some exploration so that you can write a one-page summary (with comparative table) about Speech Recognition performances of various VAs.

Answer 1:

Article [1] discusses the different types of Virtual assistants of Siri, Alexa and Google Assistant. The article comes to the conclusion that while Siri is thought to have the most accurate speech recognition technology, Alexa has the most third-party skills and services accessible and is compatible with most devices. Many people appreciate Google Assistant for how effectively it works with other Google services and products, such as Google Maps and Google Calendar.

Amazon's virtual assistant, *Alexa*. It draws attention to the fact that Alexa is compatible with a wide range of smart home gadgets and outside services, making it the most adaptable virtual assistant in terms of interoperability. Additionally, Alexa has a sizable number of "Skills" that expand upon its capabilities. Alexa gives the flexibility to perform home automation, media, shopping, news and weather, calling and messaging.

The most conversational virtual assistant is said to be *Google Assistant*. It is renowned for its proficiency with natural language processing and capacity for contextual understanding, which enables users to communicate with it in an effortless manner. Numerous gadgets, including smartphones, smart speakers, and others, support Google Assistant. Google Assistant provide functionality in various places such as home automation, entertainment, directions, reminders, third-party integration, calling, etc.

Samsung's virtual assistant, *Bixby*[8]. Bixby is incorporated into the Samsung Experience UI and is only available on Samsung smartphones. Users of Samsung devices consider Bixby to be a great

alternative since it offers close interaction with other Samsung services and features. Bixby, a virtual assistant from Samsung, has been improved with the aid of SensorFlow and Google Cloud TPUs (Tensor Processing Units)[11]. Bibxy provides functionality in reading emails, callimg, text messages, reminder, and scheduling.

Apple's *Siri* is built into the iOS operating system and therefore only accessible on Apple devices. It is referred to be a multifaceted assistant with an intuitive user interface and many features, such as the ability to send messages, make phone calls, create reminders, and more[1] Like Apple Carplay. Along with it, Siri can perform various task on command such as Calling and messaging, scheduling and reminders, navigation, search, home automation, etc.

Microsoft developed the virtual assistant *Cortana[10]*, which is largely utilized on Windows-based devices. Taking its name from an AI character in the well-known Halo video game series, Cortana was given that name. To provide consumers a more complete experience, Cortana interfaces with other Microsoft services including Bing search, Windows Calendar, and Microsoft To-Do. Cortana helps to performs various task such as calling, travel, random tips, maps, settings, scheduling, and more on commands.

Virtual Assistant	**Number of Languages**	**Accuracy**	**WER**	**Other Metrics**	**Wake Words**
Siri	21 [3]	95% [6]	Not Available	False Accept and reject rate	Hey Siri
Alexa	8 [2]	95% [6]	Not Available	Acoustic echo cancellation	Alexa, Amazon, Echo
Google Assistant	44 [20]	86% [7]	4.9 % [9]	Personalized recognition of speech	Hey Google, Ok Google
Cortana	8 [3]	Not Available	5.1% [9]	Eavesdropping aspect	Hey Cortana
Bixby	8 [5]	Not Available	Not Available	Absence of open NLP and AI	Hi Bixby

Comparison of Siri, Alexa, Google Assistant, Cortona, and Bixby

The table above describes the comparison of 5 different types of Virtual assistants Siri, Alexa, Google Assistant, Cortona and Bixby in regard to accuracy, WER, Wake word, Number of Languages spoken and other metrics. The accuracy of Siri and Alexa is highest at around 95% followed by Google Assistant. Whereas the WER of Google Assistant is less than that of Cortana.

Google Assistant is one that recognizes the most number of languages (44), following Apple's Siri (21 Languages). Bixby recognizes 8 languages which are similar to the language recognized by Cortana.

	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%

Comparison of various virtual assistants from April 2017 to July 2018 (Image Source)

[4] contrasts the capabilities and characteristics of Siri, Cortana, Google Assistant, and Amazon Alexa, four well-known virtual assistants. The significant distinctions between the four virtual assistants are highlighted in the essay, including their platforms, speech recognition, natural language processing, and interaction with other smart home device capabilities. The article also lists some of the distinctive characteristics of each virtual assistant, including Alexa's extensive range of third-party integrations, Cortana's ability to manage schedules and reminders, Google Assistant's capacity for detailed question-and-answer responses, and Siri's integration with Apple's ecosystem. The article's conclusion asserts that although each of the four virtual assistants has advantages and disadvantages, the ideal option ultimately relies on the user's particular requirements and preferences.

Question 2 – Wake Word

We ended the lecture on wake words by saying that wake words might eventually disappear and that we could invoke responses from our VA without calling them directly.
TO DO:

- Give some ideas of the impact it could have on our communication with our VA is there were no more wake words.
- Can you find additional blogs or articles on the topic? What do they say?

Answer 2:

There would be several impacts both positive as well as negative if there would have been no wake-

- There may be more background noise or speech that the Virtual Assistant needs to filter out, making it more difficult for the VA to understand the user's purpose.
- Unintended answering, It might happen that two people are having a conversation among themselves and VA started responding without even being asked to do so.

- Not answering at all, suppose a person is asking VA something, since there is no wake word detection it might respond at all.
- It might confuse VA, If there are more than two people in the room talking to each other, and in between their conversations someone asks the VA something, to which the VA didn't respond because it doesn't know when the speaker started to talk to it because of the absence of wake word.
- It may cause users to worry about their privacy and security because the VA would always be monitoring and maybe recording their conversations.
- The absence of wake words could enhance the user experience by creating the impression that they are conversing with a human assistant rather than a VA. This could make the user more at ease and encourage regular use of the VA.

There are some articles that discuss the impact VA could have on our communication with our VA if there were no more wake words:

In the article "Hey Alexa, No More: A New Function Deemphasizes Alexa's Wake Word"[22], it is discussed how Alexa, Amazon's virtual assistant, has a new feature in some regions. The function seeks to lessen the significance of Alexa's "wake phrase," With the new functionality, users won't need to speak "Alexa" first; they may simply say a command to start Alexa. With this modification, Alexa should be easier and smoother to use, and users won't need to keep using the "wake word" to communicate with the assistant as often. The functionality is being introduced gradually and will soon be accessible to all users.

The article[23] emphasizes that as virtual assistants become increasingly integrated into our everyday lives, the usage of the wake word can lead to embarrassing occasions, particularly in public or social settings, and raises privacy issues because virtual assistants are constantly listening. The author believes that the next phase in the growth of intelligent assistants is to adopt alternate approaches, such as gesture-based controls and voice biometrics, in lieu of predefined wake phrases. In addition to facilitating a more smooth and more secure engagement with virtual assistants, this change would provide an additional layer of privacy protection.

Article [24], discussed the discomfort of using a lengthy wake word, such as "Hey Siri," which has prompted Apple to explore alternatives, such as the "compressed" wake word described in the patent. It intends to enhance the user experience by discovering a shorter and more natural way to activate Siri.

Question 3 – Diarization

One topic we didn't have a chance to talk about during out lecture is speaker diarization. TO DO:

- Read this blog by IBM: New advances in speaker diarization

- Summarize the article.
- Give some ideas why diarization would be important for VAs.

Answer 3:

As per the blog [25], recognizing "who spoke when" is known as automatic speaker diarization. The authors provide a novel neural network-based speaker diarization technique that, in terms of accuracy and computational effectiveness, beats state-of-the-art approaches. The suggested technique employs a speaker encoder to extract speaker-discriminative characteristics from audio segments, and then a clustering layer to identify the various speakers for each segment. On several benchmark datasets, the authors demonstrate that the suggested technique performs better than other speaker diarization methods.

Their novel method makes use of neural networks for leveraging uncertainty information and acoustic embeddings. In addition, they put forth a cutting-edge method for measuring the size of spectral clusters. The blog[25] describes the entire dialization procedure in detail. Diarization is done by cutting the audio into segments which have speaker characteristics which have only the voice speaker's voice and then clustering is performed. For the purpose of speaker embedding they used both TDNN-based x-vectors[26] as well as LSTM-based d-vectors[27]. For clustering schemes, speaker similarity between speech segments is essential. They built a neural network (NN) to determine speaker similarity between cluster pairings in order to account for duration-dependent within-speaker variability. They give the neural network a pair of audio embeddings together with the matching durations. For estimating the number of speakers they went beyond Eigen analysis and performed Temporal response analysis, which is basically the eigenvector multiplied by the similarity matrix to get a vector(Temporal Response). They evaluated their results on the CALLHOME-500 corpus[28], Results were expressed in terms of diarization error rate (DER), which is the percentage of time that is incorrectly ascribed, and produced a state-of-the-art DER of 5.1%, which essentially means that their method performed better than other published works.

Diarization is Important because:

- By differentiating the voices of different speakers, the virtual assistant can evaluate the speech patterns, accents, and word pronunciations of each speaker, enabling more accurate speech recognition and personalizing the experience for each user. This results in more natural and instinctive interaction.
- VAs are able to acquire a better knowledge of the context of a conversation by Diarization
- Speaker diarization enables virtual assistants to instantly identify the person who interacted with VA, allowing for improved processing and faster answers.
- Diarization is a useful tool for preventing the disclosure of private information during conversations that would otherwise be overheard, such as medical or financial consultations.

Question 4 – Safety

There are different contradicting studies about the safety of using speech recognition in cars, for car-integrated VAs.
TO DO:

- Find 2 studies (or blogs discussing the studies)
- Summarize/compare their findings.
- Give your own opinion on the matter. What do you think?

Answer 4:

Voice-Activated Technology Is Called Safety Risk for Drivers[29]

The NY Times article[29] describes how voice-activated in-car systems have grown in popularity as a way for drivers to conduct activities like texting, emailing, and navigating without needing to physically interact with their gadgets. According to the report, even when drivers are not physically handling their gadgets, these systems can still be distracting and are not always dependable. The AAA Foundation for Traffic Safety study discovered that utilizing voice-activated devices while driving can increase a driver's cognitive workload and distractions, potentially increasing the likelihood of accidents.
The article also notes that there are no requirements for these systems' design or performance and that the National Highway Traffic Safety Administration (NHTSA) does not yet regulate them. According to the article, several automakers are beginning to add warnings to their owner's manuals advising drivers not to use the systems while driving.
While using voice-activated in-car technologies is meant to make driving safer, the article's conclusion notes that much more research is still needed to fully understand their effects on driving safety. The essay urges the creation of performance criteria and rules for the creation and application of these systems as part of a more thorough approach to regulating them.

In-Car Speech Recognition: The Past, Present, and Future[30]

The article "The Present and Future of In-Car Speech Recognition" article offer a thorough analysis of the present situation and prospective future applications of speech recognition technology in the automobile sector. It starts off by going through recent developments in speech recognition, such as improved accuracy and the incorporation of virtual assistants like Apple's Siri, Google Assistant, and Alexa from Amazon. According to the report, these advancements have made it possible to operate a number of in-car features—including navigation, audio playback, and climate control—hands-free. The possibility for context-aware features and multilingual support to be incorporated into in-car speech recognition technology is then covered by the author. With the use of context-aware technologies, the virtual assistant could comprehend the driver's context and offer pertinent

assistance. Drivers might utilize the system in their preferred language with multilingual support, opening it up to a larger audience.
The article concludes with an optimistic forecast for the development of in-car voice recognition technology and how it might completely alter the driving experience. The author thinks that as technology develops, everyone's driving experience will become safer, more practical, and more tailored.
Article[29] majorly mentions the negative aspects of speech recognition in cars, for car-integrated VA and contradicts Article[30], which presents a positive outlook for the future of in-car speech recognition technology and its potential to revolutionize the driving experience.

Our Opinion:

From our point of view, it has both pros and cons. It will be easier for drivers to operate many parts of their vehicles, such as climate control, phone calls, and music controls, if technology is used responsibly and in accordance with safety features and requirements. Due to the need to gather and store a person's voice data, the usage of speech recognition technology also poses privacy issues. The ability to operate many components of a car without taking one's hands off the wheel will have positive effects on functionality and use. Additionally, speech recognition technology enables people with impairments to have greater access to a vehicle's features and operations. This may increase the appeal of driving and increase its accessibility for more individuals.

Question 5 – Scaling issues in ASR

In Adam Coates videos (see both references in my slides), he mentions the problem of scaling end-to-end approaches. Often, access to large amounts of data and high-performance GPUs makes the difference between various systems these days, and not the actual ideas behind the systems.
TO DO:

- Summarize the issue.
- Give your opinion on what this scaling issue represents for the research field of Speech Recognition. For example, what is the accessibility of university research labs versus private labs to such data and computing power?

Answer 5:
In the videos, he discussed scaling up the deep learning-based approach for an end-to-end Deep learning speech engine. The Key to their approach is the application of HPC techniques, resulting in a 7x speedup over their previous system [31]. For scaling up basically he discussed 2 topics:

1) Data
2) Computation Power

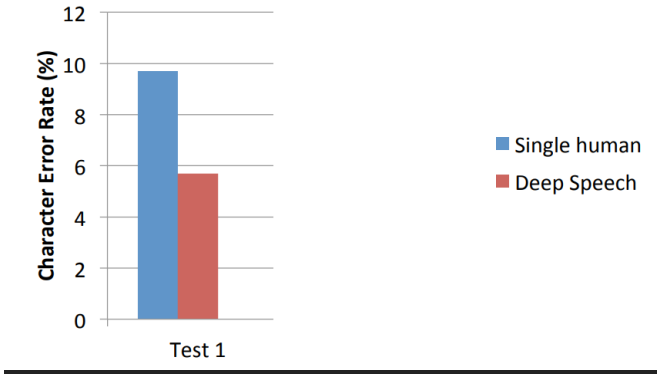
I) Data:

Adam Coates started by discussing that transcribing speech data is not cheap and roughly 1$(USD) per minute. Also, the type of speech data matters and it very much depends on the application for example "*The style of Speech*" - Like reading, conversation etc. "*Issues*", like Echo, Noise mic quality etc. and more importantly the type of application for which the data will be used. Adam also discussed the data augmentation which they performed using additive noise, as the data with actual noise along with the clean voice is costly, thus adding noise to the clean data and synthesizing a huge amount of data.

II) Computation Power:

Adam mentioned that more data will help if the model is computationally efficient. Adam also mentioned that 1 experiment on TitanX would take around 1 month to run and if you want to run different models then it won't be much beneficial, that's why to increase the computational power Use more GPUs with data parallelism and setting Minibatches upto 1024 seem useful. Putting infrastructure together and putting it on the server by getting access to servers like Buying servers (8 TitanX) reduces the training time to a week which is very significant. He also discussed different ways to use GPUs, at Baidu Research they used *Synch.SGD*. There is a need to optimize single-GPU code, by calculating time if drastically underperforms, as the library that has been imported might have hit the edge case of one of the imported libraries. Code optimization can be used for optimized computational power to handle edge cases. Also, for numerical optimization Batch normalization and SortaGrad plays a crucial part in better and easy computation as the same amount of length of Batch size makes computation efficient.

The result of the experiment is shown below in the Image.



Scaled-up Models in Mandarin (Image Source)

Results demonstrate and validate the worth of end-to-end using deep learning techniques for speech detection in various contexts and also conclude that it performs better than a Single Human.

University research labs versus private labs:

University research laboratories lack vast datasets and computational resources compared to commercial labs like technology businesses or research organizations. Funding restrictions and

institution research emphasis might limit data collecting and computational infrastructure resources. University research laboratories may have access to specialized databases and research communities for specific projects. For example, getting a server to incorporate more GPUs to improve computation power is not much likely to be invested by the University due to limited funding.
Contrast, commercial labs have access to considerably larger datasets and more potent computational capabilities, particularly those in the technology sector like access to GPUs. These businesses are able to make considerable investments in data collecting and computing infrastructure because they have ample financial resources and a focus on creating cutting-edge technologies. Private laboratories could also collaborate with other institutions, including universities, to get access to specialized databases and research communities.

In conclusion, commercial laboratories often have greater resources available for large-scale deep learning projects, even if both university research labs and private labs have their strengths and limits in terms of access to data and processing capacity.

Question 6 – Evaluation

In Word Error Rate, the cost of deletion, insertion and modification is the same. Imagine a text "the cat is lazy" and System 1 recognizes "the cat its crazy", System 2 recognizes "the clap is crazy" and System 3 recognizes "all cat is hazy". All systems would get a WER of 2/4.
TO DO:

- Give your opinion on the example above.
- Do you think the WER is appropriate or not?
- Find some articles/blogs which suggest other measures or that suggest adaptation to WER. Find at least 2 measures that you will explain.
- Summarize what you find and provide the sources.

Answer 6:
WER is the Word Error Rate. It is used to compute the error rate of our ASR (Automatic Speech Recognition) system and determine how accurate the ASR system is. It computes by taking three factors into consideration: substitutions, insertions, and deletions.
Substitution (S): It is the number of words it replaces during speech recognition; For Example, the cow is spelt as mow.
Insertions (I): It is the number of words it automatically adds up when they aren't present in real; For Example, SAT is transcribed as essay tea.
Deletion (D): The number of words deleted during speech recognition but present in real is referred to as deletions. For Example Cheese is very good is turned into Cheese is good [12].

The Word Error Rate (WER) is the summation of substitution (S), Insertions (I) and Deletion (D) divided by the total number of words spoken [13].

WER = (S+I+D) / No of words Spoken

Let's suppose there are 30 total words, and the ASR system made 4 substitutions, 3 insertions, and 3 deletions. As a result, 10 changes (S + I + D) result in WER = 10/30 = 0.333 i.e, 33% percent of WER.
The lower WER indicates that the ASR system is more accurate and efficient in speech recognition, whereas higher WER infers a less accurate and less efficient ASR system.

Scenario:
Sentence: the cat is lazy
System 1 recognizes: the cat is crazy
System 2 recognizes: the clap is crazy
System 3 recognizes: all cat is hazy
All systems would get a **WER of 2/4**

In this scenario, System 1 would be most efficient than the System 2 and 3, whereas System 3 is better than System 2 even though the WER is same for all the system. I feel this way because the WER just gave a number based on the three criteria (Substitution, Insertions and Deletion) but there are other factors as well which should be taken into consideration such as:

- What substitution has been considered
- What deletion has been done.
- What insertions are performed.
- Does substitution, deletion and insertions have affected the meaning of the sentence.

In the above scenario, we can observe that all the System has a WER rate of 2/4. But, System 1 has generated a result which is highly efficient and correlated with the real sentence, the meaning hasn't been altered. On the other hand, System has changed the noun of the sentence that is changing the cat and translated it to clap, therefore impacting the meaning of the sentence. Similarly, System 3 has changed the verb which has also changed the meaning. Therefore, System 1 has performed better than System 2 and 3 while having the same WER.

WER is not appropriate because it doesn't take many factors into account. The above scenario is one of example how only the WER is not capable of deciding the accuracy or performance measure of the ASR Systems. Moreover, it doesn't take the following factors into account:

- What could be the source of error:
    - Recording quality
    - Microphone quality
    - Pronunciation of the speaker

    - Noise
    - Abrupt words
- It doesn't determine the change made has changed the meaning of the sentence or not.

Therefore, WER alone is not appropriate to determine the accuracy and performance of the ASR System.

There are other performance metrics that can be used to evaluate the ASR system [14]:

1. **Perplexity:** It is a measure of how well a language model is able to predict the next word in a sequence of text. It is commonly used to evaluate the quality of a language model and to compare the performance of different models. A higher perplexity value signifies that the model is less certain of its predictions and more likely to come up with unexpected findings, while a lower perplexity value shows that the model is better at predicting the subsequent word.
2. **Levenshtein distance:** A string metric termed as Levenshtein distance calculates the distances between two words. It does this by figuring out how many characters must be altered to get from one word to the other in order to determine how different they are.
   For instance, The distance from word kitten to word sitting would be three since getting there would require three substitutes (changing k to s, changing e to i, and inserting g).
   Levenshtein distance ties back to WER since it tracks insertions and deletions, like many of the metrics used to assess ASR. It is beneficial because it offers a more thorough view of the changes being made than simply how many changes occur overall.

Therefore, WER is a good metric to determine ASR system robustness, but it should be considered with other factors/performance measures as well. Only WER is not enough to determine the accuracy of the ASR system.

Question 7 – Datasets

For large-scale evaluation of speech recognition software, some datasets exists.
TO DO:
Find 2 different datasets that you think could be appropriate for testing voice assistants. Describe them and compare their characteristics (make a comparative table).

Answer 7:

1. **CommonVoice dataset by Mozilla**
   CommonVoice is a large-scale resource of crowdsourced speech data that is publicly available and created by Mozilla. It offers a variety of high-quality speech data, the CommonVoice aims to make it easier and faster for programmers to train and test speech

recognition algorithms. The dataset on the common voice website has a unique MP3 and matching text file for each entry. Several datasets around 26,119 recorded hours also include demographic information like age, sex, and accent that can be used to improve speech recognition software's accuracy. There are presently 17,127 validated hours across 104 languages in the dataset [15].

Some of the features of the common voice Dataset are:

1. It is Large-Scale Dataset; therefore developers can train and test on huge datasets and models can evaluate speech recognition systems with high accuracy and precision.
2. Data is collected by crowdsourcing; therefore, it is vast and very diverse. As a result, the dataset is representative of real-world speech, including a variety of accents and languages.
3. CommonVoice supports multiple languages, including English, French, German, and several others.
4. CommonVoice is an open-source project, therefore it can be accessed freely. As a result, programmers are able to use the dataset for research and development and contribute to the project by adding new speech data or enhancing the quality of the already-existing data.

Common Voice (Source)

2. **Google Speech Commands Dataset**

The Google Speech Commands Dataset is developed by TensorFlow and AIY teams to provide a speech recognition example by using the TensorFlow API. There are 65,000 clips in the dataset that last for one second each. Each clip includes one of the 30 words, including "yes," "no," "up," and "down," which are pronounced by different subjects. Its purpose is to test how well keyword-spotting algorithms work, and it can also be used to see how well voice assistants can recognize specific commands. The Google Speech Commands dataset is widely used in the speech recognition research field as it has a lot of high-quality speech data [16].

Some of the features of the Google Speech Commands dataset are:

1. The Google Speech Commands dataset is a large and diverse dataset for training and evaluating keyword spotting systems. It contains hundreds of hours of speech data.
2. The dataset is created by Google, it is of high quality and free of any issues that may occur when using crowdsourcing speech data.
3. The Google Speech Commands dataset is limited to English-speaking accents; therefore, it is ideal for developing and testing keyword identification software in English-speaking regions.
4. The Google Speech Commands dataset is particularly suited for testing keyword detection systems because it contains brief one-second audio clips of spoken commands.
5. The Google Speech Commands dataset is a significant resource for comparing and assessing the effectiveness of keyword detection systems because it is widely used in the speech recognition research community.

CommonVoice dataset by Mozilla	**Google Speech Commands Dataset**
CommonVoice dataset is collected through crowdsourcing.	The Google Speech Commands dataset is generated by Google.
CommonVoice dataset is much larger than the Google Speech Commands dataset, as it consists of 10000 hours of speech data.	It consists of 65,000 clips in the dataset that last for one second each.
It is a multi-language dataset.	It is limited to the English Language.
CommonVoice dataset is intended for training and testing voice assistants.	It is specifically designed for testing keyword spotting systems
CommonVoice dataset contains a wide range of speech data and long audio files (conversational, read, and speech).	It consists of one-second audio clips only.
Since it is crowdsourced speech data, it is of low quality.	It is generated through Google; therefore, it is a very high-quality dataset.
As it consists of a variety of speech data, it is a more real-time and real-world dataset.	Google Speech Commands dataset is designed specifically for testing keyword spotting systems and may not reflect real-world scenarios as closely.

CommonVoice dataset is openly available to the public	Google Speech Commands dataset requires registration and access to Google's Cloud Speech API

Comparison of CommonVoice dataset by Mozilla and Google Speech Commands Dataset

Question 8 – Open Source software

For doing Speech Recognition, some open source software exists. The following site (https://fosspost.org/open-source-speech-recognition/) mentions 10 of them. Explore two of them.
TO DO:

- Describe the software found.
- Try to do a tutorial to be able to test the software.
- Does they seem easy to use, or are you still unable to do anything after some explorations?
- Put the advantages/disadvantages of each software in a comparative table.

Answer 8:

1. **Kaldi**

Kaldi is a C++-based open-source voice recognition program and is available under the Apache Public License. It was started in 2009. Kaldi is also used as a component in other speech-recognition software. It supports Windows, macOS and Linux. It offers a selection of instruments and algorithms for tasks connected to voice recognition, such as:

- Acoustic model development
- Decoding (generating transcriptions from audio)
- Feature Extraction (converting audio into a form that can be used as input to a speech recognition system)

Due to its modular architecture, Kaldi is simple to expand upon and modify to accommodate new tasks and languages. It has been used to create speech recognition systems for various languages such as English, Spanish, German, and Mandarin [17].

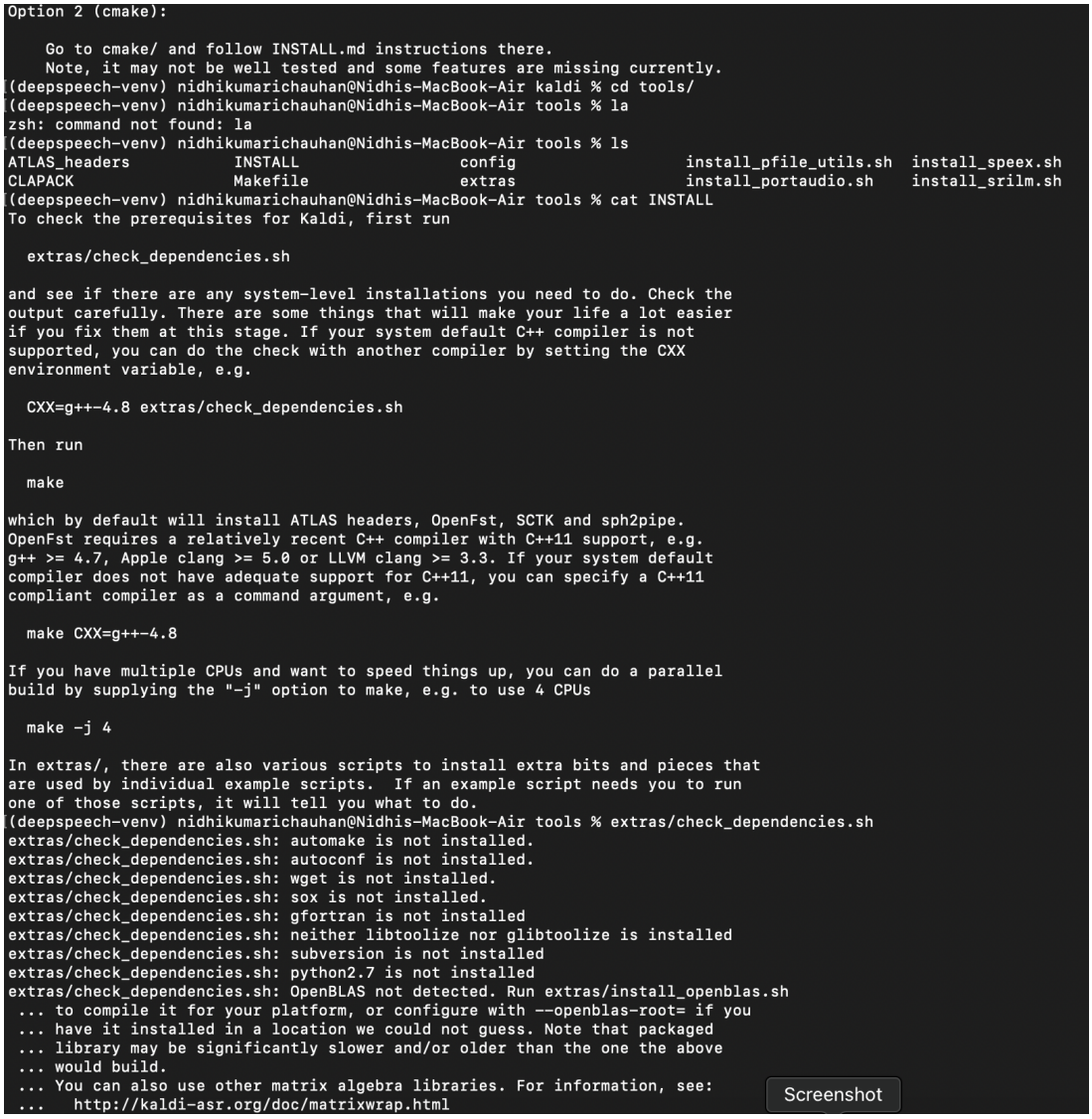
The Kaldi toolbox has a huge and active user base that offers a lot of information and assistance to individuals who are interested in using it for speech recognition and related activities.

Installation Guidelines are mentioned in [21].

Firstly, since the Kaldi is primarily hosted on GitHub. So, we need to git clone Kaldi from GitHub - https://github.com/kaldi-asr/kaldi.git

1. Cd Kaldi-asr
2. Check for INSTALL.md
3. Check for all the requirements needs to install Kaldi.
4. If the requirements are set, we can move forward and install Kaldi.

5. Else, we need to first meet and fulfil the system requirements and then move forward to install Kaldi.



Installation Error

It might be easy to use but it requires a lot of dependencies which our system wasn't able to fulfil. Therefore, we weren't able to install the Kaldi.

Dependencies required:

1. Automake - install Successfully
2. Autoconf - install Successfully
3. Wget - install Successfully
4. Sox - install Successfully
5. Gfortran - Error
6. Libtoolize/glibtoolize - Install Successfully

7. Subversion - Error
8. Python 2.7 - higher version is installed
9. OpenBLAS - Error

Advantages and Disadvantages of Kaldi:

Advantages	Disadvantages
Kaldi is open-source, available to users for free, and allows for customization and modification.	Although Kaldi has a large user base, the documentation can be minimal and challenging to read, making it difficult for users to understand how to use the toolkit efficiently.
It has a large and active user base, which gives access to a wide range of information and resources.	It requires a lot of dependencies, which need to be fulfilled before installing it.
Kaldi is easily included with other programs and databases, enabling a smooth workflow.	Kaldi is difficult to use and install, therefore making it difficult for new programmers to run.
Kaldi is a flexible tool for speech recognition as it supports multiple languages	Kaldi has fewer resources, making it more difficult to follow and install.
Kaldi is an effective tool for voice recognition jobs since it is built to handle huge amounts of data and perform effectively in real-world scenarios.	As an open-source toolkit, Kaldi needs to be updated and maintained so that it stays up to date with best practices and new technology. This can take a lot of time and may require technical knowledge.
Kaldi has a simple way to train speech recognition models, which makes it easy for users to build and change models to fit their needs.	As an open-source toolkit, Kaldi needs to be updated and maintained so that it stays up to date with best practices and new technology. This can take a lot of time and may require technical knowledge.

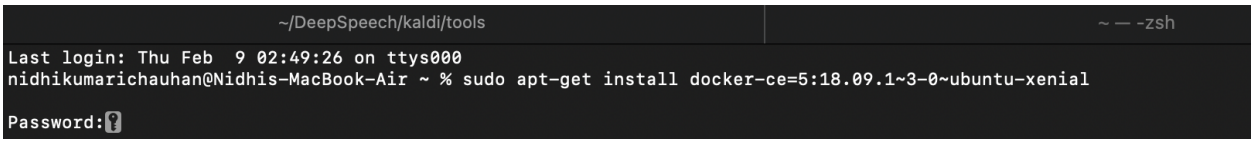
Advantages and Disadvantages of Kaldi

2. OpenSeq2Seq

Open2OpenSeq is a sequence-to-sequence model in the form of a neural network architecture used for various applications like machine translation, text summarization, and speech recognition. The library offers a high-level API for creating and training sequence-to-sequence models and is built on top of the well-known deep learning framework TensorFlow. In order to help you get started quickly, it supports a broad variety of model architectures, including attention-based models, and it comes with a lot of pre-trained models and training examples. Additionally, the library offers

tools for tracking the training procedure and assessing model performance, as well as distributed training support.
It offers a high-level API for constructing and refining sequence-to-sequence models, making it simple for users to get started and test out various model architectures. Due to these capabilities, OpenSeq2Seq is a strong and adaptable tool for creating and refining sequence-to-sequence models. This makes it popular in both academia and industry for a wide range of activities.
It can be installed by performing the steps mentioned in the Installation Guide.

However, installation of OpenSeq2Seq is not quite easy on MacOS. The commands mentioned in the above-mentioned links require a password to enter which is not mentioned on the website.



Installation Error

Along with it, it also requires a lot of dependencies to fulfil. TensorFlow and other libraries are among the dependencies that OpenSeq2Seq depends on. If you already have these libraries installed on your computer, different versions that don't work well together could make the installation difficult. Along with it, it is mentioned that it worked with previous versions of MacOS, not the new versions.

Advantages and Disadvantages of OpenSeq2Seq:

Advantages	Disadvantages
OpenSeq2Seq offers a flexible framework for designing custom neural network designs for various NLP tasks.	OpenSeq2Seq can be difficult for users who are unfamiliar with deep learning or natural language processing because these fields must be understood to use the platform properly.
OpenSeq2Seq offers pre-trained models that may be adjusted for particular use cases for a variety of NLP activities, including text summarization, machine translation, and question answering.	Because OpenSeq2Seq can use a lot of memory during training, it can be challenging to train models on big datasets or tackle challenging NLP tasks.
OpenSeq2Seq is based on the well-known deep learning framework TensorFlow and makes use of its features to offer users a powerful NLP platform.	Despite the fact that OpenSeq2Seq has an active user and developer community, getting Starting with the platform can be challenging for new users.
The flexibility of OpenSeq2Seq's design lets users add unique parts to their models, like	Since OpenSeq2Seq is based on TensorFlow, users must grasp TensorFlow to some extent

extra layers, attention mechanisms, and more.	in order to use the platform properly.
OpenSeq2Seq is open-source software, therefore, allowing users to obtain the source code, use it without charge, and participate in its development.	Users may find it challenging to connect their models to other programs or systems while using OpenSeq2Seq because it lacks a standardized API.

Advantages and Disadvantages of OpenSeq2Seq

References:

[1] Alexa-vs-siri-vs-google
[2] Language-support-voice-assistants-compared
[3] Cortana_(virtual_assistant)
[4] Siri-cortana-google-assistant-amazon-alexa-face-off
[5] Bixby(software)
[6] Speech-recognition-accuracy-history.
[7] Speech-to-text-transcript-accuracy-rate-among-leading-companies
[8] Bixby
[9] Does-word-error-rate-matter
[10] what-is-cortana
[11] Samsungs-bixby-performance-improved-with-google-cloud-tpus-and-sensorflow
[12]What-is-wer-what-does-word-error-rate-mean
[13]Word-error-rate
[14]Key-metrics-and-data-for-evaluating-speech-recognition-software
[15] Commonvoice.mozilla-datasets
[16] Launching-speech-commands-dataset
[17] Tutorial-kaldi
[18] OpenSeq2Seq-installation
[19] OpenSeq2Seq-Github
[20]Google-assistant-can-now-interpret-44-languages-on-smartphones
[21] Installing-Kaldi
[22]Hey-alexa-no-more-a-new-feature-deemphasizes-alexa-s-wake-word
[23]ts-time-smart-assistants-evolved-beyond-fixed-wake-words-like-hey-google
[24]Tired-of-saying-hey-siri-apple-working-to-shorten-wake-word-for-personal-assistant
[25]New Advances in Speaker Diarization
[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech

and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[27]Y. Higuchi, M. Suzuki and G. Kurata, "Speaker Embeddings Incorporating Acoustic Conditions for Diarization," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7129-7133, doi: 10.1109/ICASSP40776.2020.9054273.

[28] Philadelphia: Linguistic Data Consortium, University of Pennsylvania

[29]https://www.nytimes.com/2013/06/13/business/voice-activated-in-car-systems-are-called-risky.html

[30]https://summalinguae.com/language-technology/the-present-and-future-of-in-car-speech-recognition/

[31] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. 1412.5567, 2014. http://arxiv.org/abs/1412.5567.

[32] Google-assistant-is-more-accurate-than-alexa-siri-and-cortana