GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Next-generation of Virtual Personal Assistants Automated Searching Techniques: Speech Assistants

A Final Project Report of Capstone Project – 2

Submitted by

Vivek Kumar

(16SCSE112024/1613112052)

In partial fulfillment for award of the degree

Of

Bachelor of Technology

In

Computer Science and Engineering with Specialization of

Data Analytics

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. M.Arvindhan

Assistant professor

JUNE-2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"NEXT-GENERATION OF VIRTUAL PERSONAL ASSISTANTS AUTOMATED SEARCHING TECHNIQUES: SPEECH ASSISTANTS"** is the bonafied work of **"VIVEK KUMAR (1613112052)"** who carried out the project work under my supervision.

SIGNATURE OF HEAD

Dr. MUNISH SHABARWAL,
PhD (Management), PhD (CS)
Professor & Dean,
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr. M.Arvindhan

Assistant professor

School of Computing Science & Engineering

Table of content

1. **Abstract**

Speech recognition Technology is one of the fast-growing engineering technologies. Speech is a natural mode of communication for people. We learn all the relevant skills during early childhood, without instruction, and we continue to rely on speech communication throughout our lives. It comes so naturally to us that we don't realize how complex a phenomenon speech is. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state. This project is designed and developed keeping that factor in mind, and a little effort is made to achieve this aim. It has a number of applications in different areas and provides potential benefits, nearly 20% people of the world ae suffering from various disabilities; many of them are blind or unable to use their hands effectively. The speech recognition system in those particular cases provide a significant help to them, so that they can share information with people by operating computer through voice input**.** Consider the Thousands of people in world

they are not able to use their hands making typing impossible. our project it for these people who can’t type ,and see ,even for those of us who are lazy and don’t feel like it Our project is capable to recognize the speech and convert the input audio into text; it also enables a user to perform operations such as open any type of thing through google**.**

Continuous speech: When user speak in a more normal, fluid manner without having to pause between word, which is referred as continuous speech.

Discrete speech: when user speak with taking rest between each word then such speech is referred as discrete speech.

2.Introduction

Definition of Speech Recognition - it also referred to as the Automatic Speech Recognition (ASR), or computer speech recognition system it is that the process which converting the speech signal to a sequence of words, which means of an algorithm implemented as a computer virus.

Voice: Voice is that the basic, common and efficient sort of communication method for people to interact with one another . At this time speech technologies are very commonly available for a limited but interesting range of task. This technology enables machines to reply correctly and reliably to human voices and supply useful and valuable services. As communicating with computer is quicker using voice instead of using keyboard, so people will prefer such system. Communication among the person is dominated by speech , therefore it's natural for people to expect voice interfaces with computer. Need to create these components, incorporating the applicable criteria that follow. This can be accomplished by developing voice recognition system: speech-to-text which allows computer to translate voice request and dictation into text. Voice recognition system: speech-to-text is the process of converting an acoustic signal which is captured using a microphone to a set of words. The recorded data can be used for document preparation. The purpose of this system is to provide a help of those people who are suffering from various

disabilities many of them are blind or unable to use their hands effectively. The speech recognition system in those particular cases provides a significant help to them so that they can share information with the people by operating computer through voice input

Speech Recognition (SR) is the ability to translate a dictation or spoken word to text. Speech Recognition known as “automatic speech recognition “(ASR), or speech to text (STT)

- Speech recognition is the process of converting an acoustic signal, captured by a microphone or any peripherals, to a set of words.
- To achieve speech understanding we can use linguistic processing
- The recognized words can be an end in themselves, as for applications such as commands & control data entry and document preparation. In the society every one either human or animals wish to interact with each other and tries to convey own message to others. The receiver for messages may get the exact and full idea of the senders, or may get the partial idea or sometimes cannot understand anything out of it. In some cases, may happen when there is some lacking in communication (i.e. when a child convey message, the mother can understand easily while others cannot)

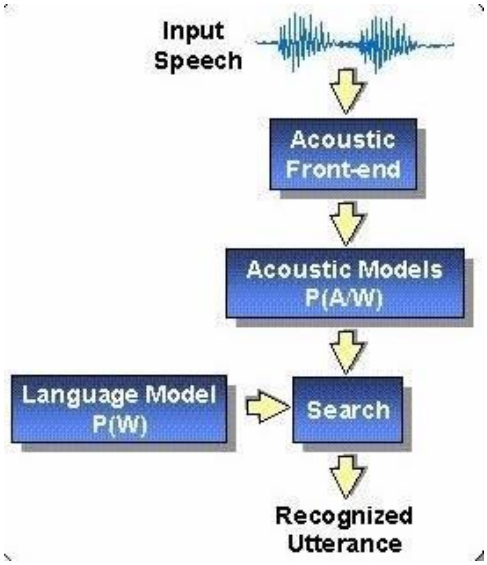


Fig.1 model of speech recognition

A.**Types of speech utterance**

Speech recognition are classified according to what type of utterance they have ability to recognize. They are classified as:

1.Isolated word: Isolated word recognizer usually requires each spoken word to have quiet (lack of an audio signal) on bot h side of the sample window. It accepts single word at a time.

2.Connected word: It is similar to isolated word, but it allows separate utterances to „run-together“ which contains a minimum pause in between them.

3.Continuous Speech: it allows the users to speak naturally and in parallel the computer will determine the content.

4.Spontaneous Speech: It is the type of speech which is natural sounding and is not rehearsed.

B. Types of speaker model

Speech recognition system is mainly categorized into two main categories based on speaker models which have following name speaker dependent and speaker independent models.

i) Speaker dependent models: These systems are designed for a particular speaker. Development of this model is very easy and more efficient but they're not so flexible.

ii) Speaker independent models: These systems are developed for sort of speaker. This system are difficult In development and fewer accurate but they're considerably flexible.

C. Sorts of vocabulary

The vocabulary size of speech recognition system affects the processing requirements, accuracy and complexity of the system. In voice recognition system: which convert speech-to-text the kinds of vocabularies are often classified as follows:

2.1 Purpose:

Nearly 20% people of the world ae suffering from various disabilities; many of them are blind or unable to use their hands effectively . The speech recognition system in those particular cases provide a significant help to them , so that they can share information with people by operating computer through voice input. Consider the Thousands of people in world they are not able to use their hands making typing impossible. our project it for these people who can't type, and see, even for those of us who are lazy and don't feel like it Our project is capable to recognize the speech and convert the input audio into text; it also enables a user to perform operations such as searching any information from any search bar.

2.2 Motivations and scope: This project has the speech recognizing and speech synthesizing capabilities though it is not a complete replacement of what we call a notepad but still a good text editor to be used through voice this software also can open windows-based software such as notepad, google chrome and etc.

3.Literature Survey

1920-1960s:

In the early 1920s machine recognition came into existence. The first machine to recognize speech to any significant degree commercially named, Radio Rex (toy) was manufactured in 1920.Research into the concepts of speech technology began as early as 1936 at Bell Labs. In 1939, Bell Labs demonstrated a speech synthesis machine (which simulates talking) at the World Fair in New York. Bell Labs later abandoned efforts to develop speech-simulated listening and recognition; based on an incorrect conclusion that artificial intelligence would ultimately be necessary for success. The earliest attempts to devise systems for automatic speech recognition by machine were made in 1950s, when various researchers tried to exploit the fundamental ideas of acoustic phonetics. During 1950s[1], most of the speech recognition systems investigated spectral resonances during the vowel region of each utterance which were extracted from output signals of an analogue filter bank and logic circuits. In 1952, at Bell laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single talker, as embodied in 10 monosyllabic words [3]. The system again relied

on spectral measurements (as provided by an analog filter bank) primarily during vowel regions. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants . They used a spectrum analyzer and a pattern matcher to make the recognition decision. A novel aspect of this research was the use of statistical information about allowable sequences of phonemes in English (a rudimentary form of language syntax) to improve overall phoneme accuracy for words consisting of two or more phonemes. Another effort of note in this period was the vowel recognizer of Forgie and Forgie constructed at 189 laboratories in 1959 in which 10 vowels embedded in a /b/-vowel/t/ format were recognized in a speaker independent manner [5]. Again a Filter bank analyzer was used to provide spectral information and a time varying estimate of the vocal tract resonances was made to deicide which vowel was spoken.

1960-1970:

In the 1960s several fundamental ideas in speech recognition surfaced and were published. In the 1960s since computers were still not fast enough, several special purpose hardware were built. However, the decade started with several Japanese laboratories entering the recognition arena and building special purpose hardware as part of their systems. On early Japanese system, described by Suzuki and Nakata of the Radio Research Lab in Tokyo, was a hardware vowel recognizer . An elaborate filter bank spectrum analyzer was used along with logic that connected the outputs

of each channel of the spectrum analyzer (in a weighted manner) to a vowel decision circuit, and majority decisions logic scheme was used to choose the spoken vowel. Another hardware effort in Japan was the work of Sakai and Doshita of kyoto University in 1962, who built a hardware phoneme recognizer. A hardware speech segmented was used along with a zero crossing analysis of different regions of the spoken input to provide the recognition output. A third Japanese effort was the digit recognizer hardware of Nagata and coworkers at NEC Laboratories in 1963. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program. One of the difficult problems of speech recognition exists in the non uniformity of time scales in speech events. In the 1960s three key research projects were initiated that have had major implications on the research and development of speech recognition for the past 20 years. The first of these projects was the efforts of Martin and his colleagues at RCA Laboratories, beginning in the late 1960s, to develop realistic solutions to the problems associated with non-uniformity of time scales in speech events. Martin developed a set of elementary time normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduce the variability of the recognition scores[9]. Martin ultimately developed the method and founded one of the first speech recognition companies, Threshold Technology, which was built, marketed and was sold speech recognition products. At about the same time, in the Soviet Union,

Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances(generally known as Dynamic Time Warping(DTW) ,including algorithms for connected word recognition.. Although the essence of the concepts of dynamic time warping, as well as rudimentary versions of the algorithms for connected word recognition, were embodied in Vintsyuk s work, it was largely unknown in the West and did not come to light until the early 1980 s; this was long after the more formal methods were proposed and implemented by others. At the same time in an independent effort in Japan Sakoe and Chiba at NEC Laboratories also started to use a dynamic Programming technique to solve the non uniformity problems.A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes. Reddy s research eventually spawned a long and highly successful speech recognition research program at Carnegie Mellon University (to which Reddy moved in the late 1960s) which, to this today, remains a world leader in continuous speech recognition systems.

1970-1980:

In the 1970s speech recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zagoruyko in Russia, Cakoe and Chiba in Japan[14], and Itakura in the united States. The Russian studies helped the advance use of pattern recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura s research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low bit rate speech coding, could be extended to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters.Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which researchers studied three distinct tasks over a period of almost two decades, namely the New Raleigh language for simple database queries, the laser patent text language for transcribing laser patents, and the office correspondent tasks called Tangora, for dictation of simple memos. Finally, at AT&T Bell Labs, researchers began a series of experiments aimed at making speech recognition systems that were truly speaker independent. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different

words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker independent patterns are now well understood and widely used. An ambitious speech understanding project was funded by the defence Advanced Research Projects Agencies(DARPA), which led to many seminal systems and technology. One of the demonstrations of speech understanding was achieved by CMU in 1973 there Heresay I system was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer.CMU s Harphy systemwas shown to be able to recognize speech using a vocabulary of 1,011 words with reasonable accuracy. One of the particular contributions from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching 190. Other systems developed under the DARPA s speech understanding program included CMU s Hearsay II and BBN s HWIM (Hear what I Mean) systems. The approach proposed by Hearsay II of using parallel asynchronous proceses that simulate the component knowledge sources in a speech system was a pioneering concept. A global blackboard was used to integrate knowledge from parallel sources to produce the next level of hypothesis. 6.4 1980-1990: Just as isolated word

recognition was a key focus of research in the 1970s, the problems of connected word recognition was a focus of research in the 1980s. Here the goal was to create a robust system capable of recognizing a fluently spoken string of words(eg., digits) based on matching a concatenated pattern of individual words. Moshey J. Lasry has developed a featurebased speech recognition system in the beginning of 1980. Wherein his studies speech spectrograms of letters and digits[97].A wide variety of the algorithm based on matching a concatenated pattern of individual words were formulated and implemented, including the two level dynamic programming approach of Sakoe at Nippon Electric Corporation ,the one pass method of Bridle and Brown at Joint Speech Research Unit(JSRU) in , the level building approach of Myers and Rabiner at Bell Labs, and the frame synchronous level building approach of Lee and Rabiner at Bell Labs. Each of these optimal matching procedures had its own implementation advantages, which were exploited for a wide range of tasks. Speech research in the 1980s was characterized by a shift in technology from template based approaches to statistical modeling methods especially the hidden Markov model approach . Although the methodology of hidden Markov modeling (HMM) was well known and understood in a few laboratories(Primarily IBM, Institute for Defense Analyses (IDA), and Dargon systems), it was not until widespread publication of the methods and theory of HMMs, in the mid1980, that the technique became widely applied in virtually, every

speech recognition research laboratory in the world. Today, most practical speech recognition systems are based on the statistical frame work developed in the 1980s and their results, with significant additional improvements have been made in the 1990s.

Hidden Markov Model(HMM): HMM is one of the key technologies developed in the 1980s, is the hidden Markov model(HMM) approach . It is a doubly stochastic process which as an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations. Although the HMM was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analysis (IDA) and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs in the mid-1980s that the technique became widely applied in virtually every speech recognition research laboratory in the world. In the early 1970s, Lenny Baum of Princeton University invented a mathematical approach to recognize speech called Hidden Markov Modeling (HMM). The HMM pattern-matching strategy was eventually adopted by each of the major companies pursuing the commercialization of speech recognition technology (SRT).The U.S. Department of Defense sponsored many practical research projects during the 70s that involved several contractors, including IBM, Dragon, AT&T, Philips and others. Progress was slow in those early years.

Neural Net: Another new technology that was reintroduced in the late 1980s was the idea of applying neural networks to problems in speech recognition. Neural networks were first introduced in the 1950s, but they did not prove useful initially because they had many practical problems. In the 1980s however, a deeper understanding of the strengths and limitations of the technology was achieved, as well as, understanding of the technology to classical signal classification methods. Several new ways of implementing systems were also proposed .

DARPA Program: Finally, the 1980s was a decade in which a major impetus was given to large vocabulary, continuous speech recognition systems by the Defense Advanced Research Projects Agency (DARPA) community, which sponsored a large research program aimed at achieving high word accuracy for a 1000 word continuous speech recognition, database management task. Major research contributions resulted from efforts at CMU(notably the well known SPHINX system)[36], BBN with the BYBLOS system Lincoln Labs, SRI, MIT and AT&T Bell Labs The SPHINX system successfully integrated the statistical method of HMM with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context dependent phone models in a sophisticated lexical decoding network. The DARPA program has continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer and the task shifting to retrieval of air travel information. At the same time, speech recognition technology has been

increasingly used within telephone networks to automate as well as enhance operator services.

1990-2000s:

In the 1990s a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error . This fundamental paradigmatic change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined and the Bayes decision theory becomes inapplicable under these circumstances. Fundamentally, the objective of a recognizer design should be to achieve the least recognition error rather than provide the 191 best fitting of a distribution function to the given (known)data set as advocated by the Bayes criterion. This error minimization concept produced a number of techniques such as discriminative training and kernel based methods. As an example of discriminative training, the Minimum Classification Error(MCE) criterion was proposed along with a corresponding Generalized Probabilistic Descent(GPD) training algorithm to minimize an objective function which acts to approximate the error rate closely. Another example was the Maximum Mutual

Information (MMI) criterion. In MMI training, the mutual information between the acoustic observation and its correct lexical symbol averaged over a training set is maximized. Although this criterion is not based on a direct minimization of the classification error rate and is quite different from the MCE based approach, it is well founded in information theory and possesses good theoretical properties. Both the MMI and MCE can lead to speech recognition performance superior to the maximum likelihood based approach . A key issue in the design and implementation of speech recognition system is how to properly choose the speech material used to train the recognition algorithm. Training may be more formally defined as supervised learning of parameters of primitive speech patterns (templates, statistical models, etc.,) used to characterize basic speech units (e.g. word or subword units), using labeled speech samples in the form of words and sentences. It also discusses two methods for generating training sets. The first, uses a nondeterministic statistical method to generate a uniform distribution of sentences from a finite state machine represented in digraph form. The second method, a deterministic heuristic approach, takes into consideration the importance of word ordering to address the problem of co articulation effects that are necessary for good training. The two methods are critically compared.

2000-2009:

a) **General:** Around 2000, a variational Bayesian (VB) estimation and clustering techniques were developed Unlike Maximum Likelihood, this

VB approach is based on a posterior distribution of parameters. Giuseppe Richardi have developed the technique to solve the problem of adaptive learning, in automatic speech recognition and also proposed active learning algorithm for ASR. In 2005, some improvements have been worked out on Large Vocabulary Continuous Speech Recognition system on performance improvement. In 2007, the difference in acoustic features between spontaneous and read speech using a large scale speech data base i.e, CSJ have been analyzed. Sadaoki Furui investigated SR methods that can adapt to speech variation using a large number of models trained based on 193 clustering techniques. In 2008, the authorshave explored the application of Conditional Random Field(CRF) to combine local posterior estimates provided by multilayer perceptions corresponding to the frame level prediction of phone and phonological attributed classes. De-wachter et.al., attempted to over-come the time dependencies, problems in speech recognition by using straight forward template matching method. Xinwei Li et.al, proposed a new optimization method i.e., semi definite programming(SDP) to solve the large margin estimation(LME) problem of continuous density HMM(CDHMM) in speech recognition. Discriminate training of acoustic models for speech recognition was proposed under Maximum mutual information(MMI) Around 2007 Rajesh M.Hegde et.al proposed an alternative method for processing the Fourier transform phase for extraction speech features, which process the

group delay feature(GDF) that can be directly computed for the speech signal.

b) DARPA program: The Effective Affordable Reusable Speech-to-Text (EARS) program was conducted to develop speech-to-text (automatic transcription) technology with the aim of achieving substantially richer and much more accurate output than before. The tasks include detection of sentence boundaries, fillers and disfluencies. The program was focusing on natural, unconstrained human speech from broadcasts and foreign conversational speech in multiple languages. The goal was to make it possible for machines to do a much better job of detecting, extracting, summarizing and translating important information, thus enabling humans to understand what was said by reading transcriptions instead of listening to audio signals .

c) Spontaneous speech recognition: Although read speech and similar types of speech, e.g. news broadcasts reading a text, can be recognized with accuracy higher than 95% using state-of-the-art of speech recognition technology, and recognition accuracy drastically decreases for spontaneous speech. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech. In order to increase recognition performance for spontaneous speech, several projects have been conducted. In Japan, a 5-year national project Spontaneous Speech: Corpus and Processing Technology was conducted . A world-largest spontaneous speech corpus, Corpus of

Spontaneous Japanese (CSJ) consisting of approximately 7 millions of words, corresponding to 700 hours of speech, was built, and various new techniques were investigated. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic speech summarization . The three analyses on the effects of spontaneous speech on continuous speech recognition performance are described in viz.,

(1) spontaneous speech effects significantly degrade recognition performance,

(2) fluent spontaneous speech yields word accuracies equivalent to read speech, and (3) using spontaneous speech training data. These can significantly improve the performance for recognizing spontaneous speech. It is concluded that word accuracy can be improved by explicitly modeling spontaneous effects in the recognizer, and by using as much spontaneous speech training data as possible. Inclusion of read speech training data, even within the task domain, does not significantly improve performance.

d) Robust speech recognition: To further increase the robustness of speech recognition systems, especially for spontaneous speech, utterance verification and confidence measures, are being intensively investigated . In order to have intelligent or human-like interactions in dialogue applications, it is important to attach to each recognized event a number

that indicates how confidently the ASR system can accept the recognized events. The confidence measure serves as a reference guide for a dialogue system to provide an appropriate response to its users. To detect semantically, significant parts and reject irrelevant portions in spontaneous utterances, a detection based approach has recently been investigated . The combined recognition and verification strategy work well especially for ill-formed utterances. In order to build acoustic models more sophisticated than conventional HMMs, the dynamic Bayesian network has recently been investigated . Around 2000, a QBPC, systems were developed to find the unknown and mismatch between training and testing conditions. A DCT fast subspace techniques has been proposed to approximate the KLT for autoregressive progress. A novel implementation of a mini-max decision rule for continuous density HMM-based Robust speech recognition is developed by combining the idea of mini-max decision rule with a normal viterbi search. Speech signal modeling techniques well suited to high performance and robust isolated word recognition have been contributed. The first robust Large vocabulary continuous speech recognition that uses syllable-level acoustic unit of LVCSR on telephone bandwidth speech is described in . In 2003, a novel regression based Bayesian predictive classification was developed for speech Hidden markov model. Wolfgang Rchichal has described the methods of improving the robustness and accuracy of the acoustic modeling using decision tree based state tying. Giuluva Garau

et.al., investigated on Large vocabulary continuous speech recognition. Xiong Xiao have shown a novel technique that normalizes the modulation spectra of speech signal. Kernel based nonlinear predictive coding procedure, that yields speech features which are robust to nonstationary noise contaminated speech signal. Features maximally in sensitive to additive noise are obtained by growth transformation of regression functions that span a reproducing a kernel Hilbert space (RKHS). Soundararajan proposed a supervised approach using regression trees to learn non linear transformation of the uncertainty from the 194 linear spectral domain to the cepstral domain. Experiments are conducted on Aurora-4 Database.

e) Multimodal speech recognition: Humans use multimodal communication when they speak to each other. Studies in speech intelligibility have shown that having both visual and audio information increases the rate of successful transfer of information, especially when the message is complex or when communication takes place in a noisy environment. The use of the visual face information, particularly lip information, in speech recognition has been investigated, and results show that using both types of information gives better recognition performances than using only the audio or only the visual information, particularly, in noisy environment. Jerome R., have developed Large Vocabulary Speech Recognition with Multi-span Statistical Language Models and the work done in this paper characterizes the behavior of such multi span modeling

in actual recognition. A novel subspace modeling is presented in including selection approach for noisy speech recognition. In subspace modeling, authors have developed a factor analysis representation of noisy speech i.e., a generalization of a signal subspace representation. They also explored the optimal subspace selection via solving the hypothesis test problems. Subspace selection via testing the correlation of residual speech, provides high recognition accuracies than that of testing the equivalent eigen-values in the minor subspace. Because of the environmental mismatch between training and test data severely deteriorates recognition performance. Jerome R. et.al., have contributed large vocabulary speech recognition with multispan statistical language model.

f) Modeling Techniques: Eduardo et.al.introduced a set of acoustic modeling and decoding techniques for utterance verification(UV) in HMM based Continuous Speech Recognition .Lawerence K et.al., discuss regarding HMM models for Automatic speech recognition which rely on high dimension feature vectors for summarizing the short time properties of speech. These have been achieved using some parameters choosen in two ways, namely i) to maximize the likelihood of observed speech signals, or ii) to minimize the number of classification errors. Dat Tat Tran have proposed various models namely,

i) the FE-HMM,NC-FE-HMM,FE-GMM,NC-FE- GMM,FE-VQ and NC-FE-VQ in the FE approach,

ii) the FCM-HMM, NC-FCM-HMM,FCM-GMM and NC-FCM- GMM in the FCM approach and iii) the hard HMM and GMM as the special models of both FE and FCM approaches for speech recognition. A new statistical approach namely the probabilistic union model for Robust speech recognition involving partial, unknown frequency band corruption are introduced by Ji Ming et.al. Jen Tzung et.al.have surveyed a series of model selection approaches with a presentation of a novel predictive information criterion for HMM selection. Yang Liu et.al. have shown that in a metadata detection scheme in speech recognition discriminative models outperform generative than predominant HMM approaches. Alba Sloin et.al. have presented a discriminative training algorithm, that uses support vector machines(SVM) to improve the classification of discrete and continuous output probability hidden markov models. The algorithm presented in the paper uses a set of maximum likelihood (ML) trained HMM models as a baseline system, and an SVM training scheme to rescore the results of the baseline HMMs. The experimental results given in that paper reduces the error rate significantly compared to standard ML training. Paper presents a discriminative training algorithm that uses support vector machines(SVMs) to improve the classification of discrete and continuous output probability hidden markov models(HMMs). The algorithm uses a set of maximum likelihood (ML) trained HMM models as a baseline system, and an SVM training scheme to rescore the results of the baseline HMMs. Paper, proposes a Fuzzy approach to the hidden

Markov model (HMM) method called the fuzzy HMM for speech and speaker recognition as an application of fuzzy expectation maximizing algorithm in HMM. This fuzzy approach can be applied to EM-style algorithms such as the Baum- Welch algorithm for hidden Markov models, the EM algorithm for Gaussian mixture models in speech and speaker recognition. Equation and how estimation of discrete and continuous HMM parameters based on this two algorithm is explained and performance of two methods of speech recognition for one hundred words is surveyed . This paper showed better results for the fuzzy HMM, compared with the conventional HMM. A novel method to estimate continuous-density hidden Markov model (CDHMM) for speech recognition is, according to the principle of maximizing the minimum multi-class separation margin. The approach is named large margin HMM. First, they showed that this type of large margin HMM estimation problem can be formulated as a constrained mini-max optimization problem. Second, they propose to solve this constrained mini-max optimization problem by using a penalized gradient descent algorithm, where the original objective function, i.e., minimum margin, is approximated by a differentiable function and the constraints are cast as penalty terms in the objective function. Ultimately paper showed that the large margin training method yields significant recognition error rate reduction even on top of some popular discriminative training methods. In the work, techniques for recognizing phonemes automatically by using

Hidden Markov Models (HMM) were proposed. The features input to the HMMs will be extracted from a single phoneme directly rather than from a string of phonemes forming a word. Also feature extraction techniques are compared to their performance in phoneme-based recognition systems. They also describe a pattern recognition approach developed for continuous speech recognition. Modeling dynamic structure of speech is a novel paradigm in speech recognition research within the generative modeling framework, and it offers a potential to overcome limitations of the current hidden Markov modeling approach. Analogous to structured language models where syntactic 195 structure is exploited to represent long-distance relationships among words the structured speech model described in this paper make use of the dynamic structure in the hidden vocal tract resonance space to characterize long-span contextual influence among phonetic units. The paper, discusses two novel HMM based techniques that segregate a speech segment from its concurrent background. The first method can be reliably used in clean environments while the second method, which makes use of the wavelets denoising technique, is effective in noisy environments. These methods have been implemented and they showed the superiority over other popular techniques, thus, indicating that they have the potential to achieve greater levels of accuracy in speech recognition rates. Paper is motivated by large margin classifiers in machine learning. It proposed a novel method to estimate continuous-density hidden Markov model (CDHMM) for speech

recognition according to the principle of maximizing the minimum multi-class separation margin. The approach is named as large margin HMM. First, it shows this type of large margin HMM estimation problem can be formulated as a constrained mini-max optimization problem. Second, it proposes to solve this constrained mini-max optimization problem by using a penalized gradient descent algorithm, where the original objective function, i.e., minimum margin, is approximated by a differentiable function and the constraints are cast as penalty terms in the objective function. The new training method is evaluated in the speaker independent isolated E-set recognition and the TIDIGITS connected digit string recognition tasks. Experimental results clearly show that the large margin HMMs consistently outperform the conventional HMM training methods. It has been consistently observed that the large margin training method yields significant recognition error rate reduction even on top of some popular discriminative training methods. Despite their known weaknesses, hidden Markov models (HMMs) have been the dominant technique for acoustic modeling in speech recognition for over two decades. Still, the advances in the HMM framework have not solved its key problems: it discards information about time dependencies and is prone to overgeneralization. Paper, has attempted to overcome the above problems by relying on straightforward template matching. It showed the decrease in word error rate with 17% compared to the HMM results. In automatic speech recognition, hidden Markov models (HMMs) are

commonly used for speech decoding, while switching linear dynamic models (SLDMs) can be employed for a preceding model based speech feature enhancement. These model types are combined in order to obtain a novel iterative speech feature enhancement and recognition architecture. It is shown that speech feature enhancement with SLDMs can be improved by feeding back information from the HMM to the enhancement stage. Two different feedback structures are derived. In the first, the posteriors of the HMM states are used to control the model probabilities of the SLDMs, while in the second they are employed to directly influence the estimate of the speech feature distribution. Both approaches lead to improvements in recognition accuracy both on the AURORA2 and AURORA-4 databases compared to non-iterative speech feature enhancement with SLDMs. It is also shown that a combination with uncertainty decoding further enhances performance. g) Noisy speech recognition: In 2008, a new approach for speech feature enhancement in the log spectral domain for noisy speech recognition is presented. A switching linear dynamic model (SLDM) is explored as a parametric model for the clean speech distribution. The results showed that the new SLDM approach can further improve the speech feature enhancement performance in terms of noise robust recognition accuracy. Jen-Tzung et.al. present a novel subspace modeling and selection approach for noisy speech recognition. Jianping Dingebal have presented a new approach for speech feature enhancement in the large spectral domain for NSR.

Xiaodong propose a novel approach which extends the conventional GMHMM by modeling state emission(mean and variance) as a polynomial function of a continuous environment dependent variable. This is used to improve the recognition performance in noisy environment by using multi condition training. Switching Linear dynamical system(SLDC), is a new model that combines both the raw speech signal and the noise was introduced in the year 2008. This was tested using isolated digit utterance corrupted by Gaussian noise. Contrary to Autoregressive HMMs,SLDC s outperforms a state of the art feature based HMM. Mark D.Skowronski proposed echo state network classifier by combining ESN with state machine frame work for noisy speech recognition. In the paper[144], authors propose a novel approach which extends the conventional Gaussian mixture hidden Markov model (GMHMM) by modeling state emission parameters (mean and variance) as a polynomial function of a continuous environment-dependent variable. At the recognition time, a set of HMMs specific to the given value of the environment variable is instantiated and used for recognition. The maximum-likelihood (ML) estimation of the polynomial functions of the proposed variable-parameter GMHMM is given within the expectation-maximization (EM) framework.

g) Data driven approach: A new approach for deriving compound words from a training corpus was proposed. The motivation for making compound words is because under some assumptions, speech recognition

errors occur less frequently in longer words were discussed along with the accurate modeling . They have also introduced a measure based on the product between the direct and the reverse bi-gram probability of a pair of words for finding candidate pairs in order to create compound words. Paper surveys a series of model selection approaches and presents a novel predictive information criterion (PIC) for hidden Markov model (HMM) selection. The approximate Bayesian using Viterbi approach is applied for PIC selection of the best HMMs providing the largest prediction information for generalization of future data. Authors have developed a top-down prior/posterior propagation algorithm for estimation of structural hyper-parameters and they have showed the evaluation of continuous speech recognition(data driven) using decision tree HMMs, the PIC criterion outperforms ML and MDL criteria in building a compact tree structure with moderate tree size and higher recognition rate. A method of compensating for nonlinear distortions in speech representation caused by noise was proposed which is based on the histogram equalization. Paper, introduces the data driven signal decomposition method based on the empirical mode decomposition(EMD) technique. The decomposition process uses the data themselves to derive the base function in order to decompose the one-dimensional signal into a finite set of intrinsic mode signals. The novelty of EMD is that the decomposition does not use any artificial data windowing which implies

fewer artifacts in the decomposed signals. The results show that the method can be effectively used in analyzing non-stationary signals.

4. Existing System

The concept of speech recognition started somewhere in 1940s, practically the first speech recognition program was appeared in 1953 at the bell labs, that was about recognition of a digit in a noise free environment. 1940s and 1950s consider as the foundational period of the speech recognition technology, in this period work was done on the of the speech recognition that is automation and information theoretic models. In the 1960's we were able to recognize small vocabularies (order of 10-100 words) of isolated words, based on simple acoustic-phonetic properties of speech sounds. The key technologies that were developed during this decade were, filter banks and time normalization methods. In 1970s the medium vocabularies order of 100-100 words using simple template-based, pattern recognition methods were recognized. In 1980s large vocabularies (1000-unlimited) were used and speech recognition problems based on statistical, with a large range of networks for handling language structures were addressed. The key invention of this era were hidden mark model (HMM) and the stochastic language model, which together enabled powerful new methods for handling continuous speech recognition problem efficiently and with high performance. In 1990s the key technologies developed during this period were the methods for stochastic language understanding, statistical learning of acoustic and language models, and the methods for implementation of large vocabulary

speech understanding systems. After the five decades of research, the speech recognition technology has finally entered marketplace, benefiting the users in variety of ways. The challenge of designing a machine that truly functions like an intelligent human is still a major one going forward.

Around 2000 a variational Bayesian clustering and estimation was developed.it was able to predicate on posterior distribution of parameter. Also Giuseppe Richardi have developed one more technology which name was automatic speech recognition and active learning algorithm for ASR.

4.1 challenges in Existing system

The current challenges of speech recognition are caused by two major factors – reach and loud environments. This calls for even more precise systems that can tackle the most ambitious ASR use-cases. Think about live interviews, speech recognition at a loud family dinner or meetings with various people. These are the upcoming challenges to be solved for next-gen voice recognition.

Beyond this, speech recognition needs to be made available for more languages and cover wide topics. Because as of now, **ASR needs a lot of data to work well** and some of it just hasn’t been collected for certain languages and topics. Without adding these, ASR systems will remain noticeably handicapped.

The use-case for voice assistants and Voice Powered User Interfaces (VUIs) is simple. They allow humans to give voice commands to machines, which these can translate into actions. As clear as the use-case appears to be, the best method for human-machine interactions is still being shaped. Naturally, this comes with challenges for speech recognition.

A. Imprecision and false interpretations

Speech recognition software isn't always able to interpret spoken words correctly. This is due to computers not being on par with humans in understanding the contextual relation of words and sentences, causing misinterpretations of what the speaker meant to say or achieve.

Comparing humans and VUIs, the speech recognition systems are lacking millennia of contextual experience and **VUIs still encounter challenges when trying to understand the semantics of a sentence.**

A. Time and lack of efficiency

We'd usually assume that computerizing a process would speed it up. Unfortunately, this is not always the case when it comes to voice

recognition systems. In many cases using a voice app takes up more time than going with a traditional text-based version.

This is mainly due to the diverse voice patterns of humans, which VUIs are still learning to adapt to. Hence, users often need to adjust by slowing down or being more precise than normal in their pronunciation.

B. Accent and local difference

VUIs are oftentimes challenged when voice inputs divert too much from the average. Especially accents can pose a big challenge. While systems are getting better there's still a big difference in **their ability to understand American or Scottish English** for example. Even a simple cold can be a reason for voice commands not to work as well as usual.

C. Background noise and loud environment

To make the most of VUIs a quiet environment helps a lot. Whenever there is too much background noise speech recognition will be challenged. Making it especially hard to use them effectively in the urban outdoors or large public spaces/offices. With the use of specific

microphones or headsets, the limitations can be decreased but it requires an additional device, which is never desirable.

Gap between machine and human speech recognition:

What we know about human speech processing is still very limited, and we have yet to witness a complete and worthwhile unification of the science and technology of speech. In 1994, presented the following 20 themes which is believed to be an important to the greater understanding of the nature of speech and mechanisms of speech pattern processing in general:

How important is the communicative nature of speech?

- Is human-human speech communication relevant to human machine communication by speech?
- Speech technology or speech science?
- (How can we integrate speech science and technology). Whither a unified theory? Is speech special?
- Why is speech contrastive?
- Is there random variability in speech?
- How important is individuality?
- Is disfluency normal?
- How much effort does speech need?

- What is a good architecture (for speech processes)?
- What are suitable levels of representation?
- What are the units? What is the formalism?
- How important are the physiological mechanisms?
- Is time-frame based speech analysis sufficient?
- How important is adaptation?
- What are the mechanisms for learning?
- What is speech good for?
- How good is speech. After more than 10 years, we still do not have clear answers to these 20 questions.

5.Proposed System

Speech recognition has its roots in research done at Bell Labs in the early 1950s. Early systems were limited to a single speaker and had limited vocabularies of about a dozen words. Modern speech recognition systems have come a long way since their ancient counterparts. They can recognize speech from multiple speakers and have enormous vocabularies in numerous languages. The first component of speech recognition is, of course, speech. Speech must be converted from physical sound to an electrical signal with a microphone, and then to digital data with an analog-to-digital converter. Once digitized, several models can be used to transcribe the audio to text Most modern speech recognition systems rely on what is known as a Hidden Markov Model (HMM). This approach works on the assumption that a speech signal, when viewed on a short enough timescale (say, ten milliseconds), can be reasonably approximated as a stationary process—that is, a process in which statistical properties do not change over time.

In a typical HMM, the speech signal is divided into 10-millisecond fragments. The power spectrum of each fragment, which is essentially a plot of the signal's power as a function of frequency, is mapped to a vector of real numbers known as cepstral coefficients. The dimension of this vector is usually small—sometimes as low as 10, although more accurate

systems may have dimension 32 or more. The final output of the HMM is a sequence of these vectors.

To decode the speech into text, groups of vectors are matched to one or more phonemes—a fundamental unit of speech. This calculation requires training, since the sound of a phoneme varies from speaker to speaker, and even varies from one utterance to another by the same speaker. A special algorithm is then applied to determine the most likely word (or words) that produce the given sequence of phonemes.

One can imagine that this whole process may be computationally expensive. In many modern speech recognition systems, neural networks are used to simplify the speech signal using techniques for feature transformation and dimensionality reduction *before* HMM recognition. Voice activity detectors (VADs) are also used to reduce an audio signal to only the portions that are likely to contain speech. This prevents the recognizer from wasting time analyzing unnecessary parts of the signal.

Fortunately, as a Python programmer, you don't have to worry about any of this. A number of speech recognition services are available for use online through an API, and many of these services offer Python SDKs.

In a typical HMM, the speech signal is divided into 10-millisecond fragments. The power spectrum of each fragment, which is essentially a plot of the signal's power as a function of frequency, is mapped to a vector

of real numbers known as cepstral coefficients. The dimension of this vector is usually small—sometimes as low as 10, although more accurate systems may have dimension 32 or more. The final output of the HMM is a sequence of these vectors. Recognizing speech requires audio input, and Speech Recognition makes retrieving this input really easy. Instead of having to build scripts for accessing microphones and processing audio files from scratch, Speech Recognition will have you up and running in just a few minutes.

The Speech Recognition library acts as a wrapper for several popular speech APIs and is thus extremely flexible. One of these—the Google Web Speech API—supports a default API key that is hard-coded into the Speech Recognition library. That means you can get off your feet without having to sign up for a service.

The flexibility and ease-of-use of the Speech Recognition package make it an excellent choice for any Python project. However, support for every feature of each API it wraps is not guaranteed. You will need to spend some time researching the available options to find out if Speech Recognition will work in your particular case. This system take audio as input and convert it into text and if it is valid input so it open the any search engine like: chrome ,internet explorer and being and this system is also used for perform task on notepad and word.it is different from existing one because it will ask in which search engine you want to search

this text if you will provide the particular one than it will search on that. Otherwise google will be by default one.

6.Uml Diagram

A.Use case diagram

A use case diagram is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as well.

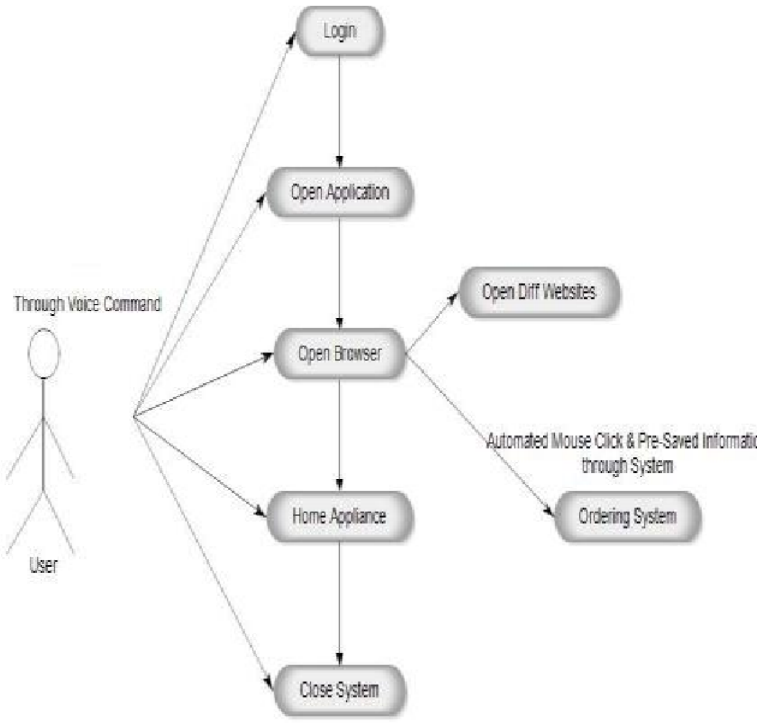


Fig-2 Use case diagram

B.SEQUENCE DIAGRAM

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart.

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

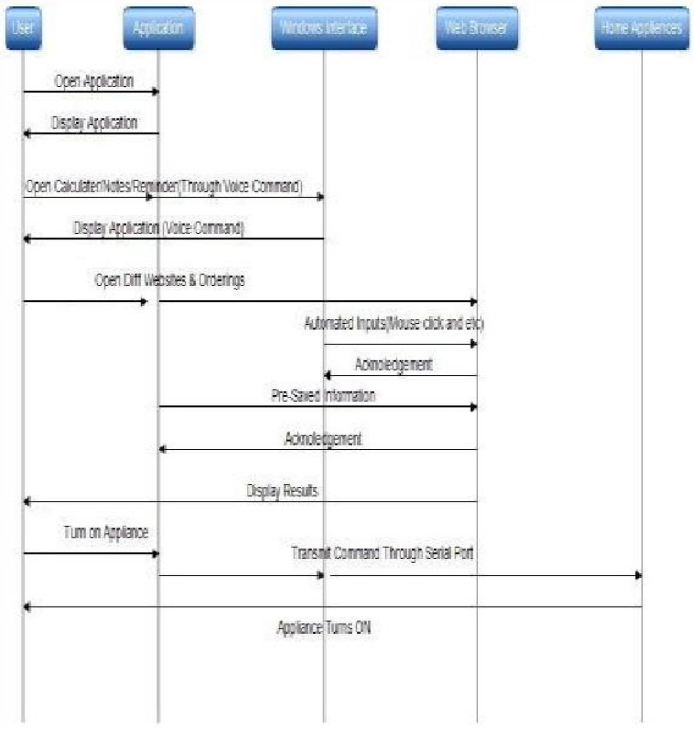


Fig-3. SEQUENCE DIAGRAM

C.ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe dynamic aspects of the system.

An activity diagram is a simple and intuitive illustration of what happens in a workflow, what activities can be done in parallel, and whether there are alternative paths through the workflow. Activity diagrams represent the business and operational work flows of a system. An Activity diagram is a dynamic diagram that shows the activity and the event that causes the object to be in the particular state.

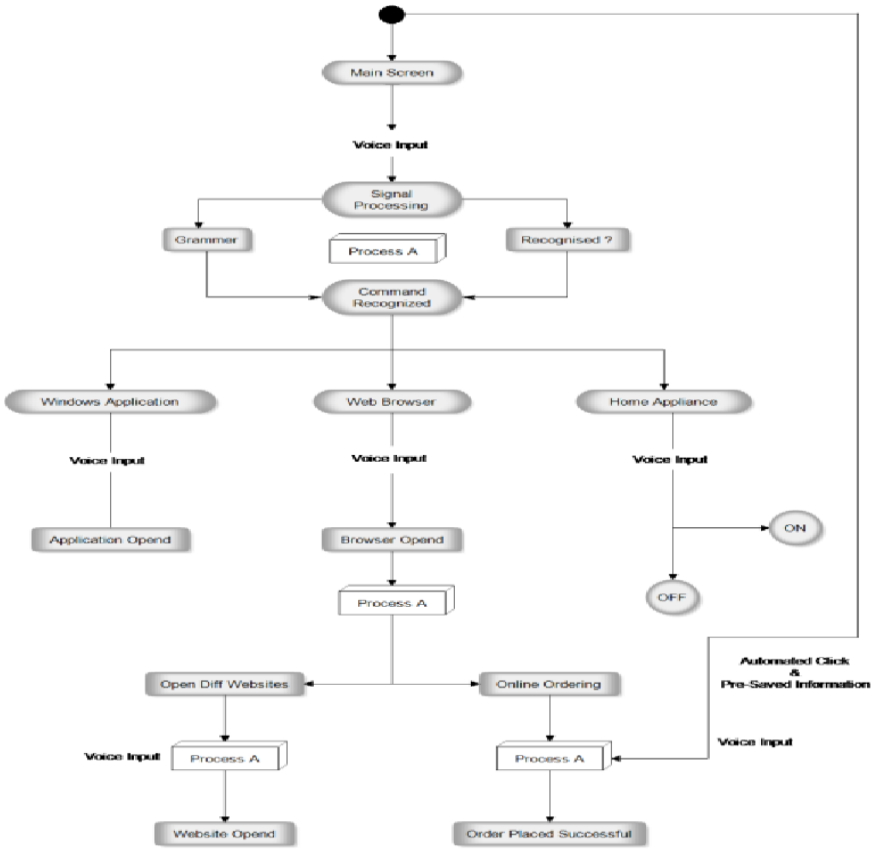


Fig-4. ACTIVITY DIAGRAM

D. E –R DIAGRAM

An ER model is an abstract way of describing a database. In the case of a relational database, which stores data in tables, some of the data in these tables point to data in other tables - for instance, your entry in the database could point to several entries for each of the phone numbers that are yours. The ER model would say that you are an entity, and each phone number is an entity, and the relationship between you and the phone numbers is 'has a phone number'. Diagrams created to design these entities and relationships are called entity– relationship diagrams or ER diagrams.

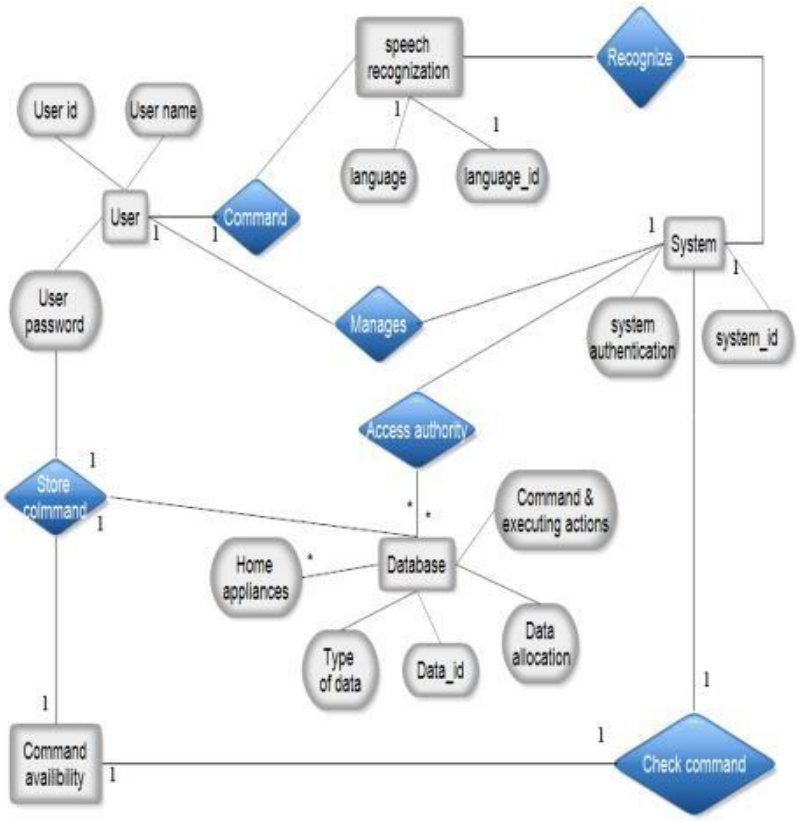


Fig-5 E –R DIAGRAM

E.FLOW DIAGRAM

The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. o Class diagrams can also be used for data modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed.

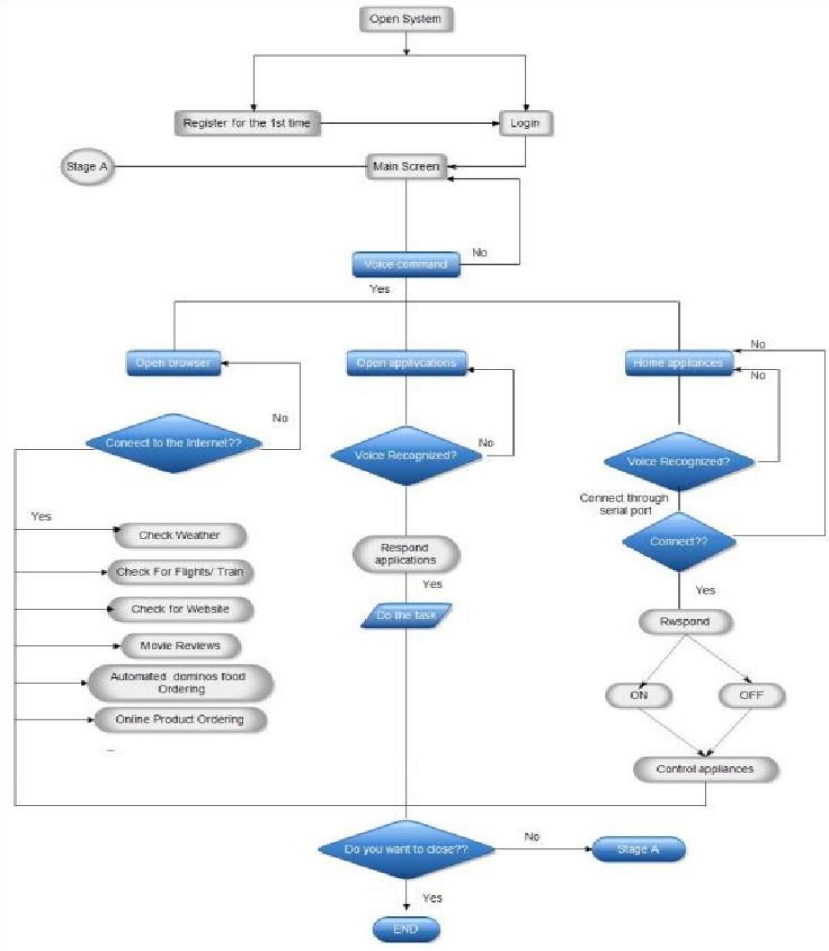


Fig-6.Flow diagram

7.Software Requirement

The software used for the development of the project is:

Operating system: Windows

Programming Language: Python (3.6)

IDE: Anaconda Navigator (Jupiter)

Package: Speech Recognition Library

Hardware components: Network communication Modem for connecting to internet, Connecting Wires.

Component	Minimum	Recommended
CPU	1.6 GHz	2.53GHz
RAM	2GB	4GB
Microphone	Mic	High quality
Sound card	Sound card	Sound card with very clear signal

8.Implementation/Architecture

Speech recognition approaches:- There mainly three approaches to speech recognition.

1. **Acoustic Phonetic Approach**
2. **Pattern Recognition Approach**
3. **Artificial Intelligence Approach**

1.Acoustic phonetic approach:

The earliest approaches to speech recognition were supported finding speech sounds and to providing appropriate labels to those sounds. This is often the thought of the acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units (phonemes) in speech which these units are broadly characterized by a group of acoustics properties that are manifested within the speech signal over time. albeit , the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds it's assumed within the acoustic-phonetic approach that the principles governing the variability are straightforward and may be readily learned by a machine.

2.Artificial Intelligence approach (Knowledge Based approach)

The Artificial Intelligence approach [97] is a hybrid of the acoustic phonetic approach and pattern recognition , approach. In this, it

exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing. In its pure form, knowledge engineering design involves the direct and explicit incorporation of experts speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge ‗ phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic

modeling. This form of knowledge application makes an important distinction between knowledge and algorithms _ Algorithms enable us to solve problems. Knowledge enable the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

3.Pattern Recognition Approach :

The pattern- Recognition approach involves two essential steps its name is, pattern training and pattern comparison. The most important feature of this approach is that it uses a well formulated mathematical framework and establishes consistent pronunciation representations, for reliable pattern comparison, from a group of labeled training samples via a proper training algorithm. A accent pattern representation are often within the sort of a speech template or a statistical model eg-(HIDDEN MARKOV MODEL or HMM and neural network).

3.1 Hidden markov model: -

Hidden markov model is very basic and important technology which is developed in approx. 1980s.It is a class of probabilistic graphical model. HMM is allow to predict the unknown sequence variable from the set of observed variable. In HMM we know the present state and we do not need

any historical information to predict the future. HMM is a doubly model means it is not observable but it can be often observed through another model that produces sequence of observations. HMM was documented and understood during a few laboratories(primary in IBM). Hidden states are the sequence of phonemes that we are looking to recognize. A phoneme based HMM for say the word 'RAT' would have /R/ /A/ and /T/ as states. In this approach, we will need to create a HMM for every word in the corpus and train it to with the utterances of the word to strengthen the model. A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters; the challenge is to determine the hidden parameters from the observable data. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions.

That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state. HMMs are more popular because they can be trained automatically and are simple and computationally feasible to use HMM considers the speech signal as quasi- static for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with, usually, mixtures of multivariate Gaussian distributions (the state output distributions). The parameters of the model are the state transition probabilities and the means, variances and mixture weights that characterize the state output distributions [10]. This uses theory from statistics in order to (sort of) arrange our feature vectors into a Markov matrix (chains) that stores probabilities of state transitions. That is, if each of our code words were to represent some state, the HMM would follow the sequence of state changes and build a model that includes the probabilities of each state progressing to another state. HMM can be characterized by following when its observations are discrete: i. N is number of states in given model, these states are hidden in model. ii. M is the number of distinct observation symbols correspond to the physical output of the certain

model. iii. A is a state transition probability distribution defined by NxN matrix as shown in equation (4).

$$A = \{aij\}$$

$$aij = p\{ qt+1 = j/qt = i\},\ 1 \leq i, j \leq Nn \quad (4)$$

$$\sum aij = 1,\ 1 \leq i, j \leq Nn$$

Where qt occupies the current state. Transition probabilities should meet the stochastic limitations B is observational symbol probability distribution matrix (3) defined by NxM matrix equation comprises

$$bj(k) = p\{ot=vk|qt=j\},\ 1<=j<=N\ ,\ 1<=k<=M$$

$$\sum bj(k) = 1,\ 1<=k<=M$$

3.2 Neural Networks (NN)-

Artificial neural network is based on the human brain we know that there are million of neurons are present in human brain. Its used for processing and neurons communicate by sending signal to each other through very complex connection. Similarly neural network also use for data processing and communicate through connection. In neural network every connection consist of weight. The value of weight can be positive or negative. Positive weights activate the neuron while negative weights inhibit it. Until now, this is the most successful and most used pattern recognition method for speech recognition. It's a mathematical model derived from a Markov Model. Speech recognition uses a slightly adapted

Markov Model. Speech is split into the smallest audible entities (not only vowels and consonants but also conjugated sound like ...). All these entities are represented as states in the Markov Model. As a word enters the Hidden Markov Model it is compared to the best suited model (entity). According to transition probabilities there exist a transition from one state to another. For example: the probability of a word starting with is almost zero. A state can also have a transition to it's own if the sound repeats itself. Markov Models seems to perform quite well in noisy environments because every sound entity is treated separately. If a sound entity is lost in the noise, the model might be able to guess that entity based on the probability of going from one sound entity to another. C. Neural Networks (NN) Neural networks have many similarities with Markov models. Both are statistical models which are represented as graphs. Where Markov models use probabilities for state transitions, neural networks use connection strengths and functions. A key difference is that neural networks are fundamentally parallel while Markov chains are serial. Frequencies in speech, occur in parallel, while syllable series and words are essentially serial. This means that both techniques are very powerful in a different context. As in the neural network, the challenge is to set the appropriate weights of the connection, the Markov model challenge is finding the appropriate transition and observation probabilities. In many speech recognition systems, both techniques are implemented together and work in a symbiotic relationship. Neural networks perform very well

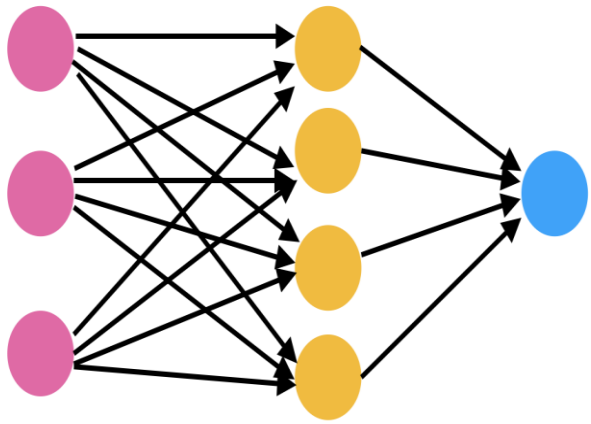
at learning phoneme probability from highly parallel audio input, while Markov models can use the phoneme observation probabilities that neural networks provide to produce the likeliest phoneme sequence or word. This is at the core of a hybrid approach to natural language understanding.

It consist three type of layer :-

1.Input layer

2.Hidden Layer

3.Output Layer



Input layer **Hidden layer** **Output layer**

Fig.7 Basic diagram of Neural network

1.Input Layer-It take the input from computer system or external environment. The main purpose of this layer is to deal with input and send to hidden layer.

2.Hidden Layer-It is collection of neuron and activation function applied on it.Its act as intermediate between input layer and output layer. Its job is process the input which get from input layer. This layer is responsible for extracting feature of input.

3.Output Layer-This layer collect and transmit all the information. The no of neurons in output layer will be directly related to which types of work that neuron network is performing.

Neural network is very similary to HMM both represent the graph because they are statistical model.Markav model uses probabilities for state transition, neural networks uses connection strength and function.The main difference between both are neural network is fundamentally parallel but Markov chain are serial.we know that Frequencies in speech.occur in parallel, while syllable series and words are essentially serial.so we can say that both this technique are very powerful during different context. As within the neural network, the challenge is to line the acceptable weights of the connection, the Markov model challenge is finding the acceptable transition and observation probabilities. Many speech recognition system for better result both this techniques are

implemented together. Neural networks perform alright at learning phoneme probability from highly parallel audio input, while Markov models can use the phoneme observation probabilities that neural networks provide to supply the likeliest phoneme sequence or word.

6.2.1 implementation of signal-preprocessing

In the previous section we have discussed the general structure of a speech recognition system. In this paper we put the main focus on the neural networks and not on the signal pre-processing, although signal pre-processing has a big impact on the performance of the speech classifier. It is important to feed the neural network with normalized input. Recorded samples never produce identical waveforms; the length, amplitude, background noise may vary. Therefore we need to perform signal pre-processing to extract only the speech related information. This means that using the right features is crucial for successful classification. Good features simplify the design of a classifier whereas weak features (with little discrimination power) can hardly be compensated with any classifier. We can divide this process on some distinctive steps like: As the neural network will have to do the speech classification, it is very important to feed the network inputs with relevant data. It’s obvious that an appropriate preprocessing is necessary in order to be sure that the input of the neural network is characteristic for every word while having a small spread amongst samples of the same word. Noise and difference in

amplitude of the signal can distort the integrity of a word while timing variations can cause a large spread amongst samples of the same word [5],[6]. These problems are dealt with in the signal preprocessing part which is composed of different sub stages: Filtering, Entropy based endpoint detection and Mel Frequency Cepstrum Coefficients. Filtering stage - samples are recorded with a standard microphone. So they contain besides speech signals a lot of distortion and noise due to the quality of the microphone or just because of picked up background noise. In this first step we perform some digital filtering to eliminate low and high frequency noise. As speech is situated in the frequency domain between 300 Hz and 3750 Hz, a bandpass filtering is performed on the input signal. This is done by passing the input signal successively trough a FIR low pass filter and then through a FIR high pass filter. An FIR filter has the advantage above an IIR filter that it has a linear phase response.

For more information on the SpeechRecognition package:

- Library reference
- Examples
- Troubleshooting page

A few interesting internet resources:

• Behind the Mic: The Science of Talking with Computers. A short film about speech processing by Google.

• A Historical Perspective of Speech Recognition by Huang, Baker and Reddy. Communications of the ACM (2014). This article provides an in-depth and scholarly look at the evolution of speech recognition technology.

• The Past, Present and Future of Speech Recognition Technology by Clark Boyd at The Startup. This blog post presents an overview of speech recognition technology, with some thoughts about the future. Some good books about speech recognition:

• The Voice in the Machine: Building Computers That Understand Speech, Pieraccini, MIT Press (2012). An accessible general-audience book covering the history of, as well as modern advances in, speech processing.

• Fundamentals of Speech Recognition, Rabiner and Juang, Prentice Hall (1993). Rabiner, a researcher at Bell Labs, was instrumental in designing some of the first commercially viable speech recognizers. This book is now over 20 years old, but a lot of the fundamentals remain the same.

• Automatic Speech Recognition: A Deep Learning Approach, Yu and Deng, Springer (2014). Yu and Deng are researchers at Microsoft and both very active in the field of speech processing. This book covers a lot

of modern approaches and cutting-edge research but is not for the mathematically faintof-heart.

Convert an audio file into text A.Steps:

1. Import Speech recognition library

2. Initializing recognizer class in order to recognize the speech. We are using google speech recognition.

3. Audio file supports by speech recognition: wav, AIFF, AIFF-C, FLAC. I used ‘wav’ file in this example

4. I have used ‘taken’ movie audio clip which says “I don’t know who you are I don’t know what you want if you’re looking for ransom I can tell you I don’t have money”

5. By default, google recognizer reads English. It supports different languages, for more details please check this documentation. . All of the magic in SpeechRecognition happens with the Recognizer class. The primary purpose of a Recognizer instance is, of course, to recognize speech. Each instance comes with a variety of setting.

All of the magic in SpeechRecognition happens with the Recognizer class.

The primary purpose of a Recognizer instance is, of course, to recognize speech. Each instance comes with a variety of settings and functionality for recognizing speech from an audio source.

Creating a Recognizer instance is easy. In your current interpreter session, just type:

```
>>> r = sr.Recognizer()
```

Each Recognizer instance has seven methods for recognizing speech from an audio source using various APIs. These are:

- recognize_bing(): Microsoft Bing Speech
- recognize_google(): Google Web Speech API
- recognize_google_cloud(): Google Cloud Speech - requires installation of the google-cloud-speech package
- recognize_houndify(): Houndify by SoundHound
- recognize_ibm(): IBM Speech to Text
- recognize_sphinx(): CMU Sphinx - requires installing PocketSphinx
- recognize_wit(): Wit.ai

Of the seven, only recognize_sphinx() works offline with the CMU Sphinx engine. The other six all require an internet connection.

A full discussion of the features and benefits of each API is beyond the scope of this tutorial. Since SpeechRecognition ships with a default API key for the Google Web Speech API, you can get started with it right away. For this reason, we'll use the Web Speech API in this guide. The other six APIs all require authentication with either an API key or a username/password combination. For more information, consult the SpeechRecognition docs.

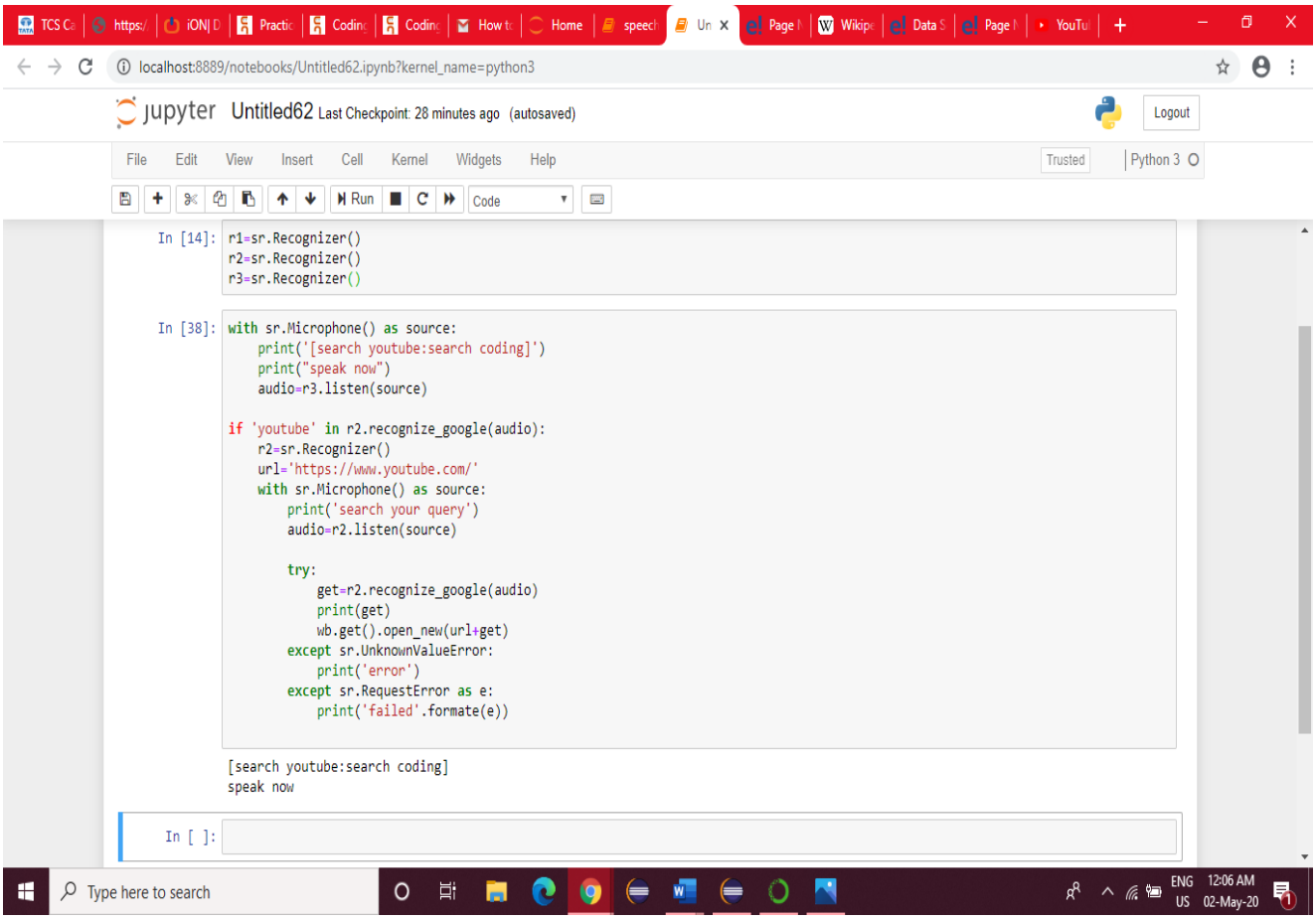
Install Library

```
$ pip install pyaudio
$ python -m speech_recognition
```

Importing packages

```
>>> import speech_recognition as sr
>> import webbrowser as wb
>> r1=sr.Recognizer()
>> r2=sr.Recognizer()
>> r3=sr.Recognizer()
```

Input-it will take voice as input


localhost:8889/notebooks/Untitled62.ipynb?kernel_name=python3
jupyter Untitled62 Last Checkpoint: 28 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [14]:
r1=sr.Recognizer()
r2=sr.Recognizer()
r3=sr.Recognizer()
In [38]:
with sr.Microphone() as source:
print('[search youtube:search coding]')
print("speak now")
audio=r3.listen(source)
if 'youtube' in r2.recognize_google(audio):
r2=sr.Recognizer()
url='https://www.youtube.com/'
with sr.Microphone() as source:
print('search your query')
audio=r2.listen(source)
try:
get=r2.recognize_google(audio)
print(get)
wb.get().open_new(url+get)
except sr.UnknownValueError:
print('error')
except sr.RequestError as e:
print('failed'.formate(e))
[search youtube:search coding]
speak now
In [ ]:
Type here to search
ENG
US
12:06 AM
02-May-20

9.CODE

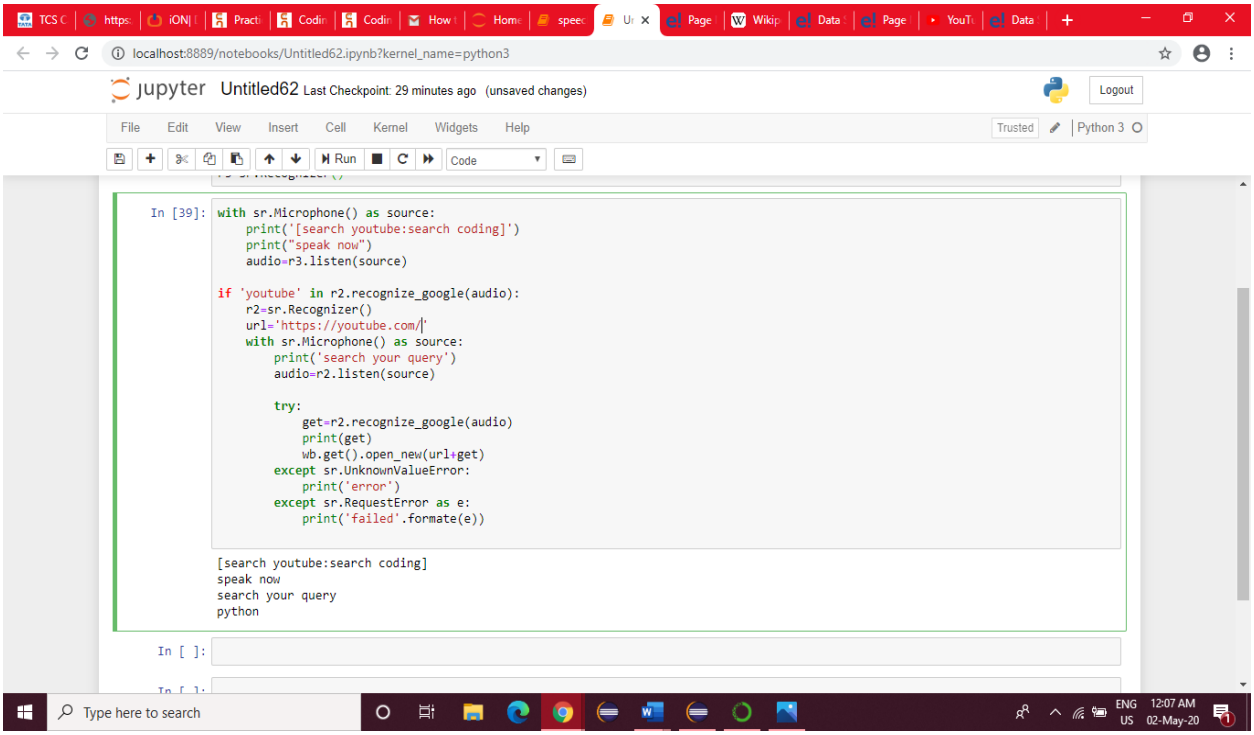
```
import speech_recognition as sr
import webbrowser as wb
r1=sr.Recognizer()
r2=sr.Recognizer()
r3=sr.Recognizer()
with sr.Microphone() as source:
    print('[search edureka:search youtube]')
    print("speak now")
    audio=r3.listen(source)

if 'video' in r2.recognize_google(audio):
    r2=sr.Recognizer()
    url='https://www.youtube.com/'
    with sr.Microphone() as source:
        print('search your query')
        audio=r2.listen(source)
```

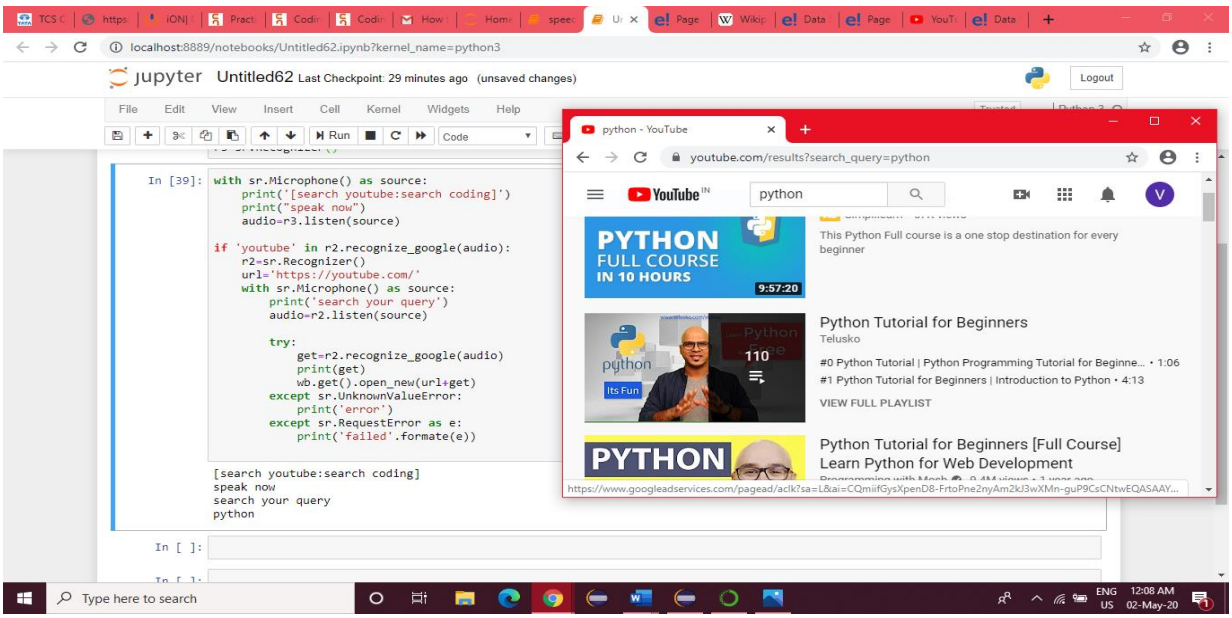
```
try:
    get=r2.recognize_google(audio)
    print(get)
    wb.get().open_new(url+get)
except sr.UnknownValueError:
    print('error')
except sr.RequestError as e:
    print('failed'.formate(e))
```

10.Output/Result

Step-1 It convert audio into text



Step-2 Than send it into search bar here google is search bar



Here our input is audio and we want to search python tutorial on YouTube so in microphone I spoke YouTube and than it will say wat you want to search in YouTube I said python so it will do open all python tutorial on YouTube.

Our neural network can predict that one likely thing “BBBB___EEE___AAA___RR”. But it also thinks that it was possible that I said “” or even “BBBB___AAA____RRRR___EEEE”.

We have lot of technology to the help of them we can remove the repeated character to single character.

- BBBB___EEE___AAA___RR becomes BB_E_A_R.
- BBBB___AAA____RRRR___EEEE becomes B_A_R_E.

We can also remove the blanks

- BB_E_A_R.becomes BEAR,B_A_R_E becomes BARER

11.Conclusion

We know that Speech is most primary mode for communication between people. So in this paper we are using various new technology for creating the system which can automate the work based on speech recognition. Simply we can say that, in this system machine communicate with human. In this system computer takes voice as a input from user and convert this into text. After this system sent this text into search engine(google, being) and find out the result. Neural network is very useful here. A small set of words could be recognized with some very simplified models. Pre-processing is most important quality on this it also provide biggest impact on neural network performance. This project is showing that neural networks can be very powerful speech signal classifiers. A small set of words could be recognized with some very simplified models. The pre-processing quality is giving the biggest impact on the neural networks performance. In some cases where the spectrogram combined with entropy based endpoint detection is used we observed poor classification performance results, making this combination as a poor strategy for the pre-processing stage. On the other hand we observed that Mel Frequency Kepstrum Coefficients are a very reliable tool for the pre-processing stage, with the good results they provide. Both the Multilayer Feedforward Network with backpropagation algorithm and the Radial Basis Functions Neural Network are achieving satisfying results when Mel Frequency Kepstrum Coefficients are used.Speech is the primary,

and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. we have also encountered a number of practical limitations which hinder a widespread deployment of application and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited. Although these areas of investigations are important the significant advances will come from studies in acousticphonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving knowledge required for natural conversation. This paper attempts to provide a comprehensive survey of research on speech recognition and to provide some year wise progress to this date. Although significant progress has been made in the last two decades, there is still work to be done, and we believe that a robust speech recognition system should be effective under full variation in: environmental conditions, speaker variability s etc. Speech Recognition is a challenging and interesting problem in and of itself. We have attempted in this paper to provide a comprehensive cursory, look and review of how much speech

recognition technology progressed in the last 60 years. Speech recognition is one of the most integrating areas of machine intelligence, since, humans do a daily activity of speech recognition. Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction. We hope this paper brings about understanding and inspiration amongst the research communities of ASR.

12.References

[1] Amos, D. (2018). The Ultimate Guide To Speech Recognition With Python. Retrieved from Real python: https://realpython.com/python-speech-recognition/

[2] Ashok Kumar, V. M. (April 2019). Speech Recognition: A Complete Perspective. International Journal of Recent Technology and Engineering (IJRTE).

[3] Geitgey, A. (2016, Dec 24). Machine Learning is Fun Part 6: How to do Speech Recognition with Deep Learning. Retrieved from Medium: https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a

[4] M.A.Anusuya, S. (nov 2009). Speech Recognition by Machine: A Review. (IJCSIS) International Journal of Computer Science and Information Security,.

[5] Reetu Kumari, R. c. (2017). SPEECH AUTOMIZATION. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 353.

[6] Zegers, P. (1998). SPEECH RECOGNITION USING NEURAL NETWORKS. *DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING*.

[7] “Speech recognition- The next revolution” 5th edition.

[8] Ksenia Shalonova, “Automatic Speech Recognition” 07 DEC 2007 Source:http://www.cs.bris.ac.uk/Teaching/Resources/COMS12303/lectures/Ksenia_Shalonoa- Speech_Recognition.pdf

[9] "Fundamentals of Speech Recognition". L. Rabiner & B. Juang. 1993. ISBN: 0130151572.

[10] http://www.abilityhub.com/speech/speech-description.htm

[11]Charu Joshi “Speech Recognition” Source: http://www.scribd.com/doc/2586608/speechrecognition.pdf

[12] John Kirriemuir “Speech recognition technologies”

[13]http://electronics.howstuffworks.com/gadgets/high-techgadgets/speechrecognition.htm

[14] https://realpython.com/python-speech-recognition/

[15] (Saxena, 2017). Artificial Neuron Networks(Basics) | Introduction to Neural Networks.https://becominghuman.ai/artificial-neuron-networks-basics-introduction-to-neural-networks-3082f1dcca8c

[16] Data mining approach and security over ddos attacks M Arvindhan, Bhanu Prakash Ande ISSN: 2229-6956 ,(ONLINE) DOI: 10.21917/ijsc.2020.0292, ICTACT JOURNAL ON SOFT COMPUTING, JANUARY 2020, VOLUME: 10, ISSUE: 02

[17] Scheming an Proficient Auto Scaling Technique for Minimizing Response Time in Load Balancing on Amazon AWS Cloud M Arvindhan, A Anand - Available at SSRN 3390801, 2019

[18] Clustering algorithm networks test cost sensitive for specialist divisions , clustering algorithm networks test cost sensitive for specialist divisions, doi: 10.21917/ijivp.2019.0298 volume: 10, issue: 02