Project

on

INTERACTIVE VOICE RESPONSE SYSTEM WITH SPEECH RECOGNITION

Submitted in partial fulfillment of the

requirement for the award of the degree of B.tech CSE

B.Tech in Computer Science Engineering

GALGOTIAS UNIVERSITY

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

Under The Supervision of Name of Supervisor:

Dewan Imdadul Islam

Submitted By:

Name – Rishav Raj	**Admission No. -** 22SCSE1010527
Name – Priya Tiwari	**Admission No. -** 22SCSE1010536
Name – Moh. Yusuf Ali	**Admission No. -** 22SCSE1010545
Name – Vaibhav Singh	**Admission No. -** 22SCSE1010550

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

IV semester, October, 2022-2023

GALGOTIAS UNIVERSITY

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

ABSTRACT

This report represents the project report on a project entitled "Interactive Voice Response System with Speech Recognition". The report discusses various methods and implementation techniques to build the system of IVR system with speech recognition and presents the results of the system being implemented. The report represents project work of developing an Interactive Voice Response (IVR) system with the capability of Speech Recognition. It is a system where the user can interact with his voice and on the basis of the user's voice input the system performs particular action as a response to the voice command. It is an approach to make use of the voice of the user to recognize the given command instead of using keyboard or other form of input. The project is helpful for various large organizations where there are many call queries or for telecommunications companies or other business ventures where user queries through their voice online or via cellular connections. The system provides an application environment whereby users are prompted to input specific voice commands and the given voice input is being processed for further task accomplishment as per the request. Making use of speech as an input signal significantly increases user experience. Unlike traditional touch-tone or key-press based IVR system the system being developed is capable to take a real time voice input from a user. Speech recognition applications are becoming more and more useful nowadays. With growth in the needs for embedded computing and the demand for emerging embedded platforms, it is required that the speech recognition systems (SRS) are available on them too. This project is a simple approach to developing a system where the speech recognition is embedded within another system to automate task using speech as input command.

TABLE OF CONTENTS

INTRODUCTION

Background

Speech probably is the most efficient and natural way to communicate with each other. Thus, being the best way of communication, it could also be a useful interface to communicate with machines and systems like IVR system. The Interactive Voice Response (IVR) system alongwith the speech recognition technology can play efficient role in providing easy and efficient customer/user service. If properly implemented it can increase the user satisfaction and offer new services. Speech Recognition has now begun to dominate the market technologyand is pushing away the traditional way of using hectic interfaces such as keyboards and mouse as input source to computer system. Voice command based applications will make life easier due to the fact that people will get easy and fast access to information. Thereforethe popularity of automatic speech recognition system has been greatly increased. The workof speech recognition further helps in establishing an easy way communication between interactive response system and users/ customers i.e. as a part of post processing of the speech recognizing process we can accomplish some computational task with such a system making voice input as a trigger to do some task within the system.

Problem Statement

Most of the works done till today on the field of IVR system has been primarily focused on the input mechanisms based on the keyboard or touch pad. In such cases it is tedious to provide the input command every time through typing of texts. This way of providing input to the computer system may be enhanced if we could provide direct speech input instead of typing. This enables in fast interaction between the system and user and therefore increases overall satisfaction of the customers. This also increases the speed of access of the information from the system. Furthermore, English language has been widely implemented in IVR systems. This has created difficulty for people while interacting with the system. Thus by implementing the Nepali voice commands it is easier to interact and provide the input to the system.

The major focus of the project being developed is the use of direct Nepali voice command for the interactive voice response system without need of typing which then further can be applicable to real world applications like call centers, customer support systems and other several organization inquiry systems.

Speech Recognition Overview

Speech Recognition (SR) is the process of extracting the string of words automatically from the speech signal, by means of an algorithm. It is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine readable format. Speech recognition is a powerful tool of the information exchange using the acoustic signal. Therefore, not surprisingly, the speech signal is for several centuries the subject of research. Speech recognition is a technology that makes a computer able to capture the words

spoken by a human with a help of microphone. These words are later on recognized by speech recognizer, and in the end, system outputs the recognized words which can also serve as input to the further systems to accomplish several task. Speech recognition is basically the science of talking with the computer, and having it correctly recognized. Speech recognition is getting the meaning of an utterance such that one can respond properly whether or not one has correctly recognized all of the words. ASR has always been considered as an important bridge in fostering better human to human and human to machine communication. In the past, however, speech never actually became an important modality in the human to machine communication. This is partly because the technology at that time was not good enough to pass the usable bar for most real world users under most real usage conditions, and partly because in many situations alternative communication modalities such as keyboard and mouse significantly outperform speech in the communication efficiency, restriction, and accuracy. In the recent years, speech technology started to change the way we live and work and became one of the primary means for humans to interact with some devices. This trend started due to the progress made in several key areas. First, Moor's law continues to function. The computational power available today, through multi-core processors, general purpose graphical processing units (GPUs), and CPU/GPU clusters, is several orders of magnitude more than that available just a decade ago. This makes training of more powerful yet complex models possible. These more computation demanding models significantly reduced the error rates of the ASR systems. Second, we can now access to much more data than before, thanks to the continued advance of the Internet and the cloud computing. By building models on big data collected from the real usage scenarios, we can eliminate many model assumptions made before and make systems more robust. Third, mobile devices, wearable devices, intelligent living room devices, and in vehicle infotainment systems became popular. On these devices and systems, alternative interaction modalities such as keyboard and mouse are less convenient than that in the personal computers. Speech, which is the natural way of human to human communication and a skill that majority of people already have, thus becomes a more favorable interaction modality on these devices and systems.

Speech Recognition Overview

Speech Recognition (SR) is the process of extracting the string of words automatically from the speech signal, by means of an algorithm. It is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine readable format. Speech recognition is a powerful tool of the information exchange using the acoustic signal. Therefore, not surprisingly, the speech signal is for several centuries the subject of research. Speech recognition is a technology that makes a computer able to capture the words spoken by a human with a help of microphone. These words are later on recognized by speech recognizer, and in the end, system outputs the recognized words which can also serve as input to the further systems to accomplish several task. Speech recognition is basically the

science of talking with the computer, and having it correctly recognized. Speech recognition is getting the meaning of an utterance such that one can respond properly whether or not one has correctly recognized all of the words. ASR has always been considered as an important bridge in fostering better human to human and human to machine communication. In the past, however, speech never actually became an important modality in the human to machine communication. This is partly because the technology at that time was not good enough to pass the usable bar for most real world users under most real usage conditions, and partly because in many situations alternative communication modalities such as keyboard and mouse significantly outperform speech in the communication efficiency, restriction, and accuracy. In the recent years, speech technology started to change the way we live and work and became one of the primary means for humans to interact with some devices. This trend started due to the progress made in several key areas. First, Moor's law continues to function. The computational power available today, through multi-core processors, general purpose graphical processing units (GPUs), and CPU/GPU clusters, is several orders of magnitude more than that available just a decade ago. This makes training of more powerful yet complex models possible. These more computation demanding models significantly reduced the error rates of the ASR systems. Second, we can now access to much more data than before, thanks to the continued advance of the Internet and the cloud computing. By building models on big data collected from the real usage scenarios, we can eliminate many model assumptions made before and make systems more robust. Third, mobile devices, wearable devices, intelligent living room devices, and in vehicle infotainment systems became popular. On these devices and systems, alternative interaction modalities such as keyboard and mouse are less convenient than that in the personal computers. Speech, which is the natural way of human to human communication and a skill that majority of people already have, thus becomes a more favorable interaction modality on these devices and systems.

Human Ear and Speech Recognition

Humans are more effective than machines at recognizing speech. This advantage for human listeners is particularly pronounced for speech that is heard against background noise, contains unfamiliar words or is degraded in other ways. Yet, automatic speech recognition (ASR) systems have made substantial advances over the past few decades and are now in everyday use by millions of people around the world. Until the performance of automatic speech recognition (ASR) surpasses human performance in accuracy and robustness, we stand to gain by understanding the basic principles behind human speech recognition (HSR). In this section we provide a brief explanation of how human hearing works and how it is modeled. We will discuss in brief the functionality of several components and try to understand the

relation of it with the speech recognition systems.

Human Hearing system

The main function of hearing system is to get information about the outside, which is carried by pressure variations in the air, that is, sound wave. Sound waves are generated by the movement or vibration of an object, that is, sound source. As the vibrating object moves out and in, the nearby air molecules create a slight increase and decrease in pressure, called condensation and rarefaction, respectively. From the pressure variations, we perceive what the sound source is and where it comes from. We perceive a sound wave, which is a continual time series signal, by the ears. We also perceive three-dimensional acoustic space by the ears, mainly because the head-related transfer function (HRTF) between a point of a sound source and the two ear entrances has directional characteristics from the shapes of the head and the pinna. The pinna significantly modify the incoming sound, particularly at high frequencies,and this is important in our ability for sound localization.After a sound wave arrives nearby,it passes through the peripheral auditory system, the outer ear, middle ear, and inner ear.

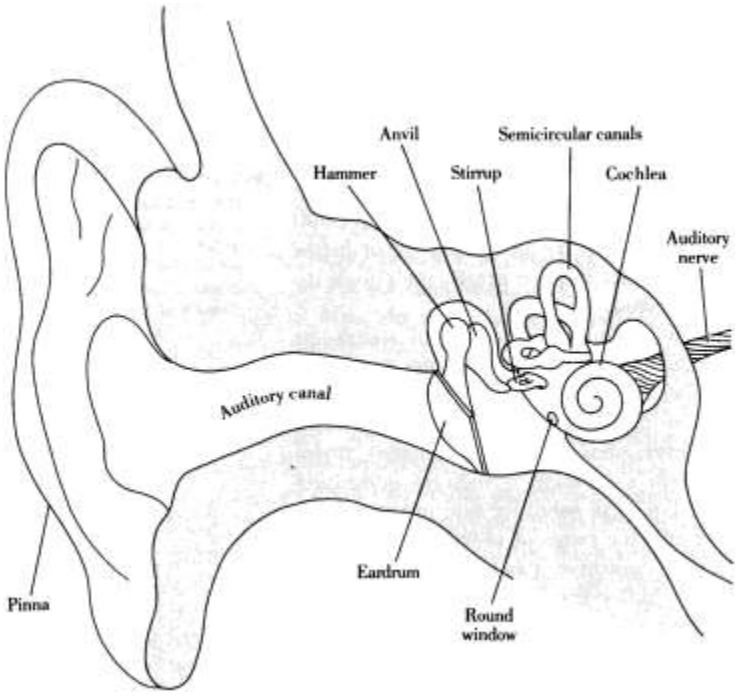

Figure Human Ear Hearing System

1. The Outer Ear

   The outer ear is the external part of the auditory system, including the pinna and the ear canal. Sound travels down the ear canal and causes the eardrum, or tympanic membrane, to vibrate. Because of the resonance of the outer ear, we are more sensitive to sound frequencies between 1000 and 6000 Hz. The pinna is the only visible part ofthe ear (the auricle) with its special helical shape. It is the first part of the ear that reacts with sound. The function of the pinna is to act as a kind of funnel which assists in directing the sound further into the ear. Without this funnel the sound waves would take a more direct route into the auditory canal. This would be both difficult and wasteful as much of the sound would be lost making it harder to hear and understand the sounds.The pinna is essential due to the difference in pressure inside and outside the ear. Theresistance of the air is higher inside the ear than outside because the air inside the ear is compressed and thus under greater pressure. In order for the sound waves to enter the ear in the best possible way the resistance must not be too high. This is where the pinna helps by overcoming the difference in pressure inside and outside the ear. The pinnafunctions as a kind of intermediate link which makes the transition smoother and lessbrutal allowing more sound to pass into the auditory canal (meatus).Once the soundwaves have passed the pinna, they move two to three centimetres into the auditory canalbefore hitting the eardrum, also known as the tympanic membrane. The function of the ear canal is to transmit sound from the pinna to the eardrum. The eardrum (tympanic membrane), is a membrane at the end of the auditory canal and marks the beginningof the middle ear. The eardrum is extremely sensitive and pressure from sound wavesmakes the eardrum vibrate. The auditory canal functions as a natural hearing aid which automatically amplifies low and less penetrating sounds of the human voice. In this way the ear compensates for some of the weaknesses of the human voice, and makes it easier to hear and understand ordinary conversation.

2. The Middle Ear

   The middle ear is the part of the ear between the eardrum and the oval window. The middle ear transmits sound from the outer ear to the inner ear. The middle ear consists of three bones: the hammer (malleus), the anvil (incus) and the stirrup (stapes), the oval window, the round window and the Eustrachian tube.The eardrum is very thin, measures approximately eight to ten millimeter in diameter and is stretched by means of small muscles. The pressure from sound waves makes the eardrum vibrate. The vibrations are transmitted further into the ear via three bones in the middle ear. These three bones form a kind of bridge, and the stirrup, which is the last bone that sounds reach, is connected to the oval window. The oval window is a membrane covering the entrance to the cochlea in the inner ear. When the eardrum vibrates, the sound waves

travel via the hammer and anvil to the stirrup and then on to the oval window. When the sound waves are transmitted from the eardrum to the oval window, the middle ear is functioning as an acoustic transformer amplifying the sound waves before they move on into the inner ear. The pressure of the sound waves on the oval window is some 20 times higher than on the eardrum.

The pressure is increased due to the difference in size between the relatively large surface of the eardrum and the smaller surface of the oval window. The round window in the middle ear vibrates in opposite phase to vibrations entering the inner ear through the oval window. In doing so, it allows fluid in the cochlea to move. The Eustachian tube is also found in the middle ear, and connects the ear with the rearmost part of the palate. The Eustachian tube's function is to equalize the air pressure on both sides of the eardrum, ensuring that pressure does not build up in the ear. The tube opens when you swallow, thus equalizing the air pressure inside and outside the ear.

3. The Inner Ear

The inner ear is the innermost part of the ear, which consist of the cochlea, the balance mechanism, the vestibular and the auditory nerve. Once the vibrations of the eardrum have been transmitted to the oval window, the sound waves continue their journey into the inner ear. The inner ear is a maze of tubes and passages, referred to as the labyrinth. In the labyrinth can be found the vestibular and the cochlea.

In the cochlea, sound waves are transformed into electrical impulses which are sent on to the brain. The brain then translates the impulses into sounds that we know and understand.The cochlea resembles a snail shell or a wound-up hose and is filled with a fluid called perilymph and contains two closely positioned membranes. These membranes form a type of partition wall in the cochlea. However, in order for the fluid to move freely in the cochlea from one side of the partition wall to the other, the wallhas a little hole in it (the helicotrema). This hole is necessary, in ensuring that the vibrations from the oval window are transmitted to all the fluid in the cochlea.

The auditory nerve is a bundle of nerve fibres that carry information between the cochlea in the inner ear and the brain. The function of the auditory nerve is to transmit signals from the inner ear to the brain.The hair fibres in the cochlea are all connected to the auditory nerve and, depending on the nature of the movements in the cochlear fluid, different hair fibres are put into motion. When the hair fibres move they send electrical signals to the auditory nerve which is connected to the auditory centre of the brain. In the brain the electrical impulses are translated into sounds which we recognise and

understand. As a consequence, these hair fibres are essential to our hearing ability. Should these hair fibres become damaged, then our hearing ability will deteriorate.

The vestibular is another important part of the inner ear. The vestibular is the organ of equilibrium. The vestibular's function is to register the body's movements, thus ensuring that we can keep our balance. The vestibular consists of three ring-shaped passages, oriented in three different planes. All three passages are filled with fluid that moves in accordance with the body's movements. In addition to the fluid, these passages also contain thousands of hair fibres which react to the movement of the fluid sending little impulses to the brain. The brain then decodes these impulses which are used to help the body keep its balance.

Sound Signal Transduction Mechanism

In the human ear the basiliar membrane is contained within cochlea that supports thousands of sensory cells which forms the cochlear nerve. It is one of the innermost part of the ear. The basiliar membrane acts as a frequency spectrum analyzer. When exposed to a high frequency signal, the basilar membrane resonates where it is stiff, resulting in the excitation of nerve cells close to the oval window. Likewise, low frequency sounds excite nerve cells at the far end of the basilar membrane. This makes specific fibers in the cochlear nerve respond to specific frequencies. This organization is called the place principle, and is preserved throughout the auditory pathway into the brain. Also the principle called volley principle is used for the transduction purpose of the sound signal arriving the human ear. Here a nerve cell on the basilar membrane can encode audio information by producing an action potential in response to each cycle of the vibration. For example, a 200 hertz sound wave can be represented by a neuron producing 200 action potentials per second. However, this only works at frequencies below about 500 hertz, the maximum rate that neurons can produce action potentials. The human ear overcomes this problem by allowing several nerve cells to take turns performing this single task. For example, a 3000 hertz tone might be represented by ten nerve cells alternately firing at 300 times per second. This extends the range of the volley principle to about 4 kHz, above which the place principle is exclusively used.

Table below shows the relationship between sound intensity and perceived loudness. It is common to express sound intensity on a logarithmic scale, called decibel SPL (Sound Power Level). On this scale, zero dB SPL is a sound wave power of $10^{-16} watts/cm2$, about the weakest sound detectable by the human ear. Normal speech is at about 60 dB SPL, while painful damage to the ear occurs at about 140 dB SPL.

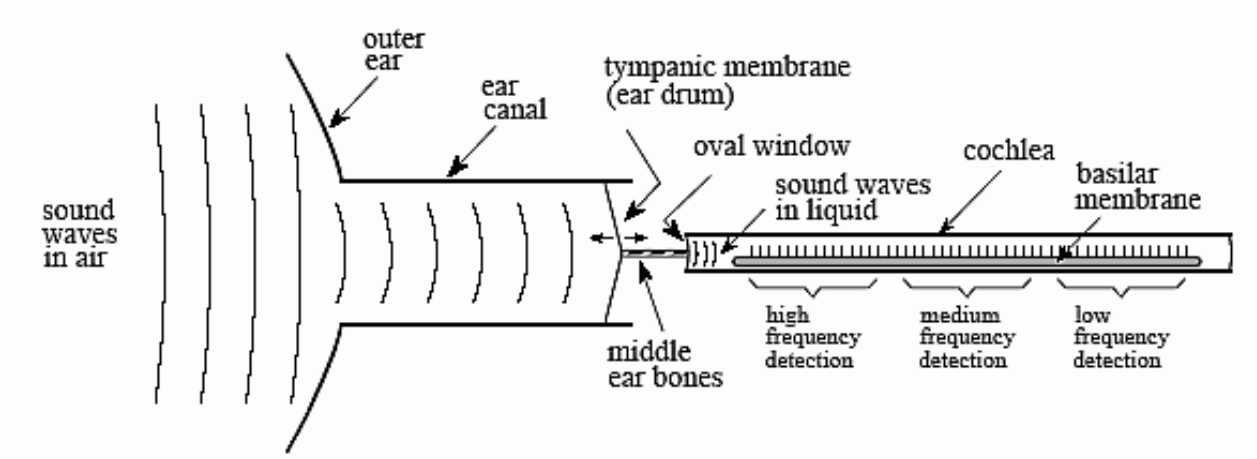

Figure Human Auditary System

The difference between the loudest and faintest sounds that humans can hear is about 120 dB, a range of one-million in amplitude. Listeners can detect a change in loudness whenthe signal is altered by about one dB (a 12 per cent change in amplitude). In other words, there are only about 120 levels of loudness that can be perceived from the faintest whisper to the loudest thunder. The sensitivity of the ear is amazing; when listening to very weak sounds, the ear drum vibrates less than the diameter of a single molecule. The perception of loudness relates roughly to the sound power to an exponent of 1/3. For example, if you increase the sound power by a factor of ten, listeners will report that the loudness has increased by a factor of about two ($10^{1/3} \approx 2$). This is a major problem for eliminating undesirable environmental sounds, for instance, the beefed-up stereo in the next door apartment. Suppose you diligently cover 99 per cent of your wall with a perfect soundproof material, missing only one per cent of the surface area due to doors, corners, vents, etc. Even though the sound power has been reduced to only 1 per cent of its former value, the perceived loudness has only dropped to about $0.011/3 \approx 0.2$, or 20 per cent. The range of human hearing is generally considered to be 20 Hz to 20 kHz, but it is far more sensitive to sounds between one kHz and four kHz. For example, listeners can detect sounds as low as 0 dB SPL at three kHz, but require 40 dB SPL at 100 hertz (an amplitude increase of 100). Listeners can tell that two tones are different if their frequencies differ by more than about 0.3 per cent at three kHz. This increases to three percent at 100 hertz. The primary advantage of having two ears is the ability to identify the direction of the sound. Human listeners can detect the difference between two sound sources that are placed as little as three degrees apart, about the width of a person at 10 meters. This directional information is obtained in two separate ways. First, frequencies above about one kHz are strongly shadowed by the head. In other words, the ear nearest the sound receives a stronger signal than the ear on the opposite side of the head. The second clue to directionality is that the ear on the far side of the head hears the sound slightly later than the near ear,

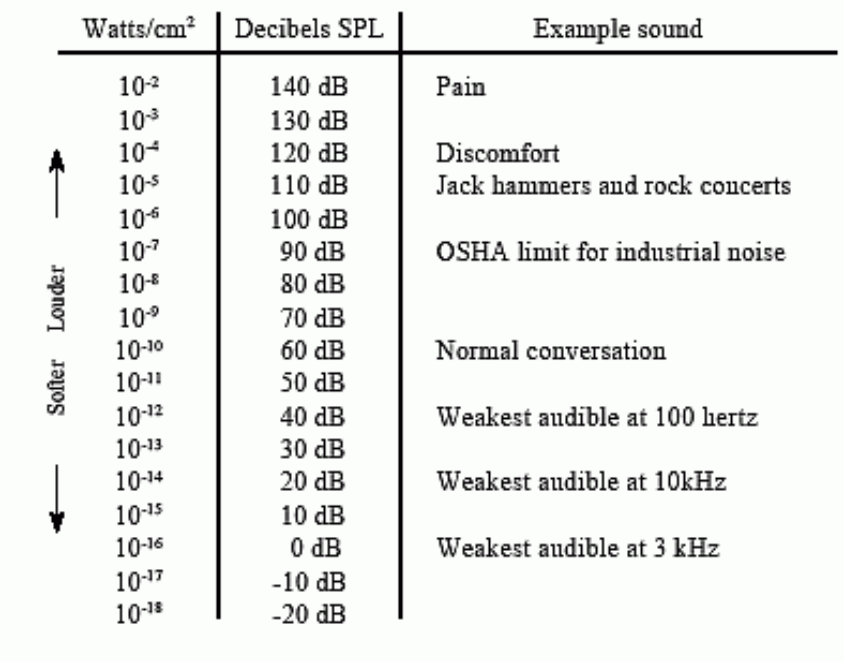

Figure Audibility range of sound by human ear at different intensity level

due to its greater distance from the source. Based on a typical head size (about 22 cm) and the speed of sound (about 340 meters per second), an angular discrimination of three degrees requires a timing precision of about 30 microseconds. Since this timing requires the volley principle, this clue to directionality is predominately used for sounds less than about one kHz. Both these sources of directional information are greatly aided by the ability to turn the head and observe the change in the signals. An interesting sensation occurs when a listener is presented with exactly the same sounds to both ears, such as listening to monaural sound through headphones. The brain concludes that the sound is coming from the center of the listener's head. While human hearing can determine the direction a sound is from, it does poorly in identifying the distance to the sound source. This is because there are few clues available in a sound wave that can provide this information. Human hearing weakly perceives that high frequency sounds are nearby, while low frequency sounds are distant. This is because sound waves dissipate their higher frequencies as they propagate long distances.

Speech Recognition System

The idea behind speech recognition is to provide a means to transcribe spoken words into written text. There exist many approaches to achieve this goal. The most simple technique is to build a model for every word that needs to be recognized.Speech signal primarily conveys the words or message being spoken. Area of speech recognition is concerned with determining

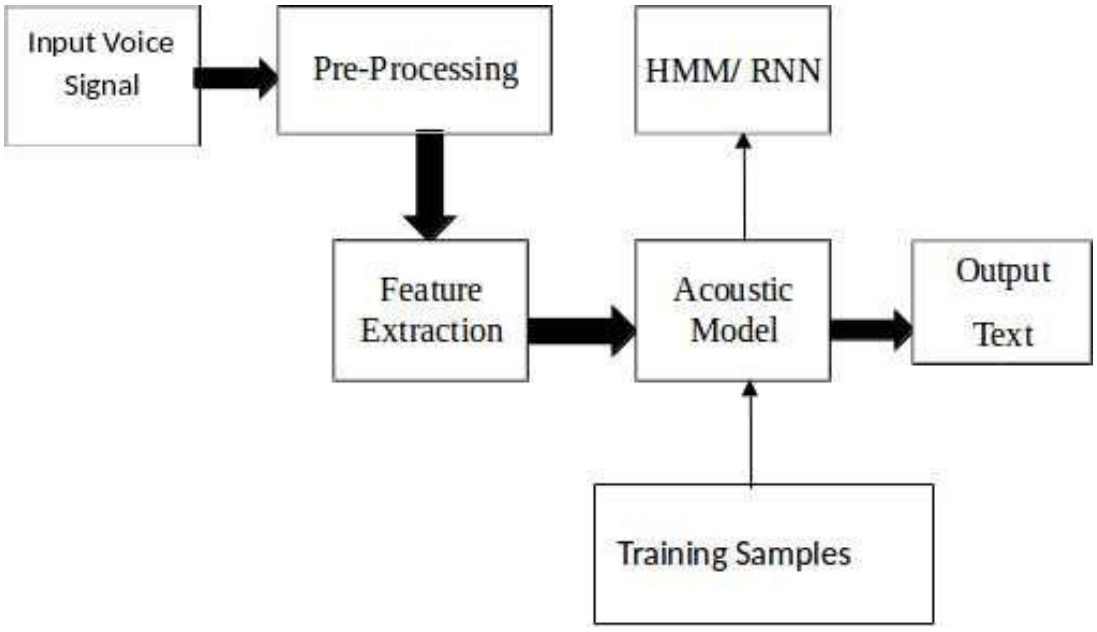

Figure: Speech Recognition System

the underlying meaning in the utterance. Success in speech recognition depends on extracting and modeling the speech dependent characteristics which can effectively distinguish one word from another. The system is a collection of several modules as shown in Figure

Input Voice Signal

The preliminary requirement of the IVR system is the input voice signal from the user such that the system may interact with the user. Most computer systems have the built in microphone facility for this purpose. Also with the help of external microphone the voice signal can be input to the system, the PC sound card produces the equivalent digital representation of received audio. During the phase of training the speech engine the recorded audio is taken and then are used for the sample generation and finally fed into the system model for training purpose and for the interactive voice response system the real time input voiceof the system user is taken using the microphone. The circumstances under which input voice signal is uttered plays important role in speech recognition i.e. the factors such as toonoisy environment, wrong utterance of word etc may diminish the performance of system. Therefore the input signal must be as clear as possible for the best results possible.

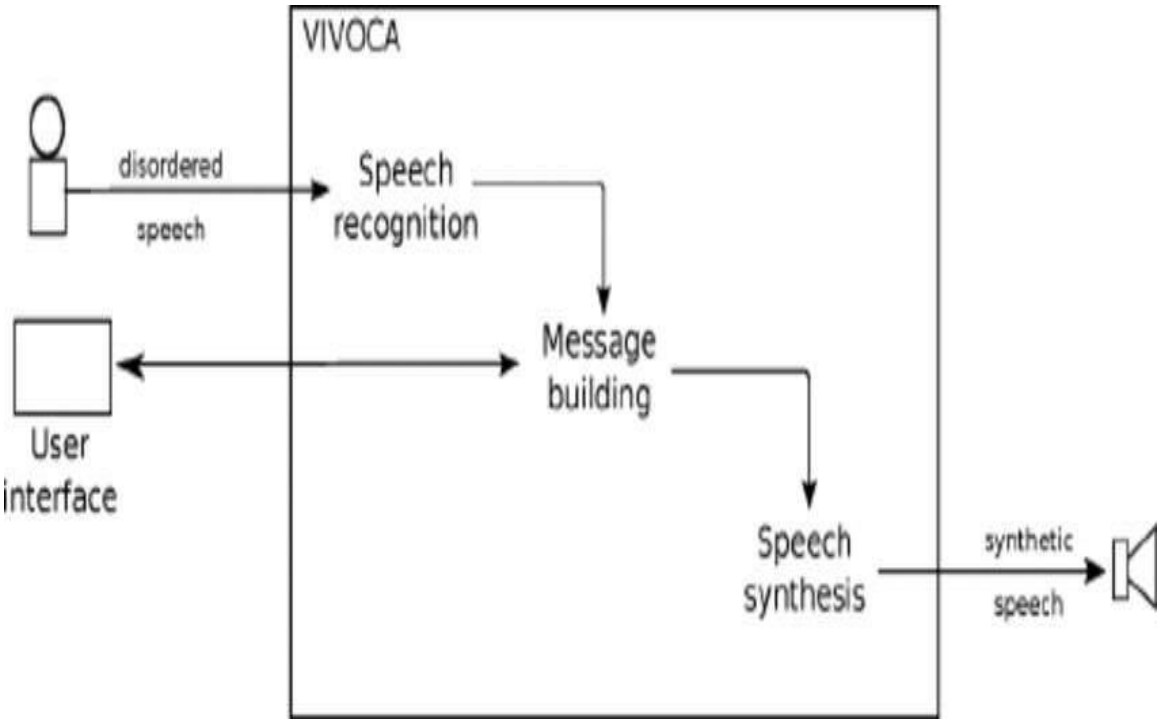

Figure: **Input Voice Signal**

Preprocessing stage

The stage of the speech preprocessing refers to the purification of the input voice signal so as to feed it into main speech recognition engine in a suitable format for best outcomes. The preprocessing stage in speech recognition systems is used in order to increase the efficiency of subsequent feature extraction and classification stages and therefore to improve the overall recognition performance. Commonly the preprocessing includes the sampling step, a windowing and a de-noising step as shown in Figure below. At the end of the preprocessing the compressed and filtered speech frames are forwarded to the feature extraction stage. These processes are discussed below in brief.

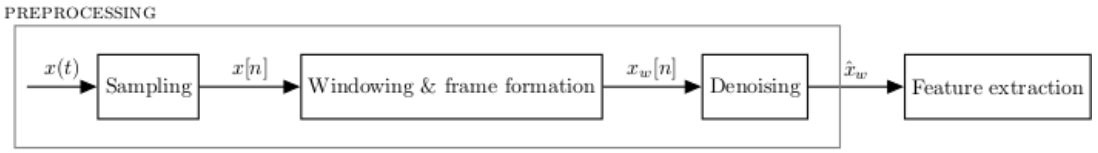

Figure: Input speech preprocessing

Sampling stage

In order that a computer is able to process the speech signal, it first has to be digitized. Therefore the time-continuous speech signal is sampled and quantized. The result is a time- and value discrete signal. According to the Nyquist-Shannon sampling theorem a time-continuous signal that is band limited to a certain finite frequency fmax needs to be sampled with a sampling frequency of at least 2 f_{max}. In this way it can be reconstructed by its time-discrete signal. Since human speech has a relatively low bandwidth (mostly between 100Hz and 8 KHz) a sampling frequency of 16 KHz is sufficient for speech recognition tasks.

Windowing and frame formation

Speech is a non-stationary time variant signal. We assume that human speech is built from a dictionary of phonemes, while for most of the phonemes the properties of speech remain invariant for a short period of time (5-100ms). In order to obtain frames we multiply the speech signal with a windowing function. This windowing function weights the signal in the time domain and divides it into a sequence of partial signals. By doing so we gain time information of every partial signal keeping in mindthat an important step of the preprocessing and feature extraction is a spectral analysis of each frame.

Denoising stage

The stage of denoising or noise reduction, also referred to as enhancing of speech degraded by noise, aims to improve the speech signals quality. The objective is to improve the intelligibility, a measure of how comprehensible speech is. Noise corrupting speech signals can be grouped coarsely into the following three classes:

- Microphone related noise
- Electrical noise (e.g. electromagnetically induced or radiated noise) and
- Environmental noise

The first two types of noise can be easily compensated by training the speech recognizers on corresponding noisy speech samples, but compensating the environmental noise is not that elementary, due to its high variability.

Noise is ubiquitous in almost all acoustic environments. The speech signal, that is recorded by a microphone is generally infected by noise originating from various sources. Such contamination can change the characteristics of the speech signals and degrade the speech quality and intelligibility, thereby causing significant harm to human-to-machine communication systems.
Noise detection and reduction for speech applications is often formulated as a digital filtering problem, where the clean speech estimation is obtained by passing the noisy speech through a linear filter. With such a formulation, the core issue of noise reduction becomes how to design an optimal filter that can significantly suppress noise without noticeable speech distortion.

Noise reduction is the crucial step in speech signal processing. Each signal is contained with some kind of noise in it which deteriotes the speech signal quality.

Noise reduction techniques depending on the domain of analyses like Time, Frequency or Time-Frequency/Time-Scale.

The Noise reduction methods are classified into four classes of algorithms: Spectral Subtractive, Subspace, Statistical-model based and Wiener-type. Some popular Noise reduction algorithms are, The log minimum mean square error logMMSE (Ephraim & Malah 1985), The traditional Wiener (Scalart & Filho 1996), The spectral subtraction based on reduced-delay convolution (Gustafsson 2001), The exception of the logMMSE-SPU (Cohen & Berdugo 2002), The logMMSE with speech-presence uncertainty (Cohen Berdugo 2002), The multiband spectral-subtractive (Kamath & Loizou 2002), The generalized subspace approach (Hu &Loizou 2003), The perceptually based subspace approach (Jabloun & Champagne 2003), The Wiener filtering based on wavelet-thresholded multitaper spectra (Hu & Loizou 2004), Least-Mean-Square

(LMS), Adaptive noise cancellation (ANC) [3], Normalized(N) LMS, Modified(M)- NLMS, Error nonlinearity (EN)-LMS, Normalized data nonlinearity (NDN)-LMS adaptation etc. Among those many methods, one of the most simple and effective is the spectral subtraction method. It is quite popular method. Spectral Subtraction method, subtracts the estimated noise from the original signal to enhance the speech recognition. The noise is estimated from the original signal itself and subtracted to the original signal, which thus improves the Signal-to-Noise ratio (SNR). It is assumed that the signal is distorted by a wide-band, stationary, additive noise, the noise estimate is the same during the analysis and the restoration and the phase is the same in the original and restored signal.

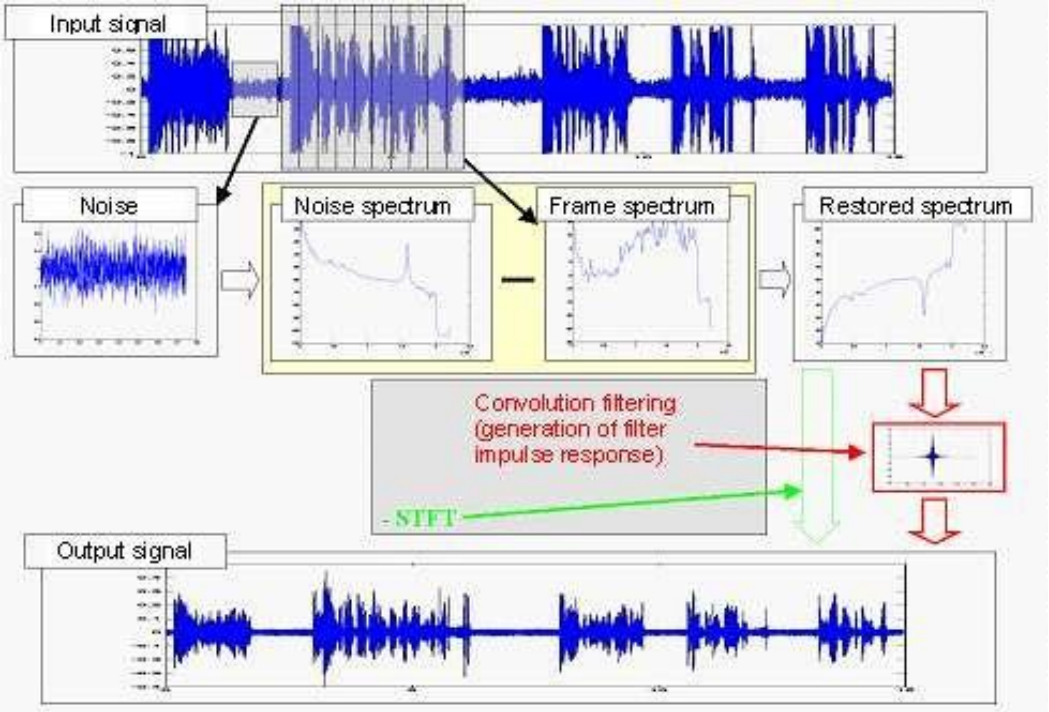

Figure: Noise Removal Process in input voice signal

1.1.1 Feature Extraction Stage

After the preprocessing step, feature extraction is the second component of automatic speech recognition (ASR) systems. It helps to identify the components of audio signals that are good for identifying the linguistic content and discarding all the other stuff such as background noise. The speech signal are slowly timed varying signals (quasi-stationary). When examined over a sufficiently short period of time, the characteristics of signal remain fairly stationary.The information in the speech signal is represented by the short term amplitude of the speech

signal.The extraction of feature vectors is based on these short term amplitude spectrum of speech signals. This component should derive descriptive features from the windowed and enhanced speech signal to enable a classification of sounds. The feature extraction is needed because the raw speech signal contains information besides the linguistic message and has a high dimensionality. Both characteristics of the raw speech signal would be unfeasible for the classification of sounds and result in a high word error rate. Therefore, the feature extraction algorithm derives a characteristic feature vector with a lower dimensionality, which is used for the classification of sounds.

There are several feature extraction techniques such as Linear Predictive Analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) etc. MFCC is the most commonly used feature extraction method in ASR. To extract a feature vector containing all information about the linguistic message, MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. Among several features generated we consider only the relevant feature set for the classification model. These generated feature set is known as feature vector which define mathematical characteristics of a speech signal. Such feature vectors act as a input to the classification models such as HMM(Hidden Markov Mode)l, RNN(Recurrent Neural Network) etc.

2.4.3.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients is a popular feature extraction technique in speech recognition. The Mel Frequency Cepstral Coefficients are the representation of a windowed short term signal derived from the Fast Fourier Transform (FFT) of the signal on a non linear mel scale of frequency, which is based on the human ear scale.

For the computation of MFCC, the speech signal is divided and framed into 20-40ms long frames. The frames are overlapped for smooth transitions. The next step is to perform Discrete Fourier Transform of the frames. FFT is used to speed up the processing. Then the frequencies obtained from the FFT are wrapped onto the mel scale. A mel is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels. The mapping between frequency in Hertz and mel scale is linear below 1000 Hz and logarithmic above 1000 Hz. The mel frequency m can be computed from frequency as

$$mel(f) = 1127ln(1 + \frac{f}{700}) \quad (2.1)$$

The mel-spaced filterbanks are computed. This is a set of 20-40(standard is 26) triangular filters that we apply to the output of DFT from earlier steps. Then log of the each energy in the filterbank is taken. The next step is to calculate Discrete Cosine Transformation (DCT) which ranges coefficients according to the significance.

2.4.3.2 Linear Predictive Coding(LPC)

Linear Predictive Coding (LPC) is a powerful speech analysis technique. The basic idea behind LPC is that a specific speech sample at the current time can be approximated asa linear combination of past speech samples.

Linear Prediction is the technique of computation of a parametric model based on least mean squared error theory. The speech signal is approximated as a linear combination of its precious samples. The obtained LPC coefficients describe the formants. The frequency at which the resonant peaks occur are called the formant frequencies. Thus, in this method, locations of the formants in a speech signal are estimated by computing the linear predictive coefficients over a sliding window and finding the peaks in the spectrum of the resulting LP filer.

2.4.3.3 Perceptual Linear Prediction (PLP)

The Perceptual Linear Prediction model describes the psychophysics of human hearing process more accurately in feature extraction process. PLP, similar to LPC analysis, is based on the short-term spectrum of speech. But, PLP modifies the short term spectrum of the speech by several psychophysically based transformations to match human auditory system.

The PLP coefficients are calculated by first carrying out N-point DFT. A frequency warping to Bark scale is applied. The critical-band power spectrum is computed through discrete convolution of the power spectrum with the piece-wise approximation of the critical-band curve. The smoothed spectrum is down-sampled at intervals around 1 Bark. The three steps of frequency warping, smoothing and sampling are integrate into a single filter-bank called Bark filter bank. An equal loudness pre-emphasis weight the filter-bank outputs. The equalized values are further processed by Linear Prediction (LP). Applying LP to the warped line spectrum computes the predictor coefficients of a signal that has this warped spectrumas a power spectrum.

METHODOLOGY

Methodology overview

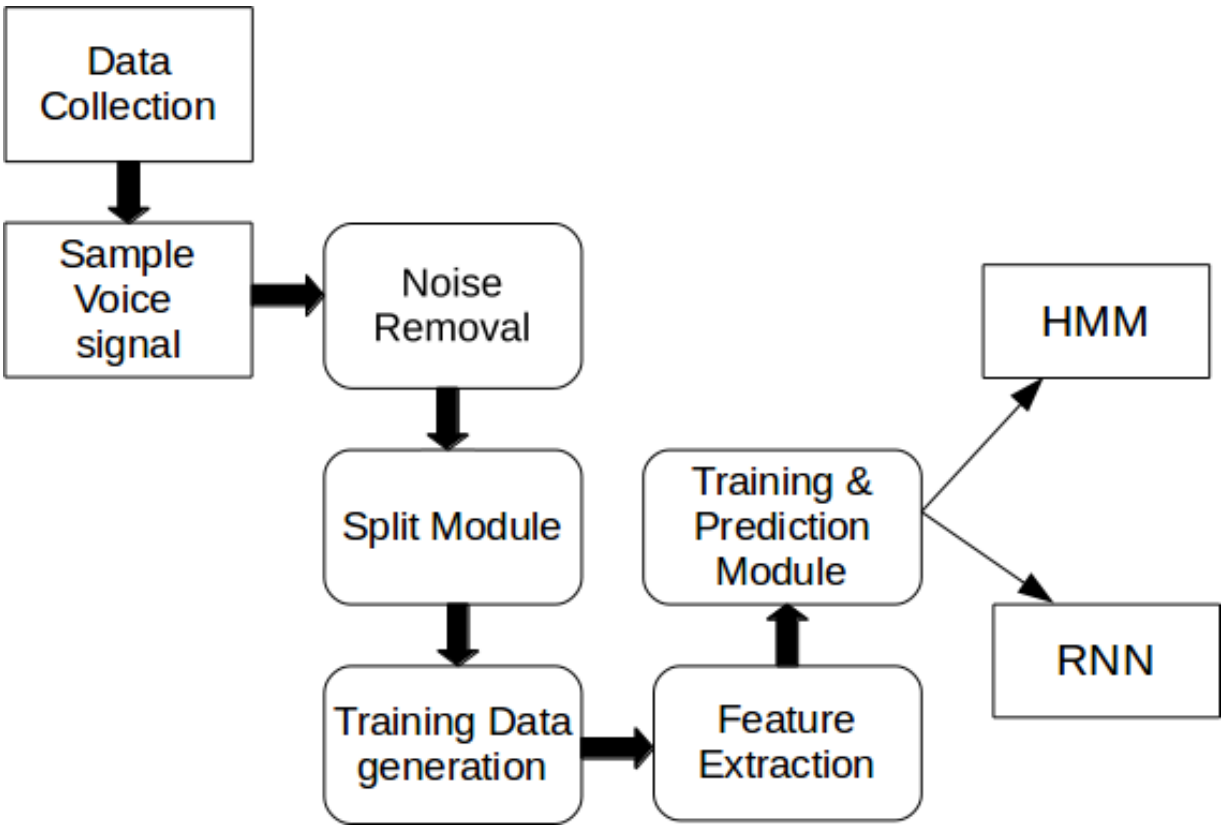

Figure : Outline of the project methodology

Before beginning the project work we went through several researches. Through the researches we learnt about several useful methodologies so that we can approach towards the project development phase. Going through several papers and books we selected suitable approach that fulfilled our needs as per our domain system. Figure 3.1 shows the block diagram that represents the overall outline of the project development methodology. The preliminary stage began with the collection of sample voice signal from different speakers using recording software tool called audacity. The recorded voice signal consist of external noises which were depreciated using the noice removal technique. After that from each recorded signals training data samples were generated using split module which makes the data samples suitable for feature extraction. Using those samples specific feature extraction technique is used to extract unique feature vectors. In our case we used MFCC for feature extraction purpose. The extractd features were used to train the prediction model. In our project we have used two approaches for prediction purpose which are HMM and RNN

approaches respectively. The details regarding the methodology is discussed in the sections below.

Data Collection

For a project data is the most important part that may be of various forms such as text, audio, video etc. In our case we required the audio data as it deals with the recognition of voice from the user to interact with system.Thus large samples of data are required for the trainingof the model and hence enhance the performance ability of the recognition model. The Data collection phase is the initial phase of the project to begin with. This stage has been divided to two phases as discussed below:

Sample Voice Collection

In collecting required data for the project we needed the Nepali spoken words as we were dealing for the interactive voice response system using Nepali speech. The sources for Nepali speech data were not available in the web so that we had to manually collect the data from the individuals The sample voice collection was done from several individual possible.

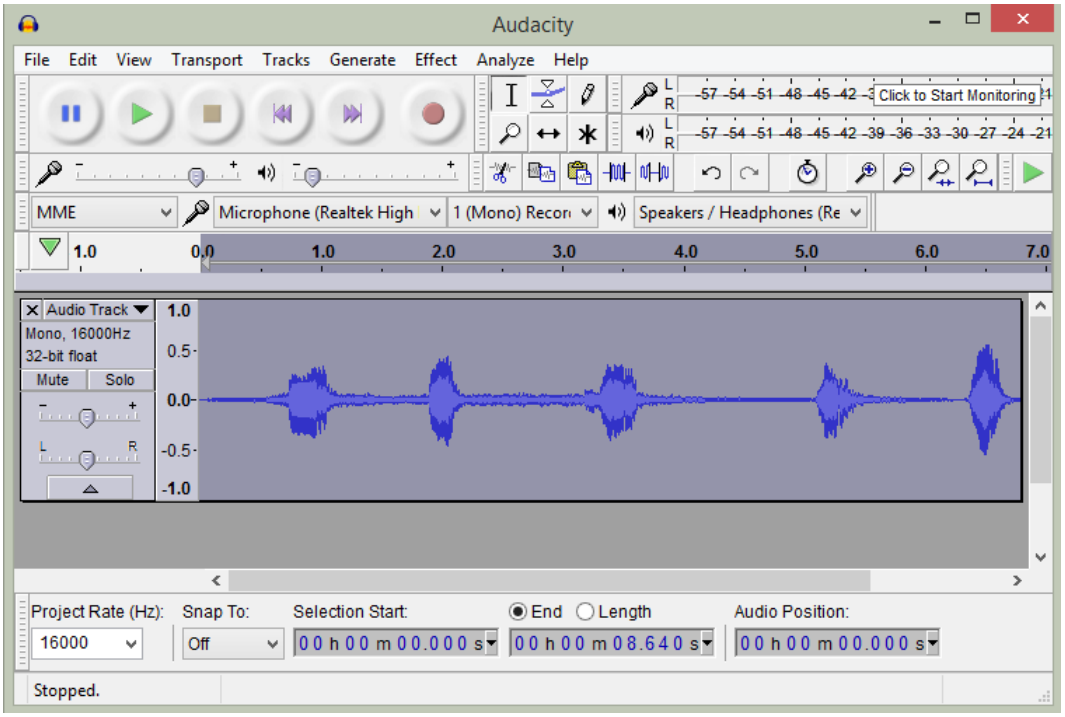

Figure : Audacity Interface for sound sample recording

For recording purpose we used a simple software called "Audacity", which is free and cross platform software Licensed under the GNU General Public License (GPL). It easily runs on several platforms such as Windows, Mac OS X/mac OS and GNU/Linux. Audacity can record live audio through a microphone or mixer, or digitize recordings from other media. With some sound cards, and on any recent version of Windows, Audacity can also capture streaming audio. Thus for collecting sound samples we developed a simple tutorial on recording the speech and collected data from several individuals and also manually recorded the voices.As our Devanagari script is very vast for the preliminary stage we are working with the simple sound samples from numbers 0 to 9. For our project we used the sound samples of frequency rate of 16000Hz recorded using the mono-channel configuration.Similarly we used 16-bit samples for the project. The sample voice collected thus were used for the training data generation.

Training Data Generation

After collecting the sample voices from multiple sources, we need to create training samples to be used in speech recognition. The sample voice contains speech signals numbered from $0 - 9$. Now these samples need to be separated into individual files each containing a digit so that it can be used as training samples.

To generate sample of each digit, we have to detect silence zones from the sample file and separate each based on those silence regions.

Now, to detect the silence zones, it 's highly depended on the audio signal. We need to detect the threshold level, below which the sound can be considered as silent. In order to detect the threshold level, first we fragment the whole audio sample into certain number of frames based on sample width. Then, we calculate the root mean square value which gives an approximate estimate of threshold level, below which the sound frames can be considered silence and above it as voiced activity.

After the detection of threshold level, we need to partition the sound sample. Based on the threshold level for a sound sample, now we detect the silent zones, and thereafter the ranges. We define a minimum silence length in second. And the audio sample is fragmented into number of sample per frame based on silence length. If a frame 's rms value is less than threshold then, the frame is considered as silent frame, and so on silent ranges are determined. These ranges can be complimented then to generate the ranges that actually contains the voice. Then the audio sample is partitioned based on these non-silence activity ranges.

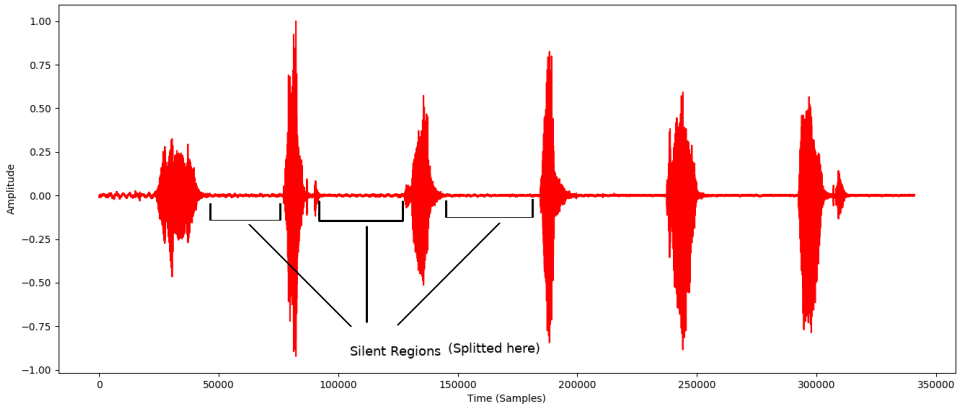

Figure : Training samples generation from recorded voice

Noise Reduction

The speech signal may contain some noise. Actually, most of the speech signals contains background noise. This noise degrades the quality of the speech signal and create hindrance in the processing the speech. Thus, a noise reduction algorithm is needed to reduce the noise level without affecting the speech signal quality. One of the noise reduction algorithm used in this project is Spectral Subtraction method. It is performed independently in the frequency bands corresponding to the auditory critical bands.

The spectral subtraction method is a simple and effective method of noise reduction. In this method, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved. It is assumed that the signal is distorted by a wide-band, stationary, additive noise, the noise estimate is the same during the analysis and the restoration and the phase is the same in the original and restored signal.

The noisy signal y(m) is a sum of the desired signal x(m) and the noise n(m):

$$y(m) = x(m) + n(m) \tag{3.1}$$

In the frequency domain, this may be denoted as:

$$Y(j\omega) = X(j\omega) + N(j\omega) \Rightarrow X(j\omega) = Y(j\omega)N(j\omega) \tag{3.2}$$

where $Y(j\omega)$, $X(j\omega)$, $N(j\omega)$ are Fourier transforms of y(m), x(m), n(m), respectively. To take

the Discrete Fourier Transform of the frame, perform the following:

$$Y_i(k) = \sum_{m=1}^{N} y_i(m)h(m)e^{-j2\pi km/N} \qquad 1 \le k \le K \tag{3.3}$$

where $h(m)$ is an N sample long analysis window (e.g. hamming window), and K is the length of the DFT. The magnitude spectrum for the speech frame $y_i(m)$ is given by:

$$M_i(k) = |Y_i(k)| \tag{3.4}$$

The phase spectrum is calculated in the following way:

$$\theta_i(k) = tan^{-1}\left(\frac{Im(Y_i(k))}{Re(Y_i(k))}\right) \tag{3.5}$$

The statistic parameters of the noise are not known, thus the noise and the speech signal are replaced by their estimates:

$$\hat{X}(j\omega) = Y(j\omega) - \hat{N}(j\omega) \tag{3.6}$$

Noise is the time-average estimate of first few frames of the speech signal.

$$\hat{N}(j\omega) = \frac{1}{K}\sum_{i=0}^{K-1} |N_i(j\omega)| \tag{3.7}$$

The noise reduced clean speech signal is then obtained by performing the inverse Fourier transform of $X(j\omega)$

The procedural Explanation of the noise removal technique of the input speech signal is discussed below:

The basic principle is as follows: if we assume additive noise, then we can subtract the noise spectrum from the noisy speech spectrum, so we are left with what should look like the clean speech spectrum. For this we need to know what the noise spectrum looks like, so we estimate it during regions of no speech (parts of the signal that contain only noise) and then assume it won't change much from frame to frame.

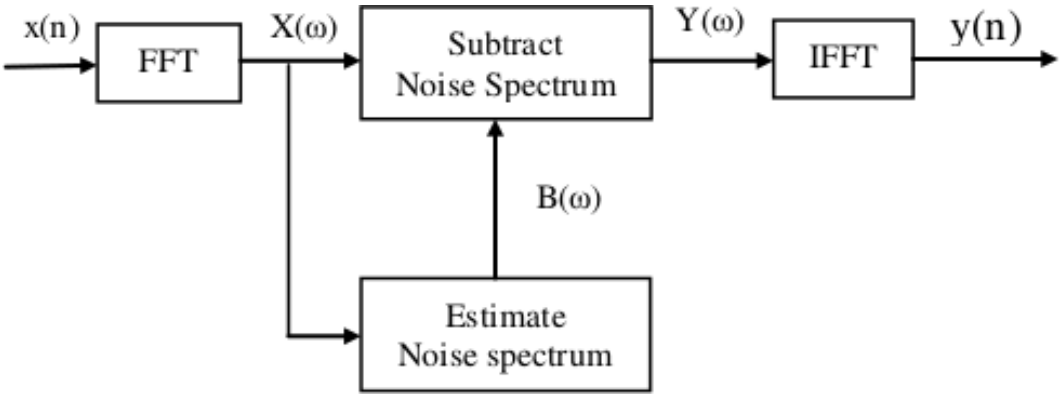

Figure : Noise removal process in speech signal

Let's assume a clean time domain signal, $x(m)$ to which we add additive noise, $n(m)$.The signal we actually see is the sum $y(m) = x(m) + n(m)$. We wish to estimate the true value of $x(m)$ given $y(m)$ only.

The first step in spectral subtraction is to frame the speech signal into short, overlapping frames. Typically frames are taken to be about 20ms long. For a 16KHz sampled audio file, this corresponds to

$$0.020s * 16,000 samples/s = 320 \quad samples$$

We then use an overlap of 50%, or about 200 samples. This means the first frame starts at sample 0, the second starts at sample 200, the third at 400 etc.

Then, we need a window, which contains certain number of frames. We used Hamming window for this project. We then take the discrete Fourier transform of each frame and extract the magnitude and phase spectrum from each. Let $y_i(m)$ be framed time domain signal and $Y_i(k)$ is its complex Discrete Fourier Transform, (DFT) where m ranges over 1-400 (if our frames are 400 samples) and i, ranges over the number of frames. Then $M_i(k)$ is the magnitude spectrum of frame i.

We used Fast Fourier Transform (FFT) in real world practice. The magnitude spectrum for the speech frame $y_i(m)$ is given by:

$$M_i(k) = |(Y_i(k)| \tag{3.8}$$

Pseudocode

```
windowed_frame = frame .* hamming(length(frame));
complex_spec = fft(windowed_frame,512);
mag_spec = abs(complex_spec);
phase_spec = angle(complex_spec);
```

Now, to get the estimate of noise, a common assumption is that the first few frames of an audio signal consist of silence.To get our noise estimate, we can take the mean of the first few or so frames. With the magnitude and noise estimate for each frame, now we proceed with spectral subtraction: subtraction of noise estimate. It can be done as:

$$\hat{M}_i(k) = \begin{cases} M_i(k) - N_i(k) & if \quad M_i(k) \geq N_i(k) \\ 0 & if \quad M_i(k) < N_i(k) \end{cases} \tag{3.9}$$

Here, $\hat{M}_i(k)$ is the estimated clean spectrum. $M_i(k)$ is the noisy spectrum. And $N_i(k)$ is the noise estimate.

Note that we restrict our estimated magnitude spectrum to be positive, since magnitude spectrum must be.

```
clean_spec =
mag_spec - n
oise_est;clean
_spec(clean_s
pec < 0) = 0;
```

Now we have an estimate of the clean spectrum for every frame. With this, we would like to reconstruct the original audio recording which will have less background noise. Forthis we use the clean spectrum $M_i(k)$ and phase spectrum from each frame that we calculatedat the beginning.

Then, the estimated clean complex spectrum for each frame $Y_i(k)$ (enh_spec) isenh_spec = clean_spec .* exp (j * phase_spec)

Now, Inverse FFT (IFFT) of $\hat{Y}_i(k)$ is calculated to reconstruct our original signal $\hat{x}(m)$ and do overlap add of the resulting time-domain frames.

Implementation of Audio Files Working:

```
def sendEmail(do, content):
    server = smtplib.SMTP('smtp.gmail.com', 587)
    server.ehlo
    server.starttls()
    server.login('vaibhav.22scse1010550@galgotiasuniversity.edu.in','gu@12345')
    server.sendmail('vaibhav.22scse1010550@galgotiasuniversity.edu.in', to, content)
    server.close

if __name__ == "__main__":
    #wishme()
    while True :
        query = takeCommand().lower()

        if 'wikipedia' in query:
            speak("Searching Wikipedia...")
            query = query.replace("wikipedia", "")
            results = wikipedia.summary(query, sentences=4)
            speak("According to Wikipedia")
            print(results)
            speak(results)


        elif 'send email to shivam' in query:
            try :
                speak("What should I say?")
                content = takeCommand()
                to = '108shivamvivek@gmail.com'
                sendEmail(to , content)
                speak("Email has been sent")
            except Exception as e :
                print(e)
                speak("Sorry Sir. I can't able to send this email."
```

Feature Extraction Stage

After the preprocessing step, feature extraction is the second component of automatic speech recognition (ASR) systems. It helps to identify the components of audio signals that are good for identifying the linguistic content and discarding all the other stuff such as background noise. The speech signal are slowly timed varying signals (quasi-stationary). When examined over a sufficiently short period of time, the characteristics of signal remain fairly stationary. The information in the speech signal is represented by the short term amplitude of the speech signal.The extraction of feature vectors is based on these short term amplitude spectrum of speech signals. This component should derive descriptive features from the windowed and enhanced speech signal to enable a classification of sounds. The feature extraction is needed because the raw speech signal contains information besides the linguistic message and has a high dimensionality. Both characteristics of the raw speech signal would be unfeasible for the classification of sounds and result in a high word error rate. Therefore, the feature extraction algorithm derives a characteristic feature vector with a lower dimensionality, which is usedfor the classification of sounds.

There are several feature extraction techniques such as Linear Predictive Analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) etc. MFCC is the most commonly used feature extraction method in ASR. To extract a feature vector containing all information about the linguistic message, MFCC mimics the logarithmic perception of loudness and pitch of human auditory system and tries to eliminate speaker dependent characteristics by excluding the fundamental frequency and their harmonics. Among several features generated we consider only the relevant feature set for the classification model. These generated feature set is known as feature vector which define mathematical characteristics of a speech signal. Such feature vectors act as a input to the classification models such as HMM(Hidden Markov Mode)l, RNN(Recurrent Neural Network.

CONCLUSION

Speech Recognition has become very important in today's world. With the advancements in technology and improvements in recognition algorithms, speech has become one of the primary source of input for many applications. Speech is the most efficient and natural way of communication. So, it is intuitive that speech recognition systems have found applications in various fields. Interactive Voice Response (IVR) systems are one of the prominent systems that have a huge potential for use of voice signals as input to the system. With this in mind, we presented an idea for the development of an IVR system with Automatic Speech Recognition (ASR). The initial objective of the project was to develop a system capable of recognizing voice signals in Nepali Language input to the IVR system. Throughout the course of the development phase, various limitations and obstacles were encountered which prompted us to develop the system capable of recognizing words corresponding to the digits of the Nepali Language. For this, we researched on various methods of speech recognition and used the findings of these researches to develop the system. The project was implemented by using algorithms like Noise Reduction, Voice Activity Detection, MFCC Feature Extraction, Hidden Markov Model and Recurrent Neural Network. The overall accuracy of the system while using HMM was around 70 percentage and while using RNN was around 80 percentage. The greater accuracy of RNN is due to the fact that, RNN do not make Markov assumptions and can account for long term dependencies when modeling natural language and due to the the greater representational power of neural networks and their ability to perform intelligent smoothing by taking into account syntactic and semantic features. Though the accuracy seems to be a bit less, the accuracy is good compared to the fact that we had such less data set available. With proper amount of data set available the project can get much higher accuracy and can be implemented.

REFERENCES

<1> Christopher Olah. Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

<2>Andrej Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks. http: //karpathy.github.io/2015/05/21/rnn-effectiveness/.

<3>Md Salam, Dzulkifli Mohamad, and Sheikh Salleh. Malay isolated speech recognition using neural network: A work in finding number of hidden nodes and learning parameters. The International Arab Journal of Information Technology, 8, 2011.

<4>Hidden Markov model. In Wikipedia. https://en.wikipedia.org/wiki/Hidden_Markov_model.

<5>Markov model. In Wikipedia. https://en.wikipedia.org/wiki/Markov_model.

<6>Dr. Jason Brownlee. Machine Learning Mastery Blog Series. http://machinelearningmastery.com/blog/.

<7>Aavaas Gajurel, Anup Pokhrel, and Manish K. Sharma. Nepali Speech Recognition, 2015.

<8>Tan Lee. Automatic Recognition of Isloated Cantonese Syllables using Neural Network. PhD thesis, The Chinese University of Hong Kong, Hong Kong, 1996.

<9>Yaushiro Takahashi, Yukihiro Nomura, Jianming LU, Hiroo Sekiya, and Takashi Yahagi. Isolated Word Recognition Using Very Simple Recurrent Neural Network Plus and Noise Compensation LPC Analysis. Technical report, Graduate School of Science and Technology, Chiba University.

<10>M. M. EL Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf, and M. A. Al-Alaoui. Arabic Speech Recognition Using Recurrent Neural Network. Technical report, Electrical and Computer Engineering Department, Faculty of Engineering and Architecture, American University of Beir