Speech Learning Through Game Object

Project thesis submitted
in partial fulfillment of the requirement for the degree of

Bachelor of Technology

By

Abhijeet Kumar Vatsha
Roll No. 16010122

Under the Supervision of
Dr. Himangshu Sarma

Department of Computer Science and Engineering
Indian Institute of Information Technology Manipur
November, 2019

Abstract

This project is an effort towards designing a platform that helps people learning and speaking of word and further language. In this platform there will be game consisting of three levels. In each level there will be object and against each object there will be audio stored in database which consist of real pronunciation of that object. The first level consists of object against alphabet [A-Z] with a simple word pronunciation stored in database. The user has to utter word if the spoken word and stored word audio matches user got point and move to the next object. Once the score is in multiple of 26 the level will change.The project can be divided into two parts. The first part contains the real time audio signal comparison i.e comparing the signal similarity of audio spoken by user to the audio stored in database.The second part consists of integration audio signal similarity system in game platform and developing a complete system.The process of executing the first part of project involves survey for speech signal property, prepossessing the data for anomalies, and extracting feature using Mel-frequency Cepstral coefficients (MFCC). The features are then used to find the similarity between the audio signal using cosine similarity. Many other experiment are also performed in order to compare audio signal. Comparing the target audio signal with IPA(international phonetics alphabet) sound stored in database by finding amplitude and applying cosine similarity on the amplitude of both signal. The diversity in speech signal make the speech signal comparison little complex. The accuracy achieved in the experiment was not very high.

Declaration

I declare that this submission represents my idea in my own words and where others' idea or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/sources in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from proper permission has not been taken when needed.

(Abhijeet Kumar Vatsha)

Date: (16010122)

Department of Computer Science & Engineering
Indian Institute of Information Technology Manipur

Dr. Himangshu Sarma
Assistant Professor

Email: himangshu@iiitmanipur.ac.in

CERTIFICATE

This is to certify that the thesis entitled **Speech Learning Through Game Object** submitted by Abhijeet Kumar vatsha (16010122), a final year B.Tech. student in the Department of Computer Science and Engineering, Indian Institute of Information Technology Senapati, Manipur is a record of an original research work carried out by him under my supervision and guidance. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Signature of Supervisor

Dr. Himangshu Sarma
Assistant Professor
Dept. of Computer Science and Engg.
(IIIT Senapati, Manipur)

Department of Computer Science & Engineering
Indian Institute of Information Technology Manipur

CERTIFICATE

This is to certify that the project thesis entitled **"Speech Learning Through Game Object"** submitted by Abhijeet Kumar Vatsha (16010122), a final year B.Tech. student in the Department of Computer Science and Engineering, Indian Institute of Information Technology Senapati,Manipur is approved for the degree of Bachelor of Technology in Department of Computer Science and Engineering.

Dr. N. Kishorjit Singh
Head of Department
Dept. of CSE
IIIT Senapati, Manipur

Signature of Examiner 1: _____

Signature of Examiner 2: _____

Signature of Examiner 3: _____

Signature of Examiner 4: _____

Acknowledgement

On the submission of my project on Indian Sign Language Recognition, I would like to express my indebted gratitude and thanks to my supervisor Dr. Himangshu Sarma, Assistant Professor Department of Computer Science and Engineering who in spite of being extraordinarily busy, spare time for guidance and keep us on the correct path. I truthfully appreciate and value his admired supervision and support from the start to the end of this project Dr. Himangshu Sarma have been great sources of motivation to us and I thank him from the core of my heart. Last but not the least I would like to thank each and every person who is involved directly or indirectly to make this project successful.

- Abhijeet Kumar Vatsha

Contents

Abstract ii

Declaration iii

Certificate iv

Certificate v

Acknowledgement vi

Table of contents vii

List of figures 0

1 Introduction 2

1.1 Speech Processing 3

1.2 History of Speech Recognition 3

1.3 ASR System 3

1.4 Motivation 4

1.5 Challenges 5

1.6 Gantt chart 6

2 Literature Survey 7

2.1 Audio Similarity Comparison Using Audio Fingerprint Algorithm 8
2.2 Audio Similarity Comparison to find interference. 9
2.3 Spectral Envelope Matching Of Audio Signal. 10
2.4 Signal Analysis based on similarity function 11
2.5 Isolated Word Automatic Speech Recognition. 11
2.6 Speech Recognition based on Zero-Crossing Features 12
2.7 Similarity-based Voice Activity Detection Algorithm 13
2.8 Audio quality assessment using structural similarity 13
2.9 Recognition Of Speaker Using GMM 14
2.10 Speech Recognition using Support Vector Machines 14
2.11 Summary 15

3 System Analysis and Design **16**

3.0.1 Life Cycle Model 17
3.1 Model Followed in Project 18
3.2 System Overview 19
3.3 Detailed Description Of System Design 20
3.4 Speech Processing 21
3.4.1 Speech Preprocessing. 22
3.4.2 Framing 22
3.4.3 Windowing 23
3.4.4 Discrete Fourier Transform (DFT) 23
3.5 Feature Extraction Techniques 23
3.5.1 Cepstral Transform Coefficients 23
3.5.2 Homomorphic Speech Processing 24
3.5.3 Linear Predictive Coding(LPC) 24

3.5.4 Mel Frequency Cepstral Coefficient . . . 25
3.6 First Approach . . . 25
3.6.1 Reading Audio File. . . . 25
3.6.2 Removing unvoiced region . . . 25
3.6.3 Normalization . . . 26
3.6.4 Framing . . . 27
3.6.5 Windowing . . . 28
3.6.6 Discrete Fourier transform (DFT) . . . 29
3.6.7 Calculating Power Spectral . . . 29
3.6.8 Compute MEL-spaced filterbank . . . 29
3.6.9 Discrete Cosine Transform (DCT) . . . 29
3.6.10 MFCC Coefficients . . . 30
3.6.11 Cosine Similarity . . . 30
3.7 Audio Similarity Comparison Using IPA . . . 31
3.7.1 IPA . . . 31
3.7.2 Approach . . . 31

4 Result Evaluation and Observation **32**
4.1 Result . . . 32
4.2 Observations . . . 34

5 Conclusion **35**
5.1 Future Scope . . . 36

List of Figures

1.1 Generic Automatic Speech Recognition (ASR) system 4

1.2 Gantt chart 6

2.1 Audio Fingerprint Framework. 8

2.2 Real Time similarity comparison [2] 10

3.1 System Design 18

3.2 System Design 19

3.3 Detailed Design 20

3.4 Basic Step Of Speech Processing 21

3.5 Preprocessing Of Speech 22

3.6 Cepstral Transform 24

3.7 Homomorphic Transform 24

3.8 Original Speech Signal 26

3.9 Signal after silence removal 27

3.10 Windowing 28

3.11 Mel-spaced filterbank 30

4.1 Result analysis using first approach. 33

Chapter 1

Introduction

Speech learning through game object is a system that provide platform for people that help in learning and speaking of word and further language. The concept is not new as these kind of platform are already developed by many of the organization. Bolo a platform for learning speech and improving reading skill has been already developed by google. The application mainly focusing on enhancing both English and Hindi reading and speaking skill of children . Apart from this there are many other application available in the market, but they are based on the concept of ASR(Automatic speech recognition). Our work is based on similar idea but the concept of implementation is totally different.The system is basically based on Integration of two thing, i.e, real time speech signal similarity system and developing a game environment. This work is completely about developing a system that can compare two speech signal in real time environment without Using a concept of ASR(Automatic speech recognition) . The work mainly focusing on finding a specific feature of a signal among different features. The extracted feature of signal then can be used to find the similarity between two audio signal. We try to build a speech recognition system that is purely based on unsupervised approach . Although there are many experiment performed till now in order to obtain the desired result. We focused mainly on two approach that is finding MFCC (Mel-frequency cepstral coefficients) feature and applying cosine similarity, cross-correlation to find similarity between signal. We also use IPA(International phonetic alphabet) in order to find similarity between signal.

1.1 Speech Processing

Processing of speech basically indicates towards speech signal study and it's methods of processing. The signal are converted into digital format as the processing of signal is done in this format only. Hence it is a part of digital signal processing (DSP). Speech processing basically includes the acquisition of signal, manipulation of speech signal, storage and transferring the output of speech signals. When a speech signal is applied as an input to a system it is called speech recognition whereas a speech producing system is termed as speech synthesis.

1.2 History of Speech Recognition

Speech recognition started in year 1960, when a company called IBM developed a framework that could understand digits and arithmetic commands like total and plus operation. Within a decade researchers in Japan and England build systems that could recognize vowels and consonants. More fruitful result came in the year 1976 when funded project laid to the formulation of an ASR system which could recognize just over 1000 words. Further improvement in speech recognition systems came due to Hidden Markov Models (HMMs), coupled with the advancement in computer technology in the mid 1980s. The favourable outcome of HMMs give proof of the work of scientist named Frederick Jelinek at Research Center, advocated to use model based on statics to interpret speech, rather practicing computer to copy the way humans learn language i.e., through sense, style, and rules of language (a general way at the time). In the year 1990, with increase in computer's computational power, the procedure shifted from statistical to syntax and semantics using Machine Learning and Neural Network. From then till now significant work has been recognized in this field from super giants like IBM, Google, Microsoft and Apple etc [12].

1.3 ASR System

Recognition of speech is a type of sample matching. Figure 1.1 describe the processing phase intricate in recognition of speech. There are two stage in supervised identification of pattern. This includes training phase and testing phase.The procedure of extraction of significant features that are relevant for classification is common to every stage. At

the time of training phase, the specification of the model is being calculated with the use of huge number training data. During the phase of testing and identification, the test pattern features matches to the model trained for every classes and then the test pattern only belong to the class to which model matches best with the pattern observed in test class [11]. The primary objective of speech recognition system is to find the

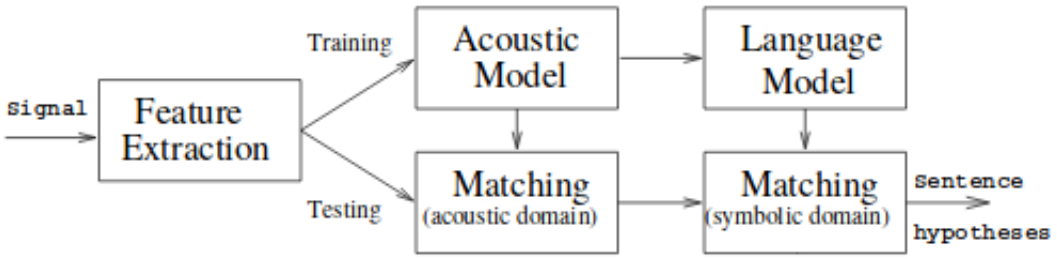

Figure 1.1: Generic Automatic Speech Recognition (ASR) system

best grouping of words subject to etymological imperatives. The sentence is comprised of words, syllables, phonemes. The acoustic model give acoustic occasion of such units which is joined with the principles to develop legitimate and significant sentences in the language. In this manner, if there should be an occurrence of speech recognition,the design coordinating stage can be seen in two spaces: representative and acoustic. In the acoustic space, an element vector compare to a little fragment of test discourse is coordinated with the acoustic model of each and each class. The section is a lot of well coordinating marks of class alongside their coordinating scores. This procedure for task of mark is rehashed for each element vector in the component vector succession processed from the test information. The resultant cross section is handled related to the language model to yield the perceived sentence.

1.4 Motivation

Language is the very basic and key requirement for staying alive in today's world. It is very difficult for kid to learn word and language through various book as they feel lack

of interest in the book and developing interest is again a big task. So to overcome this challenge game is beautiful way of learning. Games offers very fun-filled, relaxing and learning atmosphere. Game-based learning exercises that encourage students listening and speaking abilities were designed in this project. With the reach of the mobile phones rising across world, the application is aimed at helping kids who are unable to go to schools or have no access to it. The work mainly focus on developing a new learning approach or environment to help the people to improve their learning and reading skill. Effective extension of this project will develop a new methodology in the society which help people in learning and speaking of language.

1.5 Challenges

The speech signal is very complex and depends on many factors. for example If comparing female to male. There is basically a remarkable difference in frequency based on gender. The frequency range of an individual, some individual have higher frequency range then other depending on several factors in their respective sound production apparatus. Apart from this there are several factor like, Phoneme to word mapping which depends on regional differences, background noise, speed of utterance. The speech signal comparison sound simple, but unfortunately, it's not. There are many factor which make this process complicated. We can take some example: consider we have a recording of a person voice recorded in a sound proof room saying "open the door". Now if we record same utterance of word of same person but in noisy environment the recording are no longer same. If we now change the room and record it in a reverberate room, the two signal are no longer same. If we record same sentence in same room but with different speech rate as we uttered the referenced one, the signal are no longer same. Age, gender, health condition are other confounding factor that influence the signal.

1.6 Gantt chart

A Gantt outline is a kind of bar graph that represents an undertaking schedule.This diagram records the errands to be performed on the vertical pivot, and time interim's on the level axis.It demonstrates the diverse period of work in the task that to be done in the give time frame. Description of Figure 1.2 In first three weeks Requirements Analysis is done,that includes data collection,data preprocessing and Project problem Analysis.from fourth to sixth week System design is done for the project.from seventh to eleventh weeks project is implemented and tested.In last two weeks from twelve to thirteen weeks project documentation is done.

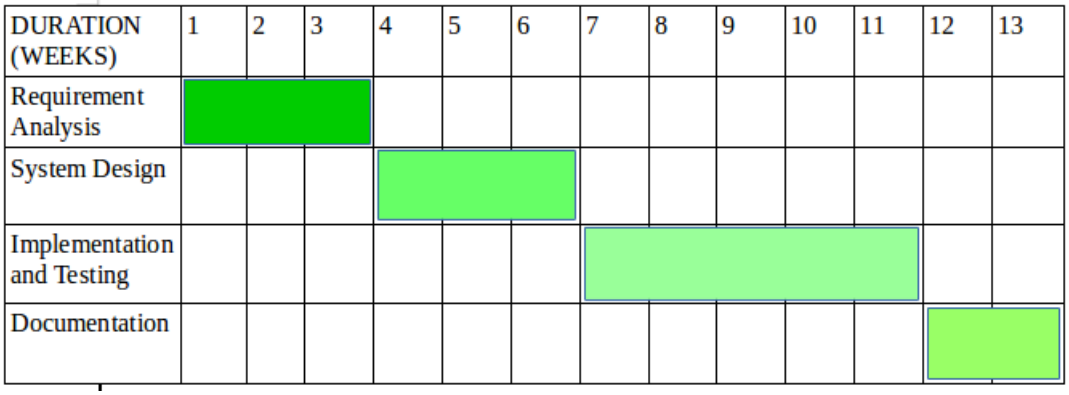

Figure 1.2: Gantt chart

Chapter 2

Literature Survey

Outline: This chapter presents the following:

1. Audio Similarity Comparison using Audio Fingerprint algorithm.
2. Real-Time Audio Similarity Comparison.
3. Spectral Envelope Matching Of Audio Signal.
4. Signal Analysis based on similarity function.
5. Isolated Word Automatic Speech Recognition.
6. Speech Recognition based on Zero-Crossing Features.
7. Similarity-based Voice Activity Detection Algorithm.
8. Audio quality assessment using structural similarity.
9. Recognition Of Speaker Using GMM.
10. Speech Recognition using Support Vector Machines.

2.1 Audio Similarity Comparison Using Audio Fingerprint Algorithm

An audio signal can be summarized on the basis of it's fingerprint or special feature. The fingerprint is based on the content and compact signature of audio signal that summarizes signal. The thought of fingerprinting have recently withdraw attention since they permit the observance of audio independently of its format and while not the necessity of meta-data or watermark embedding.Comparison means finding the relative degree of similarity-based out of some characteristics between two things. The things which are compared need to be on the same ground, following the same basic rules and Audio Comparison is no different. In this work, the fingerprints are generated from each audio files and compare them based out of them. Practically every audio fingerprints are based on spectrogram feature . A spectrogram is an approximate decomposition of the signal over time and frequency. It is created by taking a short window of time of the signal, and then performing a Fourier transform that decomposes that window over its frequencies.y repeatedly performing this calculation for subsequent windows of time, we find the frequency composition of the audio as time progresses. [1]

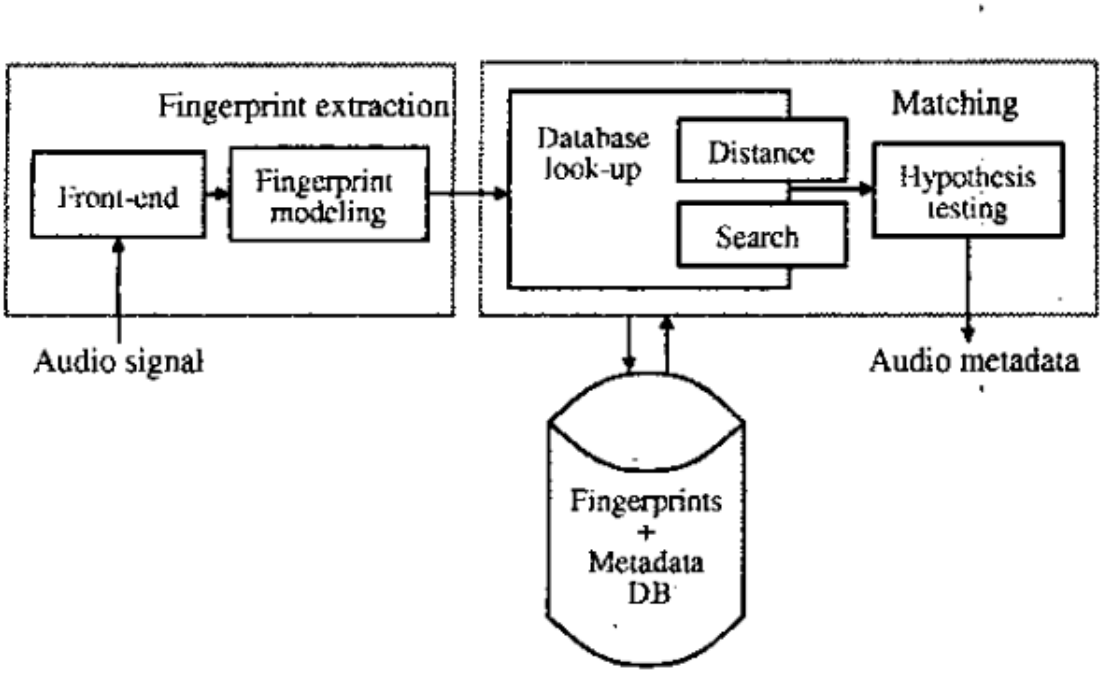

Figure 2.1: Audio Fingerprint Framework.

Given a fingerprint derived from a recording, the matching algorithmic rule searches a information of fingerprints to find the most effective match. The simplest way of comparing fingerprints, that's a distance, is thus required. Since there are high variety of comparison and computing distance are often costly too, thus we tend to need ways that speed up the search. It's common to seek out ways that use a less complicated distance to quickly discard candidates and also the additional correct however expensive distance for the reduced set of candidates. The basic steps followed by author in this approach are as followings:

1. Preprocessing

2. Framing and overlap

3. Linear Transforms: Spectral Estimates

4. Feature Extraction

5. Fingerprint Model

6. Comparing Similarity Using different algorithms.

2.2 Audio Similarity Comparison to find interference.

The projected system was develop to seek out the out the sources of the interference within the Air traffic management (ATC) communication band. This work presents a similarity comparison technique which will create automatic interference distinguishing potential. The audios that is taken as input are divided into three bands. The results of subtractions between 2 audios of constant band are less if the 2 audios are similar. To induce a outcome, minimum of two similar results from three bands are thought of positive. First, we have a tendency to extract envelops from each audios that are accustomed compare. straightforward subtraction of 2 envelops can cut back a subtraction result if 2 envelops area unit similar. Otherwise, a subtraction result are raised. Since this algorithmic program processes in time domain, thus it's straightforward to calculate and uses short computation time[2].Two audio signal present at input referred to as single audio S(t) and mixed audio M(t). Normalize signal will be filtered into 3 frequency bands, low band (below 340 Hz), mid band (340 to 3,400 Hz) and high band (upper 3,400 Hz).

The common value is calculated from the absolute normalized information of every band. This whole method is thought as envelope extraction. Each band of single audio is now subtracted with each band of mixed audio and subtraction result is obtained.

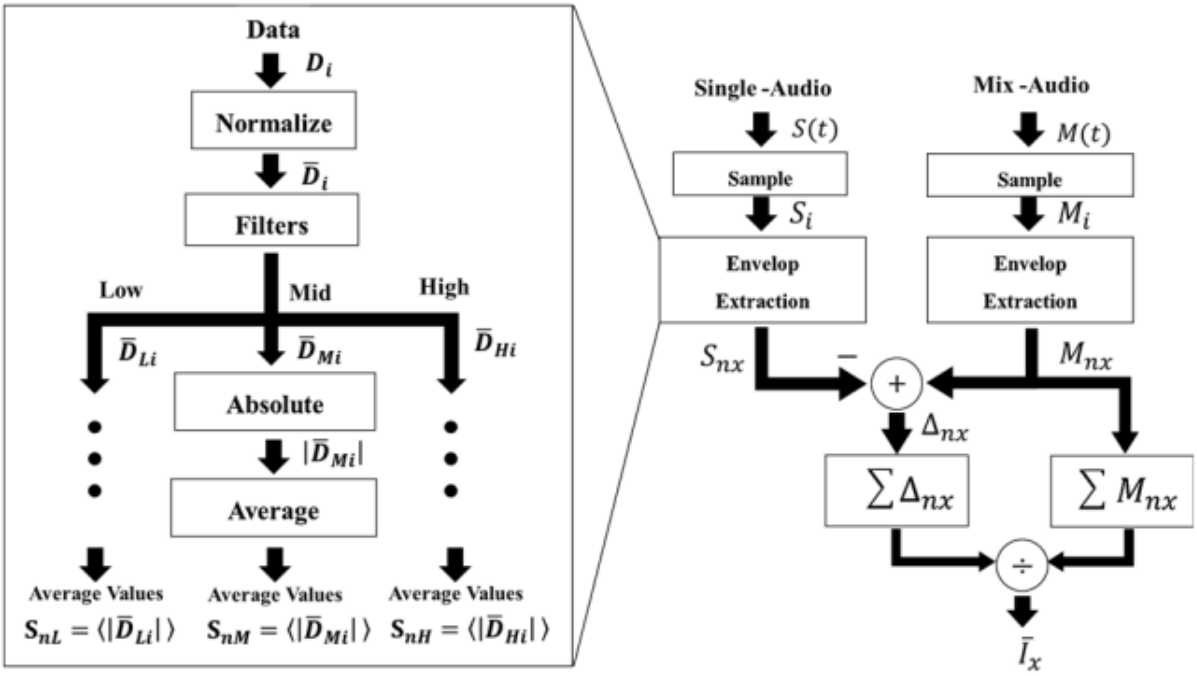

Figure 2.2: Real Time similarity comparison [2]

2.3 Spectral Envelope Matching Of Audio Signal.

This work is about defining some new or unique matrix that will find out similarity between two audio signal containing music inside it. These matrix acquire low computational cost and low sensitivity to acoustic loss as these matrix are fulfill spectral flatness criteria. The accuracy of this thesis is obtained by checking whether we can obtain two musical signal from same instrument. When the referenced signal and target signal are not of same acoustical property then in that case proposed matrix do compare using standard spectral property. After this mfcc feature of referenced as well as target audio signal is obtained and the obtained features are then passed on to radial basis function . The function then find the similarity between two audio signal based on the feature

given to it. [3]

2.4 Signal Analysis based on similarity function

This work is all about comparing two audio signal which contain music inside it. The comparison is totally based on the content of music that the audio signal is carrying of. First the feature is extracted from the audio signal using some k-mean clustering approach from the spectrum of song. This obtained signature or feature then can be used to find the similarity between the signal using some similarity finding algorithm. In this thesis they used Earth Mover's Distance, which compare by making histogram. In this work, the author has encountered followings steps:

1. First divide the signal into small frames.

2. Obtain spectral representation for each frames. Although many representation are possible but we try to select the spectrum in which signal which are similar are close to each other.

3. Now from a song or audio signal we try to cluster those frames which are similar. The similar frame will be clustered in one group and rest in another cluster. The number of cluster for each audio signal are fixed.

4. As the feature is obtained for each signal we try to find out similarity between these audio signal on the basis of there features. The algorithm used used in this thesis is EMD(Earth Mover's Distance)[4].

2.5 Isolated Word Automatic Speech Recognition.

This work present the idea about a supervised and trained system called ASR which will convert the in the information contained by signal in to text and words, so that we can use the general approach for further processing. They are first trying to obtain the best feature from the signal, here there are using mfcc feature. Then they are using some real time similarity finding algorithm which is dynamic time Wrapping (DTW).

This algorithm give the best possible alignment in real time. so this is use for feature matching. After performing all these step they are using a classifier know as k Nearest neighbour(KNN)[5]. In this article, the author gone through the following steps:

1. PreProcessing of signal.

2. Framing and Windowing.

3. Fast Fourier Transform (FFT).

4. Feature Extraction.

5. Applying DTW(Dynamic Time Wrapping).

6. Applying classification algorithm.

2.6 Speech Recognition based on Zero-Crossing Features

The thesis focused on developing a system that will recognize the voice of a person which it belong. This system has wider range of application in our day to day life such as phone lock, then door lock, google assistant and so many as such. They are first converting the audio signal information into text , then this text is used for further processing. The project is divided in three phase training , testing and recognition phase. First the feature is extracted from the audio signal or the voice of speaker then the feature extracted is stored in the database. When targeted audio comes the feature is extracted from that audio using the same technique used earlier for feature extraction. The extracted feature from target audio is the compared with feature already stored in the database of system. In this thesis, we are not considering mfcc feature . They are extracting Zero-crossing features from each voiced sample. Zero-crossing rate defines the number of time signal changes it's sign inside a frame within a certain time duration. The unvoiced part of signal has more zero crossing rate as compared to voice. So we can also remove voiced and unvoiced part using zero crossing rate. This also help in end point detection [6].

2.7 Similarity-based Voice Activity Detection Algorithm

This work is concerned about the accuracy of speech recognition system. The accuracy is very important aspect which affect the performance of such kind of system. In order of this a new algorithm has been proposed in this thesis to develop a well functioning voice activity detection algorithm. The goal of thesis is to finding the endpoint very precisely in the environment containing noise. They first find the mfcc feature of audio signal by first diving it into frame and apply all the required preprocessing. Euclidean distance is applied on the feature of test frame and noise. Apart from this correlations coefficients of test frame and background is obtained. Thus the result obtained through this approach is better than the traditional one. Correlation coefficient is better than that of euclidean distance[7].

2.8 Audio quality assessment using structural similarity

This work is mainly concerned about the quality assessment of the audio signal. For this the structural similarity concept is used. This concept was originally developed to asses the quality of image. The author focus mainly toward two different implementation of structural similarity. These are as followings:

1. The first research talk about applying the structural similarity concept to audio sequence having short and fixed time domain.

2. The second study reflect about the audio signal which break into non redundant time frequency using structural similarity.

3. The accuracies of both the study are compared in the end.

2.9 Recognition Of Speaker Using GMM

This work is concerned about developing a system which can take input as audio signal and classify the signal to which it belong according to some basic features. So this is basically about recognizing the voice of people and classify it accordingly. The first step in this is feature extraction of voice or audio signal . The feature basically extracted is mfcc from the audio signal . There are more feature like zero crossing rate , timber quality , pitch and many more, but for this thesis we are considering only the mfcc feature. Once the feature is extracted from signal it is stored in database. The obtained feature is then stored in database. When the target audio signal comes the same approach is applied to find the feature. Once the feature is obtained we need to find the similarity between the feature stored in database and the feature obtained . The maximum similarity guarantee maximum matching of voice. Here Gaussian mixture model is applied on feature of each audio signal to create a unique identity for each voice[9].

2.10 Speech Recognition using Support Vector Machines

In this work they are focusing on developing a model of speech recognition that is based on mfcc and lpc feature. This thesis is based on special type of dataset. First the feature is extracted from the given dataset. SVM algorithm is applied to all the extracted feature and result is encountered. The result obtained is matched with various result. It shows that result obtain through this approcah along with radial base function is better than the previous approaches[10].

2.11 Summary

From the state of art, it is concluded that all the reported work are not comparing signal directly. The idea of comparing signal without any supervised or machine learning approach is completely new. This work is completely based on unsupervised approach of comparing signal in real time. From the state of art it is found that most of the work has been done in speaker identification or music identification. Not a single work is reported there which will compare the real time audio signal based on the content it is encoding in itself, i.e, word or sentences. This is completely a new concept of comparing signal which in future reduces the computational cost as well reduces the time complexity. This will also develop a platform which help people in learning words and it's correct pronunciation . Further development of the project will produce a game application

Chapter 3

System Analysis and Design

Outline: This chapter presents the following:

1. System Overview.
2. Detailed Description Of System Design.
3. Development Life-Cycle.
4. Workflow Diagram.
5. Algorithms.

System Design give the basic overview of a system .i.e how system will look like and what is the basic structure, segment or the module of system . It also gives the overview about the interface of our design.
System Analysis describe about the proper requirement of the system . It describe the workflow or the data flow inside the system,. It also conclude about the software and hardware requirement of the system i,e the algorithm used to develop a system. It also discuss the feasibility of the project.

3.0.1 Life Cycle Model

Waterfall model is design approach for developing any software or product.It uses sequential approach.it has many phase and each phase has its import ants.generally the phase of waterfall model is less iterative in nature.waterfall model is mainly used by big companies for large project.since waterfall model works in linear sequential flow it is also called linear-sequential cycles model.

Sequential Phases in Waterfall Model:

1. **Requirements :** First phase of waterfall model is requirements phase.Main aim of this phase is to know about the system in details like what will be the input and output of software or product,and also know what are things required for developing this.

2. **System Design :**After the requirements is gathered and specified,next phase is design phase.based on the problem defined in previous phase system architecture is prepared that satisfy the problem of the software or product.This system design now can be coded.

3. **Implementation:**Based on the system design,system is developed taking small unit at a time.then each unit is separately tested this is also called unit testing.same way each part of system architecture is coded and tested.

4. **Integration and Testing:**In this phase each small unit is merged and then tested,this is also called integration's testing.Integrating testing is done,whole system is tested for any test case this is called system testing.

3.1 Model Followed in Project

In a viable programming advancement venture, the traditional cascade model is difficult to utilize. So,Iterative cascade model can be thought of as consolidating the important changes to the traditional cascade model to make it usable in viable programming improvement ventures. It is practically same as the established cascade model aside from certain progressions are made to build the effectiveness of the product advancement. The iterative cascade model gives criticism ways from each stage to its previous stages, which is the primary distinction from the established cascade model.Input ways presented by the iterative cascade model are appeared in the figure below. At the point when blunders are recognized at some later stage, these criticism ways permit revising mistakes submitted by software engineers amid some stage. Feedback path is very useful it helps in knowing and correcting the error.one problem with this is that there is no feedback path for feasibility study[13].

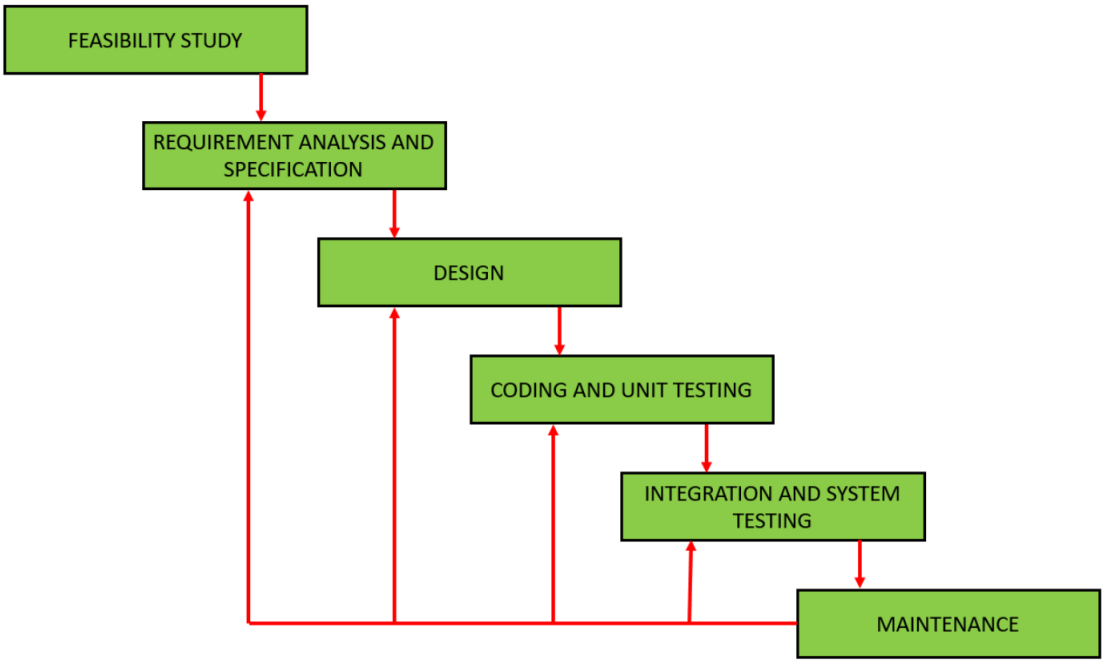

Figure 3.1: System Design

3.2 System Overview

System Design describe the complete overview of the system . The project consist of a game which consist of three level. In this project there is integration of two module. The first module consist of comparison of audio signal similarity and second module consist of game platform . This thesis is now mainly focusing on first module i.e Finding signal similarity. The input to this module is a audio signal containing some meaningful information. This signal is compared with the target signal containing same information . Module will generate some similarity score. If the score generated is above than threshold signal will be similar in this case. Else signal will not similar if the score generated is less then similarity score.

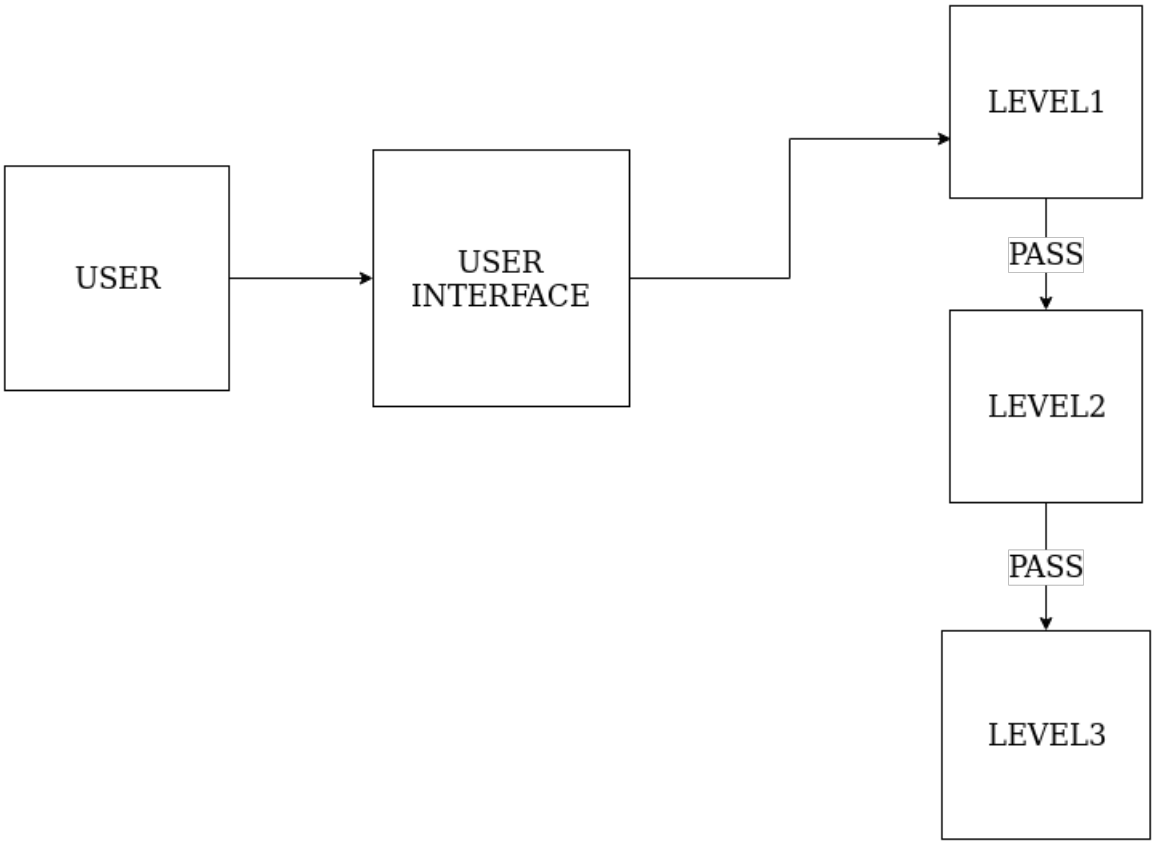

Figure 3.2: System Design

3.3 Detailed Description Of System Design

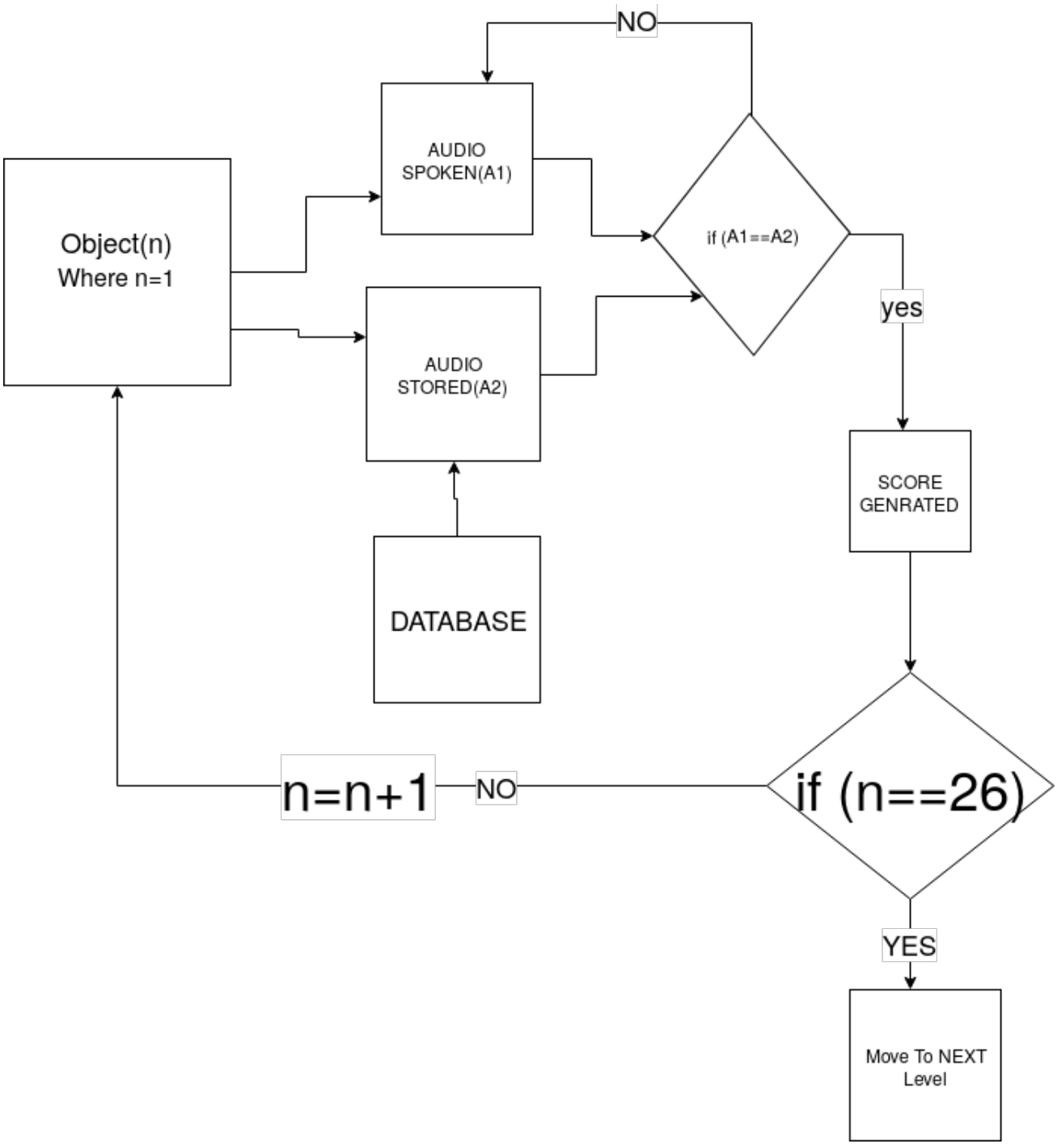

Figure 3.3: Detailed Design

The thesis is mainly focusing on developing a game that will help people in learning words and it's pronunciation . There will be three level in game. The above figure mainly describe about the level. In each level there will be object and against each object there will be audio stored in database which consist of real pronunciation of that object. The first level consists of object against alphabet [A-Z] with a simple word pronunciation stored in database . The user has to speak each word if the spoken word and stored word audio matches user got point and move to the next object . Once the score is in multiple of 26 the level will change. As the level increases the toughness of word also increases. The next level contain more complex word and some points for each correct pronunciation of word. In this way complexity of words increases as the level increases.

3.4 Speech Processing

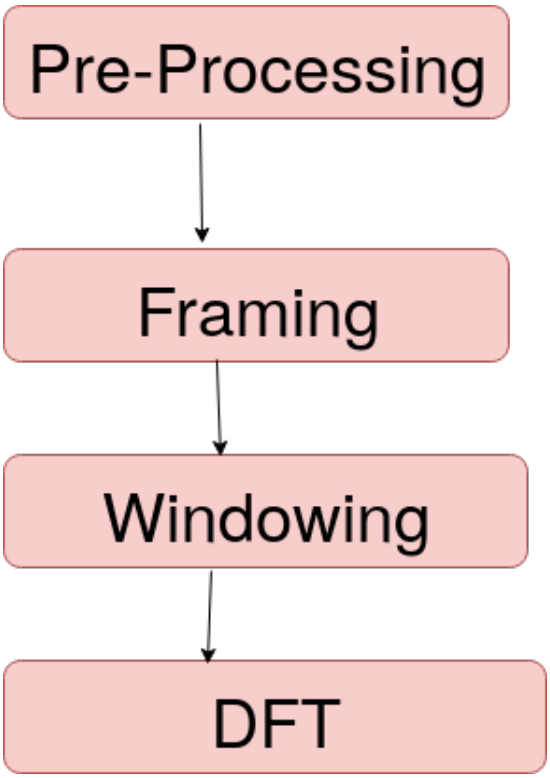

Figure 3.4: Basic Step Of Speech Processing

3.4.1 Speech Preprocessing.

Audio signal contain several type of artifact which generates because of several reason. In this section we do some preprocessing on the signal to obtained a qualitiful signal, so that signal can be used for further feature extraction and other processing. preprocessing mainly contain removal of noise and other unusual substance from the signal. It also include the separation of voiced and unvoiced sound. Removing the unvoiced part is important for further processing of signal.

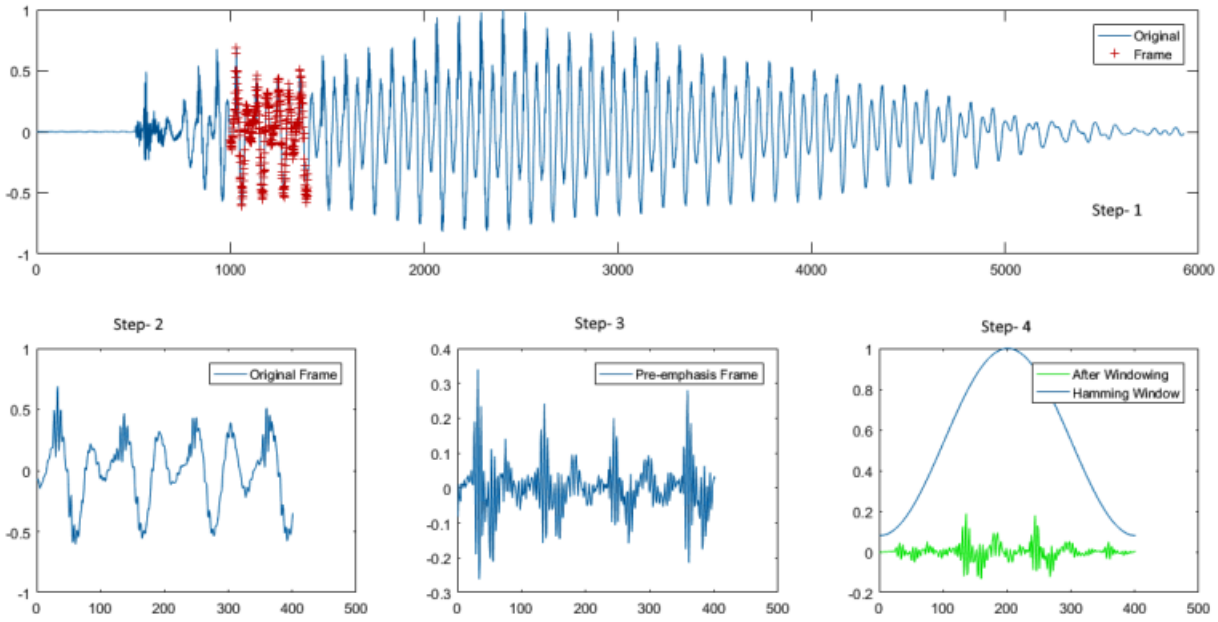

Figure 3.5: Preprocessing Of Speech

3.4.2 Framing

Signal is variant in it's complete domain. It changes continuously in time. So the signal is not suitable for further processing in it's original format. It is assumed that signal is constant in short period of time. So in that frame of time we can apply all the processing. So framing can be defined as breaking the signal in to shorter frame. Basically we divide the frame in a duration of 20 to 30 ms. Condidering the frame in much shorter duration may miss the important information. When we consider the signal for larger duration then there will be chance of getting signal which is changing continuously. So framing help in obtaining a stationary part from complete signal.

3.4.3 Windowing

Windowing is basically used to remove the discontinuity at the edge of frame. when we frame the signal it produces discontinuity at edge of signal. So to remove discontinuity from the signal windowing is applied. From the literature survey it is found that hamming window is popular window used in speech processing.

3.4.4 Discrete Fourier Transform (DFT)

The Discrete Fourier Transform (DFT) is very important tools in Signal as well as Image Processing. The DFT first calculate the frequency spectrum of signal. It is mathematical approach to transform the signal from time domain to frequency domain. In frequency domain we can do all the spectral analysis, so this transformation is very important.

3.5 Feature Extraction Techniques

This is one of the important phase of speech processing. As speech contain different features, some feature are important for further processing where as some are useless. So selecting a useful feature from the set of feature is very important. The important feature of signal includes pitch, timber quality, amplitude etc, where as background noise, emotion are the useless content.There are many approach to obtain feature. These includes mfcc, zero crossing rate, and energy barrier.

3.5.1 Cepstral Transform Coefficients

In cepstral transformation We take an input signal and then take it's discrete Fourier transformation and then applying logarithmic to do homomorphic decomposition of signal . Logarithmic filter transform non linearly combined signal into additively combined signal . Once the filtering is done we apply inverse dft to get back the signal in the original form.

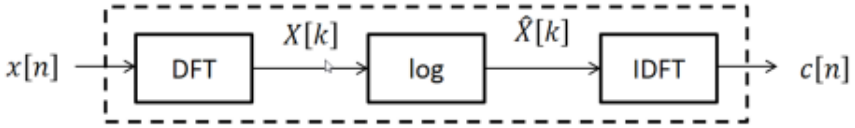

Figure 3.6: Cepstral Transform

3.5.2 Homomorphic Speech Processing

Homomorphic system for speech processing lies in the capability of transforming non linearly combined signals to additively combined signal so that linear filtering can be performed on them. In homomorphic transform there are two phase one is direct transformation of signal from non linear combined to additive and applying different type of filter. After applying all the linear filter we can get back the signal into it's original form by applying inverse dft.

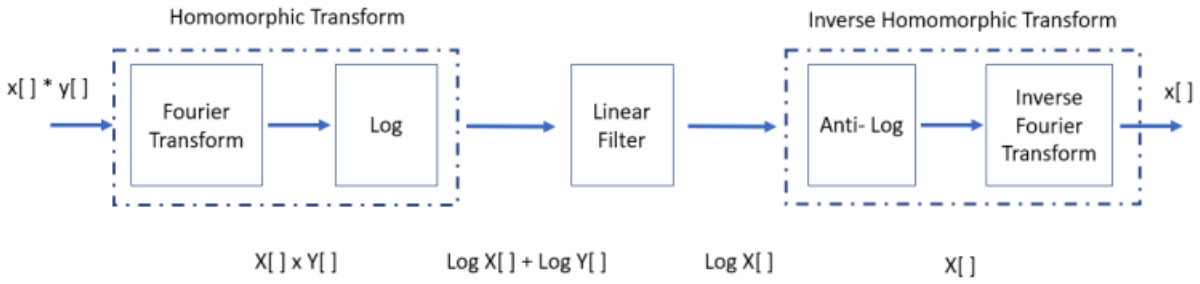

Figure 3.7: Homomorphic Transform

3.5.3 Linear Predictive Coding(LPC)

The Cepstrum contains a variety of benefits (applying filtert to seperate from source,compactness,and many more). These coefficients are highly responsive to numerical accuracy. So we generally convert the LPC coefficients to cepstrum coefficients.

3.5.4 Mel Frequency Cepstral Coefficient

This is one of the most common approach in finding the feature in case of speech signal. As the previous research claimed that it is one of the most important way of extracting feature from the signal. It follows a series of steps. These steps are as followings:

1. Signals are sliced into shorter frame.
2. Obtain the power spectrum for each frame.
3. Filter banks are applied to these spectrum.
4. Logarithmic filter are taken for all the filter banks.
5. Apply DCT to logarithmic filter result.
6. Consider the coefficient between two to thirteen discard the rest.

3.6 First Approach

This is the first approach of the work . We first do all the preprocessing of the signal then try to find out the appropriate feature of the audio signal. We then applied some similarity algorithm to find the similarity between signal. The steps followed are explained below.

3.6.1 Reading Audio File.

In this approach we first read the referenced signal and target signal. It has sampling rate(Fs) = 16000 Hz and Sampled data size = 24000 samples of referenced signal, where as target signal has sampling rate = 16000 Hz and Sampled data size = 25600.

3.6.2 Removing unvoiced region

It is also called as silent region. Unvoiced part present in the speech signal is useless and hence increase computational cost. After silence removal data size of referenced signal is 24000 samples where as data size of target signal is 25600.

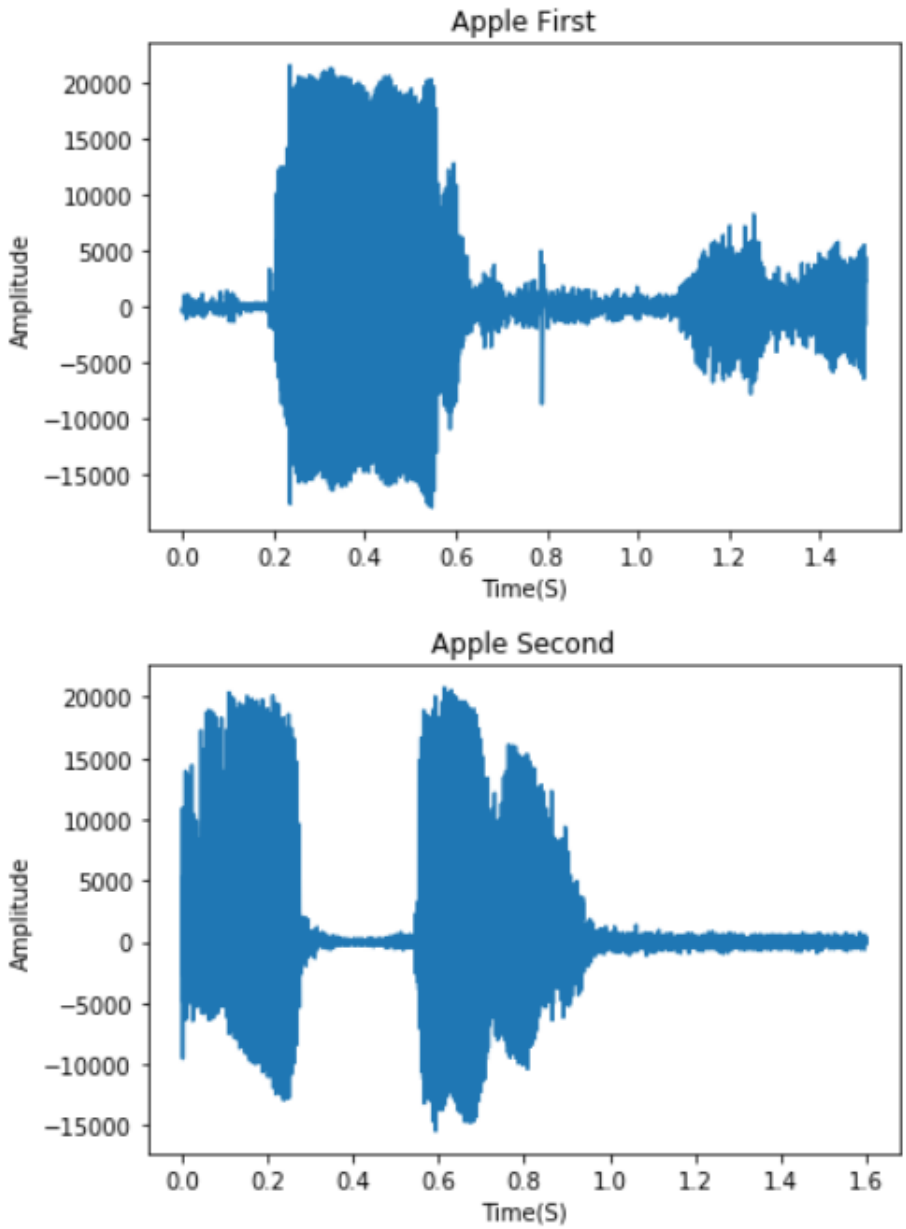

Figure 3.8: Original Speech Signal

3.6.3 Normalization

The signal need to be normalize when we want to compare it with respect to some signals. Normalization means scaling the signals in same level. If we normalize the signals in power level, that means all the signals have same power now.

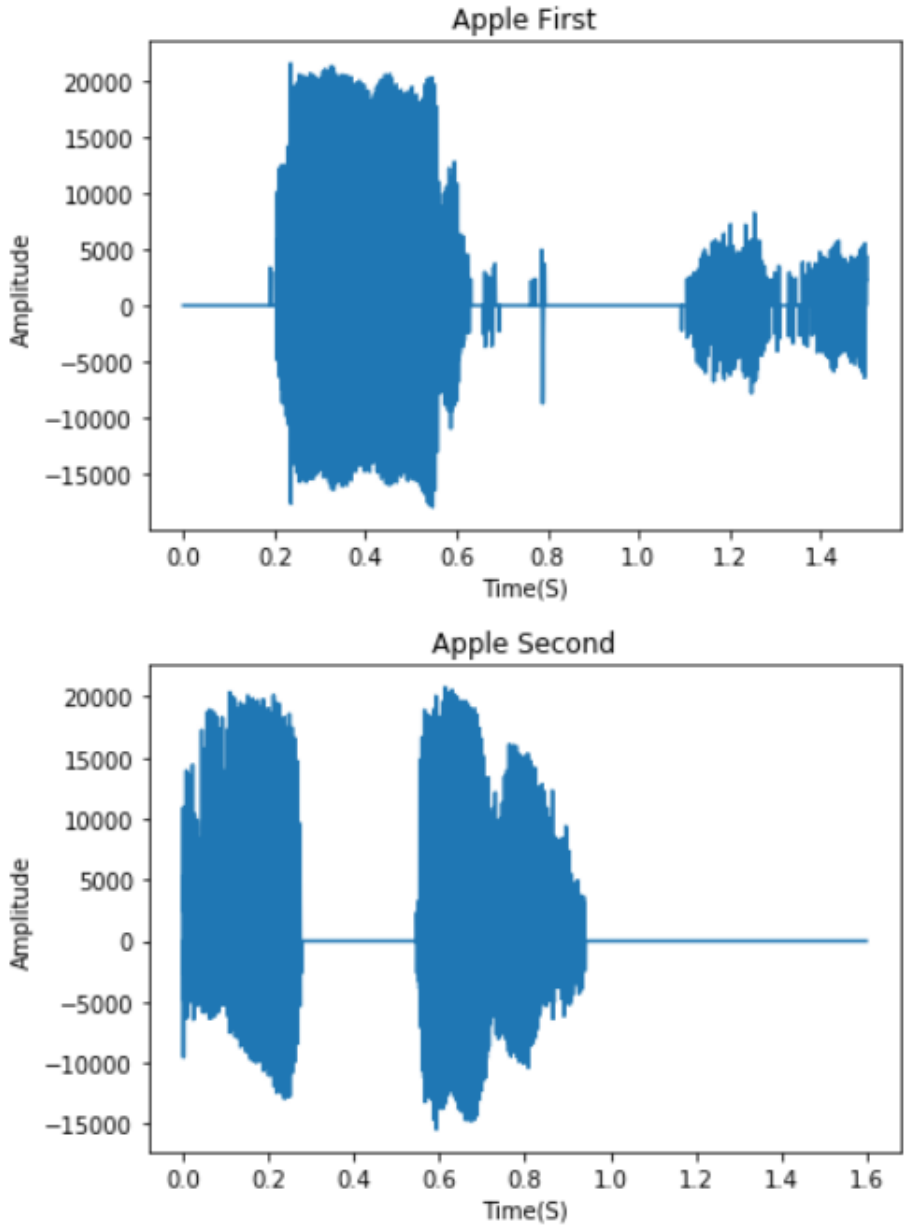

Figure 3.9: Signal after silence removal

3.6.4 Framing

Signal is variant in it's complete domain. It changes continuously in time. So the signal is not suitable for further processing in it's original format. It is assumed that signal is constant in short period of time. So in that frame of time we can apply all the processing. So framing can be defined as breaking the signal in to shorter frame. Basically we divide the frame in a duration of 20 to 30 ms. Condidering the frame in much shorter duration may miss the important information. When we consider the signal for larger duration then there will be chance of getting signal which is changing continuously. So framing

help in obtaining a stationary part from complete signal.

$$frameduration = 25{*}10^{-3}sec$$

$$frameduration = 25{*}10^{-3}sec$$

$$framesize = samplingrate{*}frameduration$$

$$= 16000{*}25{*}10^{-3}$$

$$= 400$$

$$Also,$$

$$datasize = no.of frames \div framesize$$

$$74800 \div 400$$

$$= 371$$

3.6.5 Windowing

Windowing is basically used to remove the discontinuity at the edge of frame. when we frame the signal it produces discontinuity at edge of signal. So to remove discontinuity from the signal windowing is applied. From the literature survey it is found that hamming window is popular window used in speech processing.

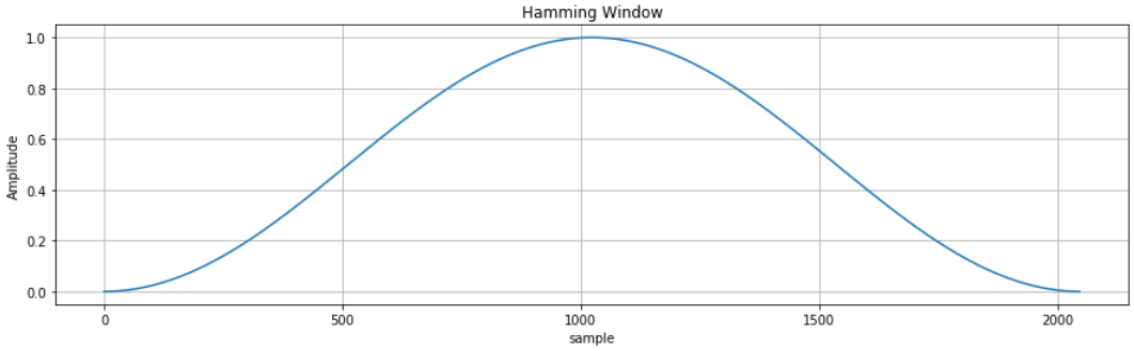

Figure 3.10: Windowing

3.6.6 Discrete Fourier transform (DFT)

Discrete Fourier transform changes the convolution in time domain to multiplication in frequency domain. It separate high frequency component from low frequency component and vice versa. It shows magnitude plot of n-point FFT where N = 2^{12} . It can be any number but N in the power of two make the algorithm run faster and larger value of N means more closeness towards DFT.

3.6.7 Calculating Power Spectral

In this section the modulus value of the complex Fourier transform is taken, and then take the square of result. This will give the strength or power of signal. The power spectral based on periodogram for the speech signal frame is given by:

$$p_i(k) = \frac{1}{N}|s_i(K)|^2$$

3.6.8 Compute MEL-spaced filterbank

We now find the MEL-spaced filterbank and then the framed audio is passed through them. That will give us information about the power in each frequency band. The filters can be created for any band of but for our thesis we will look on the entire band of sample. What special with the MEL-spaced filterbank is the spacing between the filters which grows exponentially with frequency. For any frequency band the filterbank can be made . Here the filterbank is computed for the entire frequency band.

3.6.9 Discrete Cosine Transform (DCT)

We need to apply dct on the output obtained from log filter bank. This step is performed because the obtained filter are overlapping each other. The filter bank obtained are quite related to each other. The DCT try to separate these energies. But we will consider only 12 coefficient out of 26 DCT coefficients. This is because the the coefficient of dct implies the high change which degrade performance. so we drop some coefficient to improve the performance.

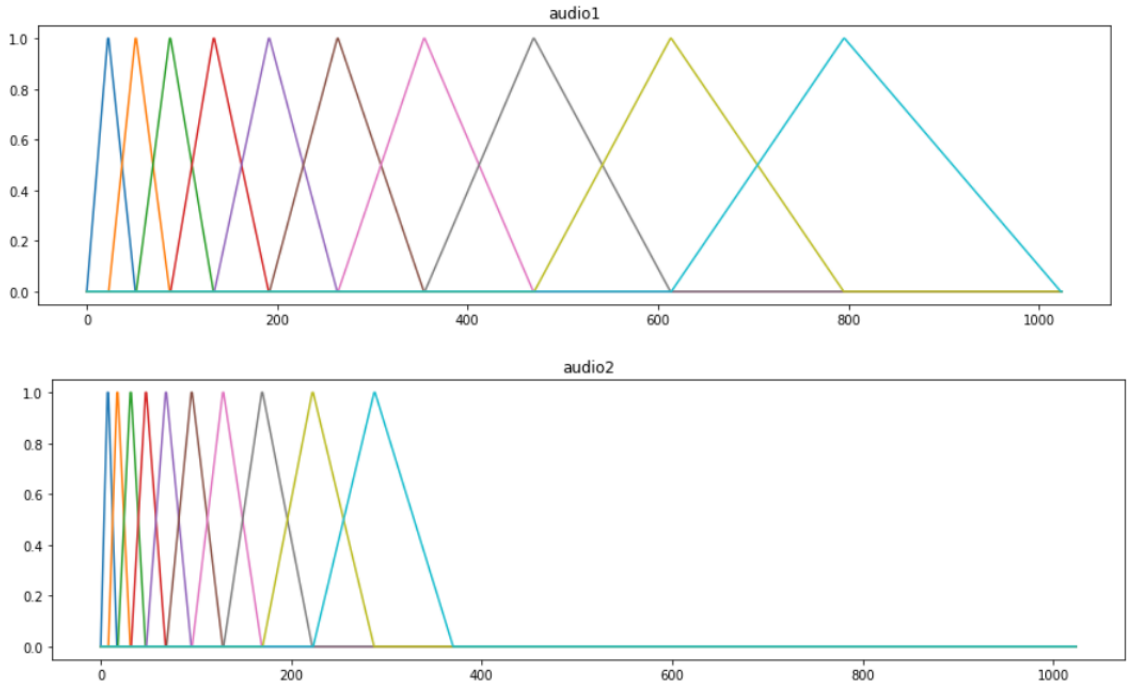

Figure 3.11: Mel-spaced filterbank

3.6.10 MFCC Coefficients

The final step of feature extraction is to generate the mfcc coefficient for both the audio . As we have 12 mfcc coefficients, so there will be 12 single order coefficients and 12 second order coefficients . When we combine these three we get a feature vector of having length 36. We can also incorporate energy as energy and energy will give rise to give a total of 36 + 3 = 39 MFCC feature vector for each frame.

3.6.11 Cosine Similarity

Cosine similarity is a metric used to determine how similarity. We try to find out the similarity score for both the audio signal . In this we first compare the total number of frame in each signal and try to equalize the frame in both the signal. When we get the equal number of frame in both the signal then we apply the cosine similarity . The accuracy in this case vary alot depending on the quality of recording. The maximum similarity score obtained in this case is 41 percent.

3.7 Audio Similarity Comparison Using IPA

In this approach we try to compare the minimum basic unit of sound called phoneme to real audio signal containing some information in it. We first do all the preprocessing used in the first approach. We then extracted the feature from sound. We considered amplitude as a feature of audio signal and phoneme signal. As amplitude change in frame within a certain time duration is zero crossing rate. This zero crossing rate is also an important feature of sound.

3.7.1 IPA

The International alphabet (IPA) is associate degree alphabetic system of phonetic notation based mostly totally on the Roman alphabet. it had been devised by the International Phonetic Association within the late nineteenth century as an even illustration of the sounds of oral communication. The IPA is employed by lexicographers, foreign language students and lecturers, linguists, speech-language pathologists, singers, actors, made language creators and translators.

3.7.2 Approach

All the preprocessing step for audio signal are same which has been used in the initial approach. In this approach first the database of International phoneme alphabet is created. For each phoneme there is a symbol correspond to them. Then we try to read all the phoneme sound present in the database one by one and storing amplitude of all the phoneme as a features. We take a targeted audio and read it and store it's . We extract the amplitude of the audio signal which work as a feature of the audio. Now we have a feature of target audio. Now we apply cosine similarity with each phoneme to target audio. we extract the symbol of phoneme which has maximum and similar matching score. The Accuracy through this approach is very poor . We get a score of 40 to 78 percent in case of phoneme which match the audio maximum. So this approach with slight modification can be explored in future scope.

Chapter 4

Result Evaluation and Observation

4.1 Result

The obtained result of this experiment can be summarized in the following manners:

1. We basically take two audio signal spoken by same person. Both the audio signal contain same word or same information . We then apply all the pip-line designed in first approach and try to check the similarity . The similarity obtained in this case is 41 percent.

2. This experiment is performed for four person and the result differ in each case. It is observed that result obtained largely depends on the quality of both the audio. In some case the result obtained for similar audio spoken by similar person containing same information fails to match miserably.

3. We perform the same experiment with the signal spoken by two different person but containing the same information. In this case system is unable to judge and the accuracy falls to 30 percent. This experiment is also performed on the voice of four person. The accuracy again vary with person to person the maximum accuracy noted was 30 percent in this experiment.

4. We try to find out the difference between two signal containing different information

spoken by same person.

5. IPA approach is giving some better result. so we can move with this approach for further experiment along with slight modification.

6. This experiment is also performed on the voice of four person . The result slightly vary depending upon the way of pronunciation.

7. The maximum similarity score in this case is 78 percent .

Cases of the experiment	Algorithm used	Accuracy
Two audio signal containing same information spoken by same person .	Mfcc along with cosine similarity.	39.687 percent
Two audio signal containing same information spoken by different person.	Mfcc along with cosine similarity.	30.56 percent
Two audio signal containing different information spoken by same person.	Mfcc along with cosine similarity.	28.67 percent
Two audio signal containing different information spoken by different person.	Mfcc along with cosine similarity.	Two much variant result in this case.

Figure 4.1: Result analysis using first approach.

The **second approach** of this work gives a better result however the approach is not completely implemented. Here the phoneme comparison is done letter by letter . The first letter of the audio containing word apple closely matches with the Close mid central rounded vowel. This sound matches with the sound of a and the match percentage in this case was 78.36. However the complete matching of each letter or each phoneme has not been done. As the recording time of the target signal and the IPA phoneme stored in the database was almost same. So the comparison done here is time dependent, but the accuracy we get for the first phoneme is better than than the first approach. So this approach can be carry forward to do comparison by breaking the target signal on the basis of phoneme energy level. However this approach required a good quality recording

of sound as it is totally based on the amplitude of given sound. The result will be more accurate if we break the target audio on the basis of energy level.

4.2 Observations

1. Since the quality of audio matter , so we need a quality recorder to record signal.
2. As audio varies at each point of time , so it is not possible to compare a signal containing specific message or words directly. So in case of signal that contain words , we need to break it on the basis of energy. As there is energy gap between each phoneme which can be concluded from the second experiment. Once this is done we can apply the same preprocessing.
3. As cosine similarity is not giving satisfied result so we can use some other mapping algorithm like DTW(Dynamic time warping) which may give the best possible alignment.
4. As mfcc feature is not giving a good result. so we can move to some other feature like zero crossing or timber quality or pitch.

Chapter 5

Conclusion

Although there was some work done priorly on signal, but all the work was directly or indirectly based on ASR and influenced by the concept of supervised approach. This thesis is dedicated towards signal comparison in real time using unsupervised approach. Audio signal comparison is done in real time without any prior training. We applied two approach for doing this. The firs approach result was not so good which can be concluded from the experimental result. We can further move with the second approach, as we get higher similarity score . Although we are comparing the whole signal with a single in spite of that the similarity score is above 50. So we can move with this approach with slight modification in future.

Modularity Our current approach to resolve this drawback may be divided into stages as shown higher than figure. The modularity within the approach encourages the actual fact that one will work on an individual basis on every of the modules to enhance their individual performance, and therefore a module will continually get replaced by an improved module to enhance the accuracy of the chain.

5.1 Future Scope

Since the quality of audio matter , so we need a quality recorder to record signal which will give better result. Also breaking the audio signal on the basis of energy level and then selecting an appropriate features from them might provide some better result. Zero crossing feature may provide better result as it can be concluded from second experiment. Use of IPA sound along with this pipline may provide better result. As cosine similarity is not giving satisfied result so we can use some other mapping algorithm like DTW(Dynamic time warping) which may give the best possible alignment. Once the Audio signal comparison system is ready we can integrate it with the game platform and obtained our proposed goal. One can then use this platform to learn words and its pronunciation.

Bibliography

[1] Pedro Cano and Eioi Batlle, Ton Kalker and Jaap Haitsma A Review of Algorithms for Audio Fingerprinting. International Journal of Computer Science and Information Technologies 5, 3 (2014), 3657–3660.

[2] D. Hand and C. Eng MIET, "The RF challenges of ATC communications", Consultant Engineer - Park Air Systems Ltd.

[3] Mathieu Lagrange, Roland Badeau, Gaël Richard. Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, Dallas, Texas, United States. pp.405–408.hal-00945296

[4] Logan, Beth and Salomon, Ariel.A Music Similarity Function Based On Signal Analysis.International Conference on Multimedia and Expo,6(june 2001)

[5] Muhammad Atif Imtiaz and Gulistan Raja,"Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW KNN",2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast).

[6] Jian Da Wu, Pang Yi Liu, Guan Long Hong, " Speech Recognition Using Zero-Crossing Features", Applied Mechanics and Materials, vol. 490-491, pp. 1287, 2014.

[7] W. Chen, Q. Hong, and X. Li, "GMM-UBM for text-dependent speaker recognition," in 2012 International Conference on Audio, Language and Image Processing, July 2012, pp. 432–435.

[8] D. Ishac, A. Abche, E. Karam, G. Nassar, and D. Callens, "A text-dependent speaker-recognition system," 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–6, 2017.

[9] S. G. Bagul and R. K. Shastri, "Text independent speaker recognition system using gmm," in 2013 International Conference on Human Computer Interactions (ICHCI). IEEE, Aug 2015, pp. 1–5.

[10] K. Aida-zade, A. Xocayev, and S. Rustamov, "Speech recognition using support vector machines," in 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Oct 2016, pp. 1–4.

[11] B.H. Juang, Lawrence R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development" Georgia Institute of Technology, Atlanta Rutgers University and the University of California, Santa Barbara .

[12] J.R. PIERCE, "Speech Recognition" from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 46, No. 4, (Part 2), 1049-1051, October 1969.

[13] Kazim Ali, "A Study of Software Development Life Cycle Process Models" International Journal of Advanced Research in Computer Science, Volume 8, No. 1, Jan-Feb 2017, ISSN No. 0976-5697.