Optimizing Human-Device Interaction through Real-Time Automated Speech Recognition and NLP

By

Tahiya Rahman Oishy

ID: 201-15-13753

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science Engineering

Supervised By

Sharun Akter Khushbu

Lecturer (Senior Scale)

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised by

Raja Tarequl Hasan Tusher

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

Daffodil International University

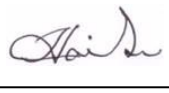
Dhaka, Bangladesh

28 December 2024

APPROVAL

This Thesis titled **"Optimizing Human-Device Interaction through Real-Time Automated Speech Recognition and NLP.",** submitted by Tahiya Rahman Oishy, ID: 201-15-13753 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on Saturday, 28th December 2024.

BOARD OF EXAMINERS



Dr. Sheak Rashed Haider Noori **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan **Internal Examiner 1**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar **Internal Examiner 2**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Shah Md Imran **External Examiner**
Project Manager
Smart Leadership Academy
Enhancing Digital Governance and Economy (EDGE) Project
Bangladesh Computer Council
ICT Division

DECLARATION

I hereby declare that; this project has been done by us under the supervision of **Sharun Akter Khushbu**, Lecturer (Senior Scale), Department of Computer Science and Engineering. Daffodil International University. I also declare that neither this paper nor any part of this paper has been submitted elsewhere for award of any degree or diploma.

Supervised by:

29.12.2024

Sharun Akter Khushbu
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

24.12.2024

Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Tahiya Rahman Oishy 24.12.2024

Tahiya Rahman Oishy
ID: **201-15-13753**
Department of Computer Science and Engineering

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project/internship successfully.

I am really grateful and wish our profound indebtedness to **Sharun Akter Khushbu, Lecturer (Senior Scale)**, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of **"Optimizing Human-Device Interaction through Real-Time Automated Speech Recognition and NLP."** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to **Prof. Dr. Sheak Rashed Haider Noori**, Head Department of CSE for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Automatic speech recognition (ASR) is a technique which enables machines to interpret, convert and translate spoken language into text. To produce a text from spoken language, ASR system receives input from the speaker and subsequently decodes the input using some patterns, algorithms or model. In this project, the research emphasized how speech recognition systems can be used to automation tasks, prioritizing the performance of both online and offline algorithms such as Google API, PocketSphinx and Vosk in various circumstances. Therefore, in current study, ASR model had been analyzed in detail where Hidden Markove Model and Gaussian Mixture Model (HMM and GMM) symbiosis set as the base of the experiment. The project was built-up on Python to execute three platforms as preliminary target and the algorithms of the platforms are Google API, PocketSphinx and Vosk. All these three platforms had been compared to find robustness and superiority, but interestingly, Vosk was conducted extensively better accuracy than Google API and PocketSphinx. An assessment platform was prepared with the voice of different age groups and considered voice, frequency-noise and word error rate (WER) to highlight the durability of these systems. The findings illustrated that Vosk beat Google API and PocketSphinx in a variety of contexts. Therefore, to overcome the problem, in current study a predefined command list was set up as a methodical foundation for the assessment of every system in automated application. Despite of the limitations, the research was provided the companionship between human and computer especially for disabled people who are facing challenges using devices. Finally, this innovation opened a new window in ASR technique due to its effectiveness with the use of real time data and the evidence of more accuracy.

TABLE OF CONTENTS

FIGURE LIST

TABLE LIST

CHAPTER 1

INTRODUCTION

In 1774, the machine of speech was invented by Wolfgang Von Kempelen which defines 230 years of the speech machines developing period. But originally Bell Laboratories is the founder of speech recognition, and it was in 1952. It worked as understanding ten digits by a single delivered speech.

1.1 Identification of Speech

Transforming oral delivered language into textual form defines the identification of speech recognition. It dismantles obstacles to communicate by allowing to understand user performance and commands, helps disables with hands-free operation, acts like private secretary. Increasing comiser power, availability of data and great functionality are happening due to recognition of voice.

1.2 Motivation

The offline voice recognition system aims to make simple, reliable and compatible life for people, especially who are far away from places where internet connection is not available. This system is also adaptable for those people who are unable to use the internet and dependent on speech can easily operate device.

1.3 Research Question

1. What is the operational mechanism of speech recognition in offline mode without internet activity?
2. What is the target of this system?

1.4 Scope of Research

- Speech to text conversion
- Live data processing
- Human-computer interaction
- Offline based engine processing
- Helpful for unable people
- Performance and accuracy improvement

1.5 Working process of Automated Speech Recognition (ASR)

This Automatic speech recognition (ASR) system functions through a microphone which decodes the delivered voice by taking energy. It decodes the data from the old format to a digital one and translates it into plain readable text. This procedure synthesizes the delivered data to sound by processing algorithm and then use natural language processing (NLP) for identifying words. The process happens by taking sound vibration my microphone and transforming it into electrical signal. This signal processes in machine component and translate that into digital signal. Then this signal passes to the voice recognizer. After recognizing sound, it converts to text. HMM and GMM help to identify the model of voice. HMM based ASR tool is used as the foundation of this evaluation.

1.6 Voice Identification Technologies

Google API is an automatic speech recognition system which enables developers to integrate Google services, such as Maps, Cloud, and AI tools, into their applications, allowing for enhanced functionality and seamless user experiences. Another one is Amazon transcribe which is a part of AWS, provides speech-to-text services for audio files and real-time streaming. It includes custom vocabulary and transcription analytics and convert accurately video, audio to text. On the other hand, Vosk is an offline speech recognition engine that processes live audio data, eliminating the need for internet connectivity. It supports over 20 languages and delivers real-time transcription. It is also capable of making it ideal for applications and offline based voice-controlled devices. For speech detection and reorganization there is a toolkit named Kaldi which processes audio data by extracting features by using HMM. Kaldi also decodes audio

inputs using trained models, integrating acoustic, language and lexical model to produce transcriptions. If discuss about another offline toolkit that is PocketSphinx which is lightweight open-source speech recognition system is a part of CMU Sphinx project. PocketSphinx is known for its customizability, portability and ability to run on different platforms, like devices embedded systems, and desktops.

1.7 Circumstances of voice recognition with systems

This study is based on offline and online platforms where the main target to prove offline platform's feasibility with low usage, accessibility with better accuracy for unable and rural people who are unaware of internet connection or cannot use the network. Device identification, using of mouse keyboard for operating and making Graphical User Interface, using of python for defining WER, utilization of computer are the tasks.

1.8 Objectives

The goal of this study is to compare online and offline platforms' better performance by using ASR.

The research aim:

- Compare the performance of online and offline platforms.
- Identify better engine for the evaluation.
- Identify accuracy level and Word Error Rate (WER) of both platforms.
- Using Real-Time data by different noise level checking frequency and accuracy.
- Evaluate accuracy levels of different age levels.

1.9 Report Layout

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: Methodology
- Chapter 4: Experimental Results and Implementation
- Chapter 5: Result analysis based on Real Time Data
- Chapter 6: Summery, Conclusion, Limitations and Future Work

CHAPTER 2

LITERATURE REVIEW

2.1 Related Works

In current civilization, development of technologies generates an imperative connection with the future by the immense uses of technologies for human welfare. Speech signal processing resumes a significant influence in the field of business, health care, industry and personal computers [1]. Automated speech recognition (ASR), natural language processing and synthetic speech are some vital speech signal processing tools. Among all, ASR technology is rising in various industrial applications. In the mobile phone market ASR is used in different applications such as interactive voice response and operator services. Also in the health care sector, ASR got acceptance in disease diagnostic reports with large number of vocabulary of words. In general, growing applications of ASR in the industrial sector promotes accessibility of news products to the market. [1,2]

Voice regards a way of communication in humans, but interestingly in modern technology-based world, people are used to communicating with each other using many gadgets in daily life. Among the gadgets, the computer is the most significant one that can be conversed between humans and machines. Tabani et al. summarized key concepts of ASR as it became essential in mobile phone industry [2].

However, in order to improve the quality of ASR, researchers are making their efforts using different technology. Some efforts have been done making non-English language-based ASR such as Malayalam [3] Russian [4] and Indian [5] and also implemented in different fields successfully. In advance ASR system, most widely used algorithm is used named as Hidden Markov Model (HMM) in order to build up acoustic model as well as Mel Frequency Cepstral Coefficient (MFCC) to shape extraction from signal input [6]. Hamdan Prakoso et al., invented Indonasian ASR using limited data set and digit corpus, means a small set of vocabulary. In this ASR authors used CMUSphinx toolkit and investigated various SNR conditions to compare existing ASR. [6]

The preliminary aim of the ASR research was to allow a device to recognize words spoken by human beings, independent size vocabulary and speech characteristics with 100% accuracy level in real-time [7].

Last years, neural networks (NNs) were challenged traditional signal processing-based solution to enhance speech and successfully gained excellent implementation on a various number of works. [8].

Another study emphasized multichannel speech enhancement in ASR to generate a comprehensive description of the state-of-the-art enrichment system [9]. A unique combination of strong signal processing with DL boosted up the performance significantly in comparison with the earlier signal processing. The focus of the study was to implement DL technology in ASR for the far field scenario. In this technique data augmentation, adaptation and refinement of the supervision is included [10,11].

There are some applications or software that use speech recognition technology to convert text into documents such as Dictate which was launched in June 2017 by Microsoft. Dictate provides additional programs like Outlook, Word and Power point through voice detection. It detects 60 languages for real time audio conversion. Some other voice detection programs are iOS, a speech identifier, Dictation tool by Apple named macOS, voice typing by Google (Web), list note applied in Android, Gboard a function of Android and iOS etc. [12,13].

For instance, Computer-Aided-Design (CAD) is a voice command receptor speech recognition technology. In CAD, the system uses special speech recognition technology which detects commands that are reduced and subsequently mapped to target the function. This process is known as CAD sematic search, which is a combination of three processes named Tokenization, Parts-Of-Speech Tagging (POS tagging) and verb-based Identification [14,15].

Now a days a huge number of applications of ASR technology are used in education and different necessary consumer products for daily life use to make livelihood easy and comfortable [16,17]. ASR is an amazing technology that can be used in education systems especially for the poor writing skilled students to improve their writing by saying words loudly without concentration of pronunciation, spelling and other mechanics of writing [18,19,20].

2.2 Comparative Overview

Voice recognition technology has emerged as a promising way to improve the quality of life for people with various impairments by providing them with a control and communication method that suits their various and particular requirements. Due to

having online platforms like Google API, the offline project played an indispensable role. There is very rare research on offline but in this research, Vosk proved how much scalability offline platforms have. By comparing Vosk and PocketSphinx, Vosk performed greater than PocketSphinx and almost like Google API. With the testing of multiple frequency and noise level of different ages, environment compels us to believe about the offline engines. All these evaluations and investigations helped me to choose these three algorithms where Vosk performed better than others.

2.3 The Extent of the Issue

Complicated issues like figuring out real-time speech patterns and accommodating differences show how complicated my issue is. To get started in the field of voice recognition for people with disabilities, I needed to collect data, choose features, build a model, test it, and put it into use. These problems can be solved by using proper toolkit.

2.4 Challenges

There are many difficulties like dealing with different voice patterns, the processing of real-time data, choosing the accurate model which supports all required languages and speaking styles and finding the right mix between freedom and customization. Also, when adding different models to the Google API, PocketSphinx and Vosk frameworks, it's important to make sure they work with each other. Because of these problems, it is hard to make speech recognition systems that work well for handicapped people without doing a lot of study, getting feedback from users, and making technological advances all the time.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodology

The advanced automation setup shown in the lower figure features a seamlessly integrated speech recognition system. The microphone acts as the main input here. These advanced systems do a great job of filtering out background noise, turning spoken words into text, and carrying out the right commands. If they don't find a match, they simply return to their starting point without any fuss.

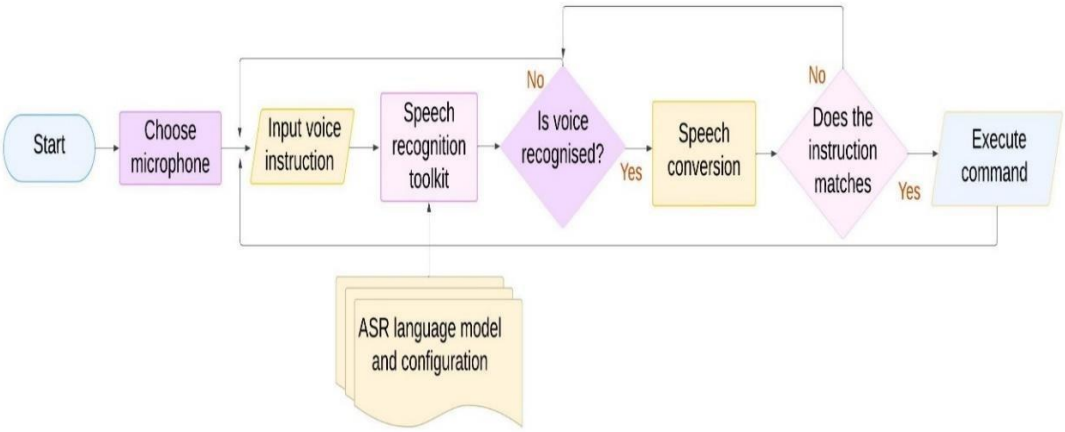


Figure 3.1: Flow Diagram of Suggested Voice Recognition System

In figure 3.1, the identified speech will be handled by speech recognition systems. At this stage, speech recognition systems will remove any ambient noise and unidentifiable talks.

Automatic Speech Recognition (ASR) is all about converting spoken words into written text. Early research in this area focused on the challenges of using statistical models, especially the combination of Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). The world of automated transcription is a delicate balance between advancing technology and understanding the nuances of language.

From figure 3.2, much of the effort of this study is on identifying efficient offline Python engines that lower Word Error Rate (WER) in voice input transcription and evaluating online and offline speech correctness.

ASR equation	$W^* \, p(W\|O) = argmax \, p(O\|W) \, p(W)$
WER equation	$WER = W + D + I/N$
Hidden Markove Model (HMM)	$Process \quad x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow x_{T-1}$ $Observation \quad O_0 \rightarrow O_1 \rightarrow O_2 \rightarrow O_{T-1}$
Gaussian Mixture Model (GMM)	$p(x) = \sum_{i=n}^{k} \emptyset \, iN(x \mid \mu_i, \sigma_i)$

Figure 3.2: The Mathematical Equation of ASR, WER, HMM, GMM

Figure 3.2 shows every basic formula used in my models. Figure 3.2 carefully shows the basic formulations underlying the models, therefore offering a basic knowledge of the essential mechanics guiding the optimization of speech-to-text transcription in the offline Python environment.

3.2 Methodology Overview

The research strategy outlined in this article begins with a careful selection of suitable platforms for offline speech recognition. The trial period is divided into two phases. The first phase involves conducting an initial experiment and analyzing the results. Based on the initial experiment, a follow-up experiment is conducted to thoroughly assess the most promising options for final implementation. The final phase focuses on developing multiple applications, each programmed with a predefined set of commands for real-time execution. The describing steps are representing in the figure 3.3 below:

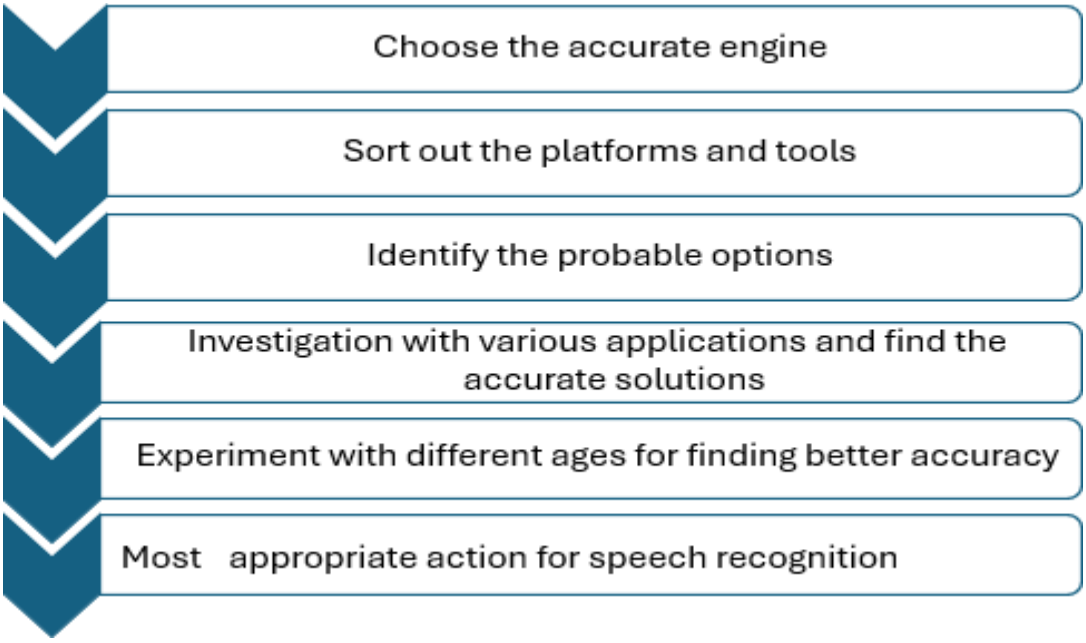


Figure 3.3: Research Methodology Steps

The journey of research has been started with a deep exploration of existing voice activated platforms. A series of experiments has been designed to improve accuracy and lower word error rate (WER) for automation purposes. By taking a quantitative approach and working within a clear research framework, we aimed to better understand and address the challenges in this field.

3.3 Work of Command List

The list of 25 commands is based on the foundation for assessing apps especially set up for specific experiments with parent command for each command on table 3.1. There are some commands taken for testing.

Table 3.1: Applications Command Declaration

Serial	Instructions	Work	Base Command
1	Start	Start application	
2	Mouse	Mouse functionalities	
3	Keyboard	Keyboard functionalities	
4	Left click	moving cursor towards left	2
5	Right click	moving cursor towards right	2
6	Double click	moving cursor to double tap	2
7	Browser	opening browser	1
8	Move left	moving cursor towards left	2
9	Move right	moving cursor towards right	2
10	Move up	moving cursor towards up	2
11	Move down	moving cursor towards down	2
12	Scroll up	Scrolling up	2
13	Scroll down	Scrolling down	2
12	Click	Single left click or select	3
13	Press enter	Enter pressed	3
14	Press escape	Enter escaped	3
15	Paste	Press cntrl + V	3
16	Copy	Press cntrl + C	3
17	Select all	Press cntrl + A	3
18	Undo	Undo last action	3
19	Tab	Moving next tab	3
20	Exit program	Exiting mouse	3
21	Show menu	Showing menu	2
22	Type	Start typing	3
23	Open notepad	Click and open notepad	2
24	Open calculator	Click and open calculator	2
25	Close window	Close current window	3

CHAPTER 4

EXPERIMENTAL RESULTS AND IMPLIMENTATION

4.1 Initial Experiment

Comparing and evaluating the accuracy between offline and online using grounded science techniques are the main tactic of my initial experiment. Word Error Rate (WER) and Speech to Text accuracy have been used while it's about taking certain speech features into consideration of environment.

4.2 Process of Circumstances

The initial strategy for this experiment is to take some real time data to compute the accuracy level by delivering in a well-controlled soundless environment with a 50-57dB noise level. A set up is made to configure three programs to the default settings to create a baseline output for each platform. A sentence is taken from Wikipedia to run the three test cases for each platform for the experiment. The average result of each platform is compared for choosing the best accuracy and lower error rate. Before the examination, uppercase is removed from the sentence, converted words to the lowercase and eliminated punctuation for determining Word Error Rate (WER).

4.3 Performance of Primary Experiment

The major experiment is among Google API, PockeytSphinx and Vosk for comparing the performance between online and offline according to their best accuracy, lower error rate and usage capabilities. These platforms are chosen because of their commendable and tremendous performance and capabilities on offline and online. The whole process is executed with some testing method by setting up proper environment where various operating systems exist. A paragraph has been chosen for better utilization from Wikipedia for correct assessment is given below:

“The electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of light with intensity inversely proportional to the square of the distance from its source region.” (ref. Wikipedia)

4.4 Experiment of Real Time Delivered Speech to Text using Google API

After using two platforms, here I am using Google API as a comparison of previous two platforms, showing the accuracy and WER of real time delivered speech. Main tactic of this platform is to measure the online accuracy and WER. Below table 4.1 is defining WER of the reference text and delivered audio.

Table 4.1: WER of Google API

Delivered Speech	**the electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of light with intensity inversely proportional to the square of the distance from its source region.**	**WER**
Test Run 1	electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of the light with Intercity inversely proportional to the square of the	25%
Test Run 2	the electromagnetic radiation emitted during a solar panel from the sun at a speed of light with the Intercity inversely proportional to the square of the distance of its resource reason	28.12%
Test Run 3	the electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of the light with intensity inversely proportional to the square of the distance from its source region	6.25%
	Average	19.79%

4.5 Experiment of Real Time Delivered Speech Converting to Text with PocketSphinx

There's a little difficulty here in measuring accuracy rate than the previous platform and word error rate is more. The prime target of this analysis is to compute the WER and accuracy rate of US language using NLP are given in table 4.2.

Table 4.2: WER output of PocketSphinx

Delivered Speech	**the electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of light with intensity inversely proportional to the square of the distance from its source region.**	**WER**
Test Run 1	that it on any cheating radiation any didn't do enough solar flare up on that alley from the sun and the speed of light we eat and eat less he proportionate to split all the distance found solace or the dna	90.62%
Test Run 2	in a home in to mean to the muslim thing for what it says on the senate does the fbi a convenience again once the proportion of it's better than these things and get subsidies in	81.37%
Test Run 3	but he intimated edition and they get here again as the the air but it's alice on the sand at the speed of light to medium density nervously a proportionate to discredit of the distance from its solar city he added	74.63%
	Averages	82.207%

4.6 Real Time Delivered Data to Text Experiment with Vosk

Basically, I have chosen the third one model of Vosk for this experiment which is shown in the below table 4.3 because of its more durability of accuracy and dynamic way of vocabulary configuration.

Table 4.3: Model Compatible with VOSK API List

	Model Name	Notes	Word error rate/Speed	Size
1	vosk-model-small-en-us-0.15	Lightweight wideband model for Android and RPi	9.85 (librispeech test clean) 10.38 (tedlium)	40M
2	vosk-model-en-us 0.22	Accurate generic US English model	5.69 (librispeech test clean) 6.05 (tedlium) 29.78(callcenter)	1.8G
3	vosk-model-en-us 0.22-lgraph	Big US English model with dynamic graph	7.82 (librispeech) 8.20 (tedlium)	128 M
4	vosk-model-en-us 0.42-gigaspeech	Accurate generic US English model trained by Kaldi on Gigaspeech. Mostly for podcasts, not for telephony	5.64 (librispeech test clean) 6.24 (tedlium) 30.17 (callcenter)	2.3G

By using a small model of Vosk, ascertaining the output and computing the WER accordingly. Table 4.4 gives the calculation of WER properly of Vosk.

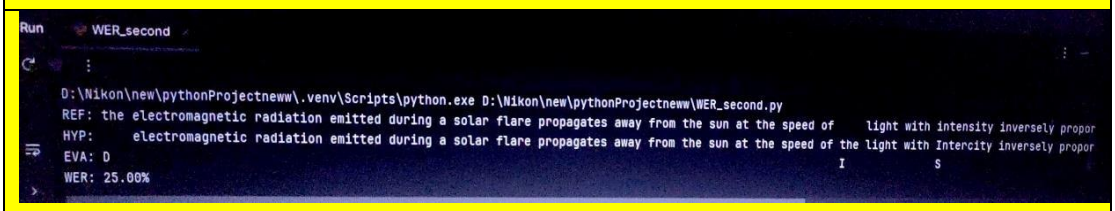
Table 4.4: Vosk WER Output

Delivered Speech	**the electromagnetic radiation emitted during a solar flare propagates away from the sun at the speed of light with intensity inversely proportional to the square of the distance from its source region.**	**WER**
Test Run 1	the electromagnetic radiation emitted during a solar flare the it's away from the sun at the speed of light with intensity inversely proportional to the square of the distance from each source reason	12.50%
Test Run 2	the electromagnetic radiation emitted doing a solar flare propagates our area from the sun at the speed of the light with intensity inversely proportional to the square all the distance from it so so region	28%
Test Run 3	the electromagnetic radiation emitted during a solar flare bravo get away from the sun at the speed of light we the intensity inversely proportional to the square of the distance from it's source region	18.75%
	Average	19.75%

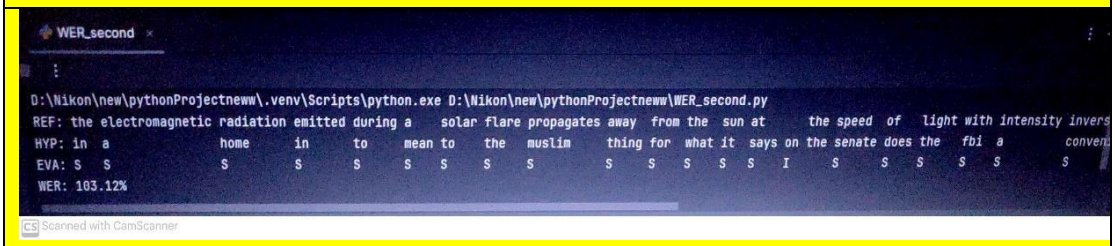
4.7 Result Analysis

Word Error Rate (WER) has been calculated using WER-In-Python tool in the upper table. Accuracy rate and WER have been calculated of three stages' test cases. Firstly, there is given a reference line in the “reference.txt” which gets implemented in the “hypothesis.txt” file. The implemented line's located hypothesis text. All the process of implementation is called WER output. The proper reference text and hypothesis are shown in figure 4.1.

Google API



PocketSphinx



Vosk

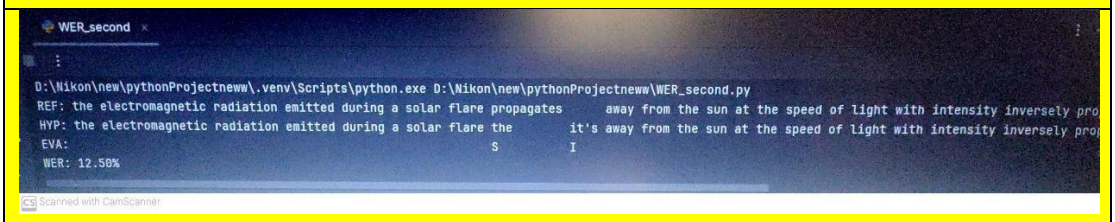


Figure 4.1: Some Sample WER Examples of Google API, PocketSphinx and Vosk

From the previous three calculations and figure 4.1, it is visible that Vosk and Google API are giving almost the same WER Average percentages than PocketSphinx. PocketSphinx has a bigger average of WER has proved on table 4.2. Here, Vosk and Google API have lower error rates and better accuracy. Vosk's average WER is 19.75% where Google API has 19.79% error rate which is almost similar. On the other hand, Google API has proved its accuracy rate and WER on table 4.1. A calculation of average word error rate (WER) and average accuracy rate is given on the table 4.5 below.

Table 4.5: Total Accuracy and WER of Google API, PocketSphinx and Vosk

Run	Google API	PocketSphinx	Vosk
1st	**25%**	**90.62%**	**12.50%**
2nd	**28.12%**	**81.37%**	**28%**
3rd	**6.25%**	**74.63%**	**18.75%**
Avg. WER	**19.79%**	**82.207%**	**19.75%**
Avg. Accuracy	**80.21%**	**17.793%**	**80.25%**

It will be more specific which platform provides better accuracy by the figure 4.2.

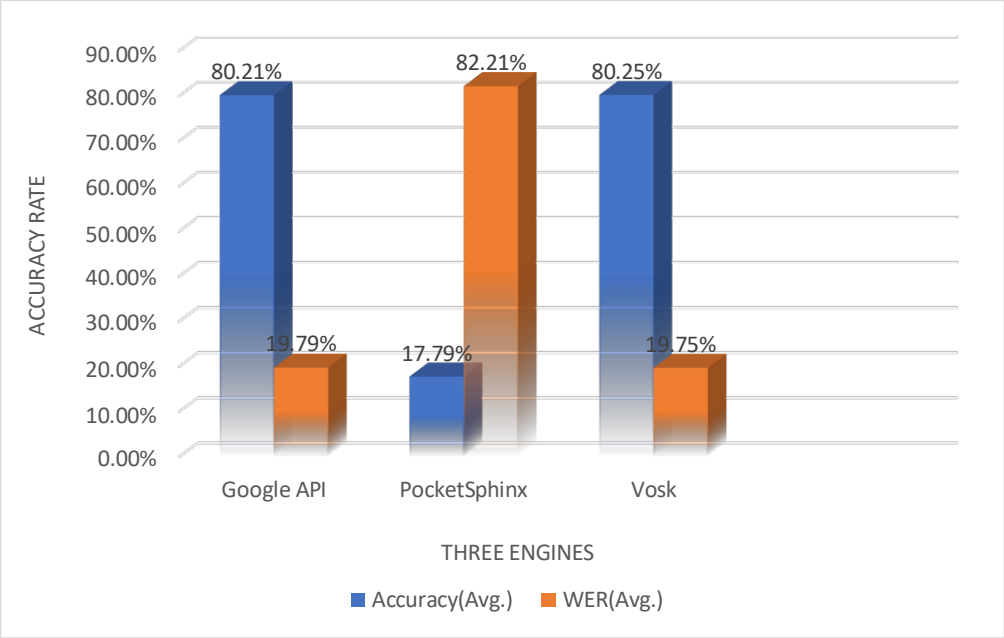


Figure 4.2: Accuracy and WER comparison of Google API, PocketSphinx, Vosk

From the above graph, it is obvious that Vosk is slightly better than Google API where there's no place for PocketSphinx because its Error Rate is very much higher than others.

4.8 Real Time Speech to Text Technique with Vosk

When the final model is selected in this phase, the project will be configured and set up to catch speech correctly in noise and execute commands appropriately. “KaldiRecognizer()” has been used as a built-in function of Vosk to recognize the process. The equation is given below:

rec = vosk.KaldiRecognizer(model, args.samplerate, phrases)

The below figure 4.3 shows how fast command can run on Google API, PocketSphinx and Vosk in 50dB to 57dB:

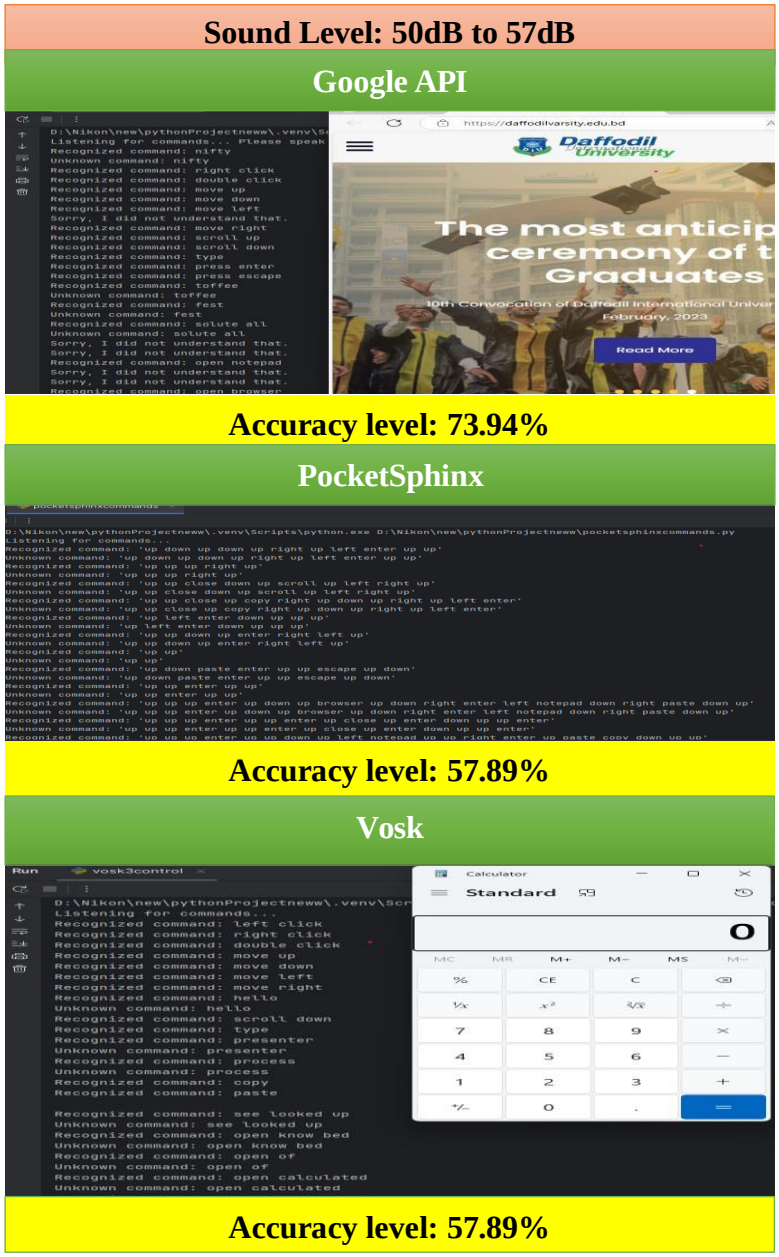


Figure 4.3: Output of 50dB to 57dB Sound Level of Google API, PocketSphinx, Vosk

The figure shows Vosk's accuracy level during high to higher sound level. When sound is 50dB to 57dB, the accuracy is 73.69% and 68.42% accuracy comes on 70dB to 87dB sound level.

4.9 Sound Level Detector

In figure 4.4, it shows an application has been used to measure the accuracy level of command during different noise levels. A “dB” meter application has been used where noise level is taken in between 50 to 57 dB and 70 to 87dB. A demo figure is given below of measurement.

Figure 4.4: Different Noise Level

Here, two distinctive noise levels have been taken for comparing the better accuracy.

4.10 Command Capture at Noise Level 50dB to 57dB

The noise level has been chosen 50dB to 57dB at the very beginning where dB defines decibels. This noise level is throughout to be normal for use in an office or at home. It is mentioned that automated operation level has a beatable noise level shown below on table 4.6.

Table 4.6: Audio Instructions on 50dB to 57dB Sound Level

Instructions	Google API	PocketSphinx	Vosk
Left click	Non-Execute	Execute	Execute
Right click	Execute	Execute	Execute
Double click	Execute	Non-Execute	Execute
Move up	Execute	Execute	Execute
Move down	Execute	Non-Execute	Execute
Move left	Execute	Execute	Non-Execute
Move right	Execute	Execute	Execute
Scroll up	Execute	Non-Execute	Non-Execute
Scroll down	Execute	Non-Execute	Execute
Type	Execute	Execute	Execute
Press enter	Execute	Execute	Non-Execute
Press escape	Execute	Non-Execute	Non-Execute
Copy	Non-Execute	Execute	Execute
Paste	Execute	Non-Execute	Execute
Select all	Non-Execute	Execute	Execute
Close window	Execute	Execute	Execute
Open notepad	Execute	Non-Execute	Non-Execute
Open calculator	Non-Execute	Execute	Execute
Open browser	Execute	Non-Execute	Execute
Accuracy	78.94%	57.89%	73.69%

The above table 4.6 shows that Vosk and Google API are giving almost the same accuracy rate where PocketSphinx is worse than others. The accuracy of PocketSphinx is 57.89% which is much lesser. Vosk and Google API get activated when any voice wave is detected, and its accuracy is much higher than others. Vosk's accuracy level is 73.69% and Google API gives 73.94% accuracy. It is very clear that there is a 0.25% difference between the accuracy level of Vosk and Google API.

Figure 4.5 will give a clearer view of these three platforms' accuracy level.

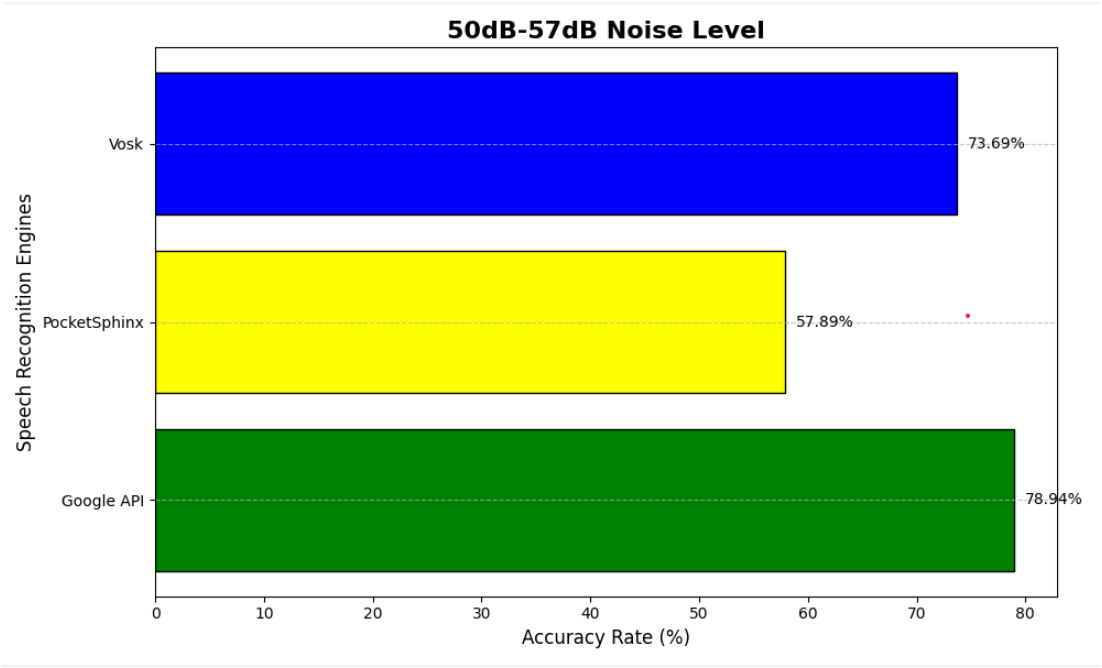


Figure 4.5: Graph of Accuracy Rate on 50dB to 57dB Noise Level

4.11 Command Capture at Noise level 70dB to 87dB

This time noise level has been increased more than before. On table 4.7, 70dB to 87dB are identified louder than normal environment. Some of the machines have been used to increase noise such as loudspeakers. The experiment has been done near a place where construction work is going on like there is noise of mixture machine, drill machine, tiles cutting machine, transport noise etc. Some of these machines have been brought purposely to the experiment. Table 4.7 represents a comprehensive breakdown of the accuracy scores observed under the conditions of heightened noise levels. It highlights how the increased noise has impacted the performance metrics, offering specific data points and analysis for better understanding. Each entry illustrates the relationship between the varying degrees of noise and the resulting accuracy, allowing for a clear assessment of the influence that noise has on performance outcomes.

Table 4.7: Audio Instructions on 70dB to 87dB

Instructions	Google API	PocketSphinx	Vosk
Left click	Non-Execute	Execute	Execute
Right click	Non-Execute	Execute	Execute
Double click	Non-Execute	Non-Execute	Non-Execute
Move up	Execute	Execute	Execute
Move down	Execute	Non-Execute	Execute
Move left	Execute	Execute	Non-Execute
Move right	Execute	Execute	Execute
Scroll up	Execute	Non-Execute	Non-Execute
Scroll down	Execute	Non-Execute	Execute
Type	Non-Execute	Execute	Execute
Press enter	Execute	Non-Execute	Non-Execute
Press escape	Execute	Non-Execute	Non-Execute
Copy	Non-Execute	Execute	Execute
Paste	Non-Execute	Non-Execute	Execute
Select all	Non-Execute	Non-Execute	Execute
Close window	Execute	Non-Execute	Execute
Open notepad	Execute	Non-Execute	Non-Execute
Open calculator	Execute	Non-Execute	Execute
Open browser	Execute	Non-Execute	Execute
Accuracy	63.16%	36.84%	68.42%

From the above 4.7 table, it is pretty much obvious that Google API and Vosk are performing well than PocketSphinx. The graph shows how Google API, Vosk, and PocketSphinx perform in recognizing speech within loud environments, specifically between 70dB to 87dB. Vosk leads with an accuracy of 68.42%, making it the best choice, especially for offline use and in noisy settings. Google API comes close with 63.16% accuracy, which is solid and generally effective, especially with its advanced cloud capabilities that can handle various accents and languages, though it relies on internet connectivity. PocketSphinx, with its 36.84% accuracy, doesn't quite match up to the other two platforms, but it's lightweight and designed for devices with limited

resources, making it handy in low-power situations. It could benefit from some extra noise reduction to improve its accuracy, though. For handling loud environments, Vosk is the most reliable choice.

By figure 4.6, it will be clearer which platforms perform better. The graph is given below:

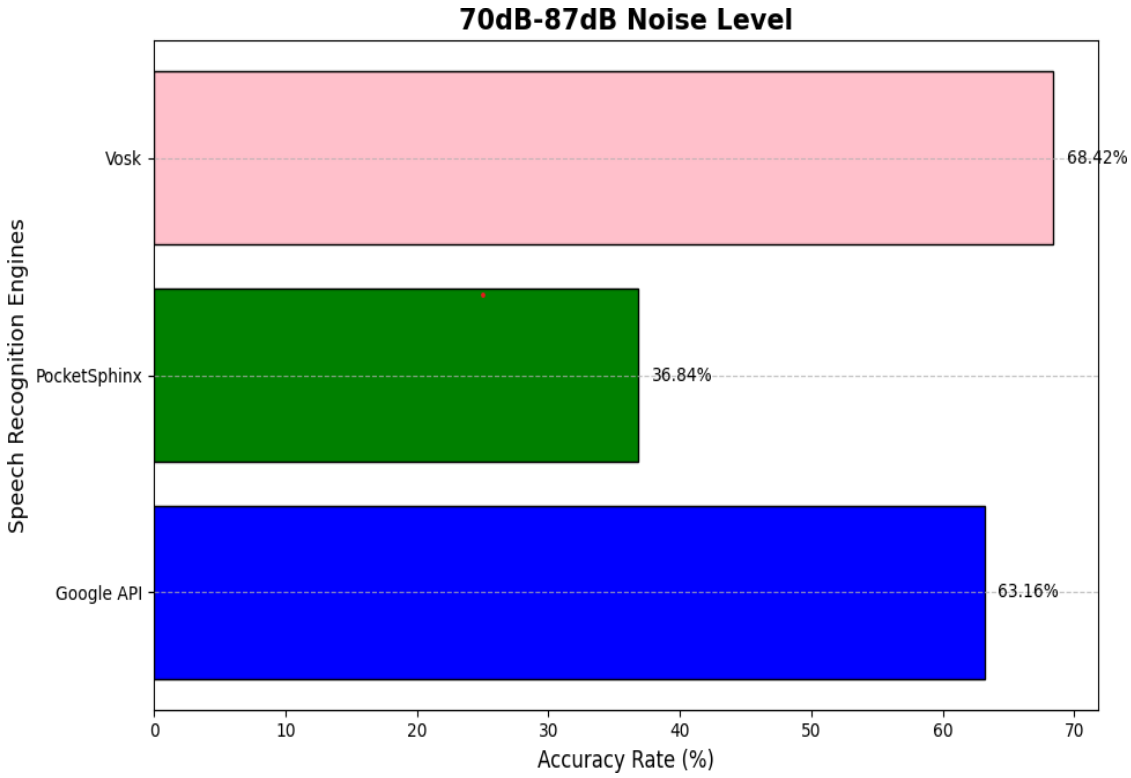


Figure 4.6: Graph of Accuracy Rate on 70dB to 87dB Noise Level

The below figure 4.7 shows how fast command can run on Google API, PocketSphinx and Vosk in 70dB to 87dB:

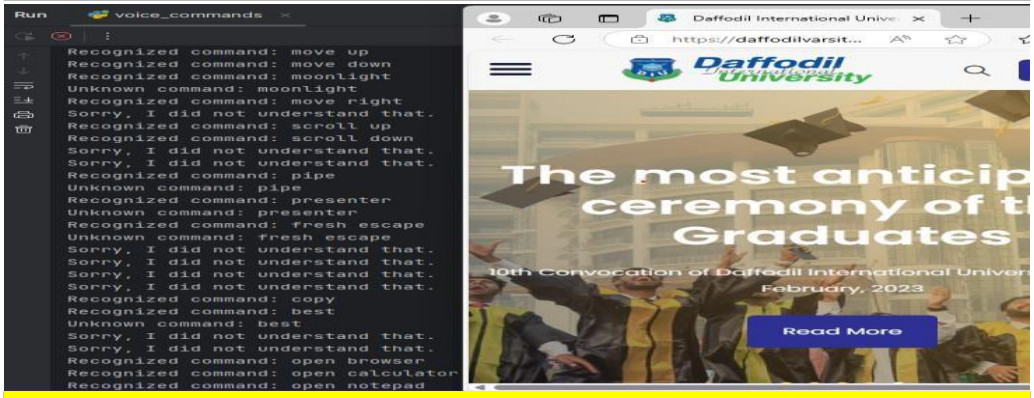
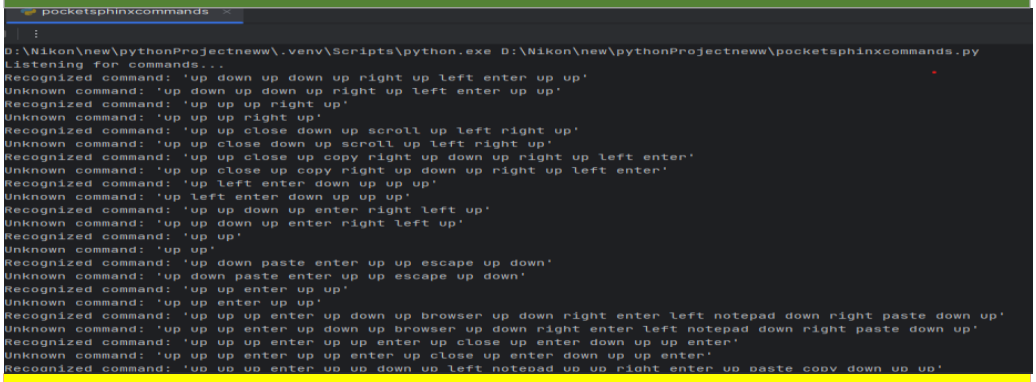
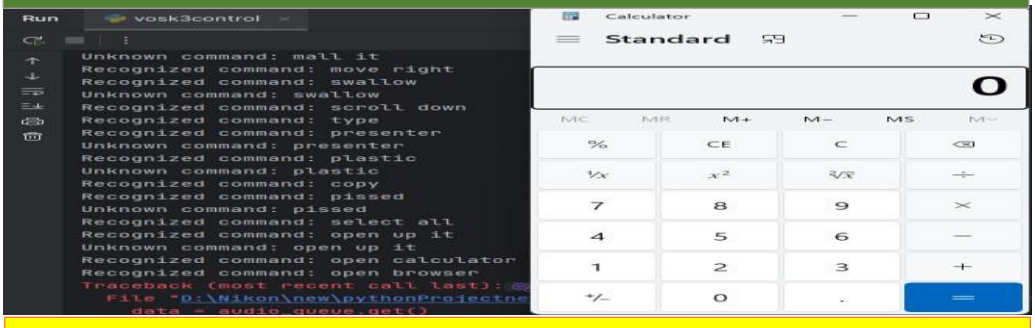
Sound Level: 70dB to 87dB
Google API

Accuracy Level: 63.16%
PocketSphinx

Accuracy Level: 36.84%
Vosk

Accuracy Level: 68.42%

Figure 4.7: Output of 70dB to 87dB Sound Level of Google API, PocketSphinx, Vosk

4.12 Result Analysis

In the above experiment, there are three types of platforms which are compared in two different noise levels. The whole experiment is on the above table and graph. The graph provides a simple, unambiguous view of how well each model performs under various noise levels. It illustrates which models remain correct and which may fail by providing a sense of how effectively each model adjusts to these fluctuating noise levels. We can rapidly understand the dependability of any model in difficult, chaotic situations, thanks to this visual picture. This clear snapshot gives a helpful overview of each model's ability to stay reliable under noisy conditions. This method not only gives the data additional depth but also increases their applicability in real-world scenarios with different background noise levels.

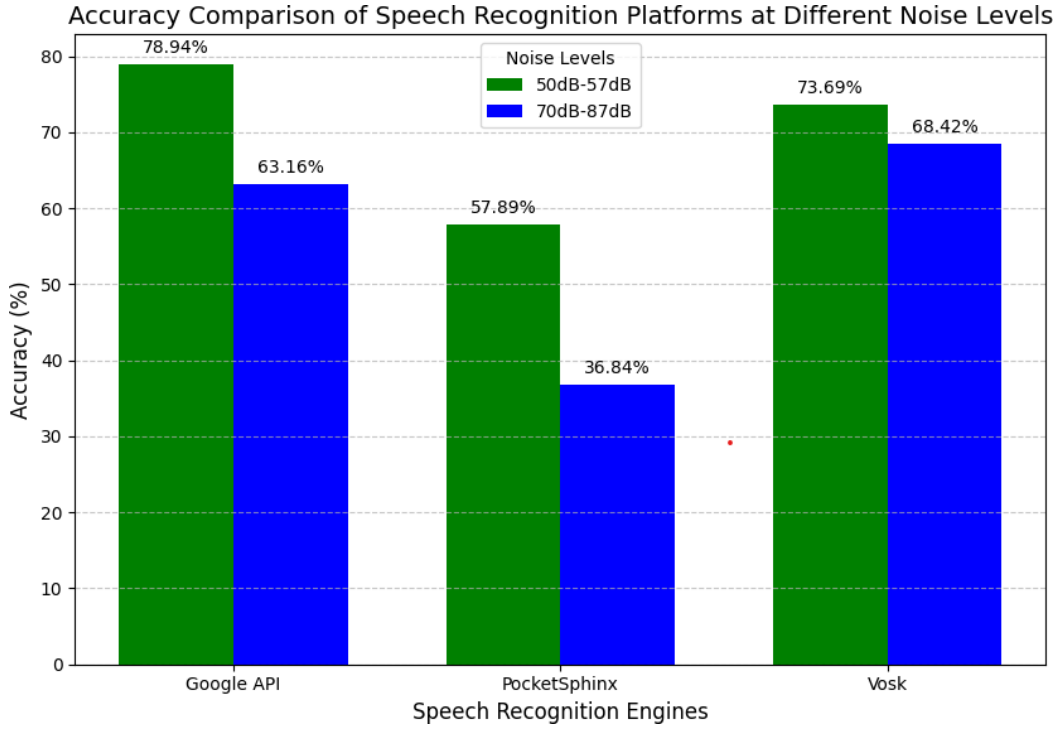


Figure 4.8: Accuracy Rate Comparison of Three Engines with Different Noise Levels

According to the upper table 4.8, it defines that green and blue columns indicate two levels of noise. Green defines 50dB to 57dB and blue indicates 70dB to 87dB noise level where Google API shows 73.94% accuracy in the level of 50dB to 57dB and 63.16% in 70dB to 87dB. On the other hand, Vosk's optimization is 73.69% on 50dB

to 57dB and 68.42% in 70dB to 87dB which are not that much difference between Google API and Vosk but PocketSphinx's optimization is very lower and declining for better performance. PocketSphinx is constantly providing lower accuracy.

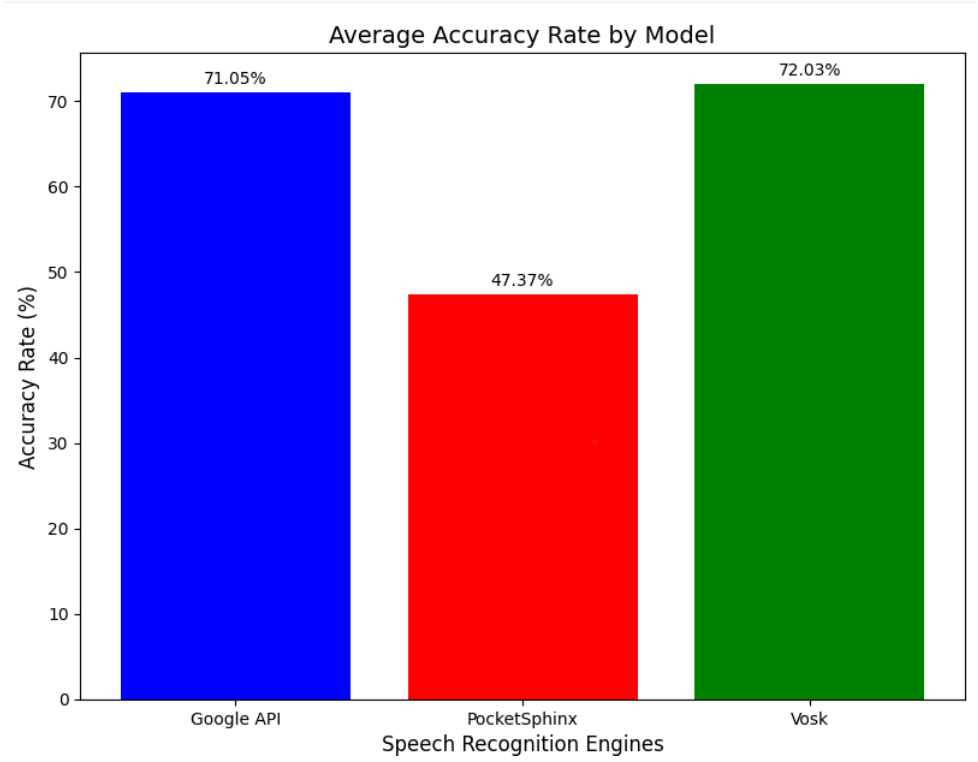


Figure 4.9: Average Accuracy Rate of Three Speech Engines

From the above graph 4.9, we get to know that, with strong precision and good noise management, Vosk outperforms the other two systems. Although it can't quite equal Vosk's performance, Google API is at least as competitive as Vosk. Contrarily, PocketSphinx performs less well in loud environments, making it less suitable for situations when accuracy is crucial. Overall, Vosk is the most superior in fighting with challenging environments.

CHAPTER 5

RESULT ANALYSIS BASED ON REAL TIME DATA

5.1 Experiment with the Data of Different Ages

Prior testing has shown that VOSK is the most accurate and appropriate choice for a variety of automation jobs. This conclusion is based on the findings of those experiments. Even though the researcher evaluated each program by utilizing voice commands, a comprehensive review must include a broad variety of ages and voices in order to accurately represent how the applications are used in the real world. The VOSK platform makes use of a clear and adaptable paradigm that is capable of swiftly correcting any faults that may occur in voice recognition.

5.2 Result Analysis

Because it doesn't need internet connectivity to utilize its voice recognition capability, this VOSK software is very adaptable and is made to run on a variety of systems. Fourteen individuals, including three children, five women, and six men, recently assisted in testing it. To test the app's accuracy, I gave voice commands at noise levels ranging from 50–57 dB to 70–87 dB to check how well it could understand instructions in different loud settings. Accuracy is just one aspect of this extensive testing; another is demonstrating how effectively the software can identify instructions in the presence of background noise. Making sure the VOSK app functions as effectively in actual circumstances as opposed to simply ideal ones is the goal.

Table 5.1 will portray 14 people's accuracy level properly and define “Yes” as executive command and “No” as non-executive command.

Table 5.1: Audio Testing of Vosk with Different Ages (50-57) dB

	Command	M 75y	M 60y	M 47y	M 39y	M 43y	M 20y	F 70y	F 50y	F 47y	F 39y	F 25y	C 14y	C 9y	C 5y
1	Left click	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
2	Right click	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3	Double click	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
4	Move up	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes
5	Move down	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	Move left	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
7	Move right	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8	Copy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
9	Paste	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
10	Scroll up	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
11	Scroll down	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
12	Close window	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13	Open Calculator	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
14	Open notepad	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
15	Open browser	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
	Accuracy	86.7 %	80%	100 %	94 %	100 %	100 %	86.6 %	86.6 %	100 %	93.3 %	100 %	100 %	80 %	73.3 %

The above chart 5.1 is defining “M”, “F”, “C” which means consecutively male, female and child of different ages who are involved in the experiment of sound accuracy test. Cross refers to the wrong pronunciation and Tik mark refers to right pronunciation. A total of 15 data have been taken. A bar chart is given to define the chart's proper result.

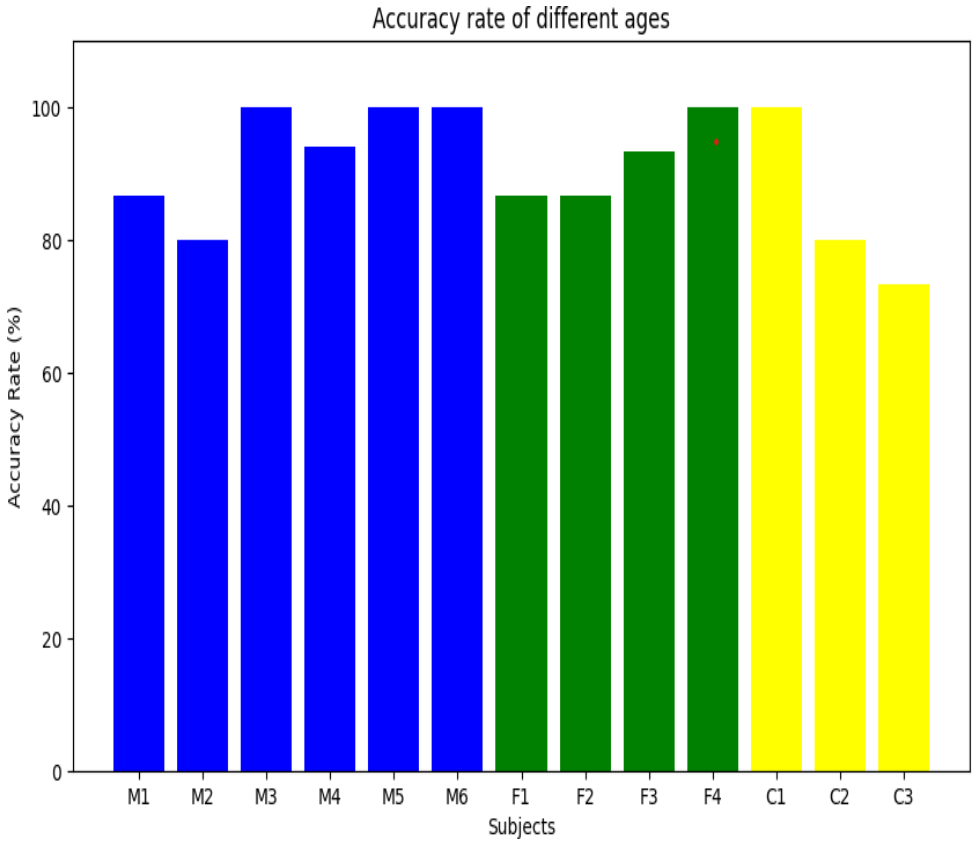


Figure 5.1: Accuracy Rate of Different Age Groups on 50dB-57dB Sound Level

From the above figure 5.1, I find M3, M5, M6 are showing almost similar accuracy and M2 is the lowest of M group. On the other hand, F4 shows the highest value from F group. F1 and F2 are the lowest. From the C group, C1 is defining the highest accuracy and C3 is giving the lowest rate. Overall, the graph shows M3, M5, M6, F4 and C1 are the most increased value in total data. M2, F1, F2 and C3 are the most decreased value. The whole accuracy rate is belonging to 50dB to 57dB noise level.

Table 5.2: Audio Testing of Vosk with Different Ages (70-87) dB

	Command	M 75y	M 60y	M 47y	M 39y	M 43y	M 20y	F 70y	F 50y	F 47y	F 39y	F 25y	C 14y	C 9y	C 5y
1	Left click	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
2	Right click	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3	Double click	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
4	Move up	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes
5	Move down	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	Move left	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
7	Move right	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8	Copy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
9	Paste	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
10	Scroll up	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
11	Scroll down	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
12	Close window	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13	Open Calculator	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
14	Open notepad	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
15	Open browser	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
	Accuracy	86.7 %	80 %	93.4 %	100 %	100 %	100 %	81 %	86.6 %	100 %	93.3 %	100 %	100 %	80 %	73.3 %

It will be easier to understand the accuracy rate properly by a graph. The graph is given below.

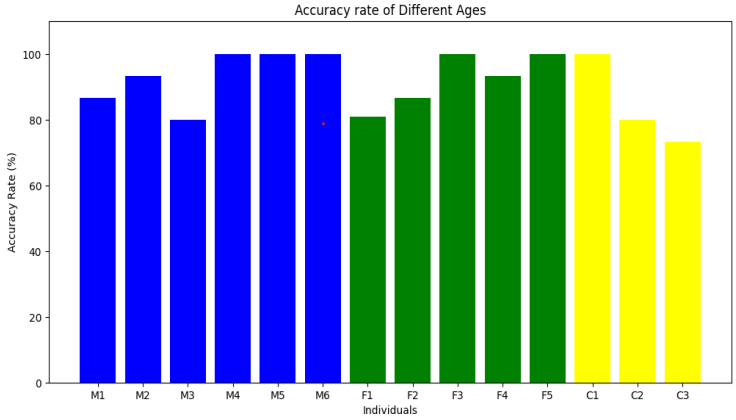


Figure 5.2: Accuracy Rate of Different Age Groups on 70dB-87dB Sound Level

From the above chart 5.2, it is visible that M3, F1 and C3 are showing the lowest value. M4, M5, M6, F3, F5 and C1 present the highest value. Overall, the graph has slightly changed from the previous graph which was based on 50 to 57dB.

5.3 CPU Utilization

CPU utilization means how much computer uses during the engines running.

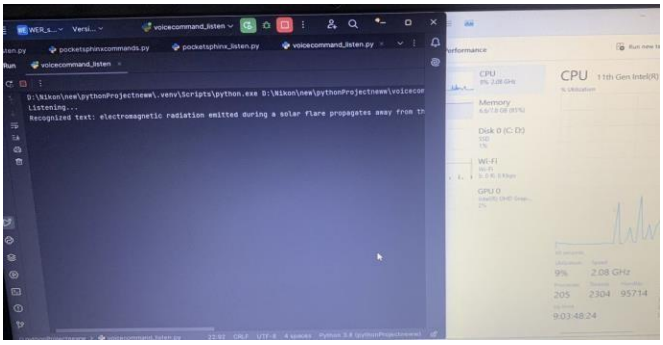


Figure 5.3: CPU Utilization with Google API

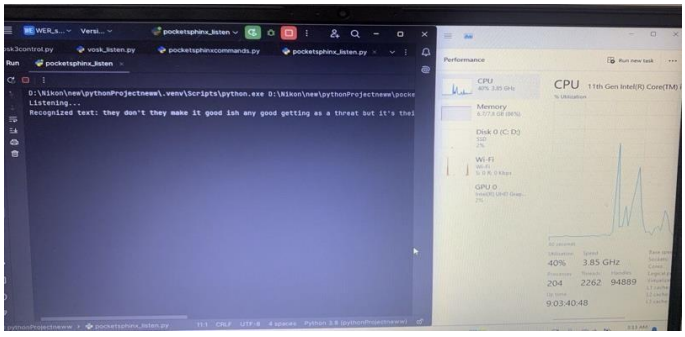


Figure 5.4: CPU Utilization with PocketSphinx

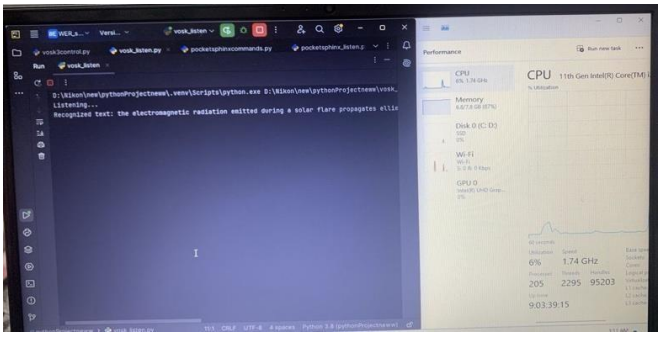


Figure 5.5: CPU Utilization with Vosk

From the figure 5.5, CPU is used 6% during recognizing speech on Vosk. On the other hand, 40% of CPU is needed for executing program on Pocketsphinx in figure 5.4 and 9% working process need in CPU for Google API in figure 5.3.

5.4 Comparative Analysis

The data which had been taken to compare with my research is given below:

Table: Number of Recognized Words

Group	Vosk Big	Vosk Small	Google STT	Total
General	1520 (82.83%)	1448 (78.91%)	1593 (86.81%)	1835
Children (4y-18y)	383 (71.72%)	348 (65.17%)	419 (78.46%)	534
Females (26y-42y)	554 (83.94%)	541 (81.97%)	592 (89.70%)	660
Males (26y-40y)	583 (90.95%)	559 (87.21%)	582 (90.80%)	641

Table: Number of Not Recognized Words

Group	Vosk Big	Vosk Small	Google STT	Total
General	315 (17.17%)	387 (21.09%)	242 (13.19%)	1835
Children (4y-18y)	151 (28.28%)	186 (34.83%)	115 (21.54%)	534
Females (26y-42y)	109 (16.06%)	119 (18.03%)	68 (10.30%)	660
Males (26y-40y)	58 (9.05%)	82 (12.79%)	59 (9.20%)	641

Figure 5.6: Data of Other Paper

The comparison between this research and other paper is given below through a graph:

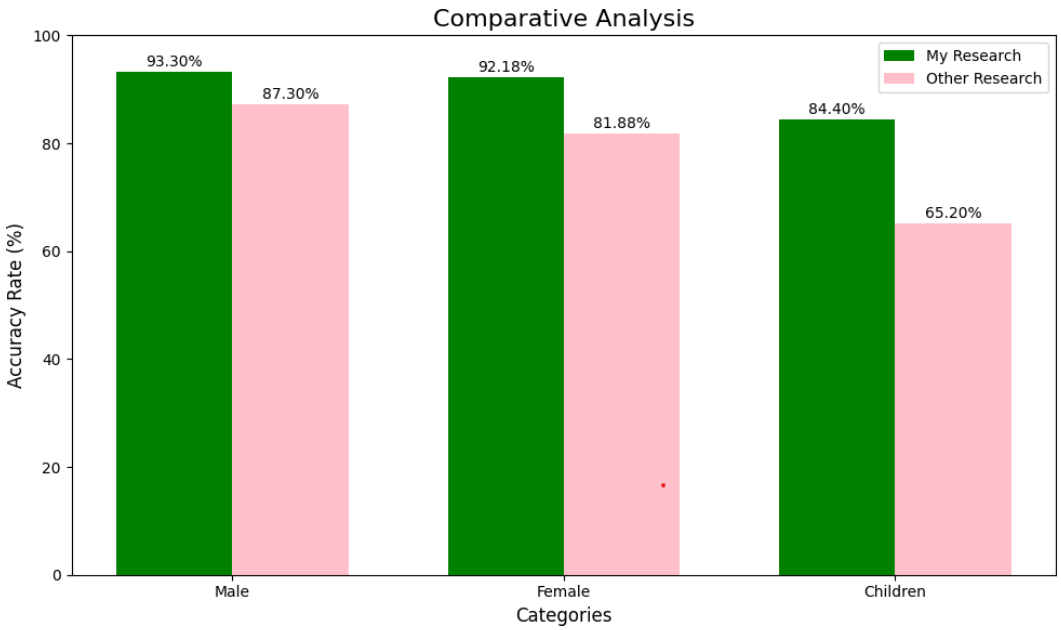


Figure 5.7: Comparative Analysis between My Research and Other Paper

In figure 5.6, the comparison of accuracy rate between this research and other paper is showing tremendously. In this research, the accuracy rate of Male is 93.3% whereas other paper has 87.3%. Female and children's accuracy rates are 92.18% and 84.4% but on other papers the rates are consecutively 81.88% and 65.2%. So, the distinction between them is massive.

5.5 Discussion

This investigation provides the biggest advantage to those rural people who are unable to reach the modern network and unaware of this. It is fruitful for disable who face difficulties in operating online devices. The main target is proving the better performance of offline platforms due to having online platform's performance. Proper evaluation, accessibility, quality, features and requirements have been provided for earning the acceptance of humans. Robust validation of the model is crucial for offline models that need significant testing to assure reliability in the actual world.

CHAPTER 6

SUMMERY, CONCLUSION, LIMITATIONS AND FUTURE WORK

6.1 Research Summery

The feasibility, adaptability and potential of online and offline ASR platforms' evaluation is the moral investigation of this research. This study is compatible for the given accessibility for unable and unaware people. Three engines were used to identify better accuracy. These models were shown to be usable offline without an internet connection after extensive testing, which included accuracy evaluations and trialed with multiple participants. The goal of the investigation is to introduce the acknowledgment of ASR in real world and better companionship of human with easy connectivity to the technology.

6.2 Conclusion

Performance of ASR depends on accuracy and speed, but accuracy depends on the vocabulary size and various types of engines and their language models. The original target of this research is to compare the flexibility of online and offline voice recognition platforms by using different engines and language models. The accuracy test occurred with real time data with multiple testing methods. In this paper, I have compared three well-known voice recognition engines including Google API, PocketSphinx and Vosk. Kaldi has been used in terms of recognizing audio and inference time. Comparison between Google API, PocketSphinx and Vosk has proven that Google API and Vosk both engines achieve similar results, but Vosk is a little bit better than Google API in terms of accuracy rate. In the future, I look forward to improving the scalability and flexibility of human computer interaction and multilingual integration. Need to ensure the time utilization of computers is less with less WER. More adaptability and compatibility can be increased between human-computer.

6.3 Limitations

This research has some limitations that there are specific language models in offline voice recognition system which are possibly applied with these engines, but all languages models are not applicable to execute. As it is an offline automation system, the application needs to be up to date which cannot be possible without internet connectivity. On the other hand, online systems cannot run out without network connectivity. So, the remote area where internet connection is rarely possible is difficult to operate device through online. The people who are unaware about digital network and disable might be in trouble in operating systems.

6.4 Future Work

This future plan of this study is to use these engines in various forms of human-device interaction by adding more languages to identify different types of emotions. Ensuring privacy in terms of using real time data and handling data are another focus. The invention of speech-controlled devices that may control by human instruction can be indispensable creation which may easily meet some of the real-world needs.

REFERENCES

[1] D.J. Vajpai and A. Bora, Industrial Applications of Automatic Speech Recognition Systems, Int. Journal of Engineering Research and Applications, 2016, 6 (3), 88-95.

[2] H.Tabani, J.M. Arnau, J. Tubella, A. Gonzalez, Performance analysis and optimization of automatic speech recognition, . IEEE Transactions on Multi-Scale Computing 46 Systems, 2017, 4(4), 847-860.

[3] A. V. Anand, P. Devi, J. Stephen and B. V. K, "Malayalam Speech Recognition System and Its Application for visually impaired people," IEEE, 2012.

[4] G. Anumanchipalli, R. Chittur, S. Joshi, R. Kumar and S. P. Singh, Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems, International Conference on Speech and Computer (SPECOM) Proceedings, Hyderabad, India, 2005.

[5] D. Vazhenina and K. Markov, "Recent Developments in the Russian Speech Recognition Technology," in IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS), 2010.

[6] H. Prakoso, R. Ferdiana, R. Hartanto, Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset, International Symposium on Electronics and Smart Devices (ISESD) 2016, 29-30.

[7] B. Wu, K. Li, M. Yang, and C.-H. Lee, A reverberation-time-aware approach to speech dereverberation based on deep neural networks, IEEE/ACM Trans. Audio, Speech, Language Process., 2017, 25 (1), 102–111.

[8] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix and T. Nakatani, Far-Field Automatic Speech Recognition, in Proceedings of the IEEE, 2021, 109 (2), 124-148.

[9] Kumar, T., Mahrishi, M., Meena, G. (2022). A Comprehensive Review of Recent Automatic Speech Summarization and Keyword Identification Techniques. In: Fernandes, S.L., Sharma,

T.K. (eds) Artificial Intelligence in Industrial Applications. Learning and Analytics in Intelligent Systems, vol 25. Springer, Cham

[10] D. A. Fraihat, Y. Sharrab, F. Alzyoud, A. Qahmash, M. Tarawneh, A. Maaita, Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments, Human-centric Computing and Information Sciences, 2024, 14:15.

[11] H. Kheddar, Y. Himeur, S. A. Maadeed, A. Amira, F. Bensaali, Deep transfer learning for automatic speech recognition: Towards better generalization, Knowledge-Based Systems,2023, 277, 110851.

[12] S. McCrocklin, A. Humaidan, E. Edalatishams, ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference, ISSN 2380-9566, Ames, IA, September 2018, 191-200.

[13] S. P. Panda, Automated speech recognition system in advancement of human-computer interaction, International Conference on Computing Methodologies and Communication (ICCMC), 2017, 302-306.

[14] S. Xue, X. Y. Kou, S. T. Tan, Natural Voice-Enabled CAD: Modeling via Natural Discourse, Computer Aided Design and Applications, 2009, 6(1), 125-136.

[15] H. Garg, S. Solanki, S. Verma, Automation and Presentation of Word Document Using Speech Recognition, 2020 International Conference for Emerging Technology (INCET), 2020, 1-5.

[16] L. Matsane, A. Jadhav, R. Ajoodha, The use of Automatic Speech Recognition in Education for Identifying Attitudes of the Speakers, 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, 2020, 1-7.

[17] N. A. S. Wijaya, A. Marduama, A. K. Nisa, A. N. Fahmi, The Implementation of Automated Speech Recognition (ASR) in ELT Classroom: A Systematic Literature Review from 2012-2023, 2023, 7 (7), 816-828.

[18] T. Y. Ahn, S. M. Lee, User experience of a mobile speaking application with automatic speech recognition for EFL learning, British Journal of Educational Technology, 2015, 47(4), 778–786

[19] M. Bashori, H. R. Van, H. Strik, C. Cucchiarini, Look, I can speak correctly: learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. Computer Assisted Language Learning, 2022, 1–29.

[20] R. R. Sehgal, S. Agarwal and G. Raj, Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems, International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, 213-218.

oyshi11

ORIGINALITY REPORT

20%	18%	10%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	en.wikipedia.org Internet Source	3%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	Submitted to Higher Education Commission Pakistan Student Paper	1%
5	Hamdan Prakoso, Ridi Ferdiana, Rudy Hartanto. "Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset", 2016 International Symposium on Electronics and Smart Devices (ISESD), 2016 Publication	1%
6	Dan Nickolai, Emma Schaefer, Paula Figueroa. "Aggregating the evidence of automatic speech recognition research claims in CALL", System, 2024 Publication	1%