



Voice Mail System Using Machine Learning

Deepika Patel¹

Dept. of CSE

Rungta College of Engineering and Technology

Raipur, Chhattisgarh

email: deepikapatel1209@gmail.com

Toran Verma²

Dept of CSE (DS)

CMR Engineering College, Hyderabad

email: toranverma.003@gmail.com

Shreya Chandrakar³

Dept of CSE

Rungta College of Engineering and Technology

Raipur, Chhattisgarh

email: shreya.chandrakar@rungta.ac.in

Abstract—

An email is a type of communication that allows users to send and receive messages. It is very useful for business communication as it allows people to exchange important and confidential information. Around 2.2 billion people globally have a vision impairment, and about 25% of the population is uneducated or not able to use basic communication technology. These individuals are mainly affected by the lack of visual perception and the knowledge about using electronic devices. A Voice Mail System is a type of communication that uses the user's spoken commands to perform various tasks. These tasks include sending and receiving messages. It takes the receiver's email address, the message's subject, and the location of the file to the recipient's address. Due to the rapid emergence of AI and machine learning, Voice Mail Systems have been able to perform various tasks without the user having to type in a keyboard. In this paper, we will develop a system that uses speech recognition to send files to every user.

1531

Keywords— *Voice Mail Speech Recognition, Email System, NLP, Machine Learning.*

DOI Number: 10.14704/nq.2022.20.10.NQ55119

NeuroQuantology 2022; 20(10): 1531-1544

1. INTRODUCTION

The Voice Based Email System was developed for individuals with visual impairments. Its goal is to perform various tasks based on the commands given by the user. In terms of typing, human uttering is more natural than saying a single word. An average person can type around 35 words in a minute, while those with mental disabilities can utter over a hundred words. The rapid emergence of AI and machine learning has made Voice Based Email System capable of performing various tasks without

requiring users to type in a keyboard. The main goal of a voice based email system is to send not only emails but also send files like pdf, image etc. This paper aims to develop a Voice Based Email System that uses speech recognition technologies to send files for every user. Block diagram of Voice Mail System is shown in Fig. 1. The main objective of the Voice Based Email System is to help the visually challenged, students, and illiterate people to grow into society. It will allow them to use email service just like other people.



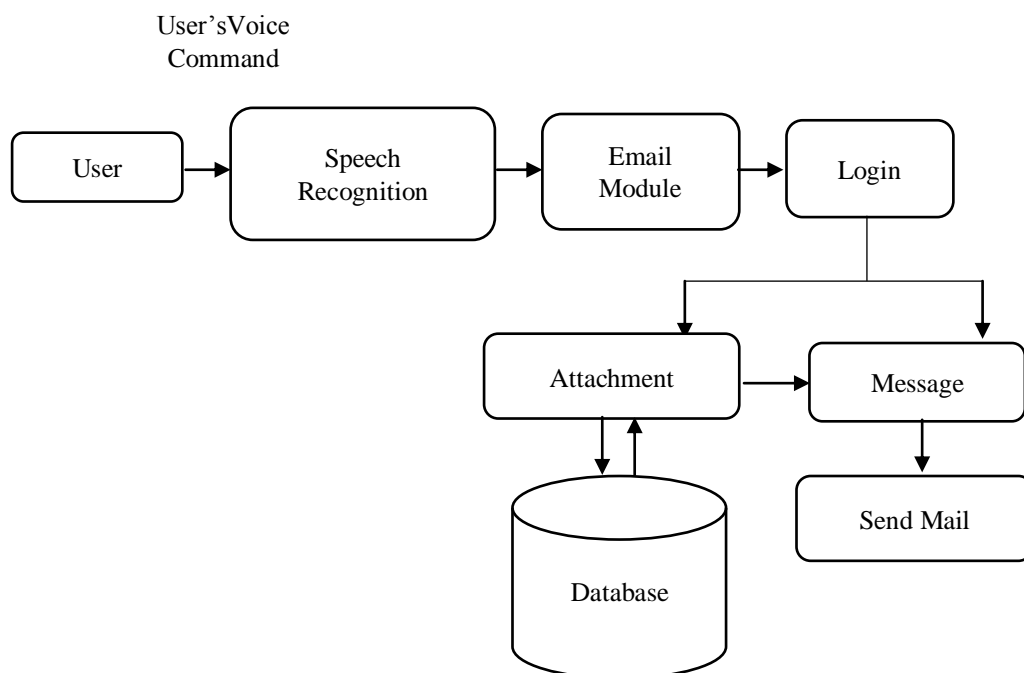


Fig 1: Block Diagram of Voice Mail System

An electronic mail is a type of communication that is sent and received between people using various electronic devices. The email system is a network of computers that handle the exchange of messages. These include programs that are used to send and retrieve messages, as well as agents that are designed to look after them. In addition to being able to contain text, messages in mail also contain audio and video data. The person who receives a message is referred to as the recipient, while the one who sent it is called the sender. The email system follows a client-server approach. When an email is sent, it is sent to a mail server, which then handles the details of the message. The outgoing mail server then checks the postage address of the message to determine where it should be sent. The Simple Mail Transfer Protocol (SMTP) used in this process. The details about the recipient are stored in the SMTP database, which is then sent to the email recipient's MTA server. The MTA chooses the appropriate destination for the mail, and it also decides which client to use, such as using POP or IMAP. Once the

recipient receives an email notification, the mail will wait until the recipient's device fetches it.

In 2011, Harsh [8] and his colleagues proposed a system that would allow users to send an email using their voice commands. The researchers focused on developing an artificial speech framework that could be used to perform various tasks. Rutuja (2018) et al.[7] created a virtual assistant that can be used by people with visual impairments to perform various tasks. It can also monitor their activities and send and receive emails. It can also recognize images using optical character recognition. Through their work, Harivans (2021) et al.[4] were able to create a system that can be used by people with visual impairments. This system can help them communicate with others and provide them with the necessary facilities to perform various tasks. In 2020, Rijwan et al.[1] developed an artificial email system that can be used to provide users with visual impairments a more accessible way to access their email accounts.



2. METHODOLOGY

Our proposed system is a web-based application that makes it easy for people to access email system. It will allow users to perform various activities to access the services that they need. The system will additionally direct them to the appropriate activities and will require them to complete the tasks in order to access the services. The work flow of voice base email system is shown in Fig. 2.

When a user calls system, it will respond and ask the user to send mail. It will then log into his mail account and ask the receiver's address, subject, and message to add any attachments. If the user doesn't respond, it will send the message and subject to the receiver's address. After searching for the file, the system will add it to the email and send it to the receiver's address. We have included three modules in this system: Text-to-Speech (TTS), Speech-to-Text (STT) and Mail communication. The Text-to-Speech module converts the instruction given by the user to speech. On the other hand, the Speech-to-Text module takes the user's speech and converts it to text. The mail communication module is used for sending and receiving emails.

The system's methodology for sending and receiving mail using attachments is built on deep learning. It uses various parts such as User Command, Noise Elimination, Speech Recognition, knowledge abstraction, response generation and Email Module to analyze and improve the performance of the

system. The system's methodology is built on knowledge abstraction, which is a process that involves extracting data from the collected information. Response generation is also done on the basis of the data generated by the knowledge process.

The entire process is divided into following sub-process:

2.1 System Introduction

In this section, the system will introduce itself using speech synthesis and will ask queries to perform task. Speech synthesis is a type of artificial intelligence that makes human speech. Text to speech (TTS) is a type of speech that is converted into a spoken voice. The quality of the voice depends on the engine used. A TTS system is usually divided into two main parts. The first one converts text into a language specification, while the second one generates a waveform.

The Hidden Markov Model (HMM)-based synthesis method is a statistical method that can be used to create a speech synthesizer. It takes into account the various components of a speech, such as the audio signal, frequency sampling, and duration. The resulting speech waves are then modeled using HMMs. The system's excitation and spectrum parameters are derived from the speech database and are modeled using context-sensitive HMMs. The synthesis phase then generates speech signals according to the text that will be synthesized.



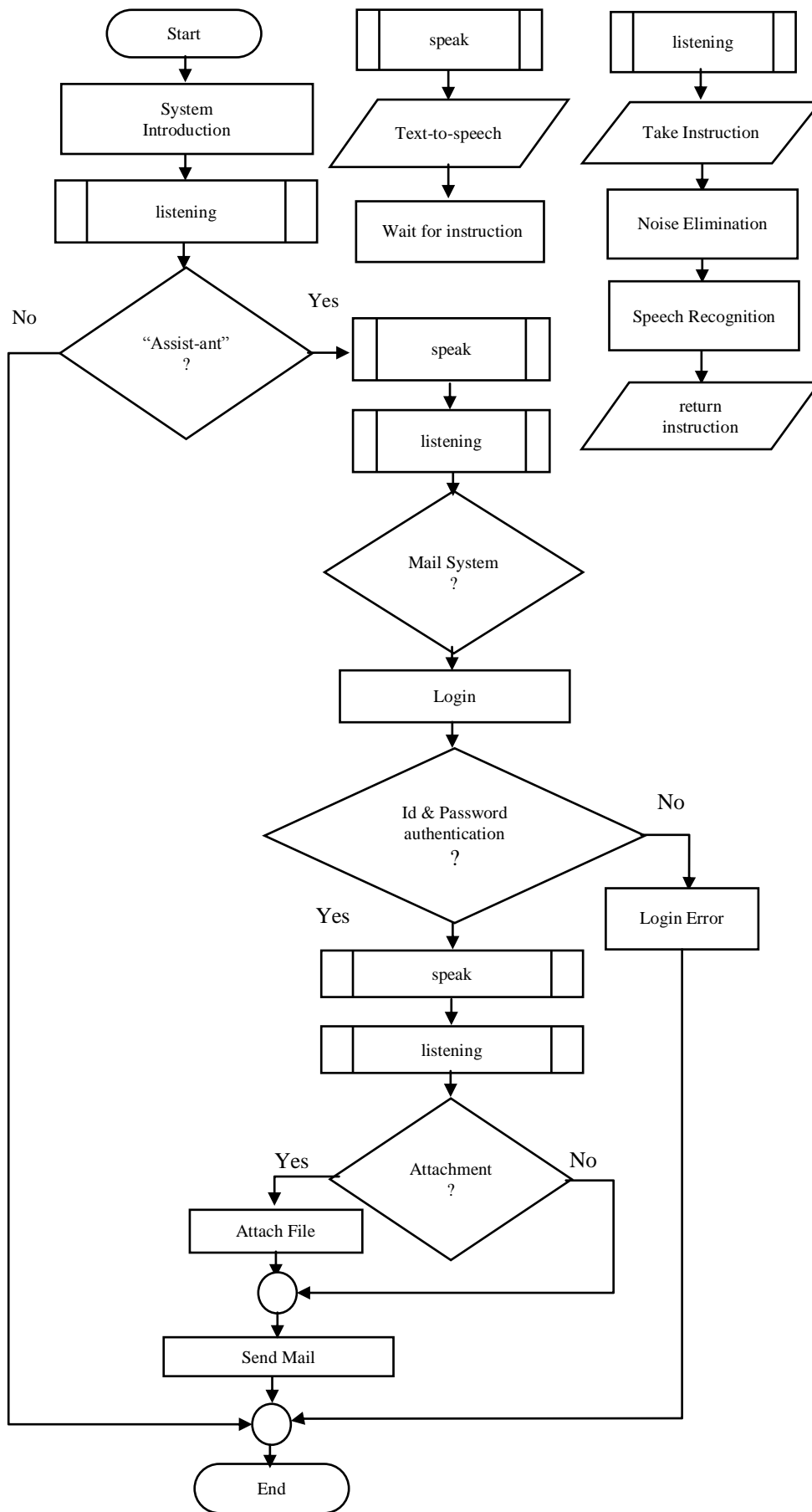


Fig. 2: Voice Mail System work flow



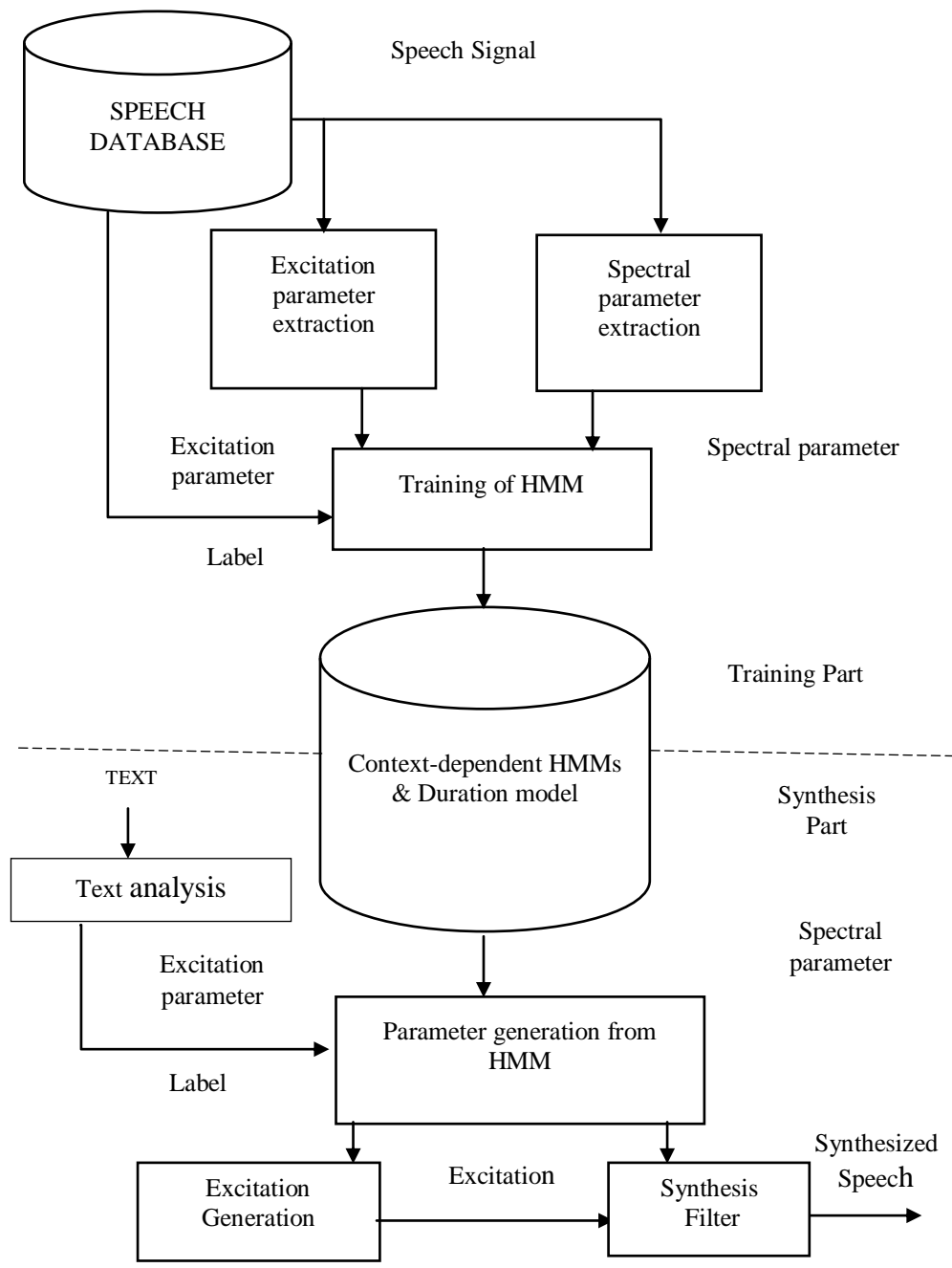
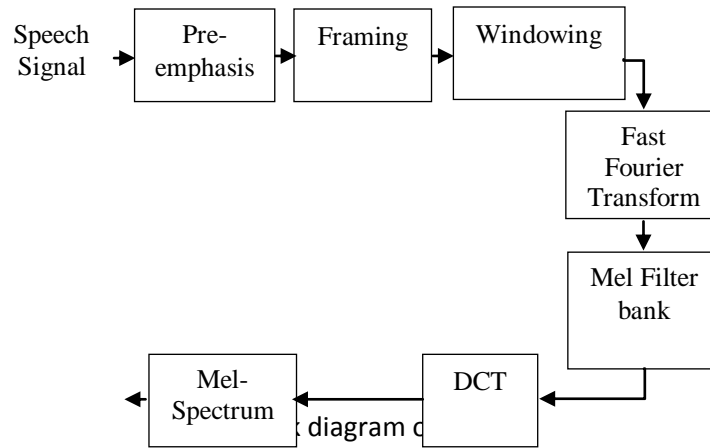


Fig. 3: The HMM based statistical speech synthesis system



The feature extraction process is shown Fig. 4.



2.1.1 Per-emphasis

The aim of this step is to attenuate the high frequencies of the audio which was stifled in the human’s audio production mechanism. It can also boost up the amplitude of high frequency signal. The speech signal $S[n]$ is sent to a high pass lower order filter defined in Eq. (1) and the equivalent Z-transform of the filter defined in Eq. (2).

$$S_y[n] = S[n] - \alpha S[n - 1] \tag{1}$$

where $S_y[n]$ is an output signal and α lies between, $0.9 < \alpha < 1$.

$$H[Z] = 1 - \alpha Z^{-1} \tag{2}$$

2.1.2 Frame blocking

A sound wave can be divided into several numbers of frames. Since each frame is individually analyzed and synthesized, it can be labeled with a single vector. This method ensures that the continuous audio signal is not lost in information. The resulting audio signal is then divided into N numbers of samples. Where each neighboring frame is separated by $(N < M)$, the sound wave is labeled as coherent. Here the first frame contains N numbers of samples and the next frame starts M numbers of the samples so the first frame and second frame overlap by $(N - M)$ numbers of samples. Correspondingly, the next frame contains $2M$ number of samples. Therefore, the first frame and the third frame are overlap by $(N - 2M)$ numbers of samples. In the similar manner the speech signal is framing so no discontinuity occurs in audio samples.

2.1.3 Windowing

A window is needed to smooth the effect of using a finite-sized segment for the subsequent feature extraction by tapering each frame at the beginning and end edges. The hamming window is multiplied with each frames so the continuity will be occurs in between the first point of the first frame and last point of the last frame.

If the signal is denoted as $s[n]$, and hamming windowing is denoted as $w[n]$, then it is mathematically defined as Eq. (3).

$$Y[n] = s[n]w[n] \tag{3}$$

2.1.4 Fast Fourier Transform

The Fast Fourier transform (FFT) is a commonly used algorithm for calculating the power spectrum and the frequency spectrum of a frame. It is also referred to as Short-Time Fourier-Transform (STFT).

2.1.5 Mel Filter Bank



The Mel spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel-filter bank shown in Fig. 5. The triangular filters is applied on a Mel-scale to the power spectrum to extract frequency bands. The location of these filters are equal and it gives a relation in between linear frequency and mel frequency and it is defined as Eq. (4) where f denotes the physical frequency in Hz , and $mel[f]$ denotes the perceived frequency. Humans do not hear sound in a linear scale so the logarithm of the magnitude spectrum computed. The considered filter bank parameters of audio signals are as follows:

Number of windows: 22
Length of each feature: 26

$$mel[f] = 2595 \times \log_{10}(1 + f/700) \tag{4}$$

2.1.6 Discrete Cosine Transform

Due to overlapped filter banks the filter energies are correlated. DCT is computed to de-correlate the filter bank energies. The spectral values from second to fourteenth are taken and 13 Coefficient will represent information regarding vocal tract features.

2.1.7 Mel Spectrum Features by Deltas and Double deltas

Speech signal is not constant from frame to frame. Features related to change in time are taken into account. The 13 delta or velocity and 13 double delta or acceleration features are obtained. The delta coefficients are obtained by Eq. (5).

$$\delta(t) = c(t + 1) - c(t) \tag{5}$$

Double delta is computed from the delta at time $t + 1$ to time t in a similar manner as we compute delta.

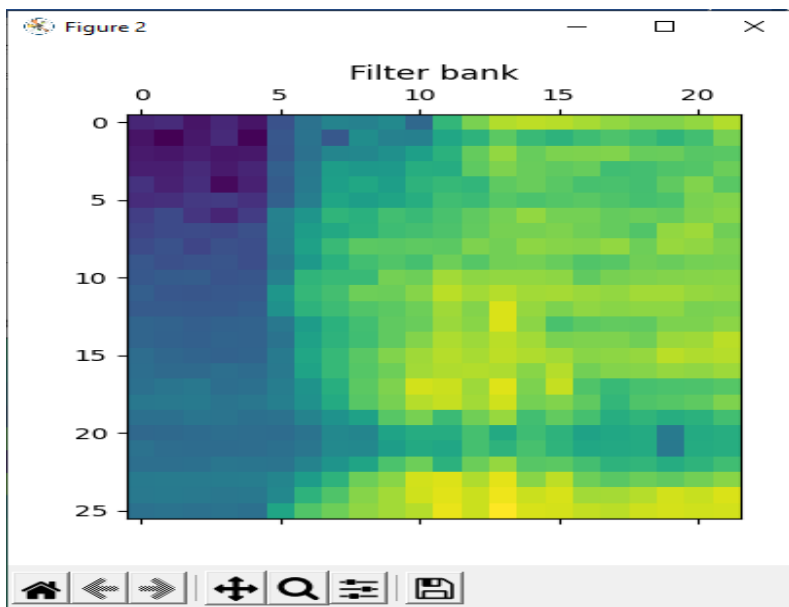


Fig. 5: Filter bank Feature Extraction from audio signal

2.2 Speech Recognition

A speech recognition system is a process that uses a device's audio to identify the words that are spoken as shown in Fig. 6. It then performs a variety of acoustic analyses to find out what the words actually mean.



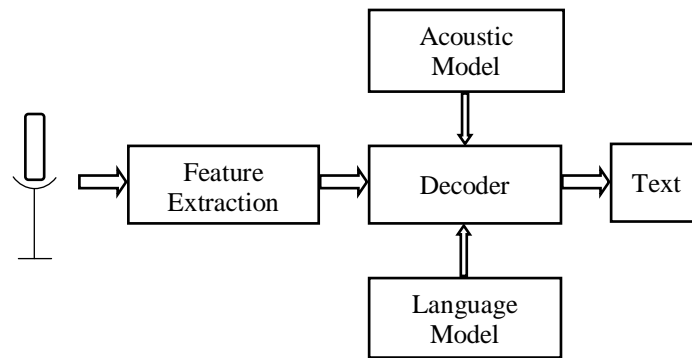


Fig. 6: *Speech Recognition*

2.2.1 Visualizing Audio Signals

The first step in creating a speech recognition system involves understanding the audio signal's structure. This process can be performed by recording the audio signal using a microphone, which is very important since machine learning cannot understand digitized audio signals. After converting the audio signal into a discrete form, it needs to perform a sampling procedure at a certain frequency. This method ensures that the signal is continuous. Since humans prefer to hear continuous audio signals over digitized ones, a graph and data extracted from the audio signal are generated as shown in Fig. 7.

Data extracted from given audio signal are as follows:

Signal shape: (9051264, 2)

Signal data type: int16

Signal duration: 205.24 seconds

2.2.2 Feature Extraction

The next step in creating a speech recognition system is to perform feature extraction. This process is very important since it involves converting the audio signal into a usable feature vector. For analysis of audio signals, feature extraction techniques are used such as LPC, MFCC etc. The feature extraction process is shown in Fig. 8. It is divided into three parts as pre-emphasis, frame blocking & windowing, and feature extraction.

2.2.3 Acoustic model

The acoustic model is the main component of a speech recognition system. It is developed to improve the system's performance by detecting the spoken phoneme. This is a minimal unit of sound that is used to distinguish between the words' meanings.

2.2.4 Language model

The language model is the single largest component that is trained on billions of words. It is developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary. For predicting correct word sequence of n^{th} likelihood word on the basis of $(n - 1)^{th}$ preceding word, $n - gram$ language model is used. The probability of occurrence $P(W)$ of a word sequence W is calculated where

$$P(W) = P(W_1, W_2, \dots, W_{n-1}, W_n) = P(W_1) \cdot P(W_2|W_1) \cdot P(W_3|W_1W_2) \cdot P(W_n|W_1W_2 \dots W_{n-1})$$

2.2.5 Lexical Model

Lexicon is developed for Word to Phoneme mapping. It is responsible for pronunciation of each word in a given language. It is used in creation of HMM. Through lexical model, various combinations of phones are defined to give valid words for the recognition.

2.2.6 Decoder

Decoder is used to find most likely word sequence given from transmitted speech waveform and acoustic model, transform it into text for future use.



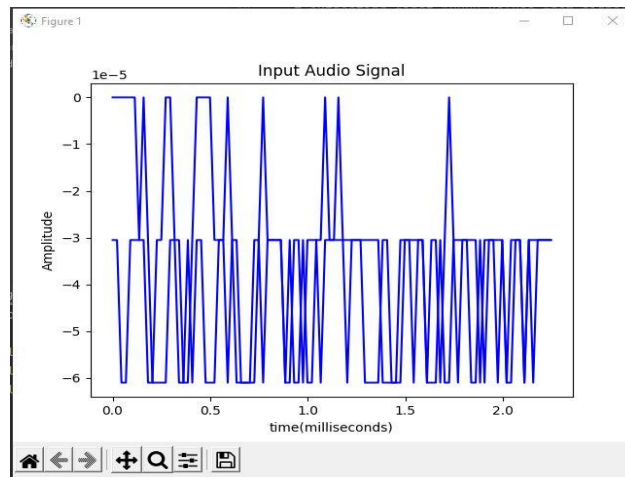


Fig. 7: Graph of recorded audio signal

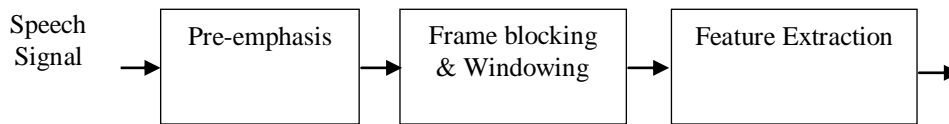


Fig. 8: Block diagram of feature extraction

2.3 Noise Elimination

In the real world, audio signal may contain channel distortions and background noise. To minimize these effects, it is important that the system uses the Noise Reduction technique before it passes the audio signal to the speech transcription system. This technique is carried out by using the Mel Frequency Cepstral coefficient (MFCC). The goal of the MFCC technique is to extract the specific details of the speech. This process can be performed by taking into account the speaker's specific parameters. The MFCC feature extraction from audio signal shown in Fig. 9. Following

MFCC parameters used for given audio signals:

Number of Windows = 22

Length of each feature = 13

2.4 Email Module

Email is one of the most valuable services on the internet at the present time. Many of the internet systems use SMTP as a method to transfer mail from one user to another. SMTP is a sending protocol and is used to send the mail while POP (post office protocol) or IMAP (internet message access protocol) are used to retrieve those emails at the receiver's side.



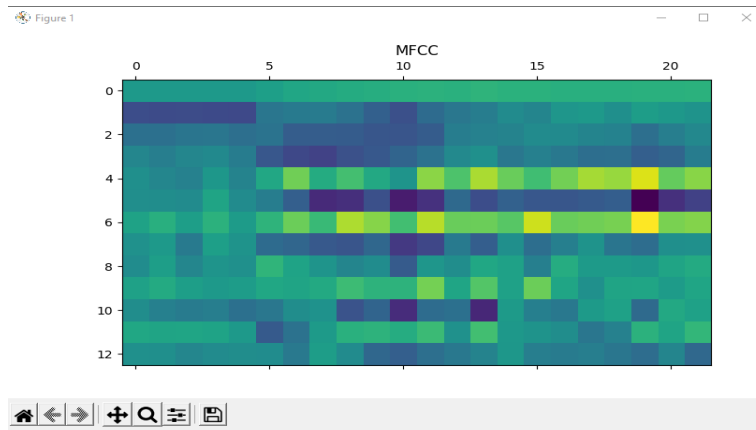


Fig. 9: MFCC Feature Extraction from audio signal.

3 RESULT AND DISCUSSION

3.1 System Introduction

Whenever system will run it will introduce itself and wait for user's command as shown in Fig. 10.

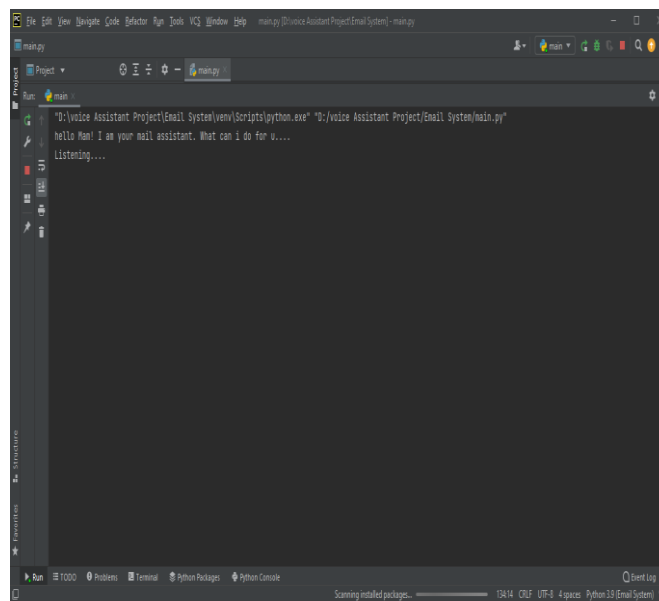


Fig. 10: System Introduction

3.2 Information gathering

The required information of receiver's and message detail from user are collected as shown in Fig. 11.

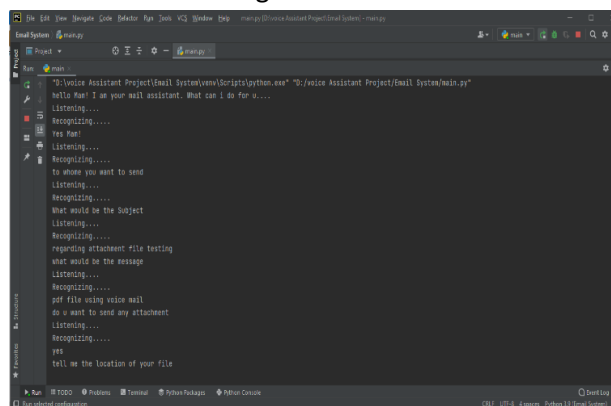


Fig. 11: Information gathering



3.3 Searching File

The location of file which is stored in the system, search for it, select it and attach it to the mail system and send it to receiver's email address as shown in Fig. 12.

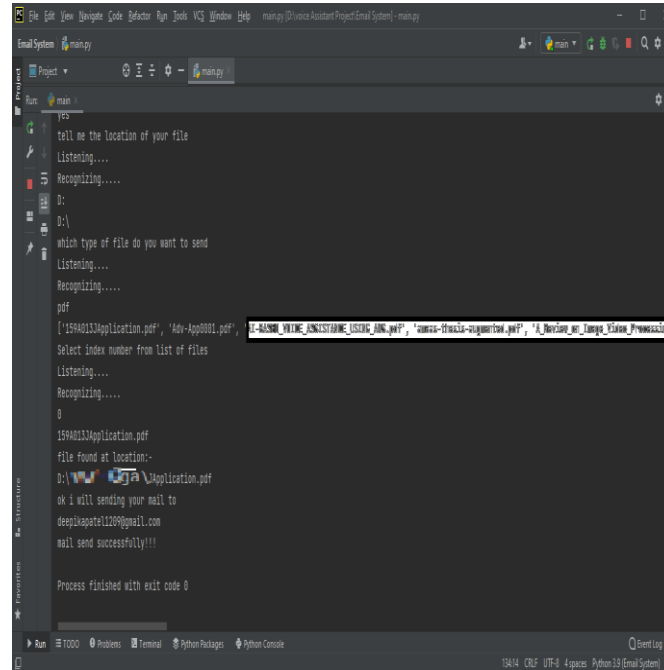


Fig. 12: Searching File in system

3.4 Receiver's mailbox

The Pop operation performed on new mail in receiver's email account which is send by user via voice command as shown in Fig. 13.

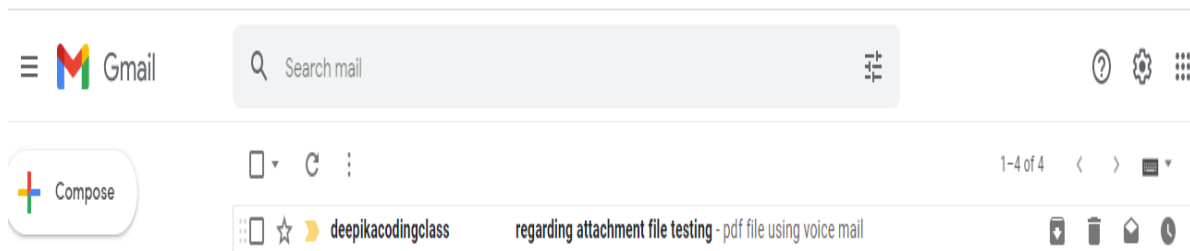


Fig. 13: Receiver's mailbox

3.5 Receive pdf file via voice command

The result of receiving pdf file via voice command is shown in Fig. 14.



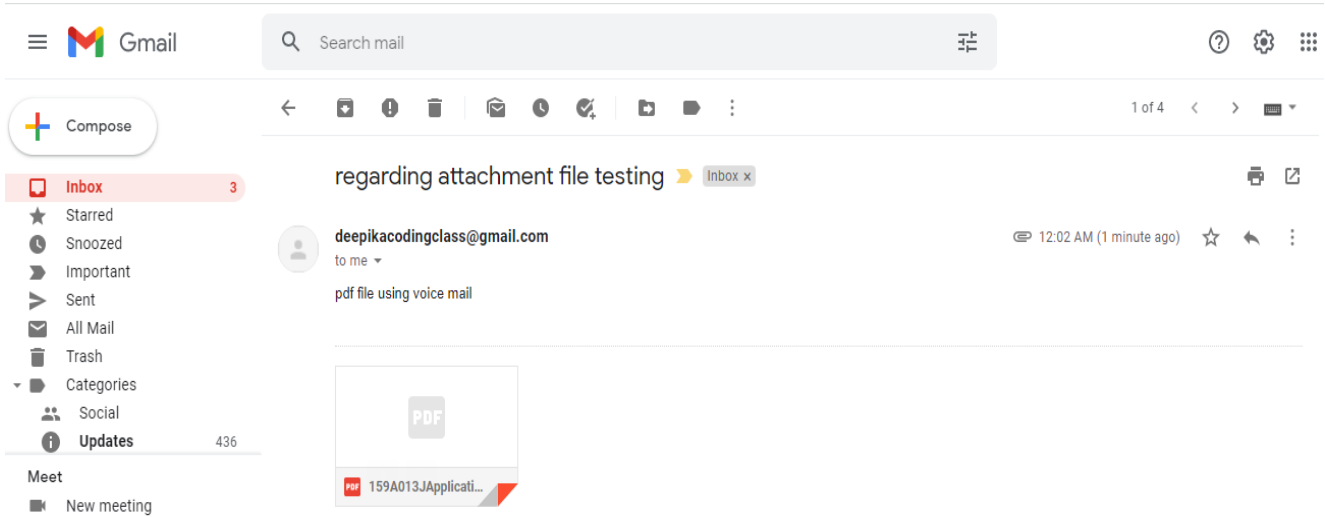


Fig. 14:Receiving pdf file

3.6 Receive image via voice command

The result of received png file via voice command is shown in Fig. 15.

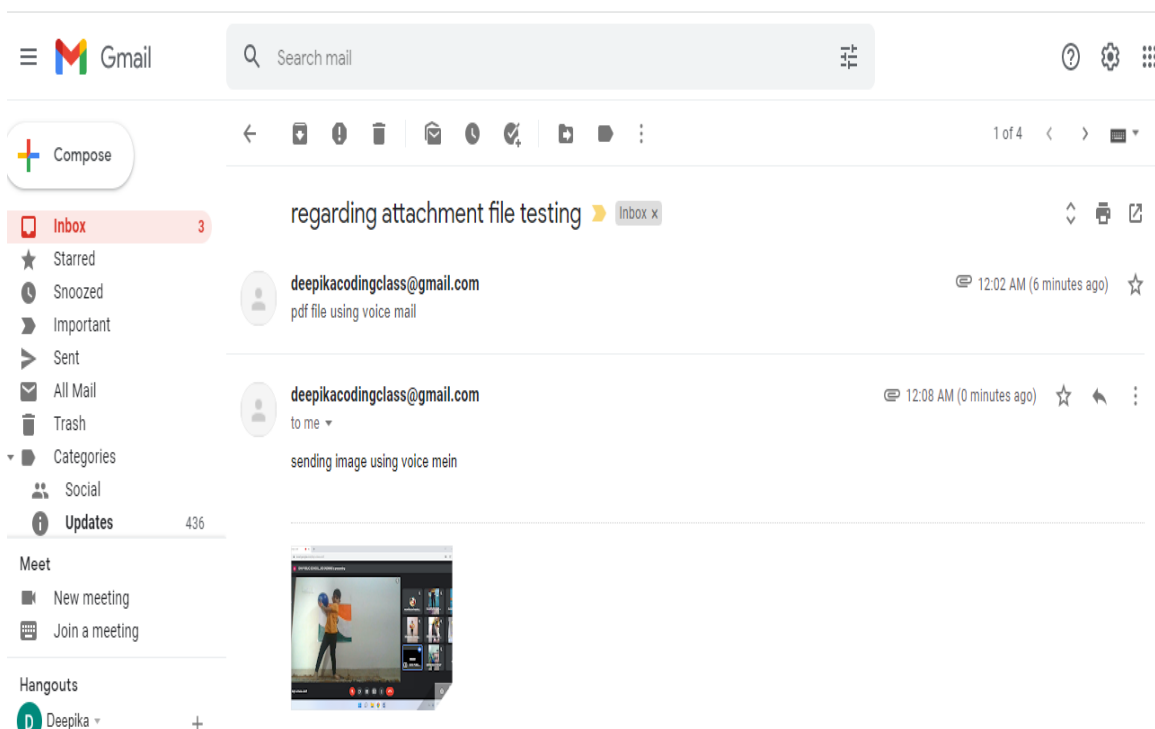


Fig. 15: Receiving png file



4. CONCLUSION

The Voice Mail System is a well-designed and user-friendly system that can be easily accessed by any age group. It can help society by allowing the visually challenged and illiterate people to grow. It also improves the features of the existing mail system. With this system, one can easily send and receive files with voice commands. The use of the keyboard has been eliminated with the Voice Mail System. Instead, the user will only require responding to the commands that are given by the system in order to perform the desired operations. For sending attachments, the user must provide the location of the file so that the system can access it.

There is wide future scope of this system. Aside from being able to send multiple attachments, the Voice Mail System can also be enhanced to allow users to access different languages. It can additionally be used to access deleted mails and spam mails. In addition, a sign language system can be integrated into the system to make it more robust and scalable.

The development of a large vocabulary speech recognition model is often hindered by the lack of training data. This issue can lead to the development of large models that are ideal for real applications. Another issue is the lack of sparseness. This is typically encountered during the training of domain-specific models.

References

- [1] Khan, Rijwan, Pawan Kumar Sharma, Sumit Raj, Sushil Kr Verma, and SparshKatiyar. "Voice Based E-Mail System Using Artificial Intelligence." International Journal of Engineering and Advanced Technology (IJEAT), Volume 9, Issue 3, 2020.
- [2] Shah, Harsh D., Amit Sundas, and Shabnam Sharma. "Controlling Email System Using Audio With Speech Recognition And Text To Speech." In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 1-7. IEEE, 2021.
- [3] Chaudhary, Bhushan.S., GunjanDhande,

Shital Salve, SanikaLohar and SonaliSasane. "Voice Based Email System For Blind Person." IJARIE, Vol-7, Issue-2 2021

- [4] Chidgopkar, Vedant, SurajJadhav, Atharva Joshi and Abhishek Khedekar. "Voice Based E-Mail System For The Blind." International Research Journal of Engineering and Technology (IRJET), Volume: 07, Issue: 04, Apr 2020
- [5] Ingle, Pranjal, HarshadaKanade, ArtiLanke. "Voice Based E-Mail System For Blinds." International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), Volume 3, Issue 1, 2016.
- [6] Mullapudi, Harshasri, ManyamDurgaBhavani, and MisraRavikanth. "Voice Based Email For Blind." International Journal of Innovative Research in Computer Science & Technology (IJRCST), Volume-9, Issue-4, July 2021.
- [7] KukadeRutuja V., Ruchita G. Fengse, Kiran D. Rodge, Siddhi P. Ransing, Vina M. Lomte. "Virtual Personal Assistant For The Blind", International Journal of Computer Science And Technology(IJCST)." International Journal of Engineering and Advanced Technology (IJEAT), Volume 9, Issue 4, October - December 2018.
- [8] Singh, Parwinder Pal, Pushpa Rani. "An Approach to Extract Feature using MFCC." IOSR Journal of Engineering (IOSRJEN), Vol. 04, Issue 08, August. 2014.
- [9] Chavan,Prajakta, Devesh Jain, Pradnya Savant, Zeba Shaikh. "VOICE BASED EMAIL SYSTEM." International Journal of Scientific & Engineering Research, Volume 9, Issue 2, February-2018.
- [10] Ranjan, Rajeev, Abhishek Thakur. " Analysis of Feature Extraction Techniques for Speech Recognition System." International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-7C2, May 2019.
- [11] Bhatt,Shobha, Anurag Jain,Amita Dev. "Acoustic Modeling In Speech Recognition: A



- Systematic Review." International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 4, 2020.
- [12] Anusuya M.A., S.K. Katti. "Speech Recognition By Machine: A Review." International Journal of Computer Science and Information Security (IJCSIS), Vol. 6, No. 3, 2009.
- [13] Ghai, Wiquis, Navdeep Singh. "Literature Review On Automatic Speech Recognition." International Journal of Computer Applications, Vol. 41, No. 8, 2012.

