Yare Speech Identification System (YSIS) to Enhance the Performance of Speech Recognition System

Shifa Fatma[1], Prof. Pankaj Raghuvanshi[2]
[1]Dept of CSE
[2]HOD, Dept of CSE
[1, 2] Alpine Institute of Technology, Ujjain

Abstract- *Yare Speech Identification System (YSIS) also known as Automatic Speech Recognition System or Voice Recognition System. YSIS is an assistant based software system which means understanding the user's voice by the system and performing any type of the task. The YSIS system we have developed is a desktop-based on python modules and libraries. The aims of this research to improve existing features of speech recognition system or adding some new features to meet the evolving needs of the users. It can help people with a variety of disabilities and also useful for people with physical disabilities who often find typing difficulties or help those with spelling difficulties, including users with dyslexia, because recognized words are almost always correctly spelled. Even blind people who couldn't see the machine can interact with it using their voice only and this will also reduce the work using the keyboard, mouse, and other input devices. YSIS system can be found in various devices, including cars, smartwatches smartphones, smart speakers, laptop, and even household appliances. They can be activated by voice commands or specific trigger phrases, allowing users to engage in hands-free and conversational interactions. These assistants have become integral parts of our daily lives, offering personalized, user-friendly and convenient assistance across various tasks and domains.*

Keywords- AI based Automatic Speech Recognition, API, Personal Computing, Python, Yare Speech Identification System (YSIS).

I. INTRODUCTION

Today's generation of speech recognition products are more affordable and user-friendly than ever before. The system helps illiterate and disabled people to perform their day-to-day tasks like displaying climate reviews, getting updates of mail, web scraping, on off devices, answering phone calls, reading newspaper, schedule appointments, play music and so on. The system makes our life easier and saves time. These system require less initial training than their predecessors and typically offer much improved accuracy. For these reasons, more and more people with special needs are considering speech recognition as an alternate method for computer access. With the use of this system, we can automate the task easily, just give the input to the system in the speech form and all the tasks will be done by it from converting your speech into text form to taking out keywords from that text and execute the query to give results to the user [1]. The most well-known automatic speech recognition techniques which are existing system in the real world called Apple Siri, Cortana, Amazon Alexa and Google voice so all of their front ends are ASR engines and so on [2,10]. Users may use for voice commands to ask their assistants questions, control home automation devices, to-do lists, making payments, send messages to anyone on WhatsApp, send SMS, reading the newspaper, getting weather updates, remainder set, automate YouTube and Chrome, set alarm, and so on[3,4]. Some research is based on visually impaired in recognizing text on real objects and provide audio feedback towards the user in real-time [5]. Some is based on voice user interface for personal computers with the help of framework to overcome the problem in integrating multiple application software APIs [6]. Automating the applications such as MS-Word, MS-PowerPoint and MS-Excel that manage documents and presentations via voice commands using python programming language [7]. Voice control system implemented by Raspberry Pi, open API and AI [8]. User to send emails seamlessly through voice command and reading email [9].

The drawback of previous (research) system is that it only works on limited functionalities such as telling jokes, send email, read email, automate document and presentation, web scrapping, news portal reading, weather forecasting, performing different calculations and so on. For this reason, we are adding some features of our system like automate website actions, automate browsing, booking tickets, online shopping, WhatsApp message, email sending and receiving, dictating, reading pdf, schedule meetings, controls the telephone calls, manages the personal activities through calendar and so on. In my project work I have chosen Python programming language with huge libraries and API's to build the AI-based Voice assistant. Our software always repeats the command that the user enters into the system, so that the user knows if they entered the correct command or not. On the other hand, it continues to listen and respond to the user's needs until the user decides to stop. The main goal of this system is improve the intelligibility of speech, enhance performance and efficiency.

II. METHODOLOGY

The idea is can we use this unlabeled data which is lying on the internet. But the fact the problem is that we don't have the transcripts for them, we don't have the labeling for them. But can we use just the audio data to make our models better. So that is the main inspiration and it will definitely reduce the fine tuning requirements. The main goal is to build these systems which can thrive on this unlabeled data and possibly make the fine tuning better.

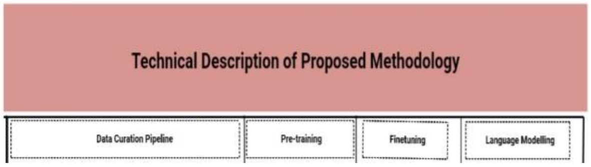


Fig. 1. Methodology of YSIS

A. *Data curation pipeline*

This data curation pipeline is a general pipeline that anyone can follow for any particular video or audio clip.So we collect unlabeled data associated it's completely raw audio but clean. In this stage, we use a clean phone sequence translated from raw data and parameters for learning the mapping between phone sequence and words.

Terms used in Data Curation Pipeline

1) **YouTube-dl:** YouTube-dl is open source program written in python and requires python interpreter to run this program. It is a free download manager for videos and audio from YouTube, Facebook, Dailymotion, Twitter and many more websites. It should run on any platform Windows, Linux and Mac.
2) **FFMPEG:** FFMPEGstands for Fast Forward Moving Picture Experts Group. It is open source software project that offers many tools for processing of audio and video files. It is platform independent i.e. Windows, Linux, and Mac.
3) **Py-webrtcvad:** RTC stands for Real-Time Communication and VAD stands for Voice Activity Detection. py-webrtcvad used to detect changes in speech audio patterns to classify a piece audio data as voiced or unvoiced. It can be useful for telephony and speech recognition free of charge via Python. The detection as a used to trigger a process.
4) **Wada.SNR:** SNR stands for signal-to-noise ratio and WADA stands for Waveform Amplitude Distribution Analysis. It's based on statistical information obtained from the amplitude distribution of a speech waveform and used to measure the level of the background noise present in a original speech signal. WADA SNR working on the assumptions of Gaussian noise and clean speech is characterized by a Gamma with a fixed shaping parameter.

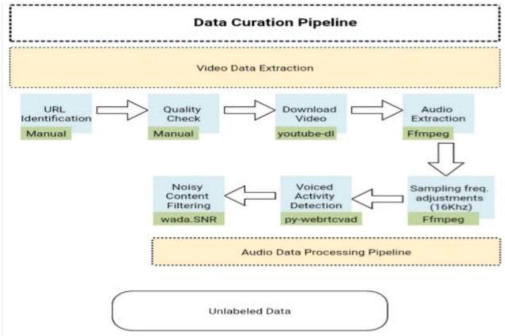


Fig.2. Performing Functions On Data Curation Pipeline

B. *Pre-training*

Now we will use Wav2vec2 to directly predict words from the audio waveform or raw data and we keep throwing at it the unlabeled data and make the model to learn some representation of the audio. Now you are just making encoder waits to learn something of the representation of that particular language. Basically is a learning process for unlabeled dataaccessible from everywhere. Also used for remove error in phone sequence or noisy data by perform augmentation method that is Replacing, Insertion, Deletion. Replacing the select phones with others, deleting the select phones and Insertion randomly sampled phones. In this stage, we are able to deploy a contrastive loss during the generation process. Using contrastive loss tries to minimize the distance between similar data representation and maximize distance between dissimilar data representation.

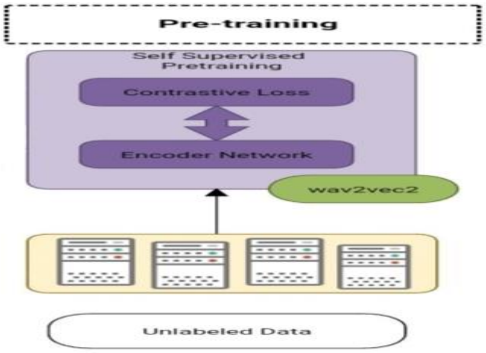


Fig.3. Performing Functions On Pre-training

C. Fine-tune

Fine tune trained on new data. Now you take pre-trained model and then add a top layer of projection or a specific language and then train the model using label data with the Connectionist Temporal Classification (CTC). So it will help in transfer learning as well as maximize that probability of valid character sequences and minimizes that probability of generating any invalid sequence. It will kind of reduce your fine tuning requirements and reduce the number of hours that you require the model to perform better.

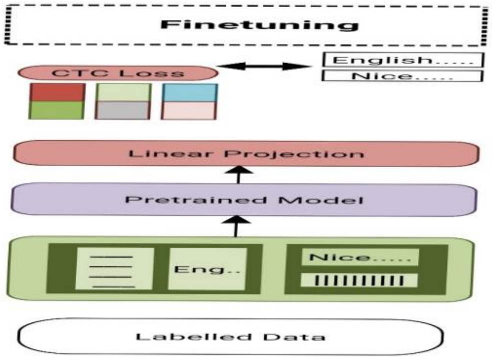


Fig.4. Performing Functions On Fine Tuning

D. Language model

This model trained on atext corpus to estimate how likely a given sentence or to predict the next word and so on. There are lots of sentences that you won't ever have seen before, so it has to be able to deal with those unseen word sequences. Back-off for unseen word sequences.Helps in mitigating errors, the spelling mistakes and the sentences incomplete at all which is taken care by this language modeling. In this stage, use CTC decoding algorithm for compress transcription to its correct length by ignoring duplicate adjacent and blanks. It also give a best path with maximum probability at every time step.

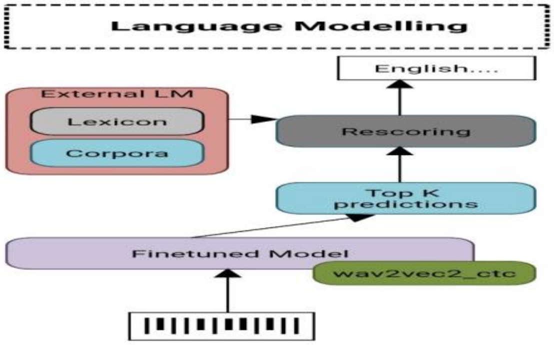


Fig.5. Performing Functions On Language Modelling

III. APPLICATIONS

A. Home Automation System

Speech Recognizer use of domestic appliances such as ovens, refrigerators, dishwashers and washing machines, lock systems, switches on and off.

B. Speech Biometric Recognition

Used to determine the stress status and Stress Control using a special kind of music.

C. Military

Speech recognizers have been operated in fighter aircraft i.e. set radio frequency, commanding an autopilot system, set the steer-point coordinates, weapons release parameters, training air traffic controllers and controlling flight displays.

D. Education

Helping students in language learning and improve reading skills.Teaching students of foreign languages to pronounce vocabulary correctly.

E. Serving the disabilities person

Speech recognition technology helps people with disabilities interact with computers more easily. Enabling people who are physically handicapped and unable to use a keyboard or mouse, can use their voices to navigate the computer and create documents.

IV. FUNCTIONS OF MY PERSONAL ASSISTANT

- Notify you as the current date, time, weather and word for the day.

- Notify you as a birthday remainder or drink water remainder.
- Tell the fact, jokes and quotes. Get latest news and show the news.
- Automate chrome, YouTube, Firefox and any web browser.
- It can find distance and direction of the route. You can also find the nearby restaurant, supermarket, Cafe.
- It can find the current location and IP address. Find the information of any phone number.
- It can search any words meaning you would like to find.
- Used for encode and decode a message.
- Empty recycle bin through the voice command.
- Make phone Calls through voice command. Sending WhatsApp messages and SMS.
- Play any music and movies on the system.
- Perform any type of calculations task and solving math's equations.
- Can open and close system files and applications.
- It can set the brightness of your monitor by a single command and change the background wallpaper of the desktop.
- Tell the Battery percentage and charge.
- It can also type for you in any application, file or document while you just have to dictate it and it will automatically type for you in the desired destination.
- It can write note or show note with date, used as a remainder.
- It can speak (read) any selected text. It can even copy, paste and select text for you.
- Record as a surveillance camera as a security purpose. Record your voice and video.
- It can take a photo for you, read pdf files, translate speech into any language, take screenshots and save it on your desktop, search images for you.
- Audio and video mute, unmute, volume up, volume down, play, pause, previous, next, skip automatically on system.
- Tell the condition of the system and internet connectivity.
- Teaching overseas students to pronounce English correctly.
  - ➢ Login any social networking site like Facebook, Instagram, telegram, LinkedIn,
  - ➢ You can also Locked, shutdown, restart of the computer.
  - ➢ Schedule meetings and appointments and make to do lists. Set alarms and timers.

ALGORITHM

Speech recognition transcription algorithm works

Start with a labeled set **S,** an unlabeled set **U** and a fixed **LM** trained on a separate text corpus.

1) Train **Mo** on **S** with SpecAugment. Set **M = Mo.**
2) Fuse **M** with **LM** and measure performance.
3) Generate labeled dataset **M (U)** with fused model.
4) Filter generated data **M (U)** to obtain **f (M(U)).**
5) Balance filtered data **f(M(U))** to obtain **b.f(M(U)).**
6) Mix dataset **b.f (M(U))** and **S**. Use mixed dataset to train new model Mo with SpecAugment.
7) Set **M = Mo** and go to 2.

A. Algorithm Description

Start with a labeled set of sentences **S** and we have an unlabeled set **U,** and we have an external language model that has been trained on a separate text corpus.

So we start with an initial model that we train on the labeled set and we use SpecAugment to already inject noise in that initial training step. SpecAugment, it's just a way of putting noise into the data.

Then we take the model that we have trained, the initial starting model, and we fuse it with the language model. And that model is sort of allowing itself to get advice from the language model when it's actually scoring sentences and doing its work.

Then we take this fused model and we take the entire unlabeled set and we attach labels. We have a prediction for it. And it goes transcribe all this stuff. And so now we have labels for all of the unlabeled data.

Then filter down that pseudo-labeled data to the ones that we're going to actually present to the student, that's the filtering step.

Now we take that pseudo-labeled data, filter data and make sure that it's balanced.

Now we've got filtered, balanced pseudo-labeled data that we merge in with the original labeled set **(S)** and, then we take a new model of the same variety that the initial one was and we train it on that larger sample and we use SpecAugment to inject noise.

And then we just go back up to the top, because student becomes the teacher and we keep iterating this process. And at each stage, we're sort of fattening up the training set and the model is getting better and we eventually at step two, we say, eventually performance isn't getting materially better and we stop. So that's basically the algorithm.

B. Terms Used in Algorithm

- **Noise injection** is by SpecAugment, a method that acts on the spectrogram of the input audio.
- **Shallow fusion** with a language model is used on the teacher network to generate better transcripts for the student network to train on.
- **Filtering** of transcripts is by fusion score and number of tokens. It is a measure of confidence that the model has in the labeling. It selects only the best ones that the teachers most confident.
- **Balancing** of utterance-transcript pairs is by token statistics. Based on just, how often the tokens from the token set actually show up.
- **Gradational filtering** is used to grow the semi-supervised dataset as the model performance improves.
- **Gradational augmentation** strength is increased with NST iterations.

V. FLOWCHART DIAGRAM

This is the flowchart of our YSIS assistant.

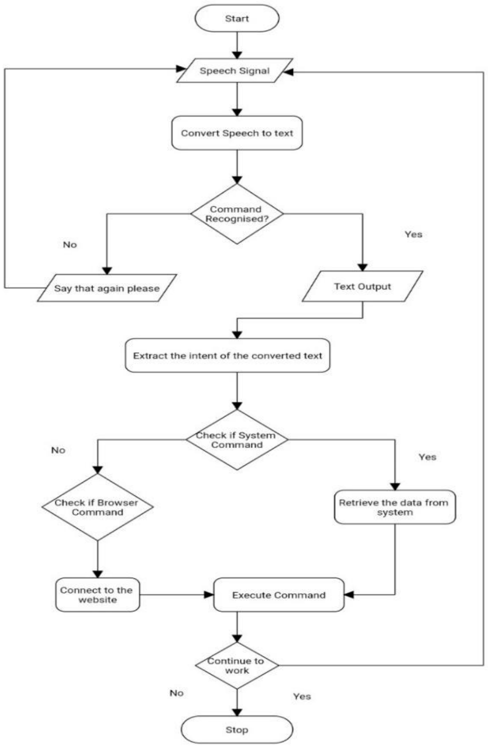


Fig. 6. Flowchart Diagram of YSIS System

- The system is activated by pressing the Start button. Once assistant is activated, it starts with greeting the user according to time, like good morning, good afternoon, good evening and also tells about the current day, time, temperature, word meaning for the day.
- After this, it will start listening to the user's voice command, through the microphone and converted into speech to text.
- If the voice command that is sent, valid a command which was saved in the database, then it will search for the particular function whether it is present in the main body or not.

- Wherever it is present it will execute the command and after successful execution it will go back to the listening mode.
- If the voice command is not valid with a command in database, then it responses with "Say that again please" and again go back to the listening mode to fetch the next instruction.
- After the process was completed, to finish the software, the user must say turn off or quit, and the loop will be broken the program is terminated.

VI. RESULT

We've covered Python-based YSIS system on Windows using available libraries and packages. Our code was implemented using Visual Studio (VS) Code Integrated Development Environment (IDEs) that make the development process smoother and more efficient YSIS system.During this research, we have found that existing system missing some useful features that's why we are adding some features of the system such as automate browsing, automate website actions, automate webpages actions, dictating, WhatsApp, automatic birthday wisher, reading pdf and so on. This system especially helpful for users with visually impaired, handicapped, speech disorder and any kind of physical disability. It's also used by people who are illiterate or have spelling difficulties.

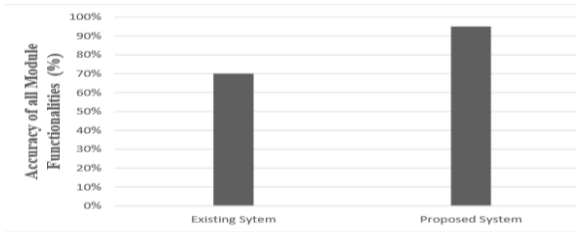
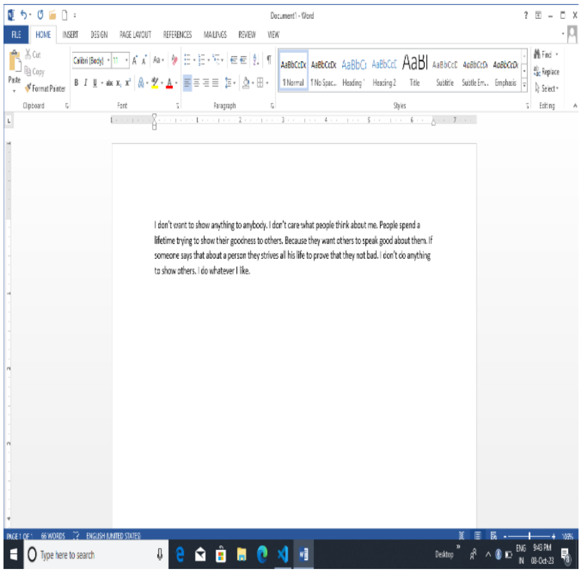


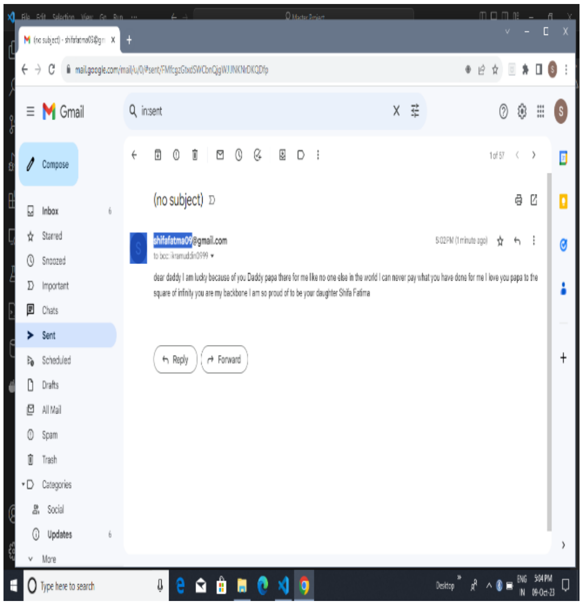
Fig.7 Comparative Analysis of Accuracy of all Functionalities

The primary goal of the project to design a system that could work as automate web browsing. To implement our system that assists users with voice commands and there are mainly follow two phases. Firstly, the ASR system takes input as speech from the user and extract the converted English phrase from the API's response. Secondly searching for the corresponding task or action associated with the voice command and then redirecting it to the Linux server with the help of HTTP Protocol. Once the server has completed the task, displaying the result on the web browser.



I don't want to show anything to anybody. I don't care what people think about me. People spend a lifetime trying to show their goodness to others. Because they want others to speak good about them. If someone says that about a person they strives all his life to prove that they not bad. I don't do anything to show others. I do whatever I like.

Snapshot of Word Document Typing Through Voice Command

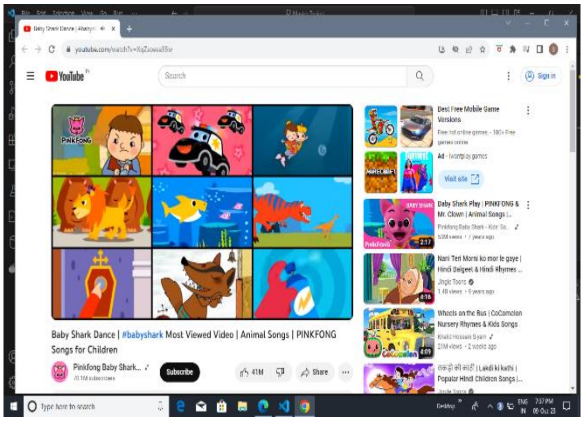


(no subject)

shifafatma09@gmail.com 5:02PM (1 minute ago)
to bcc: ikramuddin0999

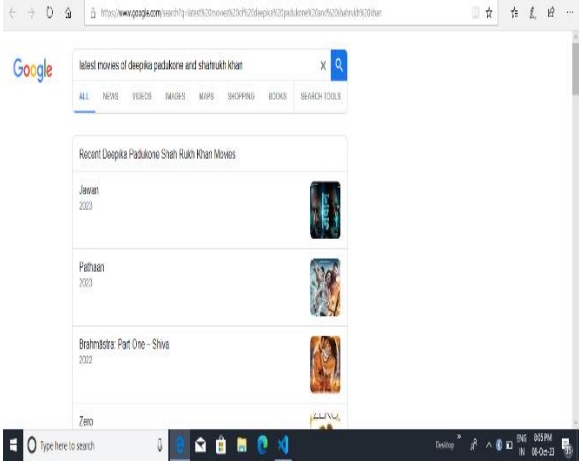
dear daddy I am lucky because of you Daddy papa there for me like no one else in the world I can never pay what you have done for me I love you papa to the square of infinity you are my backbone I am so proud of to be your daughter Shifa Fatima

Snapshots of Send Email Through Voice Command



Baby Shark Dance | #babyshark Most Viewed Video | Animal Songs | PINKFONG Songs for Children

Snapshot of Automate YouTube



latest movies of deepika padukone and shahrukh khan

Recent Deepika Padukone Shah Rukh Khan Movies

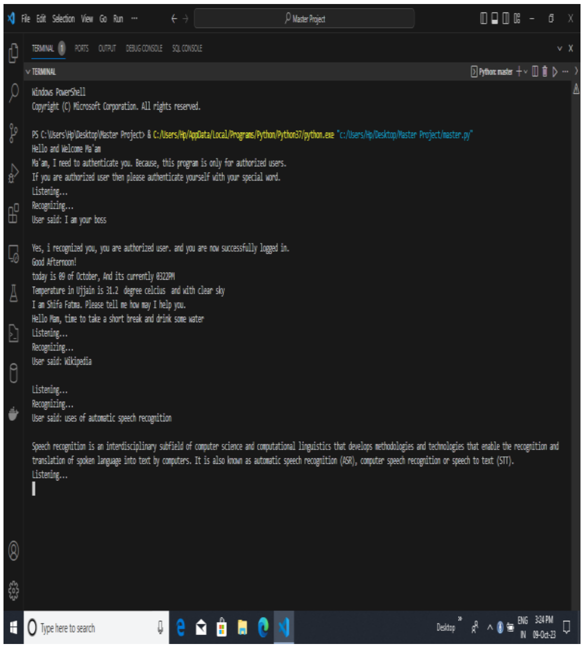
Jawan
2023

Pathaan
2023

Brahmāstra: Part One – Shiva
2022

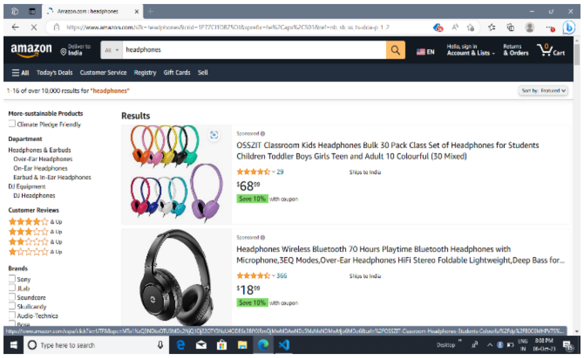
Zero

Snapshot of Automate Google Search

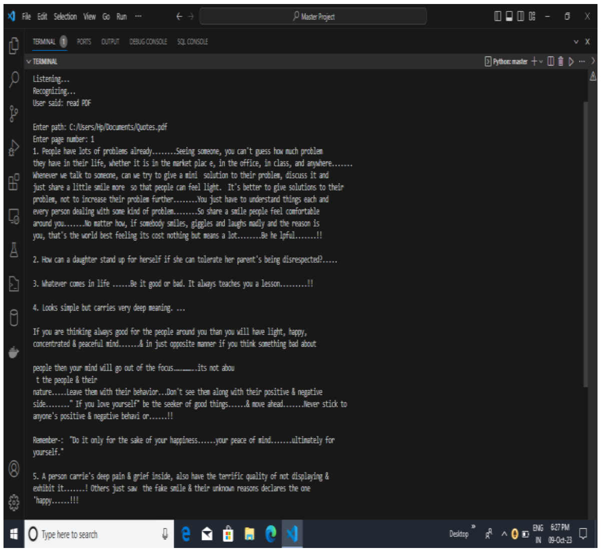


```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\Hp\Desktop\Master Project> & C:/Users/Hp/AppData/Local/Programs/Python/Python37/python.exe "c:/Users/Hp/Desktop/Master Project/master.py"
Hello and Welcome Ma'am
Ma'am, I need to authenticate you. Because, this program is only for authorized users.
If you are authorized user then please authenticate yourself with your special word.
Listening...
Recognizing...
User said: I am your boss

Yes, i recognized you, you are authorized user. and you are now successfully logged in.
Good Afternoon!
today is 09 of October, And its currently 0322PM
Temperature in Ujjain is 31.2  degree celcius  and with clear sky
I am Shifa Fatma. Please tell me how may I help you.
Hello Mam, time to take a short break and drink some water
Listening...
Recognizing...
User said: Wikipedia

Listening...
Recognizing...
User said: uses of automatic speech recognition

Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that develops methodologies and technologies that enable the recognition and
translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT).
Listening...
```

Snapshot of Authentication Based System & Search Wikipedia

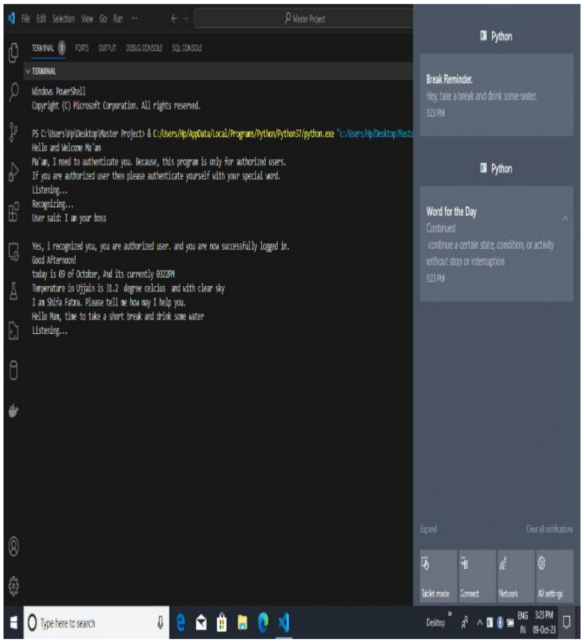

Amazon.com : headphones
amazon
Deliver to India
headphones
EN
Hello, sign in Account & Lists
Returns & Orders
0 Cart
All Today's Deals Customer Service Registry Gift Cards Sell
1-16 of over 10,000 results for "headphones"
Sort by: Featured
More-sustainable Products
Climate Pledge Friendly
Department
Headphones & Earbuds
Over-Ear Headphones
On-Ear Headphones
Earbud & In-Ear Headphones
DJ Equipment
DJ Headphones
Customer Reviews
& Up
Brands
Sony
JLab
Soundcore
Skullcandy
Audio-Technica
Results
Sponsored
OSSZIT Classroom Kids Headphones Bulk 30 Pack Class Set of Headphones for Students Children Toddler Boys Girls Teen and Adult 10 Colourful (30 Mixed)
29
Ships to India
$68.99
Save 10% with coupon
Sponsored
Headphones Wireless Bluetooth 70 Hours Playtime Bluetooth Headphones with Microphone,3EQ Modes,Over-Ear Headphones HiFi Stereo Foldable Lightweight,Deep Bass for...
366
Ships to India
$18.99
Save 10% with coupon
Type here to search
Desktop
ENG IN
8:08 PM
06-Oct-23

Snapshot Search Headphones on Amazon


File Edit Selection View Go Run ···
Master Project
TERMINAL 1 PORTS OUTPUT DEBUG CONSOLE SQL CONSOLE
TERMINAL
Type here to search
Desktop
ENG IN
6:27 PM
09-Oct-23

```
Listening...
Recognizing...
User said: read PDF

Enter path: C:/Users/Hp/Documents/Quotes.pdf
Enter page number: 1
1. People have lots of problems already........Seeing someone, you can't guess how much problem
they have in their life, whether it is in the market plac e, in the office, in class, and anywhere.......
Whenever we talk to someone, can we try to give a mini  solution to their problem, discuss it and
just share a little smile more  so that people can feel light.  It's better to give solutions to their
problem, not to increase their problem further........You just have to understand things each and
every person dealing with some kind of problem........So share a smile people feel comfortable
around you.......No matter how, if somebody smiles, giggles and laughs madly and the reason is
you, that's the world best feeling its cost nothing but means a lot........Be he lpful.......!!

2. How can a daughter stand up for herself if she can tolerate her parent's being disrespected?.....

3. Whatever comes in life ......Be it good or bad. It always teaches you a lesson.........!!

4. Looks simple but carries very deep meaning. ...

If you are thinking always good for the people around you than you will have light, happy,
concentrated & peaceful mind.......& in just opposite manner if you think something bad about

people then your mind will go out of the focus...........its not abou
 t the people & their
nature.....Leave them with their behavior...Don't see them along with their positive & negative
side........" If you love yourself" be the seeker of good things......& move ahead.......Never stick to
anyone's positive & negative behavi or......!!

Remember-:  "Do it only for the sake of your happiness......your peace of mind.......ultimately for
yourself."

5. A person carrie's deep pain & grief inside, also have the terrific quality of not displaying &
exhibit it.......! Others just saw  the fake smile & their unknown reasons declares the one
'happy......!!!
```

Snapshot of Reading Pdf Files



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users\Hp\Desktop\Master Project> & C:/Users/Hp/AppData/Local/Programs/Python/Python37/python.exe "c:/Users/Hp/Desktop/Maste
Hello and Welcome Ma'am
Ma'am, I need to authenticate you. Because, this program is only for authorized users.
If you are authorized user then please authenticate yourself with your special word.
Listening...
Recognizing...
User said: I am your boss

Yes, i recognized you, you are authorized user. and you are now successfully logged in.
Good Afternoon!
today is 09 of October, And its currently 0322PM
Temperature in Ujjain is 31.2  degree celcius  and with clear sky
I am Shifa Fatma. Please tell me how may I help you.
Hello Mam, time to take a short break and drink some water
Listening...
```

Break Reminder.
Hey, take a break and drink some water.
3:23 PM

Word for the Day
Continued
continue a certain state, condition, or activity without stop or interruption
3:23 PM

Snapshot of Water Remainder & Word of Day Functions

VII. CONCLUSION

In our project we have implemented and design many functions using python. This assistant currently works as an application based and performs basic tasks like sending message to user mobile, WhatsApp messages, sending e-mails, YouTube automation, webbrowser automation, music stream, weather updates, showing latest news, voice based typing notes and documents, gathering information from Google,water remainder, birthday remainder and birthday wisher, read pdf files and so on. This system we have automated many tasks with single line command.It is a beneficial thing to have in your home. It could help with many tasks and save you a lot of time. It can reduce the workload of human activities or the daily activities. The modular design of this project makes it more flexible and easily to add additional features without disturbing existing system functionalities. They also have lot of information than they are connected with internet. We hope this paper brings about inspiration, our methodology and better understanding amongst the research communities of YSIS.

REFERENCES

[1] S. Subhash, P. N. Srivatsa, S. Siddesh, A. Ullas and B. Santhosh, "Artificial Intelligence-based Voice Assistant," *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, London, UK, 2020, pp. 593-596.

[2] K. N., R. V., S. S. S. and D. R., "Intelligent Personal Assistant - Implementing Voice Commands enabling Speech Recognition," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2020, pp. 1-5.

[3] D. S. Zwakman, D. Pal, T. Triyason and V. Vanijja, "Usability of Voice-based Intelligent Personal Assistants," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 652-657.

[4] D. S. Zwakman, D. Pal, T. Triyason and C. Arpnikanondt, "Voice Usability Scale: Measuring the User Experience with Voice Assistants," *2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, Chennai, India, 2020, pp. 308-311.

[5] E. Marvin, "Digital Assistant for the Visually Impaired," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 723-728.

[6] P. Dabre, R. Gonsalves, R. Chandvaniya and A. V. Nimkar, "A Framework for System Interfacing of Voice User Interface for Personal Computers," 2020 3rd

International Conference on Communication System, Computing and IT Applications (CSCITA), Mumbai, India, 2020, pp. 1-6.

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

[8] T. -K. Kim, "Short Research on Voice Control System Based on Artificial Intelligence Assistant," 2020 International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, 2020, pp. 1-2.

[9] S. Noel, "Human computer interaction(HCI) based Smart Voice Email (Vmail) Application - Assistant for Visually Impaired Users (VIU)," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 895-900.

[10] A. M. Klein, A. Hinderks, M. Schrepp and J. Thomaschewski, "Measuring User Experience Quality of Voice Assistant," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), Seville, Spain, 2020, pp. 1-4.

[11] J. Zhao and W. -Q. Zhang, "Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models," in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1227-1241, Oct. 2022.

[12] https://encord.com/blog/data-curation-for-computer-vision/
https://www.mdpi.com/1424-8220/23/2/870