A Review on Automatic Speech Recognition System

Shifa Fatma[1]& Prof. Pankaj Raghuvanshi[2]
[1]Dept of CSE
[2]H.O.D., Dept of CSE
[1, 2] Alpine Institute of Technology, Ujjain

Abstract- *ASR is one which accurately translates spoken utterances into text. So text can be either in terms of words or it can be a word sequence, or it can be in terms of syllables, or it can be any sub-word units or phones, or even characters.This assistant developed for Windows users and it based on desktop assistant using python module, inbuilt functions and libraries. With the help of this assistant there will be no need to write the commands again and again for performing a particular task. Once the whole system is ready to use, we can perform many tasks in the easiest ways. It can understand and carry out the audio instructions given by the user. We don't have to worry about using input devices like the keyboard and mouse, so we'll use them less. It saves the user a lot of time and also reduce the hardware cost and space taken by it.People who are blind, elderly, or physically disabled can engage with equipment via the speech recognition system. ASR system can be found in various devices, including cars, smartwatches smartphones, smart speakers, laptop, and even household appliances.*

Keywords- AI based Automatic Speech Recognition, API, Desktop Assistant, Personal Computing, Python.

I. INTRODUCTION

Automatic Speech Recognition (ASR) technological advancement is increasing day by day, revolutionizing the way we interact with machines and systems through speech. Earlier, automatic speech recognition system performed very few tasks. But, Nowadays most of the electronic gadgets come along with automate system. Including the used of machine learning, python programming, artificial intelligence, deep learning and few more technologies have revolutionize the field of advanced that we can perform any type of task. ASR allows us to convert audio into usable structured data. Typically in the form of a readable transcript.

You have a person or an audio source saying something textual. And you have a bunch of microphones which are receiving the audio signals.And you pass it to an Automatic Speech Recognition system, whose job is now to infer the original source transcript that the person spoke or that the device played. It's a very natural interface for human communication, you don't need a mouse or a keyboard, so it's obviously a good way to interact with machines [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, sending email, 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]. Voice control system implemented by Raspberry Pi, open API and AI. Using this techniques users to design their own system through the speech recognition interface and various modules [8]. User to send emails seamlessly through voice command and reading email. The proposed to design a mechanism which converts Speech To Text for email sending and convert Text To Speech for reading emails. The application develop with Google Web Kit API which is used for speech processing and speech recognizing [9].

II. OBJECTIVE

The main objective of speech recognition is to recognize WHO is speaking. The purpose of speech recognition understanding and comprehending WHAT was spoken. It is used to identify a person by analyzing their tone, voice pitch, and accent. The main objective of this research, is to specify that this ASR system will be helpful for blind people to operate the system just by speech.

The research have revealed the fact that existing system missing some features. For this reason, we are adding some features of the system like automate website actions, automate browsing, dictating, reading pdf and so on. The main goal of this system is improve the intelligibility of speech, enhance performance robustness, efficiency and man machine communications. The ASR system makes our life easier and save times.

III. LITRATURE SURVEY

In this paper presented on Artificial Intelligence based Voice Assistant for visually disabled persons. Voice assistants are all written in python programming languages and using a predefined libraries. This project will recognize voice respond according to the user. This assistant currently performs certain tasks like weather updates, tell news today, find location, Google search, playing song and movies on YouTube, sending mail, translate, shutdown and restart system, take screenshot and turning on/off smart phone applications[1]. In this paper, IPA is similar to chat bots. This assistant is provided authentication. User needs to be login id and password and the only means to access it is by logging into the system. The system would ask for various operation such as edit, read, paste, copy, and other similar operations that shall be performed inside it once the application opens. The system uses the ferment synthesis, a type of speech synthesizer for responding to the user through voice command [2]. In this paper presented on Usability of Voice-based Intelligent Personal Assistants. The primary objective behind this study was to check the suitability of SUS for measuring the usability in a voice-only context. From a design perspective there is a difference between speech and GUI interfaces, and the same is reflected in the usability scenario too [3]. In this paper, Voice Usability Scale for measuring the user experience with Voice Assistant. This based on themes of information quality and relevance, semantic intelligence, and user satisfaction. It can be used for the improvement of product and services as well as evaluation and research purposes. The main goal to deploy product a valid and reliable for evaluating voice assistant [4].

The research of this paper, in order to solve the problem, we introduced DAVID a digital assistant application aim to help the visually impaired in recognizing text on real-world objects and provide an audio feedback in real-time. It utilizes voice user interface technology such as speech recognizing and speech synthesis as the means of interaction through voice input [5]. In this paper presented on A Framework for System Interfacing of Voice User Interface for Personal Computers. The research in the domain have been focused around natural language processing, speech to text, method for commands execution etc. The applications such as voice based email, playing music, dictation and so on are being researched in isolation. Some research is based on navigation using voice while some is based on mouse and keyboard control using voice. Integration of the voice for the entire operating system has been less studied [6]. In this paper presented on Automation and Presentation of Word Document Using Speech Recognition. This system is trained to work with documents and presentations via voice commands. This system developed in python programming language with modules. It's also performed services of music player, checking date and time, Google searching engine, Wikipedia searching engine, YouTube, mailing, camera, help menu, MS-Word and MS-PowerPoint services. This system mainly focused to provide an effortless way of eliminating the need to type which will result in fast and error-free work [7].

This paper proposes a voice control system using Raspberry Pi and open API artificial intelligence (AI). In this paper, Google Assistant services used as the Open API for send voice commands. Voice data sent to Google server are converted to text after being recognized and the data are subsequently sent to the conditional auto-run mode IFTTT service. For IFTTT, one recipe performs one operation. One Recipe created by Trigger and Action sends a PUT request in JSON format to the API server, which has already been set by itself through Webbooks. The API server exchanges data via WLAN communication and Raspberry Pi. The received PUT request updates the PIN status associated with Raspberry Pi. This system allows users to implement their own system through the speech recognition interface and using some modules [8].

In this paper, presented Vmail for visually impaired users to send emails seamlessly through voice commands without any dependency. This application is develop a mechanism which converts Speech to Text (STT) for email composing and also converts Text To Speech (TTS) for reading emails. The user can completely write and edit email using their voice with less effort [9].

In this paper, Measuring User Experience Quality of Voice Assistants. This article presents a new approach to the flexible evaluation of voice assistance systems. In this article, three scales for measuring the UX aspects of voice assistance systems are response behavior, response quality and comprehensibility. These three new scales are designed to capture the interaction, or more precisely, the UX aspects of the user's communication with the VAs. These voice communication scales can be combined with otherscales of the UEQ+ framework to create a product-related questionnaire [10].

IV. WORKING OF ASR

Speech is the first converted from physical sound to electrical energy using a microphone and then to digital data using an analog to digital convertor. This digital data can be converted to text using algorithms like Neural Networks or Hidden Markov Models.

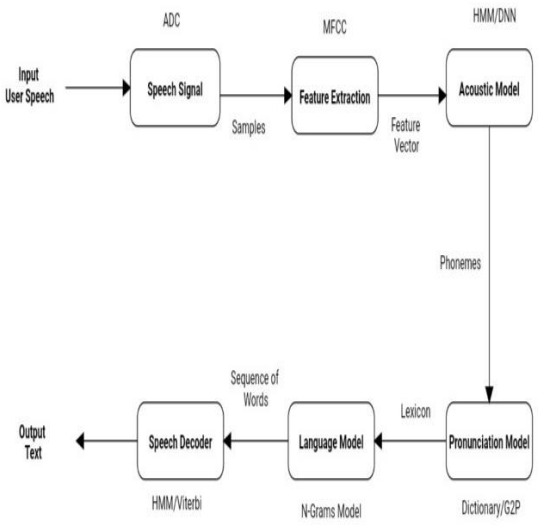
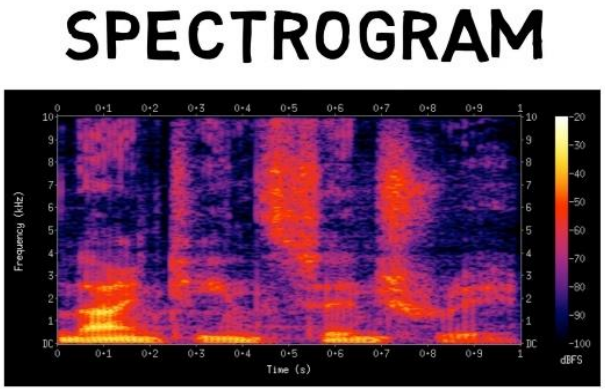


Fig.1. Working Models of ASR

A. Speech Signal (Acoustic Analysis):

The first component is acoustic analysis, which sees the raw speech waveform and converting it into some discrete representation or features. Because you can't really work with a speech signal. So, you sample it and generate these discrete samples. It also filters the sound to remove unnecessary noise, and normalizes the sound and speed of the speech, to match the prerecorded samples in the machine's device. Actually, we can name that also a microphone. Microphone converting a sentence from analog to digital. The microphone generates a spectrogram, now we will use spectrogram for a visual representation of the frequency content over time in an audio recording.In the speech recognition process we actually need three elements of sound: its frequency, its intensity and the time it took to make it. Therefore we will be using a super complex formula called Fast Fourier Transform to convert the graph of spectrogram.

The spectrogram shows the frequency on the vertical axis or y axis (as in high pitches versus low pitches) and the time on the horizontal axis or x axis. And the colors are actually the energy that you use to make the sound. The areas that are brighter signify high frequencies, and the areas that are darker signify low ones. Each of these samples are typically of the order of 10 to 25 milliseconds of speech frames (acoustic frame).



GRAPH OF SPECTROGRAM

B. Features Extraction:

The features are extracted from the discrete sample (speech frames). Feature extraction is performed on each acoustic frame. Features extraction with Mel Frequency Cepstral Coefficient or MFCCs. Used for reduce the dimensionality of our data and squeeze noise out of system and widely used to represent the content of each audio utterance.

C. Acoustic Model:

Acoustic model creates the basic unit of phoneme from features. The way that this has worked is you first take the audio and feed it an acoustic model which is actually going to produce a phonemic transcription of this particular audio recording.Each word can be represented as a sequence of phonemes. So you have a probabilistic model, which maps sequences of feature vectors (acoustic feature) to a sequence of phonemes. This mapping are not giving you a one best sequence. It give a priority distribution of phone sequences. And these priority distributions are mixtures of Gaussian. Acoustic Models are powered by classical Hidden Markov Models (HMM)**.** The transition between phones and the corresponding observable can be modeled with the HMM. So it gives us probability of phonemes.

D. Pronunciation Model:

The acoustic module, which produces phone sequences. So now, we eventually want to get a word sequence, from the speech utterance. So this is just an intermediate representation, these phones. So now we use these large pronunciation dictionaries. So this the model, which provides a link between these phone sequences and words.A pronunciation model that connects phonemes together to form words.

E. Language Model:

The language model contains the information about the possible or probable sequences of words which then form sentences and also improves the accuracy of the predicted transcription. There are two ways to formalize language models. Firstly,Deterministic models are defined by the grammar rulesand Secondly, Statistical models are probability occurrence of sequence of words (unigram, bigram, trigram). The language model is very crucial because it can be used to disambiguate between similar acoustics.

F. Decoder:

Searches for the best possible word sequence among all possible word sequences. In this stage, we perform search graphs technique by Hidden Markov Models. This process in speech recognizer find a most probable sequence of words whose corresponding models (acoustic model, pronunciation model, language model) best match the input feature vector sequence. The Viterbi decoder is commonly used as a decoding process for determine the most probable sequence of words or phonemes by backtracking through the trellis. This process involve searching the one with the highest probability and evaluating their likelihood using the models. It's also improve the efficiency, readability and accuracy of the final transcription.

V. PERFORMANCE EVALUATION

Evaluate the ASR quality by WER

The Word Error Rate (WER) is a mainly used for metric to evaluate the performance of speech recognizer, there are different metrics available. The most popular such metrics are the word error rate and the word accuracy. Its measures the percentage of errors in the transcriptions generated by the Automatic Speech Recognition system. The assistant will generate its own transcription for each utterance. WER is calculated by comparing the recognized words with the reference (ground truth) transcription of the same utterance.
In order to calculate a word error rate we first have to count the number of substituted words **S**, the number of inserted words by the speech recognizer **I** and the number of deleted words by the recognizer **D** and divide that sum by the overall number of words in the reference transcription. Now in order to determine whether a word is really counted as a substitution, insertion or deletion, we need to first perform analignment. And this is usually done through dynamic programming algorithms which punish the certain types of errors, substitutions, insertions and deletions in a certain way. There are standardized methods for calculating this alignment and for then determining a word error rate.Multiply the result by 100 to express the WER as a percentage.

The formula for calculating the WER is as follows:

$$\frac{(\text{Total (S)} + \text{Total (I)} + \text{Total (D)})}{\text{Total Reference Words (N)}}$$

Where,

S is the number of word substitutions errors, which occur when a recognized output is different word from the corresponding word in the reference.

D is the number of word deletions errors, which occur when a recognized output is missing word from the corresponding word in the reference.

I is the number of word insertions errors,which occur when a recognized output is extra word present from the corresponding word is absence in the reference.

N is the total number of words in the reference transcription.

The numerator **(S + D + I)** represents the total number of errors, and divided by the denominator **(N)** represents reference transcription, provides the error rate as a percentage.

For example, if the reference transcription contains 100 words and the system makes 5 substitutions, 3 deletions, and 2 insertions, the WER would be:

WER = (5 + 3 + 2) / 100 = 0.1 or 10%

The overall WER was observed as 10%.
The word accuracy is actually 1 minus this word error rate for a individual word recognizer.
Accuracy = 100 – WER%

Lower WER values which means better performance and accuracy, as it means fewer errors produced in transcription and closer to the reference transcription. In the other hand higher WER values which means more errors in the transcription.

VI. CHALLENGES

A. Major challenges in speech recognition

- Acoustic variability - the same phonemes pronounced in different contexts will have different acoustic realization.
- High intra-speaker variability: when the same speaker pronounces phonemes differently.
- High inter-speaker variability: differences between speakers.

B. Disordered speech and speech recognition

- Production of a critical mass of phonemes may be limited.
- A lack of consistency of articulatory patterns results in highly variable acoustic representation over time.
- Severely dysarthria speech too distant from acoustic models generated from typical speech.
- However, even people with moderate to severe dysarthria speech are understood by close friends and family!

C. Main research challenges

How to build a well-performing speech recognizer from sparse data?

{¡}For each individual it's tiring to record many hours of data
{¡¡} Difficult to get transcriptions (which words were said)
{¡¡¡} Hardly any found data (e.g. from databases, TV, YouTube) to supplement with

High need for personalization

{¡} Speaker-independent models do not generally work well.
{¡¡} We can't assume to 'pool' data from other speakers with dysarthria.

VII. MODULES USED IN AUTOMATIC SPEECH RECOGNITION SYSTEM

Several libraries and packages that can aid in implementing these system, such as:-

Speech Recognition

The main focuses of speech processing is to provide an interaction between a human and a system. It's allows machine to understand human language. It is a process of understanding the words, phrases and sentences that are spoken by human beings. And then convert the spoken words into written format. It's also plays an important role in home automation, artificial intelligence, etc.

Smtplib

Its deals to connect to the mail server to send email. There are 3 steps involved - initialize, send mail (), quit. Emails consisting of only text body without any subject or any attachments.

OS Module

It's provides a portable way of using operating system dependent functionalities like creating folder, removing folder, retrieving the content, identifying the directory, and so on.

DateTime

The Date-time module to work with dates and times. For the user wants a remainder or alarm at a certain time. This module will help to access the time and look that the task is done according to the user's time condition.

Wikipedia

It's possible for the ASR to process the queries on Wikipedia and display the results to users. The number of lines of information to be displayed can be set manually.

Webbrowser

This module is useful for webbrowser controller. It's allows the system display web information to users. For example; - If the user says as a input "open YouTube" the system directs to the web browser and open YouTube in the web browser and retrieves the data for the user.

PyAudio

PyAudio is required if and only if you want to use microphone input (Microphone). It allows to play sound, record audio streams on the variety of platforms.

gTTs

gTTS stands for Google Text-to-Speech. It converts specified text into audio, which we can save as an mp3 file. It also supports many languages such as English, Hindi, French, German and much more.

Pyttsx3

This module is used for text to speech conversion, its works offline. It's provide two voices first is female and the

second is male which is provided by “sapi5” for windows and also provides run and wait functionality. It determines how much time the system will wait for another input of user.

JSON

JSON stand for JavaScript Object Notation. It is popular data format for representing a structuring data. It is mainly used for transferring and storing data in web applications. This module is a part of the built-in python package.

VIII. CONCLUSION

This review paper presented the automatic speech recognition system with application, challenges, performance evaluation, methodologies, issues and technique. Different ways available for implementation of this system and it's based on modification features. Automatic speech recognition is useful in many fields like health care, home appliances, education, military, telephony and many more. It can help people with a variety of disabilities and also useful for people with physical disabilities who often find typing difficult 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. It can reduce the workload of basic human activities or the daily activities and this will also reduce the work using the keyboard, mouse, and other input devices. It could help with many tasks perform easily and save you a lot of time. This ASR system currently works online and its performs basic tasks given by users like playing music and videos, weather updates, searching Wikipedia, telling latest news, opening desktop application, managing mails and so on.

Our main aims of this research to improve and advance existing features of speech recognition system or adding more new features to meet the evolving needs of the users. Overall this paper covers reviews on all aspects of automatic speech recognition system and in future with the help of this review we can add new features and functionality to improve the existing system. It's also work on any languages using different approaches that gives better performance than existing work done on it.

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] 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] 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] 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] 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] 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] https://upload.wikimedia.org/wikipedia/commons/c/c5/Spectrogram-19thC.png