NEW HORIZON
COLLEGE OF ENGINEERING
New Horizon Knowledge Park, Ring Road, Marathalli
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade, Accredited by NBA

"VOICE RECOGNITION"

A MINI PROJECT REPORT

Submitted by

PRAKRITI SHARMA K P [1NH18IS078]

Under the guidance of,

Dr. R J Anandhi

Head of Department (HOD), ISE, NHCE

In partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

FOR

COURSE NAME: MINI PROJECT

COURSE CODE: 19ISE49

NEW HORIZON
COLLEGE OF ENGINEERING
New Horizon Knowledge Park, Ring Road, Marathalli
Autonomous College Permanently Affiliated to VTU, Approved by AICTE & UGC
Accredited by NAAC with 'A' Grade, Accredited by NBA

CERTIFICATE

Certified that the project work entitled Auto-Attendance System Using Face Recognition carried out by Ms. PRAKRITI SHARMA K P, bearing USN 1NH18IS078, a bonafide student of 4th semester in partial fulfillment for the award of Bachelor of Engineering in Information Science & Engineering of the Visveswaraiah Technological University, Belagavi during the year 2019-20. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated. The project report has been approved as it satisfies the academic requirements in respect of Mini Project work prescribed for the said Degree.

Name & Signature of HOD and guide	**Name & Signature of Principal**
Dr. R J Anandhi	Dr. Manjunatha

Examiners :

Name	**Signature**
1. ……	……..
2. ……	……

VOICE RECOGNITION

ORIGINALITY REPORT

30%	%	30%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Rabiner, Lawrence, and Biing-Hwang Juang. "Speech Recognition by Machine", Electrical Engineering Handbook, 2009. Publication	3%
2	"Stochastic Approaches", Elsevier BV, 1990 Publication	2%
3	Parnyan Bahrami Dashtaki. "An Investigation into Methodology and Metrics Employed to Evaluate the (Speech-to-Speech) Way in Translation Systems", Modern Applied Science, 2017 Publication	2%
4	"Knowledge-Based Approaches", Elsevier BV, 1990 Publication	2%

5	Otsuki, Katsutoshi. "A study on automatic transcription and indexing of broadcast news =Hoso nyusu no jido onsei ninshiki to indekushingu ni kansuru kenkyu", DSpace at Waseda University, 2009.	2%
	Publication	
6	Atul D. Narkhede, Milind. U. Nemade. "Isolated digit recognition using wavelet transform and soft computing technique: A survey", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016	1%
	Publication	
7	Minh-Son Nguyen, Tu-Lanh Vo. "Vietnamese Voice Recognition for Home Automation using MFCC and DTW Techniques", 2015 International Conference on Advanced Computing and Applications (ACOMP), 2015	1%
	Publication	
8	"Interface Devices and Systems", Industrial Robots Programming, 2007	1%
	Publication	
9	"Connectionist Approaches", Elsevier BV, 1990	1%

Publication

---

10 Nair, Rekha, and Nirmala Salam. "A reliable speaker verification system based on LPCC and DTW", 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014.

Publication

1%

---

11 M. Gomathy, K. Meena, K. R. Subramaniam. "Classification of speech signal based on gender: a hybrid approach using neuro-fuzzy systems", International Journal of Speech Technology, 2011

Publication

1%

---

12 Madhu Singh, Amrutha Reddy Konala. "Speech controlled 3D robot to arrange things", 2018 International Symposium on Devices, Circuits and Systems (ISDCS), 2018

Publication

1%

---

13 Arijit Ghosal, Suchibrota Dutta. "Automatic male-female voice discrimination", 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014

1%

Publication

14	Madhavi R. Repe, S.D. Shirbahadurkar, Smita Desai. "Prosody Model for Marathi Language TTS Synthesis with Unit Search and Selection Speech Database", 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, 2010 Publication	1%
15	Cristina Romero-Gonzalez, Jesus Martinez-Gomez, Ismael Garcia-Varea. "Spoken language understanding for social robotics", 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), 2020 Publication	1%
16	"Speech Technology", Springer Science and Business Media LLC, 2010 Publication	1%
17	Sadaoki Furui. "Speech-Based Interfaces", Elsevier BV, 2007 Publication	1%

18	Filipe Neves Dos Santos. "A collaborative, non-invasive hybrid semantic localization and mapping system (HySeLAM)", Repositório Aberto da Universidade do Porto, 2014. Publication	<1%
19	Tee Wilkin, Ooi Shih Yin. "State of the art: Signature verification system", 2011 7th International Conference on Information Assurance and Security (IAS), 2011 Publication	<1%
20	Rohit K. Dubey, Tyler Thrash, Mubbasir Kapadia, Christoph Hoelscher, Victor R. Schinazi. "Information Theoretic Model to Simulate Agent-Signage Interaction for Wayfinding", Cognitive Computation, 2019 Publication	<1%
21	Hua-An Zhao, Kazuma Uchida. "A de-noising algorithm for voice recognition with low SNR", 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), 2016 Publication	<1%

22	Stevan Milinkovic. "Pure software-based speech recognition for OS-less embedded systems", 2014 3rd Mediterranean Conference on Embedded Computing (MECO), 2014 Publication	<1%
23	"Global Trends in Information Systems and Software Applications", Springer Science and Business Media LLC, 2012 Publication	<1%
24	Dewan Tanvir Ahmed, Shervin Shirmohammadi. "A decision support engine for video surveillance systems", 2011 IEEE International Conference on Multimedia and Expo, 2011 Publication	<1%
25	S. Supriya, S. M. Handore. "Speech recognition using HTK toolkit for Marathi language", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017 Publication	<1%

26	Daming Wei. "Walking stability analysis by age based on Dynamic Time Warping", 2008 8th IEEE International Conference on Computer and Information Technology, 07/2008 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

ABSTRACT

Voice is the basic, common and efficient form of communication method for people to interact with each other. Today speech technologies are commonly available for a limited but interesting range of task. This technologies enable machines to respond correctly and reliably to human voices and provide useful and valuable services. As communicating with computer is faster using voice rather than using keyboard, so people will prefer such system. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect voice interfaces with computer. 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.

ACKNOWLEDGEMENT

Any project is a task of great enormity and it cannot be accomplished by an individual without support and guidance. I am grateful to a number of individuals whose professional guidance and encouragement has made this project completion a reality.

I have a great pleasure in expressing my deep sense of gratitude to the beloved Chairman **Dr. Mohan Manghnani** for having provided me with a great infrastructure and well-furnished labs.

I take this opportunity to express my profound gratitude to the Principal **Dr.Manjunatha** for his constant support and management.

I am grateful to **Dr. R J Anandhi** , Professor and Head of Department of ISE, New Horizon College of Engineering, Bengaluru for his strong enforcement on perfection and quality during the course of my project work.

I would like to express my thanks to the guide **Dr. R J Anandhi** Senior Assistant Professor, Department of ISE, New Horizon College of Engineering, Bengaluru who has always guided me in detailed technical aspects throughout my project.

I would like to mention special thanks to all the Teaching and Non-Teaching staff members of Information Science and Engineering Department, New Horizon College of Engineering, Bengaluru for their invaluable support and guidance.

PRAKRITI SHARMA K P

1NH18IS078

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

API - Application Programming Interface
STT - Speech to Text
TTS - Text to Speech
SR – Speech Recognition
ASR – Automatic Speech Recognition
HMM - Hidden Markov Model
VAD - Voice Activity Detector
MAP - Maximum Posterior Probability
FFT – Fast Fourier Transform
DTW - Dynamic Time Warping
VQ - Vector Quantization
I/o - Input Output

CHAPTER 1

INTRODUCTION

Voice is the basic, common and effective form of communication for people to communicate with each other. Currently, speech technologies are usually available for a limited but interesting range of tasks. These technologies allows machines to respond to human voices correctly and reliably and to provide useful and valuable services. Because communication with the computer is faster using voices than using the keyboard, people prefer it over system. Communication between people is dominated by spoken language, so it is natural for people expect voice connections to be on the computer. This can be achieved through the development of a speech recognition system: speech to text, which allows the computer translate voice request and dictation into text. Speech recognition system: speech in text is the process of translates an acoustic signal captured with a microphone into a set of words. The recorded data may be used for document preparation.

Have you talked to your computer? I mean, did you really talk to your computer? Where did you really say what you said and then do something? If you do, you have used the technology known as speech recognition. With speech recognition, you can provide input to a system with your voice. Just like clicking the mouse, tapping the keyboard, or pressing a key on the phone's keyboard, you provide voice over voice, speech recognition can provide speaking input. In the world of computers, you drop a microphone to do this.

1.1 What was the motivation for the project?

In modern civilized societies, communication between human discourses is one of the most common methods. Ideas formed in the speaker's mind are communicated through speech in the form of words, phrases and sentences through apply applicable grammar rules. Speech is the most important means of communication between people and also the most natural and effective way to exchange information between people in speech. When classifying the speech with voice, voice and silence (VAS / S), an elemental acoustic segmentation of speech, essential for speech, can to be considered. The failure in individual sounds called phonemes of this

technique can be almost identical to the sounds. Most information in the digital world consists of letters of the alphabet made up of human speech available to some who can read or understand accurate language. Language technologies can provide solutions in the form of common interfaces so that digital content can reach the masses and facilitate the exchange of information by different people who speak different languages. These technologies play an important role in multilingual societies such as India, which has about 1652 indigenous dialects / languages. Conversion of speech to text receives input from the microphone the speech form and then it is converted into text form displayed on the tabletop. Speech processing is the study of speech signals and the various methods used to process them. In the process, several applications such as coding, speech synthesis, speech recognition and speech recognition technologies; speech processing. Among the items mentioned above, speech recognition is the most important. The main purpose of speech recognition is to converts the acoustic signal obtained from a microphone or telephone to generate a set of words. In order to extract the linguistic information transmitted by the sound waves we have to use computers or electronics. This process is performed for various applications, such as security devices, home appliances, cell phones, ATM machines and computers. The overview of this article deals with different methods of converting speech to text, namely useful for different languages, such as the Phonem method for Graphem, conversion to Hindi, Kannada or English based on HMM methods for speech synthesis etc.

CHAPTER 2

SPEECH RECOGNITION

Speech recognition (also known as automatic speech) Recognition (ASR) or computer speech recognition) is the process of converting speech recognition (also known as automatic speech) Recognition (ASR) or computer speech recognition) is the process of converting a speech signal into a sequence of words, by an algorithm implemented as a computer programming a speech signal into a sequence of words, by an algorithm implemented as a computer program.

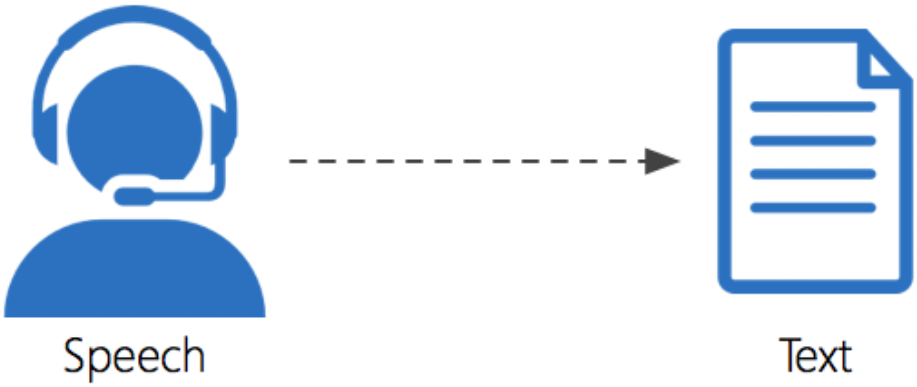

Fig. 2.1 – Speech to text

With speech recognition, you can provide input to a system with your voice. Just like clicking the mouse, tapping the keyboard, or pressing a key on the phone's keyboard, you provide voice over voice, speech recognition can provide speaking input. In the world of computers, you drop a microphone to do this.

2.1 Overview

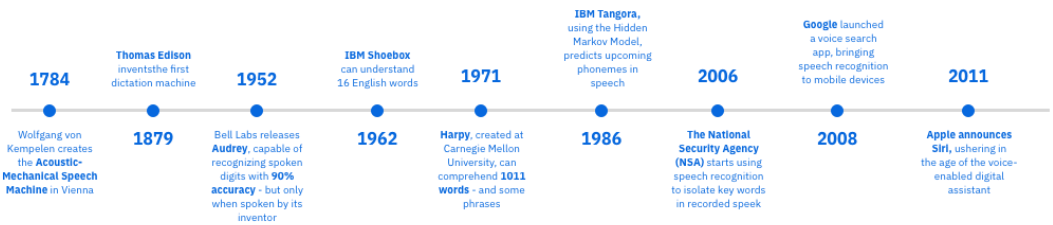

Fig. 2.1.1 – History of Speech Recognition

Speech recognition has its origins in research conducted at Bell Labs in the early 1950s. The first systems were limited to a single speaker and had a limited vocabulary of around a dozen words. Modern speech recognition systems have come a long way since their former colleagues. They can recognize the speech of several speakers and have a huge vocabulary in many languages.

The first component of speech recognition is, of course, speech. Speech must be converted from physical audio into an electrical signal with a microphone and then digital data with an analog-to-digital converter. Once digitized, different models can be used to transcribe the sound into text.

Most modern speech recognition systems are based on what is known as the Hidden Markov Model (HMM). This approach works on the assumption that a speech signal, on a sufficiently short time scale (for example, ten milliseconds), can be reasonably approximated as a stationary process - that is, a process in which statistical properties do not change over time.

Fig. 2.1.2 – Current market of SR.

In a typical HMM, the speech signal is divided into 10 millisecond fragments. The power spectrum of each fragment, which is essentially a graph of signal strength as a function of frequency, is mapped to a vector of real numbers known as cepstral coefficients. The size of this vector is usually small - sometimes as small as 10, although more accurate systems can be 32 or larger. The final output of HMM is a series of these vectors.

To decode speech into text, groups of vectors correspond to one or more phonemes - a fundamental unit of speech. This calculation requires training as the sound of a phoneme varies from speaker to speaker and even varies from one expression to another through the same speaker. A special algorithm is then applied to determine the probable word (s) that produce the given sequence of phonemes.

One can imagine that this whole process can be computationally expensive. In many modern speech recognition systems, neural networks are used to simplify the speech signal using techniques to transform resources and reduce dimensionality before HMM recognition. Speech activity detectors (VADs) are also used to reduce a sound signal to only those parts that are likely to contain speech. This prevents the detector from wasting time analyzing unnecessary parts of the signal.

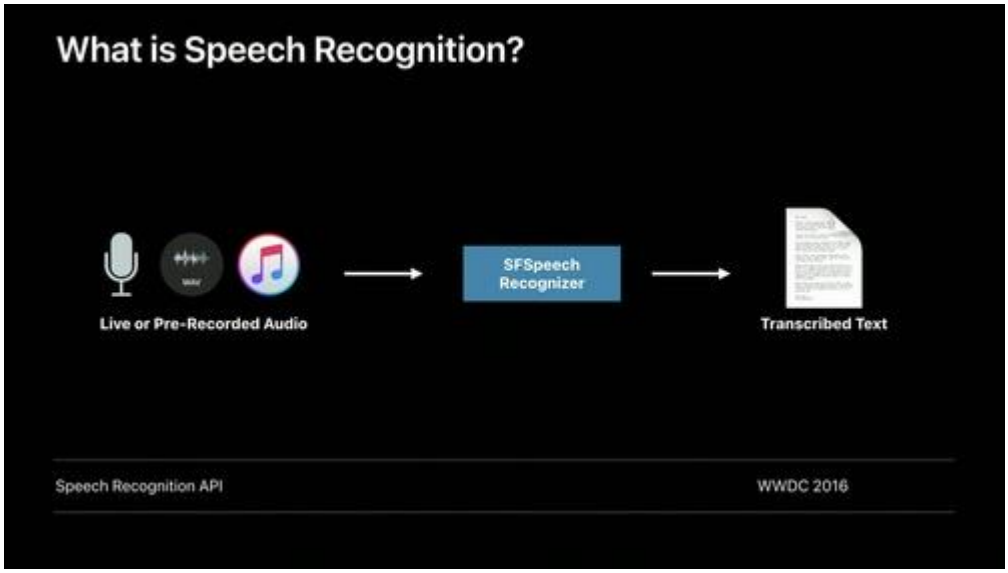

Fig. 2.1.3 – Speech Recognition API.

Fortunately, as a Python programmer, you don't have to worry about this. Several voice recognition services are available online for use via an API, and many of these services offer Python SDKs.

2.2 Basic model for speech recognition

There are two basic algorithms for speech recognition:

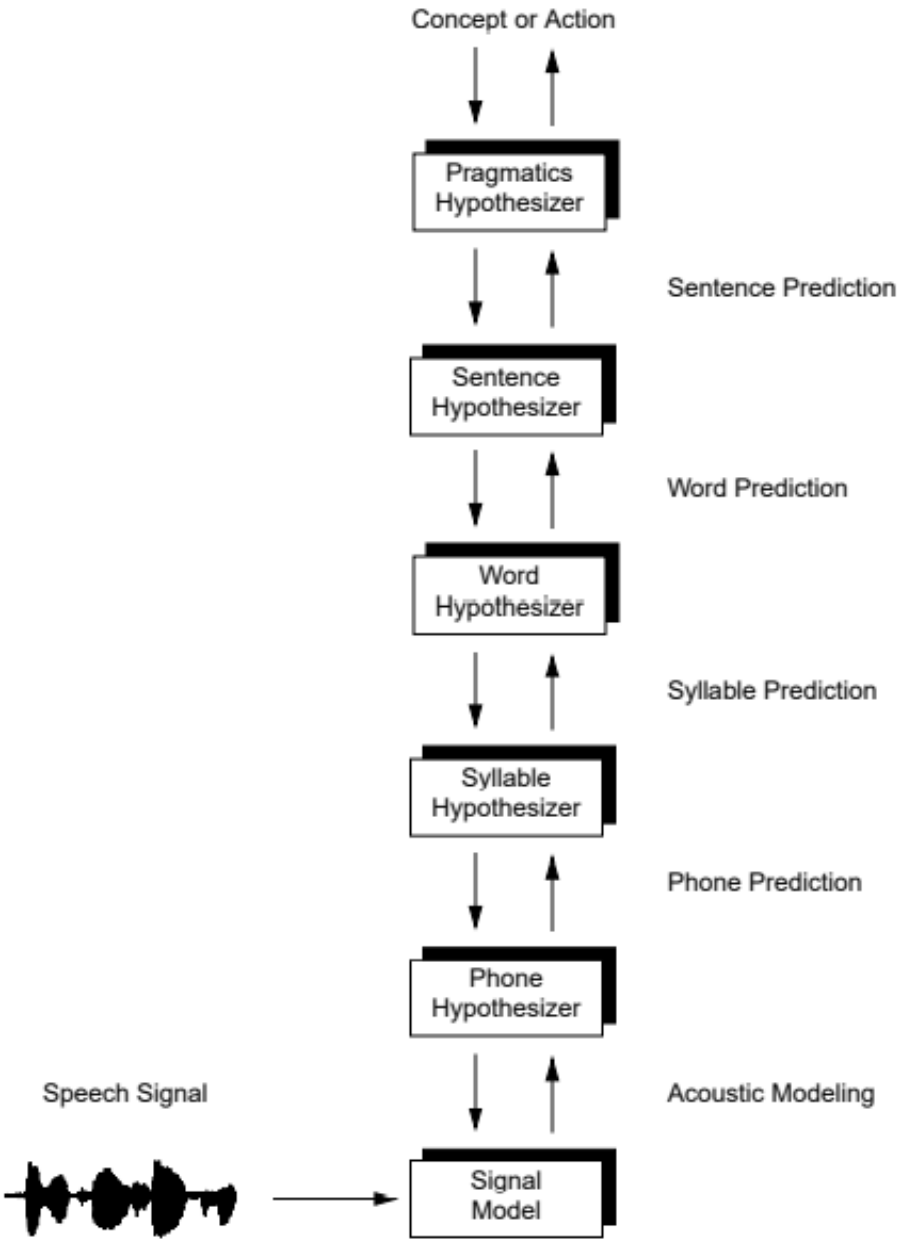

Fig. 2.2.1 – Generic Approach

Speech processing and communication research for the largely motivated by people's desire to build mechanical models to mimic people's verbal communication abilities. Speech is the most natural way to be human communication and speech processing was one of the most exciting areas of signal processing.

Speech recognition technology enabled the computer Follow and understand human voice commands speech. The main purpose of the area for speech recognition is develop techniques and systems for speech input on the machine. Speech is the most important means of communication between while.

For reasons ranging from technological curiosity on the mechanisms of mechanical realization of humans speech ability to automate simple tasks requires interaction between people and machines and research. Automatic speech recognition by machines has drawn a great deal sixty years of attention.

Based on great progress in statistical speech modeling, automatic speech recognition systems nowadays find widespread application in tasks that need it man-machine interface, such as automatic call processing telephone networks and information systems based on consultations provide up-to-date travel information, share prices, weather reports, data entry, voice dictation, access to information: travel, banks, assignments, avoidance, car portal, speech transcript, persons with disabilities supermarket (blind people), railway reservations etc.

Recognition technology is increasingly used within the telephone networks to automate and improve the operator services. This report analyzes the key highlights of the past six decades in the research and development of automated systems speech recognition in order to provide a perspective. Although there are many technological advances there are still many research questions that need to be addressed needs to be addressed.

The following figure shows a mathematical representation of speech recognition system in simple equations containing final unit, model unit, language model unit and search unit. The recognition process is shown below:

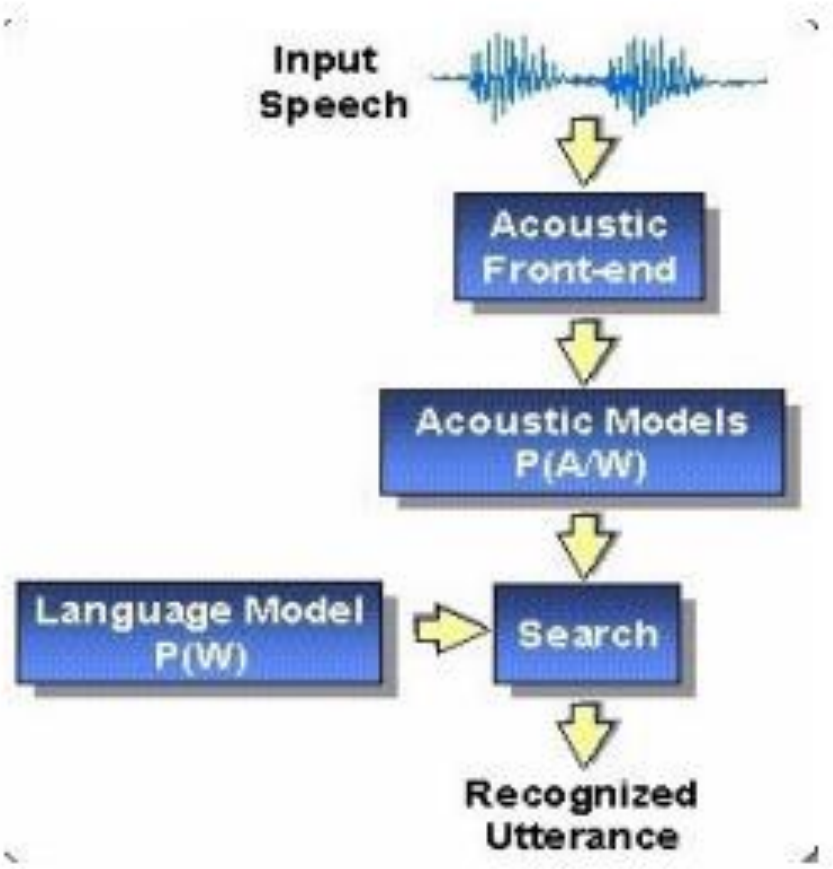

Fig. 2.2.2 – Basic Model for Speech Recognition.

The standard approach to great vocabulary is continuous speech recognition is to assume a simple probability speech production model by which a specified word drive, W, produces an acoustic observation sequence Y, with probability P (W, Y). The goal is then to decode the word string, based on the acoustic observation sequence, so that the decoded chain has the maximum posterior probability (MAP).

$$P(W/A)= \arg\max_{wW} P(W/A) \quad \text{…….. (1)}$$

Using Baye's rule, equation (1) can be written as:

$$P(W/A)=[P(A/W)P(W)]/P(A) \quad \text{……..(2)}$$

Since P (A) is independent of W, the MAP decoding rule is of the equation (1):

W=argmaxw P(A/W)P(W) ……(3)

The first term in equation (3) P (A / W) is commonly called the acoustic model, as it increases the probability of a sequence of acoustic observations, subject to the word rope. Therefore, P (A / W) is calculated. For great vocabulary, speech recognition systems, it is necessary to create statistics templates for subword units, building word templates these models of units of subwords (using a lexicon for describe the composition of the words) and then postulate sequences and evaluate the probabilities of the acoustic model via standard blending methods. The second term in comparison (3) P(W) is called a language model. Describes the probability associated with a postulated sequence of words. These language models can also be syntactic semantic limitations of language and the task of recognition.

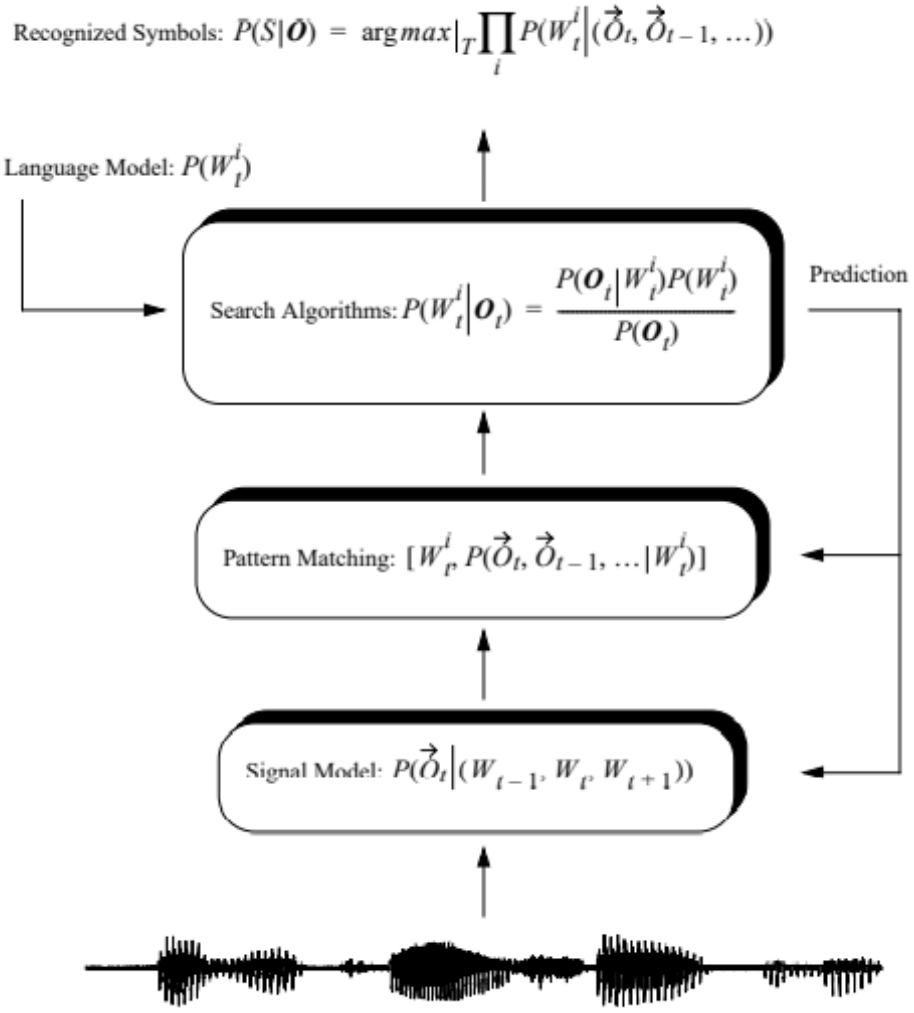

Fig. 2.2.3 – Based on HMM model.

2.3 Types of speech

Speech recognition systems can be divided into different classes, describing the different statements has the ability to recognize. These classes are classified as:

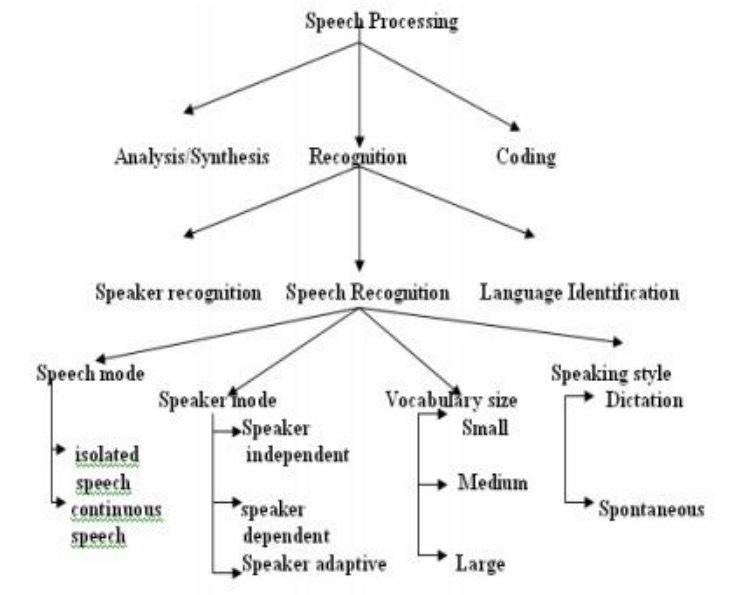

Fig. 2.3.1 – Speech Recognition Classification.

Isolated words:

Isolated word recognition usually requires every statement silence (lack of sound signal) on both sides of the preview window. Accept simple words or a single statement in time. These systems have "Hear / Don't Hear", true it requires the speaker to wait between expressions (usually processing during breaks). Isolated expression can be a better name for this class.

Related words:

Linked (or more correctly 'connected' word systems/ expressions) are similar to isolated words, but can separate you pronounced to be "combined" with a minimum break between them.

Continuous speech:

Continuous speech recognition allows users to barely speak of course, while the computer determines the content. (Basically it is dictated by a computer). Identifiers with continuous

speech functions are some of the most difficult to create because they use special methods to determine limits of expression.

Spontaneous speech:

At a basic level, it can be considered a natural speech sounding and not practiced. An ASR system with spontaneous speech ability must be able to handle a variety speech sources, such as words performed together, "anm" and "ahs", and even small waiters.

Voice identification:

Some automatic speech recognition has the ability to recognize specific users through the unique characteristics of their voices (voice biometrics). If the speaker claims to have a certain identity and the voices are used to verify that statement, this is called verification or verification. On the other hand, identification is the task of determining the identity of an unknown speaker. On a speaker verification is a 1: 1 match in which a speaker's voice is linked to the example (also called a "voice impression" or "voice model"), while speaker identification is a 1: N correspondence in which the voice is compared with the N models. There are two types of voice verification / identification system:

- Dependent on the text:

If the text needs to be the same for registration and verification, it is called text-dependent recognition. In a text-dependent system, questions can be asked among all speakers (for example: a common passphrase) or exclusive. In addition, the use of shared secrets (for example: passwords and PINs) or knowledge-based information may be unoccupied to create a multi-factor authentication scenario.

- Independent of text:

Text-dependent systems are used primarily for speaker identification, because they are more or less needed for speaker collaboration. In this case, the account and the text duration test are different. In fact, enrollment can take place without the user's knowledge, as is the case with many forensic applications. Because text-independent technologies do not compare what was said during registration and authentication, authentication applications also tend to use speech recognition to determine what the user spends at the point of authentication.

CHAPTER 3

SYSTEM REQUIREMENTS

The project has been developed with python with its extremely diverse and useful modules. Python code is simple, compact, and robust. The code developed with Python makes the implementation of the project extremely efficient.

3.1. Why do we use Python for Voice Recognition?

Voices are made up of thousands of features and fine wavelengths that require to be matched. Using Python for voice recognition breaks the process of detecting and recognizing voices into several smaller processes, each of which makes the process of Voice Recognition using Python the latest trend in Machine Learning techniques. SpeechRecognition is a module that Python uses to search for voices in an audio clip.

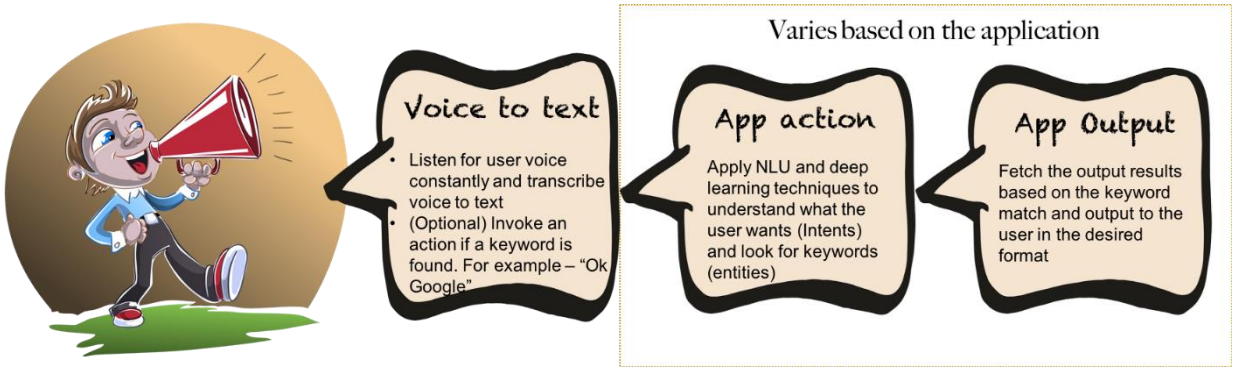

Fig. 3.1.1 – Application overview based on python.

Google's speech-to-text allows developers to convert audio to text by implementing powerful neural network models in an easy-to-use API. The API recognizes over 120 languages and variants to support its global user base. You can enable voice commands and controls, transcribe audio call sounds, and more. It can process real-time streaming or pre-recorded audio using Google's machine learning technology.

3.2. System Requirements

Table 3.1: System Requirements

COMPONENTS
LAPTOP(USED LENOVO IDEAPAD 330)
Microphone
PYTHON (3.X VERSION)
PYTHON MODULES: SpeechRecognition GoogleSpeechAPI PyAudio PyDub
Operating System: Windows XP & above, MacOS, Linux 5.3.1 & above
RAM: 512 MB at least
Memory: 5 MB in secondary storage
Working, stable internet connection
An IDE

3.3 Python Packages

In programming, a module is software that has specific functionality. For example, if you are creating a ping pong game, a module is responsible for the logic of the game and another module is responsible for drawing the game on the screen. Each module has a different file that can be edited separately.

Such built-in modules are called packages. Some packages we would be using in this project are as follows:

⇨ **SpeechRecognition**

Speech recognition has various applications, from automatic transcription of speech data (such as voice mail) to interaction with robots via speech. We would do it with the library class SpeechRecognition.

The first step, as always, is to import the necessary libraries. In that case, we just need to enter the speech recognition library that we load using pip command.

```
import speech_recognition as speech_recog
```

To convert speech to text in the only class we need, it is the Recognition class of the speech recognition module. Depending on the underlying API used to convert speech to text, the Recognition class has the following methods:

knowledge_bing () : Uses the Microsoft Bing speech API
recognize_google () : use the Google Speech API
recognize_google_cloud () : use the Google Cloud Speech API
knowledge_houndify () : uses SoundHound's Houndify API
knowledge_ibm () : use the IBM Speech to Text API
knowledge_sphinx () : use the PocketSphinx API

Among all of the above methods, the detect_sphinx () method can be used offline to translate speech into text.

To recognize speech from an audio file, we need to create an object of the AudioFile class of the speech recognition module. The path of the audio file you want to convert to text is passed to the constructor of the AudioFile class. Run the following script:

```
sample_audio = speech_recog.AudioFile('E:/Datasets/my_audio.wav')
```

In the code above, update the path to the audio file you want to transcribe.

We will use the detect_google () method to transcribe our audio files. However, the method of accept_google () requires AudioData speech recognition object module as parameter. To convert our audio file into an AudioData object, we can use the Recognizer class recording () method. We need to pass the AudioFile object to the recording method, as shown below:

```
with sample_audio as audio_file:
    audio_content = recog.record(audio_file)
```

If you check the audio_content variable type now, you will see that it has the speech_recognition.AudioData type.

```
type(audio_content)
```

Gives:

```
speech_recognition.AudioData
```

Now we can simply pass the audio_content object to the Recognizer () object's method recognize_google () and the audio file will be converted to text. Start the following script:

```
recog.recognize_google(audio_content)
```

Gives:

```
'Bristol O2 left shoulder take the winding path to reach the lake no closely the size of the gas tank degrees of
fice 30 face before you go out the race was badly strained and hung them the stray cat gave birth to kittens the
young girl gave no clear response the meal was called before the bells ring what weather is in living'
```

The above output shows the text of the audio file. You can see that the file was not 100% correctly transcribed, but the accuracy is reasonable.

CHAPTER 4

DESIGN MODULES

The project has been implemented with a number of functions/modules. The discussion of the flow of the project has been discussed in this chapter.

4.1 Components of the system

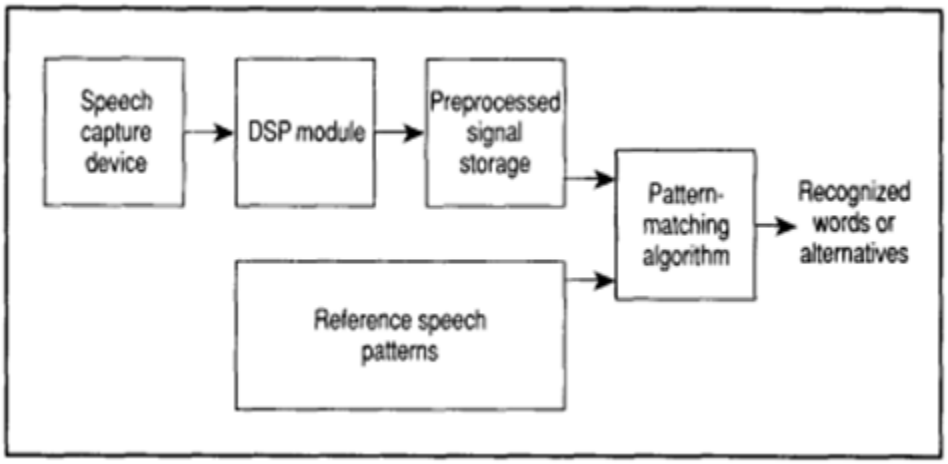

Fig. 4.1.1 – Components of Speech Recognition.

A voice recorder:

It consists of a microphone that converts the sound wave signals into electrical signals, and an analog-to-digital converter, which collects and digitizes the analog signals to obtain the discrete data that the computer can understand.

A digital signal module or processor:

Performs the processing of the raw speech signal, such as conversion into the frequency domain, restores only the necessary information, etc.

Pre-processed signal storage:

Pre-processed speech is stored in memory to perform additional speech recognition tasks.

Reference Speech Patterns:

The computer or system consists of predefined speech patterns or models already stored in memory to be used as a reference for correspondence.

Pattern matching algorithm:

The unknown speech signal is compared to the speech reference norm to determine the actual words or the word standard.

4.2 Working of the project

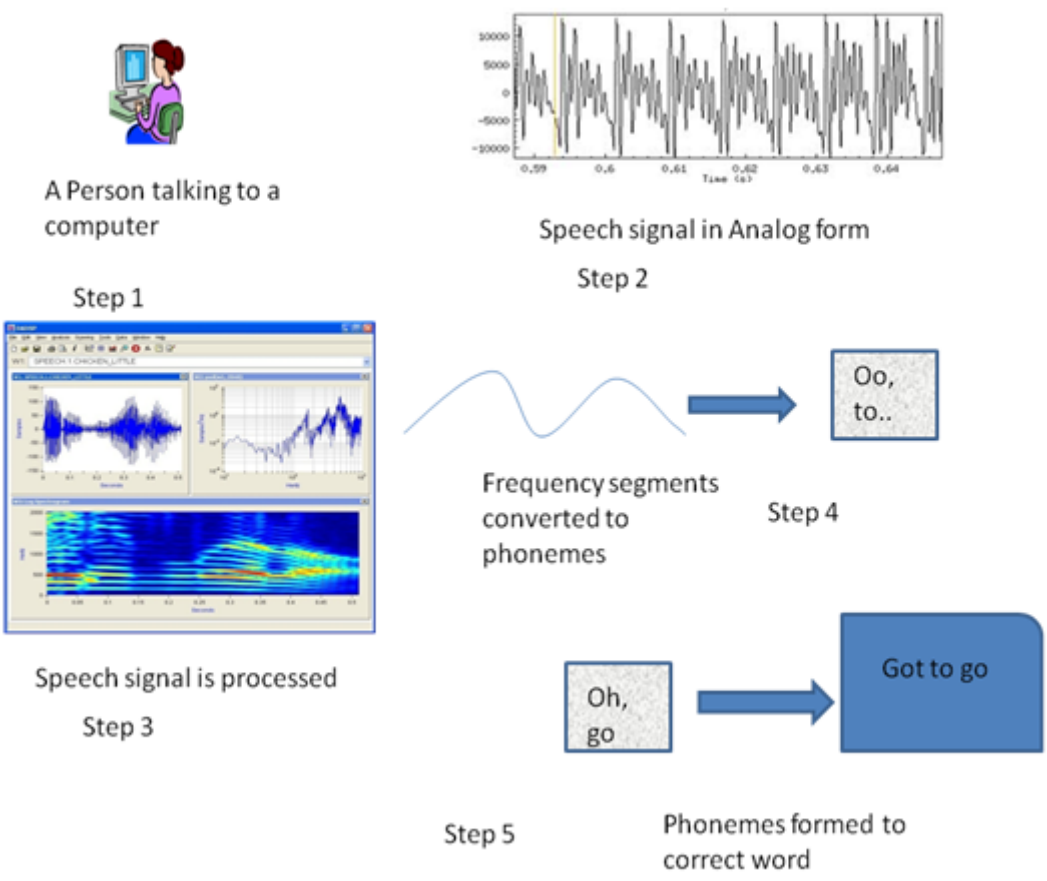

Fig. 4.2.1 – Working of Speech Recognition.

The project is designed to work in the following way:

- ⇨ A speech can be seen as an acoustic waveform, that is, the information of the signal that carries the message. A normal person with limited movement speed by their artists (speech organs) can speak at an average rate of 10 sounds per second. The average information rate is about 50 to 60 bits / second. In fact, this means that only 50 bits / second of information are needed in the speech signal. This acoustic waveform is converted into analog electrical signals by the microphone. The analog to digital converter converts this analog signal to collect digital samples through accurate wave measurements at separate intervals

- ⇨ The digitized signal consists of a flow of periodic signals sampled at 16,000 times per second and is not suitable for performing the actual speech recognition process, as the pattern cannot be easily found. To extract the real information, the signal in the time domain is converted into a signal in the frequency domain. This is done by the digital signal processor using the FFT technique. In the digital signal, the component is analyzed every 1/100 of a second and the frequency spectrum of each component is calculated. In other words, the digitized signal is segmented into small parts of frequency ranges.

- ⇨ Each segment or frequency graph represents the different sounds that people make. The computer matches the unknown segments with the phonetics stored in the specific language. This is done by the various approaches to speech recognition.

4.3 Approaches to Speech Recognition

There are basically three approaches to speech recognition. They are:

- Acoustic phonetic approach
- Pattern recognition approach
- Artificial intelligence approach

Approach	Representation	Recognition Function	Typical Criterion
Acoustic phonetic approach	Phonemes/segmentation And labeling	Probabilistic lexical access procedure	Log likelihood ratio
Pattern recognition approach			
• Template	Speech samples, pixels & curves	Correlation, distance measure	Classification error
• DTW	Set of a sequences of spectral vectors	Dynamic warping	Dissimilarity measure
• VQ	Set of spectral vectors	optimal algorithm	Euclidian distance
• Statistical	Features	Clustering functions (code book) Discriminant functions	Classification error
Neural network	Speech features/perceptrons/ Rules/units/procedures	Network function	Mean square error
Support vector machine	Kernel based features	Maximal margin hyperplane, Radial basis function classifier(fitting functions)	Minimizing a bound on the Generalization error.
Artificial Intelligence approach	Knowledge based		Word error probability

Table 4.3.1 – Approaches to Speech Recognition.

4.3.1 ACCOUSTIC PHONETIC APPROACH

The first approaches to speech recognition were based on find speech sounds and provide appropriate tags those sounds. This is the basis of acoustic phonetics approach (Hemdal and Hughes, 1967), which postulates there are finite and distinct phonetic units (phonemes) spoken language and that these units are widely characterized by a set of acoustic properties that are in the speech signal over time. Although the acoustic properties phonetic units is very variable, both with speakers and with neighboring sounds (the so-called co-articulation effect), the acoustic-phonetic approach assumes that the rules apply adjusting the variation is simple and can be easily learned by a machine.

The first step in the acoustic process phonetic approach is a spectral analysis of speech combined with a feature detection that converts the spectral measures for a set of

functions that describe the broad acoustic properties of different phonetic units. In the next step is a segmentation and labeling phase in which the speech
The signal is divided into stable acoustic regions, followed by attach one or more phonetic labels to each segment leading to the characterization of the phonemes network in the region speech. The last step in this approach is to try to valid word (or sequence of words) from the phonetic label strings produced by segmentation in labeling. At the validation process, linguistic restrictions on the task (that is, the use with vocabulary, syntax and other semantic rules) to access the word decoding lexicon based on the phonemes grid. The acoustic phonetic approach was not yet is widely used in most commercial applications.

4.3.2 PATTERN RECOGNITION APPROACH

The Pattern Matching Approach involve two significant steps that is, pattern training and pattern comparison. The essential. The characteristic of this approach is that it uses a well-formulated method mathematical structure and consistent speech pattern representations for reliable pattern comparison from set of training samples labeled using a formal training algorithm.

A speech pattern representation can be in the form of a speech model or statistical model (for example, a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller) then a word), a word or a phrase. Compare standards approach phase, a direct comparison is made between unknown speeches (the speech to be acknowledged) at each possible pattern learned in the training stage up to determine the identity of the unknown according to the friendliness of pattern matching. Pattern Matching approach has become the predominant method of speech recognition in the last six decades. A schematic block diagram of the pattern recognition is presented below. There are two methods in this, namely model approach and stochastic approach.

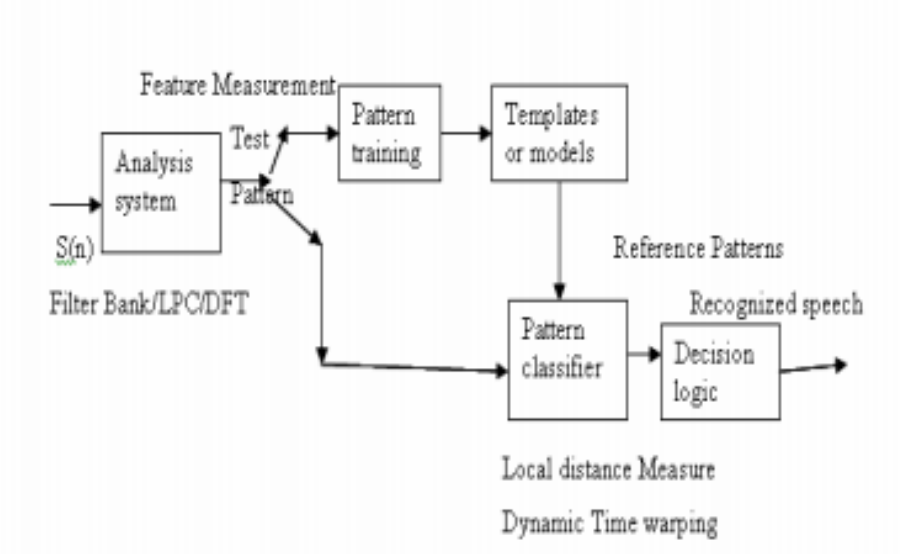

Fig. 4.3.2.1 – Pattern approach recognizer block diagram.

Template based approach:

Template-based approach for speech recognition provided a family of techniques that advanced the subject significantly over the past six decades. The underlying idea
It's simple. A collection of prototypical speech patterns is stored as reference standards that contain the dictionary of candidate's words. Recognition is then through corresponds to an unknown spoken pronunciation in each reference models and the category of best matching pattern. Templates are usually for whole words construction. It has the advantage of making mistakes because of segmentation or classification of minor acoustic sounds variable units such as phonemes can be avoided. In turn, each the word must have its own complete reference form; model preparation and adaptation become excessively expensive or impractical, as the vocabulary becomes larger than a few hundred words. A key idea in the model method is to derive a typical idea order of the speech frames of a pattern (one word) by means of average procedure and rely on the use of local spectral measure the distance to compare patterns. Another central idea is use some form of dynamic programming to temporarily align patterns to explain differences in conversation rates speech as well as about the same repetitions of the word being spoken.

Stochastic Approach:

Stochastic modeling implies the use of probability models for handling uncertain or incomplete information. By the speech recognition, uncertainty and incompleteness arise

many sources; for example, confusing sounds, speaker variability, contextual effects and homophonic words. Therefore, Stochastic models are a particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite markov model and a set outbound distributions. The transition parameters in Markov chain models, temporal variability, while the parameters in the output distribution model, spectral variability. These two types of variables are at the heart of speech recognition. Compared to the model-based approach, hidden Markov modeling is more common and has a firmer math Foundation.

A model-based model is simply a continuum HMM density, with identity covariance matrices and a slope limited topology. Although models can be trained
In some cases, they do not have the likely formulation of HMMs and typically outperform HMMs. In comparison with knowledge-based approaches; HMMs allow easy integration sources of knowledge in a composite architecture. A negative side effect of this is that HMMs do not provide much insight in the recognition process. Because of this, it is often difficult analyze the errors of an HMM system in an attempt to improve your performance. However, careful incorporation knowledge has greatly improved the HMM systems.

Dynamic time warping

It is an algorithm for measuring agreement between two sequences that can vary in time or speed. For example, there may be similarities in walking patterns detected, even if the person was walking slowly in a video and if he's faster, he walked faster, or even if there were accelerations and delays during the observation course. DTW was applied to the video, audio and graphics, in fact, any data that can be converted in a linear representation it can be analyzed with DTW.

A little well-known app is automatic speech recognition, or handle different speech speeds. Generally, DTW is a method by which a computer can find an ideal combination between two lines (for example, time series) with certain restrictions. Strings are not linear in "deformation" temporal dimension to determine a measure of your agreement

independent of certain non-linear variations in time dimension. This sequence adjustment method is used regularly the context of hidden Markov models. An example of the restrictions imposed on the compliance of the lines are in the monotony of the mapping in the temporal dimension. Continuity is less important at DTW than at other pattern matching algorithms; DTW is an algorithm especially suitable for missing sequence information, as long as there are segments long enough match to prevent. The optimization process is carried out using dynamic programming, hence the name.

Vector quantization (VQ)

It is often applied to ASR. It is useful for speech encoders, i.e. efficient data reduction. Since baud rate is not a major issue for ASR, the utility of VQ lies here in the efficiency of using compact code books for reference models and codebook researcher instead of more expensive evaluation methods.

For IWRM, every word in the vocabulary get its own VQ code, based on the training sequence of several repetitions of the word. The speech is evaluated by all codebooks and ASR select the word whose codebook is delivers the shortest measured distance. In basic VQ, codebooks does not have explicit time information (for example, the timing of phonetic segments in each word and their relative duration ignored) as codebook entries are not ordered comes from any part of the training words. Some however indirect duration tips are preserved because the codebook entries are chosen to reduce the average distance between everyone exercise frames and frames, corresponding to segments (for example, vowels) occur more frequently in training data.

These segments are therefore more likely to specify code words than less frequent consonant photos, especially with small ones code books. However, the code words exist for constant tables because these frames would contribute to large frames distances to the codebook. Often there are a few code words sufficient to represents many frames during relatively stable portions of vowels, which may suggest more code words dynamic sections of words. That relative emphasis that VQ using voice transitions can be an advantage over other ASRs comparison methods for vocabulary of similar words.

4.3.3 ARTIFICIAL INTELLIGENCE APPROACH

The Artificial Intelligence approach is a hybrid of acoustic phonetic approach and pattern recognition approach. It explores the ideas and concepts of acoustic phonetics and pattern recognition methods. Knowledge-based approach use information related to languages, phonetics and spectrogram. Some speech researchers have developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While the model is based approaches have been very effective in designing a variety speech recognition systems; they provided little insight on the processing of human speech and thus make the analysis of errors and improving the knowledge-based system is difficult. At the on the other hand, a lot of linguistic and phonetic literature provides insight and understanding to human discourse processing. In its pure form, the AI project involves the direct and explicit incorporation of experts knowledge of speech in a recognition system. This knowledge is generally derived from a careful study of spectrograms and is recorded using rules or procedures. Pure knowledge Engineering is also motivated by interest and research specialized systems. However, this approach has been limited success, largely due to the difficulty of quantifying specialists knowledge. Another difficult problem is the integration of many levels of human knowledge, phonetics, phonotactics, lexical access, syntax, semantics and pragmatics.

Alternatively, the combination of independent and asynchronous functions resources optimally remain an unsolved problem. At the more indirect ways, knowledge was also used to guide design the models and algorithms of other techniques such as model customization and stochastic modeling. This form the application of knowledge makes an important distinction between knowledge and algorithms. Algorithms allow us to solve problems. Knowledge allows algorithms to work best. This way of improving the knowledge-based system contributed significantly to the design of everyone who was successful reported strategies. Plays an important role in choosing appropriate input representation, the definition of units of speech or the design of the recognition algorithm itself.

Connectionist Approaches (Artificial Neural Networks):

The artificial intelligence approach tries to mechanize recognition procedure according to the way a person applies intelligence in visualization, analysis and characterization of speech based on set measured acoustic properties. Among the techniques used use an expert system in this class of methods (for example, neural network) that is phonemic, lexical, syntactic, semantic and even pragmatic knowledge for segmentation and labeling and using tools such as artificial NEURAL NETWORKS to learn the relationships between phonetics happenings. The focus in this approach was particularly on knowledge representation and knowledge integration sources. This method is not widely used systems. Connectionist speech modeling is the latest development in speech recognition and becoming the subject of lots of controversy. In connectionist models, knowledge or restrictions are not contained in units, rules or procedures, but spread across very simple computers units. Uncertainty is not modeled as probability or probability density functions of a single unit, but according to activity pattern in many units. Computer units are simple in nature and knowledge is not programmed in any individual unit profession; on the contrary, it is in the connections and interactions between coupled processing elements. Because the style of calculation that can be performed by the networks of these units has any similarity to the computer style in the nervous system. Connectionist models are also mentioned neural networks or artificial neural networks. Just like that, parallel distributed or massive distributed processing processing are terms used to describe these models.

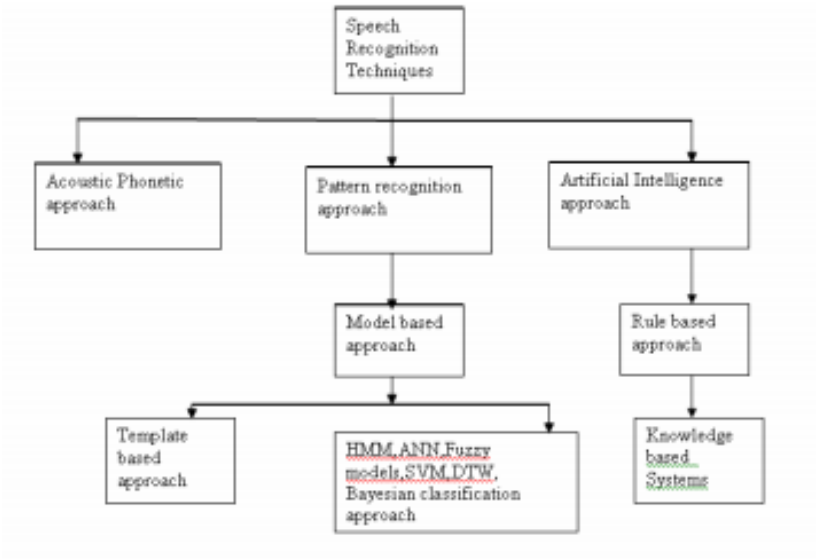

Fig. 4.3.1 – Taxomy of the approaches to SR.

4.4 Nature of the project

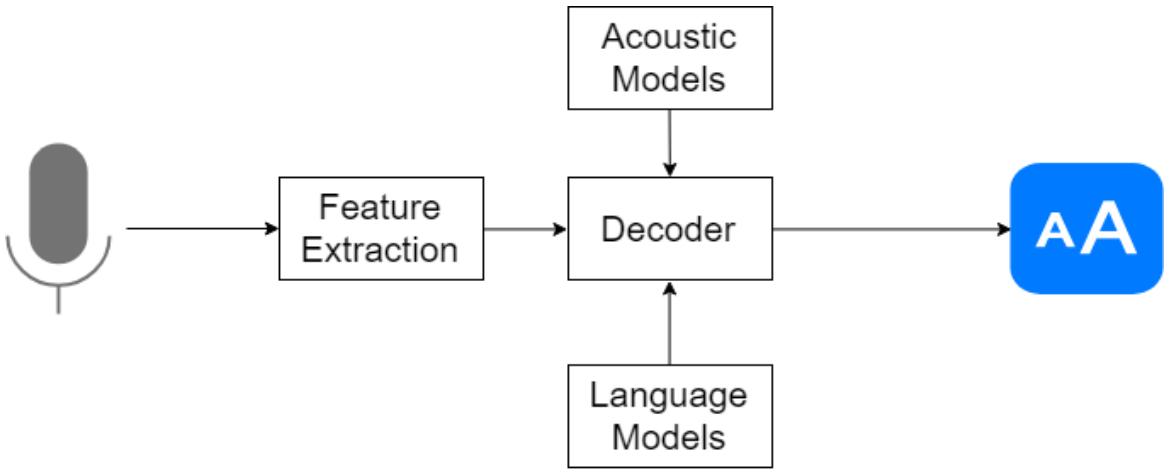

Fig. 4.4.1 – Nature of the project

The image above show the principle of speech recognition very clearly.

It is based on the acoustic and language modeling algorithm. Now the question is, what is acoustic and language modeling?

Acoustic modeling represents the relationship between linguistic speech units and sound signals.

Language modeling aligns sounds with sequences of words to differentiate those that sound similar.

Any speech recognition program is evaluated using two factors:

1. Accuracy (percentage error when converting spoken words to digital data).
2. Speed (the degree to which the program accompanies a human speaker).

CHAPTER 5

CODE AND IMPLEMENTATION

The project uses different functions to implement the Speech Recognition System. The code and the flow of the functions has been discussed in this chapter.

5.1 Module 1 Implementation

The first module helps declare the language the speech to text project will work in. For now, the project has multiple Indian language options like Hindi, Kannada, Bengali, Malayalam, Marathi, Urdu, etc. There is also a segment to implement this model for any language whose Google language code is known.

```
while(1):
    engine.say("Choose the language you want to speak in")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("1. English")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("2. Hindi")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("3. Kannada")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("4. Bengali")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("5. Malayalam")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("6. Marathi")
    engine.runAndWait()
    time.sleep(0.2)
    engine.say("7. Urdu")
    engine.runAndWait()
```

Fig. 5.1.1 – Module 1 implementation.

5.2 Module 2 Implementation

This module involves the assigning of language codes based on the language chosen by the user.

```
if n==1:
    lang='en'
elif n==2:
    lang='hi-IN'
elif n==3:
    lang='kn-IN'
elif n==4:
    lang='bn-IN'
elif n==5:
    lang='ml-IN'
elif n==6:
    lang='mr-IN'
elif n==7:
    lang='ur'
elif n==8:
    lang=input("Enter the google language code of the language you want to see
    ")
elif n==0:
    exit(0)
```

Fig. 5.2.1 – Module 2 Implementation.

5.3 Module 3 Implementation

This module is the heart of the project. This module involves hearing the audio input(through the mic here) and then producing the text output in the chosen language.

```
with sr.Microphone() as source:
    engine.say("Mic testing..")
    engine.runAndWait()
    audio = r.adjust_for_ambient_noise(source)
    print("Say something")
    audio = r.listen(source)
    engine.say("Time is over. Thanks.")
    engine.runAndWait()
try:
    print("You said: ' " + r.recognize_google(audio, language=lang)+"'")
    time.sleep(5)
except LookupError:
    engine.say("Could not understand audio. Do you want to try again?")
    engine.runAndWait()
engine.say("Do you want to continue?")
engine.runAndWait()
y=int(input("Enter 0 to quit"))
if y==0:
```

Fig. 5.3.1 – Module 3 Implementation

CHAPTER 6

RESULTS

The current results of the project are being discussed in this chapter.

6.1 Current Implementation of the Project Results

❖ **OUTPUT 1:**

In this output, the language chosen is English and the speech is also in English.

```
import speech_recognition as sr
```

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

```
with sr.Microphone() as source:
    print("SAY SOMETHING");
    audio = r.listen(source)
    print("TIME OVER, THANKS")


try:
    print("TEXT: "+r.recognize_google(audio));
except:
    pass;
```

```
SAY SOMETHING
TIME OVER, THANKS
TEXT: New Delhi is the capital of India
```

Fig. 6.1.1 – Same language I/O.

❖ **OUTPUT 2:**

In this output, the language chosen is English but the speech given as input through the mic is in Hindi.

```
import speech_recognition as sr
```

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

```
with sr.Microphone() as source:
    print("SAY SOMETHING");
    audio = r.listen(source)
    print("TIME OVER, THANKS")


try:
    print("TEXT: "+r.recognize_google(audio));
except:
    pass;
```

```
SAY SOMETHING
TIME OVER, THANKS
TEXT: main ja raha hoon
```

Fig 6.1.2 – Different language I/O.

❖ OUPUT 3:

In this output, the language chosen by the user is “Hindi” and the speech input through the mic is also in “Hindi”.

```
import speech_recognition as sr
```

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

```
with sr.Microphone() as source:
    print("SAY SOMETHING");
    audio = r.listen(source)
    print("TIME OVER, THANKS")


try:
    print("TEXT: "+r.recognize_google(audio, language = 'hi-IN'));

except:
    pass;
```

```
SAY SOMETHING
TIME OVER, THANKS
TEXT: मैं जा रहा हूं
```

Fig. 6.1.3 – Native language chosen.

6.2 FUTURE IMPLEMENTATIONS

The project can be implemented in the future in the following ways:

1. Help the disabled

A portable synthetic speech device, powered by a battery, can be used by a person with a voice impairment to express his words. The device has a specially designed keyboard that accepts the input and converts it into essential speech in a blink of an eye.

2. Learning resource for the visually impaired

The most important fact to listen to is an important skill for blind people. Blind individuals trust their ability to listen or listen quickly and efficiently for information. Students use their audience to obtain information books on tape or CD, but also to determine what's going on around you.

3. Games and education

Synthesized speeches can also be used in many educational institutions, both in the field of study and in sports. If the teacher can tired at some point, but a computer with speech synthesizer can learn all day with the same performance, efficiency and precision.

4. Telecommunications and multimedia

STT systems enable telephone access to vocal information recovery systems can be placed by the user's voice (using a speech recognition) or by the phone keyboard. Synthesized speech can also be used to send short text messages on cell phones.

5. Man-machine communication

Speech synthesis can be used in different types of human machine interactions and levels. For example, in warning, synthesized alarm systems, watches and washing machines can be used to provide more accurate information current situation. The speech signals are much better than the warning lights or the bells.

6. Voice-activated email

Voice-activated email uses speech recognition and speech synthesis technology to give users access to their email any phone. The subscriber dials a phone number to access a polling portal and collect their email messages, press a few keys and say maybe a phrase like "Receive my email". The speech synthesizer software converts the email text a voice message played over the phone. Voice-activated email is especially useful for mobile workers, because it allows them to easily access your messages from anywhere (as long as they can) telephone), without having to invest in expensive equipment, such as overlapping computers or personal digital assistants.

CHAPTER 7

CONCLUSION

In this report, we discuss the topics relevant to the development of STT systems. Conversion from speech to text is possible. It looks effective and efficient to its users if it offers natural speech and is done through several modifications to it. The system is useful for deaf and dumb people to interact with other people in society. Text to speech synthesis critical research and application area in the field of multimedia interfaces. Important references for literature related to endogenous speech signal variations and their importance in automatic speech recognition also provides implementation of speech-to-text conversion using Python. A database was created from the different domain words and syllables. The desired speech is produced by the concatenative synthesis of the speech. Speech synthesis is beneficial for people with mental disabilities. This article has a clear and simple overview of speech-to-text (STT) system operation in a step-by-step process. The system returns the input data of voice mic, process this data and convert it to the text format displayed on a computer. By engaging in research, speech becomes more effective and difficulties and emotions are natural.

BIBLIOGRAPHY

[1]. https://medium.com/@rahulvaish/speech-to-text-python-77b510f06de

[2]. https://www.researchgate.net/figure/Working-of-Speech-Recognition-process_fig1_281684270

[3]. https://www.elprocus.com/understanding-voice-recognition/

[4]. http://www.andybrain.com/qna/2007/10/03/voice-recognition-and-hardware-system-requirements/

[5]. https://cloud.google.com/speech-to-text

[6]. https://www.geeksforgeeks.org/speech-recognition-in-python-using-google-speech-api/

[7]. https://www.scribd.com/doc/130376790/Speech-Recognition-Seminar-Report#download&from_embed

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

[9]. https://en.wikipedia.org/wiki/Speech_recognition