T.C.

ALTINBAS UNIVERSITESI

Department of Electric and Computer Engineering

A REAL-TIME SPEECH RECOGNITION SYSTEM BY USING ADVANCED DIGITAL SIGNAL PROCESSING IN MATLAB

Khaled Seghayer Elgantri

M.Sc. Thesis

Asst. Prof. Dr. Dogu Cagdas ATILLA

ISTANBUL 2018

A Real-Time Speech Recognition System by Using Advanced Digital Signal Processing In MATLAB

KHALED SEGHAYER ELGANTRI

Electrical and Computer Engineering

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2018

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dogu Cagdas ATILLA

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Asst. Prof. Dogu Cagdas ATILLA	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Cagatay AYDIN	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Aysel ERSOY YILMAZ	Faculty of Engineering, Istanbul University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of .....

Asst. Prof. Cagatay AYDIN

Head of Department

Assoc. Prof. Oguz BAYAT

Director

Approval Date of Graduate School of Science and Engineering: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

KHALED SEGHAYER ELGANTRI

[Signature]

DEDICATION

To my parents, my brothers, my wife, my children and my friends who had the incentive to complete my studies and get a master's degree

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor for their help and moral support so that I can complete my message and not give up any information I need. Thank you very much, Professor: Asst.Prof Dogu Cagdas ATILLA

In the end I also do not forget to extend my thanks and appreciation to my parents, my brothers, my wife, my children, my teachers, who have been credited for reaching this stage of my educational life and my friends. Thank you very much.

ABSTRACT

A Real-Time Speech Recognition System by Using Advanced Digital Signal Processing In MATLAB

KHALED ELGANTRI

M.S. Electrical and Computer Engineering, Istanbul Altınbaş University,

Supervisor: Asst. Prof. Dogu Cagdas ATILLA

Date: [May 2018]

Pages:47

Voice recognition has become essential approach as a part of many evolved technologies and systems; it was adopted by different technology manufactures for different purposes. Modern technologies, which utilize speech recognition systems, need more accurate algorithms to perform true speech recognition function. In this project, a very efficient speech recognition system has been proposed for better outcomes as compared to the classical methods. Advanced signal processing techniques are adopted to analyze the recorded speech, which measures the required parameters prior to performing any recognition task. The speech was recorded using MATLAB by setting the recorder of two channels with 32-bit quality and hence, this speech is used as a sample for constructing algorithm.

The system involves a noise removable section to remove segments without sound so that lesser time will be taken for carrying out speech recognition. Auto-correlation was calculated for matching the recorded signal with the test signal, and then, pitch period was determined assuring the speaker's identity. The results were observed by performing speech recognition to know whether the system can identify the talk contents and the speaker identification, so access is allowed if the same speaker with same content inputs his sound. For this purpose, MATLAB software has been utilized for constructing the system and results were found.

Keywords: Hidden Markov Model (HMM). Automatic Speech Recognition(ASR). Speech Recognition(SR). Universal Asynchronous Receiver-Transmitter(UART).

TABLE OF CONTENTS

Pages

LIST OF FIGURES xi
LIST OF ABBREVIATIONS xii
1 Introduction 1
1.1 Background 1
1.2 History 1
1.3 Types of speech recognition 2
1.4 Problem Statement 2
1.5 Research Objectives 3
1.6 Thesis Organization 4
2 Literature Survey 5
2.1 Background 5
3 Methodology 16
3.1 Introduction 16
3.2 Importance of speech 16
3.3 The importance of voice 16
3.4 Types of speech recognition 18
3.5 Cross Correlation (CC) 18
3.5.1 Autocorrelation (AC) 19
3.6 Project overview 19
3.7 Sampling 20
3.8 Signal Energy and Power 21
3.9 Noise and disturbances 22
3.10 Frequency representation of signals 23
3.10.1 Fourier Transform 23
4 System Implementation 25
4.1 Introduction 25
4.2 Speech recording 25
4.3 Voice signal characteristic 26
4.3.1 Voice and silence determination 26
4.3.2 Feature determination 26
4.4 Algorithm structure 28

4.5 System strength/comparative concept 31
5 Results and Discussion 33
5.1 Outline 33
5.2 Validation 33
5.3 Signal representation 34
5.4 External test input 36
5.5 Recognition procedure 37
6 Conclusion 39
7 References 40

LIST OF FIGURES

Pages

Figure 3-1: Structure of a speech recognition system.....19
Figure 3-2: Sampling frequency vs. original signal frequency.....20
Figure 3-3: Thermal noise.....22
Figure 3-4: Gaussian noise waveform 23
Figure 3-5: Fourier Transform depicted for signal where the right side is a real part, and the left side is the imaginary part of the final signal in the frequency domain.....24
Figure 4-1: Recording algorithm in Matlab. 25
Figure 4-2: A speech signal (a) with its short-time energy (b) and zero crossing rate 27
Figure 4-3: A phoneme with its pitch cycle marks (in red). 28
Figure 4-4: Voice signal slotting.....28
Figure 4-5: SRS algorithm. 29
Figure 4-6: Silence removing algorithm. 30
Figure 4-7: Auto correlation and matching filter of voice signal. 31
Figure 5-1: Recorder signal as database input 34
Figure 5-2: Signal after recording with no silent parts 35
Figure 5-3: Two-sided Fast Fourier Transform 36
Figure 5-4: The external second input of speech. 37
Figure 5-5: The first reference input with auto-correlation results.....38

LIST OF ABBREVIATIONS

HMM.....	Hidden Markov Model
ASR…….	Automatic Speech Recognition
SR……..	Speech Recognition
DSP……..	Digital Signal Processing
FFT……	Fast Fourier Transformer
DCT……...	Discrete Cosine Transform
MDCT……..	Modified Discrete Cosine Transform
DTW……	Dynamic Time Warping
MFCC……...	Mel-Scale Frequency Cepstral Coefficients
UART……	Universal Asynchronous Receiver-Transmitter
LCD……	Liquid-Crystal Display
DFT……...	Discrete Fourier Transform
VSC……	Voice Signal Compression
DWT…….	Discrete Wavelet Transform
VQ……..	Vector Quantization
SVM……	Support Vector Matching
GSM……...	Global System for Mobile
CC……...	Cross Correlation
AC……...	Autocorrelation
SRA……...	Speech Recognition Algorithm
ZCR……..	Zero Crossing Rate
VRS…….	Voice Recognition Scheme
RAM…….	Random-Access Memory
ADC……..	Analogue to Digital Conversion
AAC……..	Advanced And Coding
WMA……...	Windows Media Audio
AVSD……..	Automatic Voice Signal Detection
GUI……...	Graphical User Interface
VQ…….	Vector Quantization
SVM…….	Support Vector Machine
VSS…….	Voice Security System

1 Introduction

1.1 Background

Speech recognition has been a recognized and well-established technology that enables computer systems to capture the human voice and collect its data with a help of a microphone. A specifically designed speech recognizer recognized the recorded data, and at the last stage, this system gives the output in the form of recognized speech/words. This speech recognition system consists of different steps, which are discussed one by one in the upcoming sections. For ideal speech recognition, the speech-recognition engine must have the capability to recognize all the utterances; however, in practical terms, its performance depends on a number of factors. They include vocabulary, number of users, and noise in the surroundings, which play a major role to assure accurate speech recognition.

1.2 History

The idea of speech recognition was presented in 1940s but the Bell Labs presented the first speech recognition system in 1952, which could only perform to some extent in complete silence. So, the foundations of speech recognition were laid in 40s and 50s because the baseline achievements were accomplished during those two decades including presenting information and automation theoretical models. During the 1960s, the speech recognition systems were improved to the extent that they could recognize small vocabulary (10 to 100 words) using basic acoustic-phonetic speech sound properties. It was an important decade because scientists and linguists joined hands to develop time normalization methods and filter banks.

During 1970s, the capacity to handle vocabulary was enhanced from 100 to 1000 words with the help of a pattern recognition method. During 1980s, the vocabulary-handling capacity was raised to the range of 1000 to unlimited; however, substantial speech recognition issues were reported and addressed. The Hidden Markov Model (HMM) was a major invention of the 1980s, which was actually a stochastic speech and language model that was very helpful to handle continuous speech recognition issues at a high pace.

During 1990s, the speech recognition research developed further when stochastic language understanding method was introduced in addition to the statistical acoustic and language models as well as several large-vocabulary speech recognition systems.

At that time after 50 years of research, speech recognition technologies were available for sale in the open market, which have so far benefitted their users in multiple ways. It was a challenge to design a machine that truly functions like an intelligent human, which is still to be achieved.

1.3 Types of speech recognition

These systems are largely divisible into classes based on their recognition power and the stored list of words they have. They are mentioned below:

1.3.1. Isolated Speech

It normally includes discussion pause between every couple of utterances, which means that it can process one utterance at a time.

1.3.2. Connected Speech

It includes connected speaking, which is just like isolated speech but in this case, the pause duration is minimum between any two utterances.

1.3.3. Continuous speech

It allows a natural speech, and it is also termed as computerized dictation.

1.3.4. Spontaneous Speech

Basically, it means natural sound or unrehearsed speech. The ASR system has the ability to recognize this kind of speech including the capability to handle extra gestures like "ums," "aahs," and stuttering sounds.

1.4 Problem Statement

In speech recognition, a very essential process is recording a speech and its conversion in digital format when digital signal-processing is applicable. Like any other technology, the target of this technology is to recognize who the speaker is. That feature is helpful and applies for assuring security because it stops chances of unauthorized access to data, equipments, networks or valuable products. Speech Recognition (SR) functions by digitally comparing the input sound signal and the source signal, which is already saved in a database of the SR system.

SR process is not very tough to understand and operate. Noise in the background may act against the recognition process. For DSP performance, the per-processing is effectively conducted. "Who is the speaker?" is the main question required to be answered by the project. Many aspects are required for evaluating how many of words a speech signal has. The accent of speaker and his language are challenging issues needed to be sorted out. From the available researches, the following challenges are obvious:

- Size: Number of word types in vocabulary, perplexity;
- Speaker: Tuned for a particular speaker, or speaker-independent? Adapting speaker's features of speech and accent;
- Acoustic environment: Competing speeches, noise, and properties of the conveying channel/medium;
- Style: Continuously spoken or isolated? Planned monologue or spontaneous conversation?

1.5 Research Objectives

This project tried to recognize the speaker's identification when he/she speaks with others. For smooth and sophisticated SR, the Matlab digital signal-processing tools box and programmable functions are very helpful.

Speakers may input their spoken language through the Matlab workstation by recording it in the digital format such as WAV or MP3. The channel type such as single channel or double channel may also be chosen with the speech rate for the best signal presentation.

A reference signal can be saved, which may be chosen for reference while performing the comparison on further stages:

- The frequency is evaluated through applying the Fast Fourier Transformation (FFT) on the signal. It also produces the information that constructs a speech signal while disturbances including microphone and circumstantial noises are linked with it.
- Digital filters must be utilized for removing noise components in the signal. Two microphones were used as the other microphone recorded the noise signal.
- After the signal settlement, we compared it with the reference database for evaluating the characteristics of the speaker.

- As a result, the system recognizes the speaker's information including the name and designation, for example: "YOU ARE: KHALED ELGNTRI, 37 AGED, DIRECTOR" or "YOU ARE NOT AUTHORIZED TO ACCESS THE SYSTEM."

1.6 Thesis Organization

This dissertation report consists of six chapters, through which, all the speech recognition processes and algorithms will be demonstrated.

Chapter one “Introduction” describes the basic definitions of SR with a problem statement of this research and its objectives.

In chapter two, “Literature Survey” has been presented, which combines the findings of relevant articles pertaining to speech recognition along with details, drawbacks and advantages of each research.

Chapter three discusses the “Methodology” of the suggested method, and the proposed methods and materials.

In chapter four “Practical Model,” the system execution using Simulink and Matlab is displayed with the settings and underlying technologies.

In chapter five “Results and Conclusions,” the outcomes monitored by this project are reported, and the conclusion has been stated with further recommendations.

Finally, chapter six “References and Publications” consists of references used in this dissertation and the research contributions.

2 Literature Survey

2.1 Background

Speech recognition technologies have been used in many projects, and they are playing pivotal roles in multiple applications and technologies such as control systems in robotics and security systems that authenticate people for accessing some place or system. Many research projects have been executed on speech recognition (SR) technologies. According to the literature, the observed researches fall into many categories such as demonstration of the SR applications and design requirements of those systems. The following approaches reported satisfactory results within this domain.

In [1], the author mentioned that SR makes use of specific words of a specified speaker and automatically recognizes and saves the information through individual speech waves. This study adds valuable information pertaining to the relevant inventions and technological advancements, which have taken place in voice recognition systems. It focuses on certain process steps, which help identifying a speaker through MATLAB. This approach first performs speech-editing and signal-degradation at the same time using Gaussian Noise. This background noise is later successfully removed using the Butterworth Filter. Moreover, this technique develops code with the help of MATLAB Program that has in-built features to match the pitch as well as format vectors of the signal with an already saved speech signal. The process of comparison does not stop here. It continues comparing the signal with other unknown speech signals to find the real and accurate match. The speech was recorded as WAV format and then, the signal magnitude was set at 30,000 points. The mentioned vector has two parts with same length in the opposite fashion. For noise removal, researchers have discussed two techniques i.e. speech degradation and speech enhancement. Because of some limitations; the current study failed to properly discuss the speech signals' common features and behavior.

As we know [2], voice recognition is a significant tool of the modern generation and it is popular in various fields for various purposes. Voice recognition technologies have made dramatic progress during the recent decades so much so that now high-performance systems and algorithms are available. Their performance is judged based on speed and accuracy while recognition accuracy is a significant and straightforward voice recognition performance parameter. This research has reviewed several voice algorithms to detect accuracy and process overheads for identifying an optimal voice recognition algorithm that can give the

best tradeoffs between processing cost (speed, power) and accuracy. Our aim is also to verify the chosen voice-recognition algorithm using MATLAB.

During our experiment, four male and female speakers spoke ten words each in a MATLAB simulation environment, which acted as reference signals for the algorithm. During the evaluating phase, all the signals were tested subject to the similar test criteria using algorithms. It was obvious from the simulation results that the Wiener Filter algorithm outperformed and fully superseded the remaining four algorithms on all grounds such as performance, power requirements for a moderately complex algorithm and its prospective implementation as hardware. Wiener's algorithms depicted 100%, 5%, and 50% accuracies with 695-867msec recognition speed range while the power range was 750-885μW.

In [3] this study, the efforts are done to identify a speaker for security purposes. Providing security or encryption for accessing the system is done with the help of a 6-digit password so a user will require that password for accessing the robot. This speech identification method compares speech signal from speaker with the signals stored in the database library. After comparing, the system identifies a speaker. Speech identification for handling or controlling any appliances or objects like Robots is possible. As an example, if a speaker says “forward,” the robot can move forward. The study involves the audio normalization that does an audio recording and then brings the average/peak amplitudes to the targeted level. Since similar gains apply in a given range, the relative dynamics and signal-to-noise ratio (SNR) normally do not change. Normalization is different as compared to dynamic range compression that applies on different gains as a recording fits in a range having definitive minima and maxima. A digital audio workstation generally provides the normalization function.The discrete cosine transforms (DCT) takes into account finite data sequence in the form of sum of cosine functions, which oscillate on varying frequencies. DCTs play a significant role for the functions of numerous scientific and engineering applications such as MP3 audios and JPEG imaging, which discard low-frequency components. They also apply in case of spectral methods, which solve partial differential equations. It is important to use cosine function in place of the sine function for compression; however, few cosine functions are required for approximating a signal. On the other hand, cosine function specifies specific boundary conditions for differential equations. Relevant transformations, which take place through modified discrete cosine transform (MDCT), are utilized in MP3, AAC, WMA, and Vorbis audio compressing.

In [4], an English dictionary has been presented that consists of meanings and accents of English words. This dictionary is voice-operated and the user inputs his/her voice by mentioning individual alphabets. When the user spells alphabets one-by-one, the dictionary provides the pronunciation and meaning of the word, which is formed with the help of alphabets spoken by the user. Latest voice recognition works with the principle of forms' recognition. This method and its algorithms are largely classified into four major classes including "Discriminant Analysis" based on Bayesian discrimination such as Hidden Markov Models (HMM), Neural Networks, and Dynamic Time Warping (DTW). Sound databases have pre-defined wave files, which play the accent and mention the meanings of words, and it is possible to use them with the help of “JAVA” and “MY-SQL.” When a wave file is imported in MATLAB, a user understands the meaning and pronunciation in audio format. It is very useful for blind persons. Our work includes vocabulary database, their pronunciations and meanings but the dictionary can be defined as per user requirement. It is extendable as and when required. This flexibility regarding size/type of dictionary provides a unique touch to our idea.

In [5], a study on project development has been presented that provides the information about control system design and use of speaker recognition codes when Matlab programming is utilized. Matlab has a very useful and user-friendly programming interface, which can be ideal for speech analyses. For the current project, we first learnt how to use Matlab programming and control system design. The fundamental speech recognition algorithm is written as a rule based on Matlab and choosing the most-likely match having pre-defined speech time frame. In this case, defining time frame is the first step in order to record the command words within 40,000msec having frequency fs = 8,000Hz.

Recording key word through “wavrecord” function is the next step, in which, magnitude value more than 0.1 is taken, and the difference is calculated. Later, the file is stored through “wavwrite” function. For storing other keywords, the process is just like before. Next, it will read the stored file at more than magnitude 0.1. Next, the difference is calculated and stored as a variable that is later matched with a pre-defined time frame when the match gives the output. Used algorithm and system-control fan speed, heater temperature and direction of robot that uses the keywords are very important in this context. It demonstrates its reliability and ease of future development. Based on the experimental outcomes, it can be claimed that the proposed algorithm is functional and utilizable with the help of voice keywords to control

home appliance and industrial robots. The proportion of correctly recognized commands has been significant. The main contribution of this study lies in keyword recognition as well as controls. The trials confirm that the proposed approach is good for keyword recognition. The proposed ASR and control systems were fully implemented.

In [6], it is obvious that the speech signal is significant for digital signal processing. In this approach, the speech sample was observed with MFCC to improve the speech characteristic representation through HMM-based training approach. Mel-Scale Frequency Cepstral Coefficients (MFCC) has been effective for extracting voice characteristics out of a voice sample while HMM is also employed for recognizing speaker-based extracted features. The main objective was speaker recognition using HMM parameters, which is then matched with the genuine voice signal. If the signal matches, the voice signal produces text with the help of MATLAB. The outcomes of the performed simulation show that the speech recognition improves as far as computational time and the system's learning precision is concerned. Here MATLAB output is interfaced with microcontroller by using UART and display through the LCD module.

In [7], the author mentioned that the speech interface with the computer is a giant step that the technology needs to take for general users. Automated Speech Recognition (ASR) will help making the technology assist the people in real life. Many speech recognition applications including aircraft direct voice inputs, speech-to-text processing, data entry, and voice dialing etc. ASR is divisible into two main categories including feature recognition and extraction. In this study we have described MATLAB feature recognition with the help of ASR neural network. This research aims to find out how neural networks are useful for recognizing isolated-word speech, which is indeed a replacement of orthodox technologies. Commonly developed methodologies are extendable for other technologies including missile tracking, sonar target recognition, and underwater acoustic signal classification. Back-propagation neural-network algorithms use input samples while the output values show a specific pattern. When these values are modified, a trained network is formed, which is used in the ASR systems.

In [8], the technique, which automatically identifies a voice signal, has been proposed in this study. AVSD is performed based on voice signal's frequency content for several samples having similar sound. An underlying objective of this approach is to automatically investigate a fake voice signal through a security system. Frequency content can be mapped in a frequency domain through finding DFT with the help of FFT algorithm. A speech signal

detection algorithm continuously computes absolute average difference between the two-adjacent voice windows and matches it with the pre-defined threshold. As per proposed algorithm, AVSD will automatically detect voice signal by males' and females' voice signals through "wav" voice signal. It will illustrate fake voice signal in case of stereo, and mono wav files having more 15-word length. For AVSD, the decided voice signal length was 1.024 seconds. AVSD supports stereo and mono signal through a wav file but if the recorded wav file is compressed to wav file format, the processing of AVSD will be better because it will take very less time for execution. For this problem, VSC (Voice Signal Compression) is already developed and published that compresses 50% source file in wav format using the same extension. The AVSD is applicable for pervasive computing.

In [9], the author stated that the actual speech-to-text converting programs convert the spoken words immediately after they are spoken. This approach introduces a unique way human-computer interaction through a specific way called as natural language processing, which is actually a speech recognition process. As a part of the current study, we recorded nine sample voices samples with the help of a microphone in order to familiarize the system with the recorded voices. Each word has some specific range of these parameters. Some words are same, but they still have some non-similar parameters like indistinct beginnings and ends, both seven and one end with similar voices but they are spelt differently. So, it sounds same sometimes, which makes system give output ‘one’ when seven is pronounced. Such type of ambiguities can be removed when samples are chosen in large numbers for one particular word. This system is noise-sensitive. This system is also sensitive as far as pronunciation is concerned. Recorded words are entered in a database but they need to be pronounced similarly during training. This system is also tone-sensitive when a word is pronounced. MFCC speech characteristics were found, and it was observed that words are distinguishable according to their associated energies. This system provides a high accuracy for text conversion. Real time outcomes are obtainable in the lab. The user was speaking to a microphone while the typed text representation appeared on the computer screen, which was the final output. It is possible to train this system for further vocabulary, speech patterns and paragraphs. Each word has its own parameters and bounded values.

In [10], the researchers presented a robust and useful process to extract the features of proposed speaker's voice that helps in speech recognition. For this, a Matlab program was developed using discrete wavelet transformation (DWT) theory. The speakers' voices were recorded and entered in the database, and DWT was allowed to calculate variables and

properties for speaker verification. When the recording time was entered, GUI pops out to confirm the entry of the voice in the database. Thus, after addition of a recorded voice, a user can test whether the system can recognize his/her voice according to his/her Sound ID. For this purpose, parameters including sample frequency, bits per sample, and the time duration are used. This system will require a user to enter another voice having set a recording time, which is normally 8 seconds. During these 8 seconds, the user will record his/her voice until the recording time is over. At that time, the system computes DWT coefficients for the recorded voice and compares the calculated outcomes with the DWT coefficients existing in the database. Experiments show that the mentioned process is very effective, and the results are satisfactory, so, the wavelet-based voice recognition and its performance are discussed and highlighted.

In [11], the author mentioned that humans find it easy to recognize familiar voices but identifying voice with the help of computer programs is quite challenging. It happens when a certain problem is addressed through algorithm development for recognizing a human voice. Saying a word in similar fashion, tone, or pronunciation is not possible when it is uttered on a different occasion. The analysis of human speech provides a different interpretation, which has different utterance speed. The current research provides details about the system that carries out voice recognition with the help of an algorithm. The voice recognition works on the principle of matching templates technique that requires a user to make a template by first recording 10 voice samples by calling a phrase, which will be the known voice. Thereafter, the voice is recorded for in-depth analysis using Discrete Fourier Transform (DFT). Because of the human speech and its nature, the data having frequencies more than 600Hz is wasted, so, when a recording converts into a specific frequency, it can be considered as a 600-dimensional Euclidean space vector. At this point, a comparison can be conducted between two vectors through normalizing those vectors and computing their difference (by component-wise subtraction in R600). Unfortunately, it is unclear what norm should be used; therefore, contrasts and comparisons between Euclidean, Taxicab, and Maximum norms have to be accomplished first. DFT was used to compare frequencies of a couple of recorded sounds as that did not change when the speech changed slightly. Chebyshev inequality helps determining two voices uttered by exactly the same person. MATLAB was used to apply and test the algorithm. The voice recognition process assures that it would recognize a speaker's enrolled voice 75% of the time.

In [12], the writer stated that humans use speech for expressing ideas, feelings, and thoughts for communication. In this world, 1% of the population has fluency problems. Stuttering is a disorder that obstructs and disrupts the fluency through repetition, pauses, unclear/prolonged utterances and interjections. Eliminating such lack of fluency would be helpful for the people with speech disorders because it will help them communicate easily. This approach proposes a system that has the ability to remove silent pauses from the negotiation and initiate the corrected and comprehensible speech. In order to increase efficiency and accuracy during the extraction process, speech signals go through pre-processing. Pre-emphasis filters speech signals using the FIR filter. Stuttering is eliminated by considering that the voiced speech possesses higher energy levels as compared to the unvoiced speech. MFCC algorithm performs feature extraction process. The VQ code book is the output of the clusters of training feature vectors of broken down speech pattern, which is saved and shifted to the database. In this process, clustering is accomplished using K-means algorithm. DTW algorithm was used to match abrupt speech with the database. Finally, the silent paused stuttered speech is corrected, and stutter-free speech is recognized. Stuttered speech correction and recognition systems have been developed. And this system clearly understands the words when a person speaks them despite speech disorder. The current system is employed only for isolated silent-pause stuttering words. It is possible to improve this system for complete sentences, and also for multi-modal stuttering.

In [13], the writer mentioned that Matlab programming interface is ideal for recognizing Hindi keywords. Matlab 7.5 is used to extract the features of Hindi keywords, for which, a database was designed. This database has eight keywords. Each keyword is entered in a database when 10 people including 8 men and 2 women speak, which result in 80 samples. The speech signal features are noted as measures of Dynamic Time Warping (DTW) and MFCC coefficients, which are used to match features during the matching processes. This approach presents the technique for detecting utterances with the help of MFCC, end-point detection, and DTW, which compares different test patterns. The recognition result is further tested to assure clean and noiseless data. This system is considered as robust when the average accuracy of the cleaned data is 97.50% whereas for noisy data, it should be at least 91.25%, which is acceptable because it is natural that the people don't mind repeating their commands at least once. It is implementable using a common microcontroller having basic dedicated memory and analog- digital converter, which accepts the spoken input. The system would be fast, small and economical for incorporating in the consumer electronics variety.

The research aims at developing a speaker-dependent, isolated-word, and limited-vocabulary speech recognition system that is small enough as a small household appliance and easy to operate.

In [14], the author states that voice recognition (VR) automatically recognizes some words of a speaker based on the information in the individual speech waves. This approach is useful both as an invention, and another VR technological development, which focuses on steps to identify a speaker through MATLAB. The wireless voice-controlled smart home systems are primarily for the benefit of the disabled and the elderly. This system consists of two major components including a VR system and a wireless device. In this case, the devices will work with the help of spoken commands such as ON and OFF. Automation of 2 devices, for example, a light and a fan were tested using four voice commands with the help of a desktop computer. The system allows a user to create a profile and automates the process of speech recognition with more than 90% accuracy. Some people operated this system by creating their profiles, and it was found that the speech recognition accuracy was 75% when the system was based on personal computer. The study outcomes show that the system is helpful for adults and physically disabled persons, who are unable to perform various activities efficiently when they are at home, and they need someone's assistance to perform those tasks.

In [15], the study reported that it is possible to link VR recognition with biometric technologies. It is provides authentication to any system based on features of voice instead of images. This research is focused on building a voice recognition system with the help of back propagation algorithms. It is possible through voice signal comparison of a speaker having recorded voice signal in a database through extracting the main voice signal features with the help of Mel-frequency Cepstral coefficients that are significant factors to achieve better recognition accuracy. The proposed scheme depends on two basic phases that each recognition system is composed of. The first phase is the extraction of features or feature analysis, which creates digital voice sources from of the existing vocabulary that is present in the voice database and it serves as an acoustic signal of a pronounced voice called as the source signal. That is actually the speaking person's name. So, every signal turns into equally-lengthy templates, and each one of them converts into attributes' vector, which extracts signal features in that template. These vectors exist in groups, and they are termed as "Features Vectors." The process repeats in case of every digital voice in a vocabulary group. The second phase is feature matching, which is also termed as the recognition process, which converts the input signal and makes it recognizable. That is called as the test signal for a

series of vectors, which convert from beginning to end during feature extraction. The mentioned features are matched with the existing ones in a database through pattern-matching method and recognition decision is given after matching with Euclidean distance function of a couple of features vector series. The first represents the vector of source signal feature while the second is the vector of test signal features. This scheme has been created with a voices database having voices of 40 people by taking 5 samples of each individual person, each of them recording a sample by pronouncing his name five times during system training.

In [16], a VR system has been presented that identifies the administrator's voice. MATLAB is used to code VR that authenticates administrator's voice. By converting the waveform into its parametric representation, the data is available to process and analyze further. Many possibilities are available to represent the speech signal parametrically in a VR system including Mel-Frequency Cepstrum Coefficients (MFCC). In this case, a recorded input voice signal is compared compare to a signal, which is saved in database through the MFCC process. The voice-based biometric systems recognize every word separately, so, when the administrator speaks his password, the system trains and stores it. During testing, speakers may utter their passwords to know whether it has a match. When MATLAB simulation is used, an output signal is obtained in the form whether the system accepted or rejected a user. From testing system results, it is obvious that the system successfully recognized the specific user's voice and rejected other users' voices. Consequently, the overall system accuracy while recognizing a user's voice was quite satisfactory despite the fact that it was a system with moderate security level.

In [17], it is obvious that languages have been complex in all eras in a real-world application. Approximately, the number of languages is 6500. It is possible to record the language in either analog or digital media such as Braille, graphic content or whistle because languages are independent of modalities. Different languages are spoken in different parts of the world. When a person speaking a different language meets another one, it is difficult to understand what he/she is talking about. Hence, detecting language is the main focus, which is spoken by another person. The databases of different languages have been recorded and maintained. PCs are hardware devices having Matlab software, which can be utilized for extracting speech features and identify a particular language. For language detection, database of different languages is created. The signal is processed with the help of a database for getting the final output, which is the final detection of a language. As a result, it was observed that spoken words of Hindi, Marathi and English were successfully detected. The system accuracy

depends on a number of unique and clear samples. If the number of samples is more, the accuracy will be higher. The neural network is used for classification, which can significantly increase the accuracy; however, the samples required per language should be minimum 500. The SVM Multi-Classifier can successfully classify more than two objects with moderate number of samples. This project's overall accuracy is 80 %.

In [18], a latest approach towards controlling the computer with voice commands using Matlab has been presented. It throws light on useful inventions and technological advancements pertaining to Matlab-based voice recognition systems. In this study, VR has been used as concept for controlling the computer. First, analog audio input is entered using a microphone that is recorded as a set of samples while MFCC feature extraction is done using voice commands, which are stored in a database. The actual input command should be implemented, so it is sent to the computer, which compares the stored database through MFCC algorithm. In case, the recording matches, the required operation is performed. This system is helpful for physically disabled persons, and also for the people who need basic computer access. In the database, audio commands including open, close, move and play are stored. When a user gives a command, it is automatically compared to the commands existing in the database. Initially, the voice input takes place with a microphone; this voice is in an analog audio signal format. The voice inputs are recorded in the analog format in the Matlab. The analog samples have to be converted into digital data. This data is stored as a digital fingerprint, which is generally used for the reference of the command. In this system, the speech commands are saved in "wav" format. The location of file or application that is to be opened is linked with that command.

In [22], a research focuses on speech recognized, and its automation with the help of speaker identification. For this purpose, wireless communication concepts are used. It explains the automation system design with the help of wireless communications as well as speaker recognition systems through Matlab coding. Straightforward and useful Matlab programming interface is an ideal choice for speech analyses and speech recognition projects. Its automation is helpful for domestic as well as industrial use. This study focuses on the wirelessly automated system design that was structured and tested. The science of speech recognition focuses on speech commands, which are available in the Matlab database, and they are matched whenever a speaker inputs his/her voice commands. MFCC algorithm has been utilized for recognizing an audio, and extracts speech features. It performs through low-power RF ZigBee transceiver wireless devices because they are normally cheap. This

automation system is intended to manage lighting, cooling heating and other functions at home or in a commercial place with the help of just voice commands. Further, if security is not a big issue then speech processor can control the appliances without speaker identification.

In [21], VSSS is designed for the users who want fool proof security to secure offices or homes. The important aspects of a security system include authenticated accessibility and appropriate authorization. Human voice comprises of voice and words. For the system to operate, the speech and the voice are mandatory to match the recorded sample. The system design also involves a GSM system for sending feedback messages to the prime member of the unit with the users' information, who is trying to access the system. This project includes amalgamation/combination of all the approaches (hybrid approach), which are helpful for achieving more accurate results. In the project, hybrid version of available methods gives better results. The voice input takes place through a microphone that is connected with the system. Authorized person/s train the system using their voice samples. The analysis of the test duration voice piece is carried by MATLAB coding. The result of the analysis determines whether accessing the system is permitted or not A GSM module is connected with the system to ensure tight security. If a system finds authorized user’s voice, the locked door will open and at the same time, a message, which contains the information of the user, will be sent to the principle member of the organization. Also, in case of intruder, the message will inform the administrator.

3 Methodology

3.1 Introduction

The voice recognition (VR) systems play a significant role to run many applications especially those, which are used for security purposes. Before adopting such a technology, many approaches should be studied such as analog-digital conversion, sampling and many other advanced technologies for digital signal processing. The impact of VR systems in many applications are discussed in this chapter along with processes within the system. In speech recognition systems, many technologies have been tested, which work with speech signals in time as well as frequency domains for extracting the principle components of a signal and its analysis for collecting the necessary information.

Usually, the information collected from speech signal processing is applied as input to run multiple control and security applications. The algorithms segregate the noise and remove the unused information within a speech signal will be discussed in the upcoming segments.

3.2 Importance of speech

Humans are distinguished among all the creatures by their capability to use tones and sound signals. Speech is an important attribute for the humans, which enables them to communicate and survive. Humans use the sound signals in routine life and every individual has his/her own tone. The variations in sound among different people exist because of their frequency difference in their vocal cords. The frequency is the only unique attribute that every human has, and that is also because of the difference in their vocal cords. The cords of music equipment such as guitar generates different tones by different actions of its cords and that usually happen when the guitar players changing the cord length by holding/pressing them from different distances. The same concept applies to human vocal cords because they can only produce the sound according to their vocal capacity with a frequency according to the capacity their cords.

3.3 The importance of voice

The sound/speech is used for communication and surviving. Moreover, the impact of a voice signal has been proven in many engineering applications and approaches. As described above, humans can be recognized by their speech so recognition-related applications including security systems are built up with the help of a voice signal. Voice signals can act

like passwords for speakers' identification. Many other applications are used besides speaker recognition; they are listed below:

- Control gadgets/devices through voice have inspired mankind for many centuries. Today, it has become possible when speech recognition (SR) systems created a niche, which is obvious in many areas of practical life. The SR accuracy has been a very significant research challenge other than noise/s, difference of speaker, variable languages, and size of vocabulary. The SR recognition system design needs attention to details including speech classification and representation, pre-processing stages, feature extraction techniques, database and performance evaluation.
- To develop a speaker identifying automated system with ability to control several devices through verbal commands requires voice command security and receiving speech commands through Matlab and Arduino.
- VR is used in robot controls, “hearing sensors,” and speech synthesizers in a mobile robot so much so that has the capability to communicate with people using natural spoken language. The SR system means a system, through which, a speaker can input his/her voice to a computer, which converts it into commands or text for executing computer functions. A hi-tech SR system makes a robot understand and execute orders. A complete SR system has sub-systems including a central control unit, computer system and SR system.
- Speech to text conversion is a needed function because speech is a significant human need as well as a convenient communication source. The human-computer interaction is called as human-computer interface. Major technological perspectives on speech-text conversion and techniques are established on every speech classification stage. The comparative study on different techniques has been accomplished. Such study includes development of human-computer interface in many languages, which comprises different techniques on every step of the SR process, and analyzes approaches to design a useful SR system. Using modern processes, methods, and algorithms, speech signals are now easy-to-process while they efficiently recognize text. In this system, internet-based speech-text engine is needed; however, the speech transfer in written format needs specific techniques, which should be accurate and comprehensible. These approaches recapitulate and work for many SR systems and cover speech-text conversion. It identifies many research application and topics that are on the forefront of this futuristic field.

3.4 Types of speech recognition

SR has three types, which depend on vocabulary, speakers, and bandwidth. These parameters have been areas of interest of researchers, and they conducted several studies on their role in the SR:

- Isolated word
- Connected word
- Continuous word
- Spontaneous word

The current research is mainly about isolated word speech recognition, which requires a user to stop/pause for some time after every utterance. This system has two phases including training and recognition phases. During the training phase, a training vector is created for every spoken word. The training vectors explore voice features to separate different words. Each training vector stores like a template that saves either a single word or a word class. They become part of a database and their function is matching during the recognition phase when a user utters a word, for which, the system is trained. The test pattern is created for a particular word, and the corresponding text string is displayed as an output using a pattern comparison technique.

3.5 Cross Correlation (CC)

Assuming the recorded speech signal of a similar/same word, and two speech signal spectrums are also similar. When cross-correlation is calculated for two same spectrums, and the results are plotted, the cross-correlation is symmetric by definition. After finding cross-correlation, the frequency spectrums of a couple of recorded speech signal, so we need to compare them. For this purpose, we find maximum value position, for utilizing the values right to maximum value and minus values on the left. The absolute value of this difference is taken to find out its mean square error using the absolute value. The cross-correlation symmetry of both the signals shows that level, on which, both the signals are matched. Moreover, the more symmetric the cross-correlation is, the smaller will be the mean square error value. By comparing mean square error for each trained word and the target word, the system decides the training word has a better match with the test signal, which depends on the least mean square error.

3.5.1 Autocorrelation (AC)

It can be found by computing the cross-correlation of a particular signal and itself. It will not correlate with the remaining/other signals. Autocorrelation algorithms measure the extent to a signal's self-correlation. Thus, a training signal's autocorrelation comparison finds out the least differences among autocorrelations.

3.6 Project overview

A speech recognition process should be implemented using Matlab to provide speaker identification services. This system is viable for the security, and it may be used for preventing un-authorized access into a particular system/data. Advance digital signal processing designs a Speech Recognition Algorithm (SRA). Accurate and practical speech recognition results in smooth and quick recognition. An authorized speaker may record his voice and the same is stored in a database directory. This voice clip will act as a reference signal for recognition. The entered signal will be analyzed by digital signal processing schemes, and hence, the principle components are detected. After recording the reference input, the program is applied to take the second input and match the same with the database. If the input is matched for the same speaker, the program will allow the access and will issue a message on the command window indicating the same; otherwise, if the input is not authorized, the program will display a message that reveals the fact. In Figure 3.1, SR system is depicted.

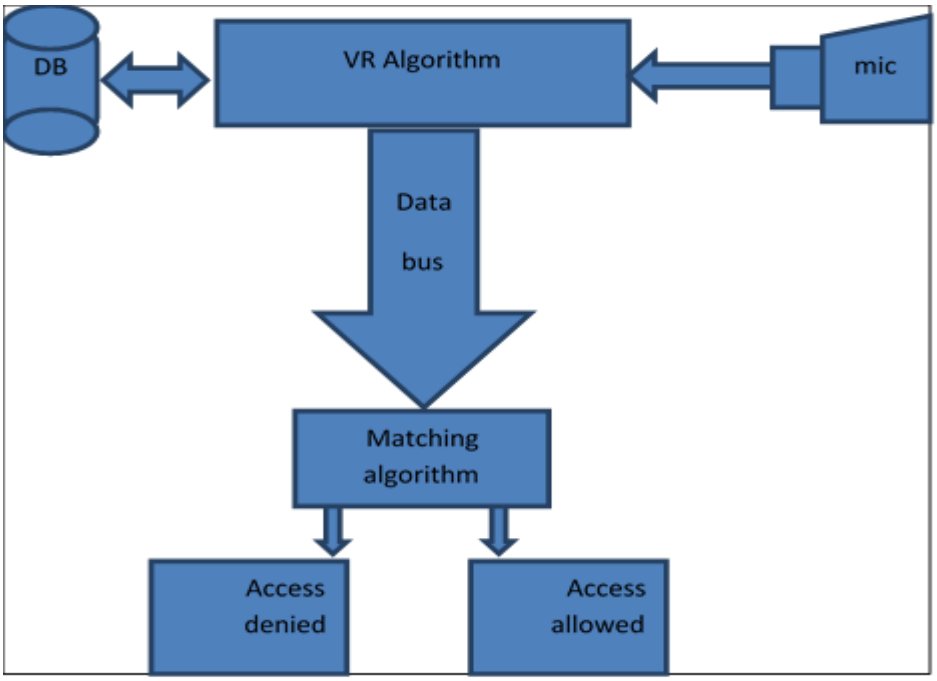

Figure 3-1: Structure of a speech recognition system

3.7 Sampling

For applying digital signal processing techniques to a speech signal, it is mandatory to convert such signal into a digital form, in which, sampling plays an important role for single processing. It involves breaking a time-domain signal in the slots having same widths as Figure 3.2 shows. The same is possible when nyquist rate (sampling theory) is used when the signal frequency is changed by sampling frequency (Fs). The sampling frequency has to be bigger than the signal frequency that was set for achieving better accuracy of signal slots.

The sampling rate is a significant attribute that helps using a speech signal, and it plays a major role in the accuracy of the analysis. Knowing a sampling frequency means identifying the location of voice information. Sampling is converting the continuous time domain signal into slots carrying the information as depicted in Figure 3.2.

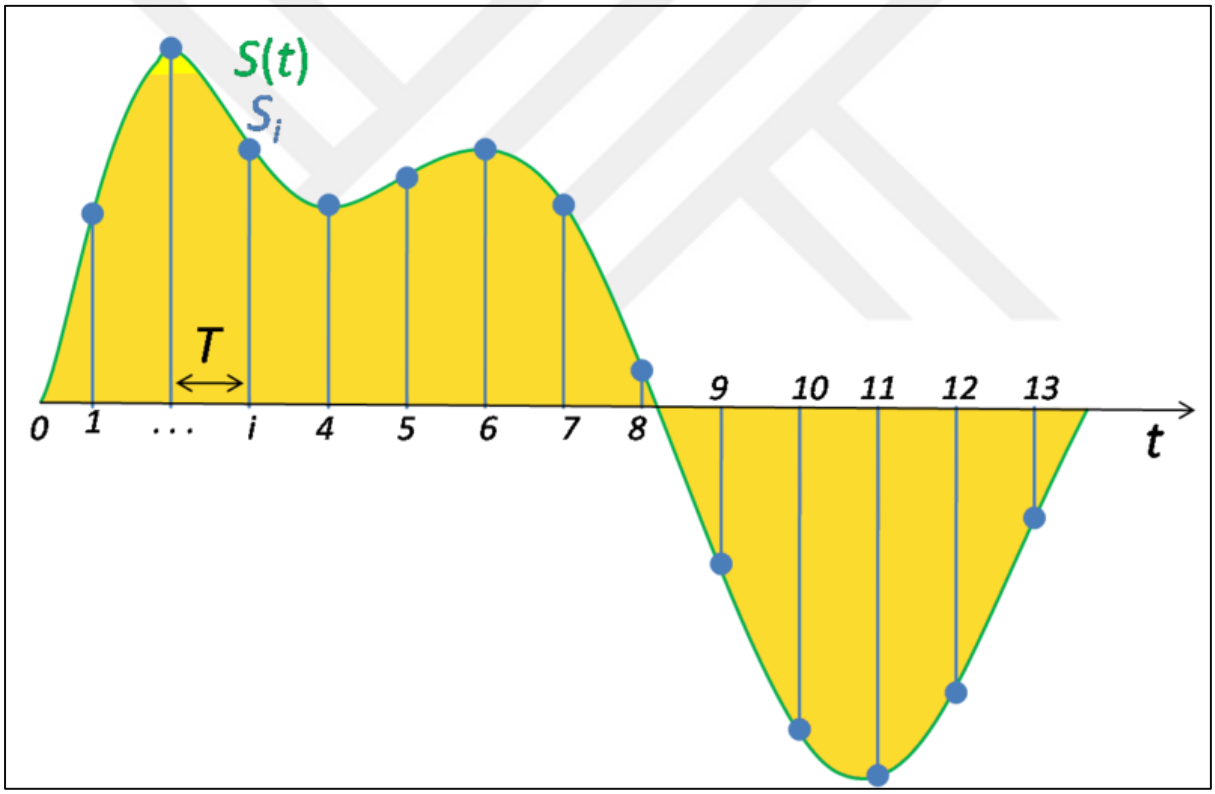

Figure 3-2: Sampling frequency vs. original signal frequency

Sampling as figure above depicts is splitting the signal's time domain into N number of slots with same length and located at a similar distance from each other along the time axis to help converting the continuous time signal into a discrete time signal [3]. The goal behind this procedure is to simplify the information detection of the said ‘signal’ and limit the noisy segments of a signal.

The sampling is conducted for functions having time, space, and other differences and the same results were obtained out of at least two dimensions. Some functions change with time, so, s(t) is a continuous function sample. Then the sampling is implemented using the continuous function value after "T" time period in seconds that serves as a sampling interval/period. This sample function takes into account a sequence such as s(nT) while n assumes integer values. The sampling frequency/rate fs shows is the average of the total samples in a second; therefore, fs = 1/T. In order to reconstruct continuous function through samples, interpolation algorithms are used. The Whittaker–Shannon interpolation formula is a mathematical alternative to the ideal low-pass filter having input of Dirac delta function, which is modulated/multiplied with sample values. In a situation, in which, a time interval of adjacent samples is constant (T) [2], the delta function sequence is termed as Dirac comb. In mathematical terms, a modulated Dirac comb equals the product of s(t) with the comb function.

The genuine mathematical abstraction is also called as "impulse sampling." A majority of sample signals do not store or reconstruct because theoretical reconstruction fidelity is considered as a customary process that reflects sampling effectiveness. The fidelity decreases because s(t) has components of frequency having less than 2 samples periodicity or when the cycles-to-samples becomes higher. Nyquist frequency means ½ number of cycles/sample × fs samples/sec = fs/2 Hz. Here, s(t) is generally a low-pass filter output, also termed as anti-aliasing filter. If the anti-aliasing filter is not used, the frequency greater than Nyquist frequency affects samples, which results in misinterpretation during the interpolation procedure.

3.8 Signal Energy and Power

They characterize signal but actually, they aren't identify power and energy. By definition, the signal power and energy means a signal x(t), which includes signals assuming complex values [4].

Signal energy of a signal x(t):

$$E = \int_{-\infty}^{\infty} |x(t)|^2 \, dt \tag{3.1}$$

Signal power of a signal x(t):

$$P = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \, dt \qquad (3.2)$$

If 0 <E< ∞, the signal x(t) will be the energy signal; however, in case of certain signals, the mentioned condition is not met. For such signals, it is considered as the power. The power of the energy signal is 0 (0 <P< ∞) while P = 0, and that energy is infinite (E = ∞). Some signals cannot be classified as either power or energy signals.

3.9 Noise and disturbances

Noise means the distortion/disturbance that exists in every electronic/electric/communication system. The property of noise is normally represented by N, where N is measure of power in a particular noise signal. Different noise/random signals are produced, which depends on the medium/internal environments of electronic devices. Noise has following categories:

- White noise can be Johnson noise, thermal noise, or Nyquist noise that takes place through charge carriers' thermal agitation. The noise can be physically and statistically derived through fluctuation-dissipation theorem.

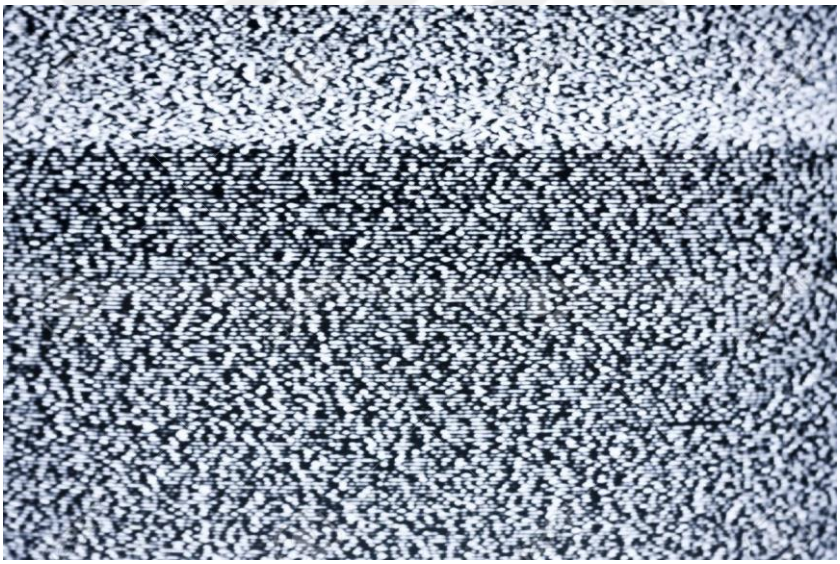

Figure 3-3: Thermal noise

- Gaussian noise: The assumed quantity that can produce when a random variable having Gaussian mean and variance apply to a transmitted signal. The reason behind its production is the channel condition, from where; a signal travels from the source to the destination. This can be quantitatively described based on the variance and standard deviation of a noisy signal [6].

$$Pn = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(n(t)-\mu)^2}{2\sigma^2}} \qquad (3.3)$$

Where, n(t) represents the grey level, μ the mean value and the σ standard deviation.

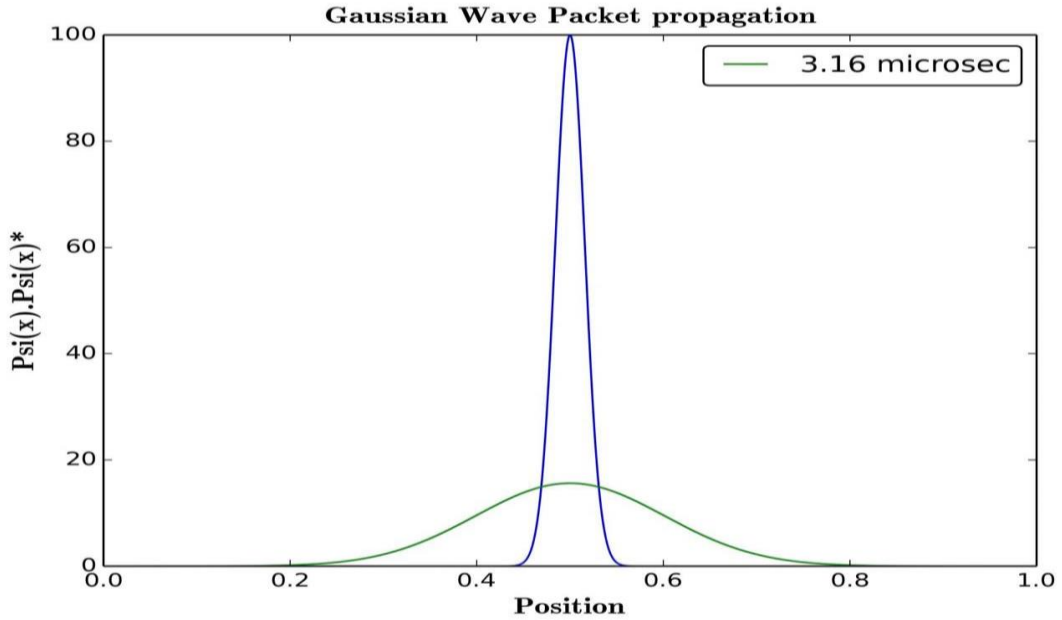

Figure 3-4: Gaussian noise waveform

- Fading and jitter: These terms apply in digital communications when the transmitter information is sent in packets. The situation, in which, the information bits are sequentially changed is the jitter phenomenon. On the other hand, the multi-path transmitted signal occurs because the environment surrounding the receiver or transmitter results in multiple copies of a signal has different signal-to-noise ratio; this occurrence is known as fading [6].

3.10 Frequency representation of signals

As demonstrated above, a signal's frequency representation is quite important to analyze that signal. The common techniques for frequency domain conversion is Fast Fourier Transformation (FFT) and wavelet transform, through which, a signal's time domain lasts for 10 seconds, and it converts into amplitude. Normally, frequencies of different components are linked with that any signal such as noise frequency can be waived-off by establishing pre-designed filtering techniques.

3.10.1 Fourier Transform

The Fourier transform decomposes a function of time (a signal) and changes into frequencies it is made up of, the way a musical chord is expressible in frequencies/pitches/constituent notes. It has complex values in terms of frequency, and if their absolute values are taken, they represent in the original function. The Fourier transform is also termed as original signal's

frequency domain representation [6]. The term Fourier transformation applies to both frequency domain representation and its mathematical operation. The Fourier transform isn't just limited to time functions but for having a unified language as well, and the original function's domain is generally considered as a time domain. Out of several important and practical functions, a function is available that can reverse it, which is called as inverse Fourier transformation/ Fourier synthesis, which recovers the genuine time function by combining contributions of different frequencies. For a continuous signal x(t), Fourier transform can be evaluated from the following:

$$X(f) = \frac{1}{T} * \int_{-\infty}^{\infty} x(t) * e^{(-jwt)}.\ dt \qquad (3.4)$$

The final signal depends on frequency (f) as shown above and illustrated in figure 5, while the genuine time domain signal is continuous over the time period (minus infinity to infinity), which is a general expression of Fourier transform whereas the actual signal could be limited to pre-defined time period i.e. -10<t<+10. The point of interest is the discrete signal Fourier transform where a signal's hash sampling rate has double frequency (F sampling = 2* F). The DFT has a link with discrete Fourier transformation, and it can be evaluated using the expression given below:

$$X[n] = \sum_{-\infty}^{\infty} x[n] * e^{-jnwt} \qquad (3.5)$$

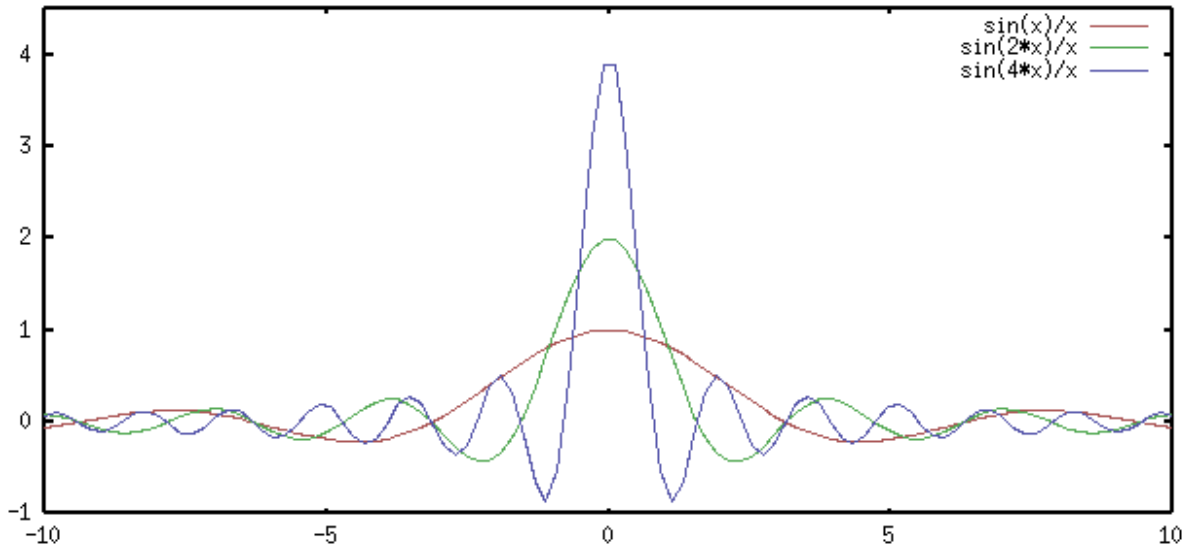

Figure 3-5: Fourier Transform depicted for signal where the right side is a real part, and the left side is the imaginary part of the final signal in the frequency domain.

4 System Implementation

4.1 Introduction

In order to assure accurate speech recognition, many sub systems should be used. Speech recording and analog-digital conversion are important stages in SRS. It is essential to have understanding about the characteristics of a speech signal in addition to the methods for detecting useful information from the said test signal, and those methods remove unused parts or noise interferences.

4.2 Speech recording

A signal is shifted to Matlab through voice recording with a microphone. Generally, high-quality recording is required for appropriately following many norms needed for the required quality. The channels of recording may be single or double (stereo/mono), and the voice signal frequency is normally 44,100 Hz and 16-bit resolution is needed for recording voice/sound signal. Figure 4.1 depicts the data input to Matlab.

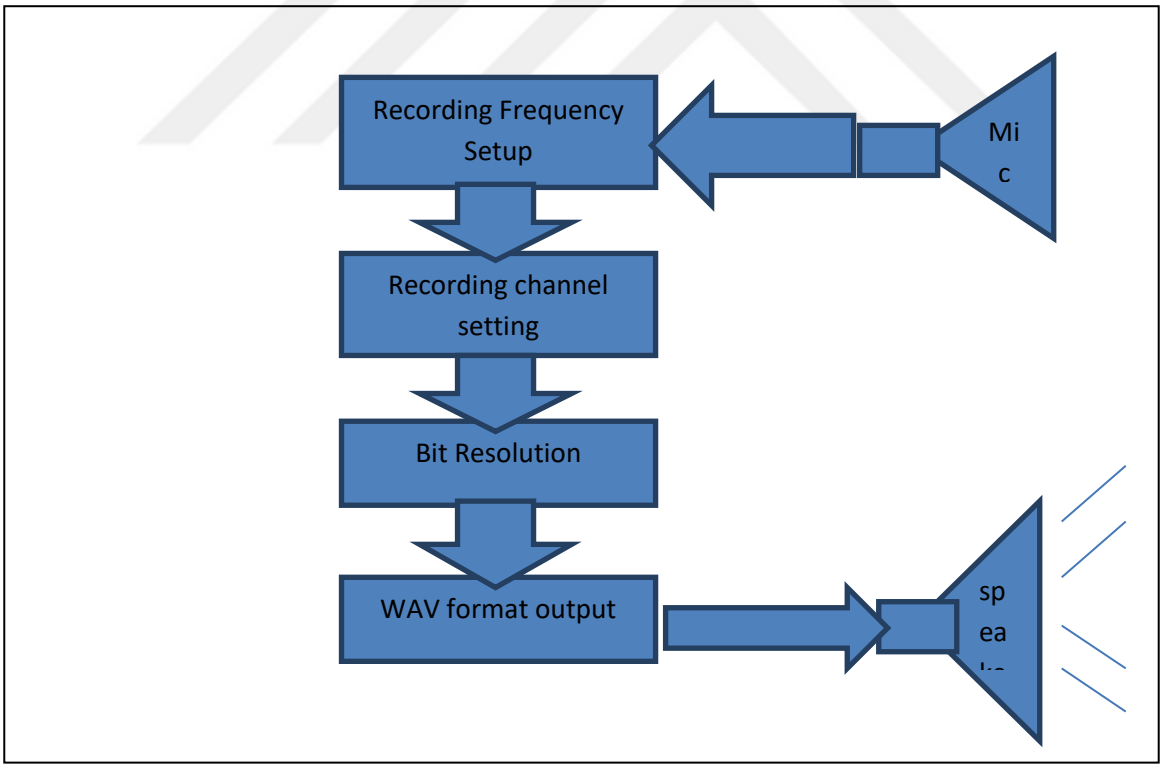

Figure 4-1: Recording algorithm in Matlab.

4.3 Voice signal characteristic

The basics of speech signals should be understood before analyzing such signals; this approach is involving the voice signal construction including the voice information detection from speech signal and noise with unvoiced segment removal. Pitch frequency is the common constant component in a voice signal because that is considered as speaker identity.

4.3.1 Voice and silence determination

Generally, speech sentence signals have two aspects; the first communicates the spoken information while the second has silence/noise that takes place between the utterances conveying no verbal information. The verbal/informative part has been further categorized in two categories:

1. The voiced speech;
2. The unvoiced speech

The voiced speech mainly consists of vowel sounds, which are produced when the air is forced out of the glottis with appropriate vocal cord adjustment. It opens and closes vocal cords that results in periodic air pulses, which excite the whole vocal tract. Psychoacoustic trials clearly show that the mentioned part has more speech information; therefore, it can characterize a speaker. Unvoiced speech comes out when the air is forced but there is obstruction or constriction in a specific place within the vocal tract (closer to the mouth-end), which creates turbulence. It is significant to distinguish among all three of them in order to conduct an appropriate speech signal analysis.

4.3.2 Feature determination

- Zero Crossing Rate: It is the rate of speech signal's crossing zero, and it provides data about its creation source. It has been shown in Figure 4.2 that the unvoiced speech possesses more ZCR as compared to a voiced speech. It happens as most of unvoiced speech energy has higher frequency as compared to the voiced speech, which is a reason behind more ZCR.
- Energy: The unvoiced spoken voices have considerably lower amplitude as compared to voiced segments. The voiced speech (Figure 7) has short-span energy values, which are considerably more as compared to the values obtained in case of unvoiced speech.

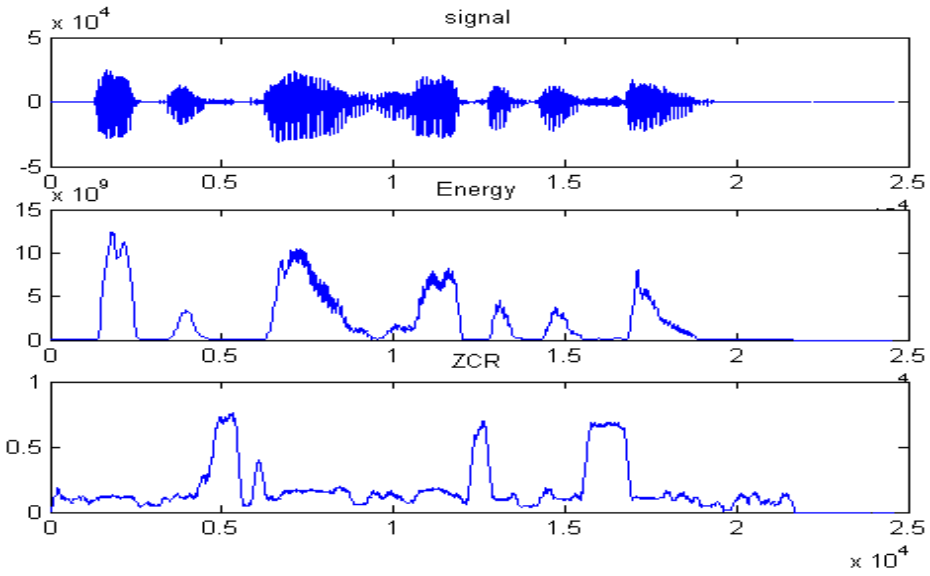

Figure 4-2: A speech signal (a) with its short-time energy (b) and zero crossing rate

In voiced speech (Figure 4.2), the short-time energy values are much higher than in unvoiced speech, which has a higher zero crossing rate.

- Cross-correlation: It is calculated for comparison between a couple of consecutive pitch cycles. The value of cross-correlation between pitch cycles is high (almost 1) in case of voiced as compared to the unvoiced speech.
- Pitch detection: A voiced speech signal is generally semi-periodic. Its fundamental time period is termed as pitch period. The mean/average of the pitch frequency, time, gains, and fluctuations vary from person to person. For speech synthesis and its signal analysis, understanding and identifying pitch is very significant. A well-known method to detect pitch is based on a reality that consecutively-occurring pitch cycles have bigger cross-correlation as compared to any two consecutive and equally lengthy speech fractions; however, their pitch cycle time is different.

The pitch detector's algorithm can be given by the following two equations:

$$\rho_\tau = \frac{(\mathrm{x}, \mathrm{y})}{\| \mathrm{x} \|. \| \mathrm{y} \|} \; ; \; \| x \| = ((x, y))^{\frac{1}{2}} \tag{4.1}$$

Figure 4.3 shows a vocal phoneme while pitch marks are shown in red.

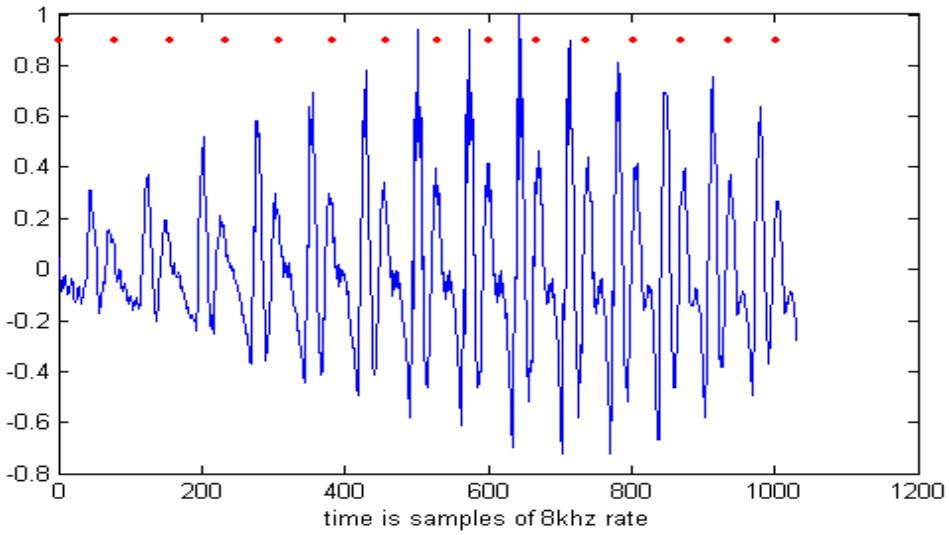

Figure 4-3: A phoneme with its pitch cycle marks (in red).

4.4 Algorithm structure

It is shown in the figure above that a speaker may enter his voice snap into Matlab using the microphone and later, the system converts the analog signal into digital format. In this case, sampling takes place as a first process occurring in the system before converting the signal into digital form; hence, the analog speech signal will appear in slots with respect to frequency for applying digital signal processing.

Figure 4.4 is depicting the entire speech recognition process, in which, the signal divides into time slots of similar width as shown in Figure 4.4.

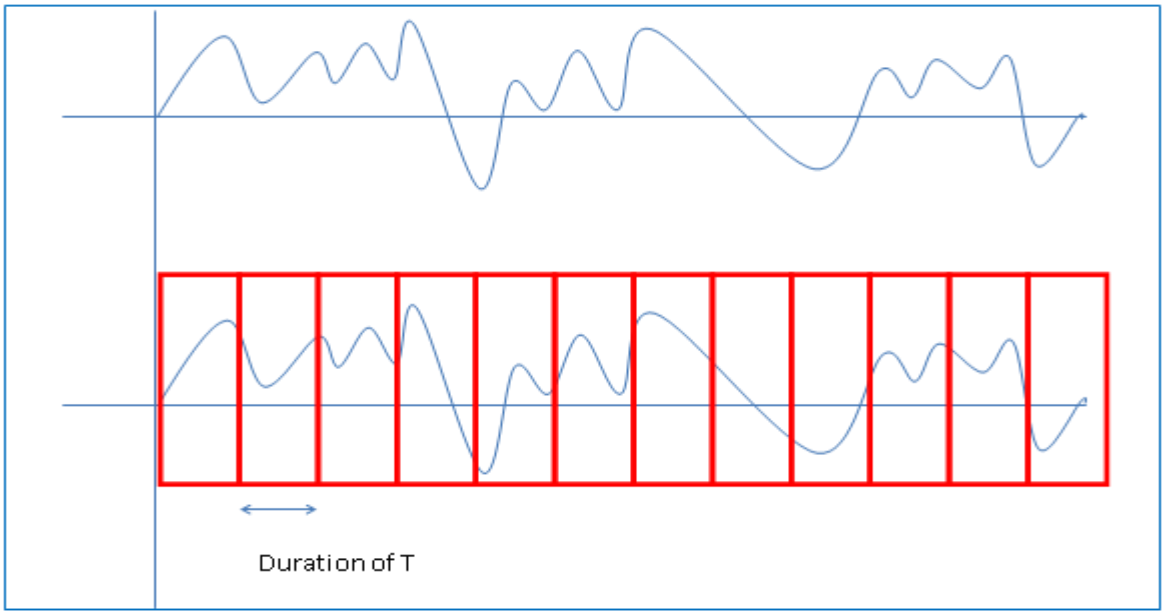

Figure 4-4: Voice signal slotting

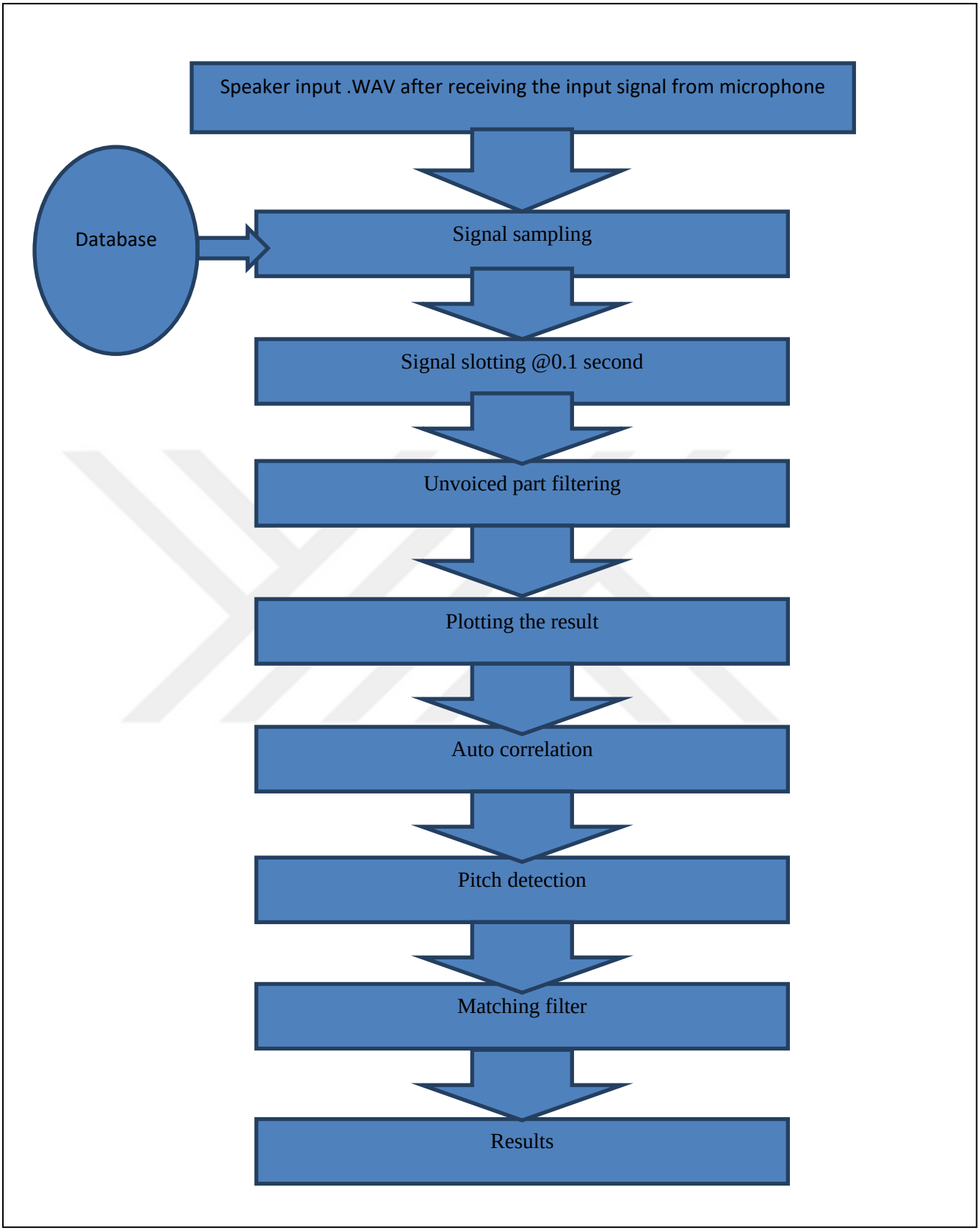

Figure 4-5: SRS algorithm.

The slots were used for measuring signal power/amplitude, which were then measured for each slot, and the minimum level of signal was identified. This level was obtained as a reference input for the next level of signal processing. The same is considered as silence, and it is removed from the signal by ending this process, which is the only remaining voiced speech segment. Figure 4.6 depicts the silence removing process.

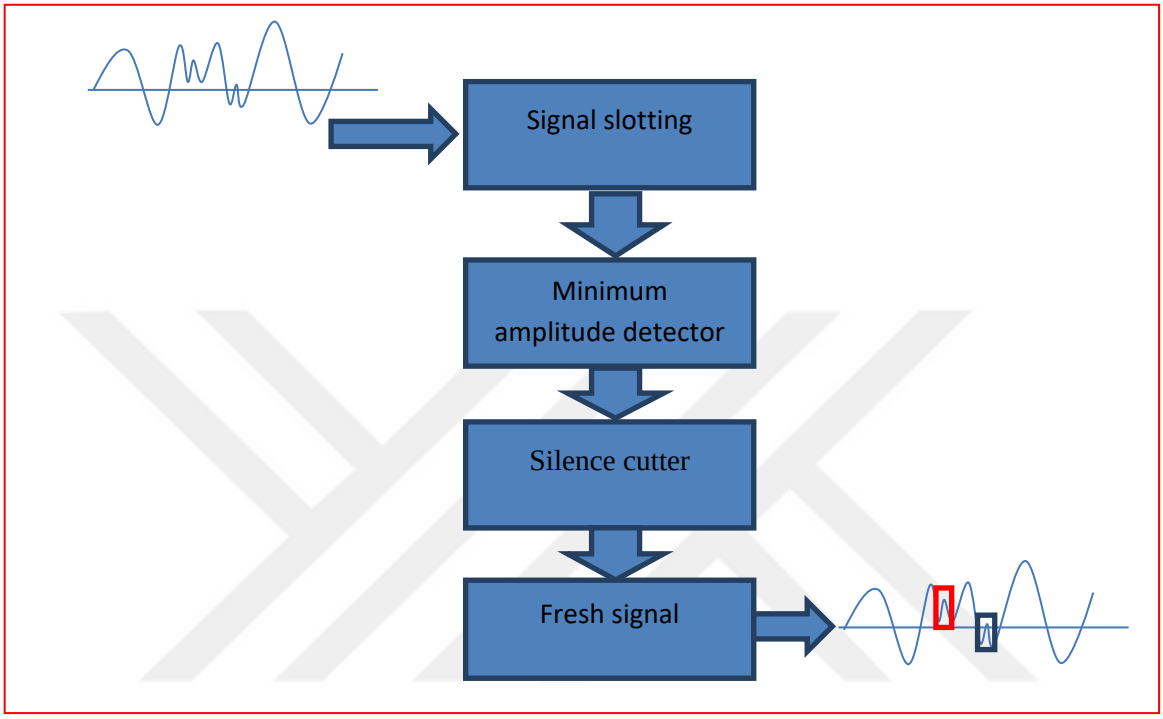

Figure 4-6: Silence removing algorithm.

As depicted in the figure, a signal, which was entered in the program, is monitored for noise and silence cancelation, so, according to the figure, the lowest amplitude is detected in each slot after comparing all signal points. Then, the minimum dot/point was detected, and it was deemed as silence, which was cancelled from all the slots by referring to the detected level (minimum amplitude among the signal samples).

Referring the resulted signal in the next procedure, the signal was correlated with a similar copy for detecting the maximum location of power, so that the signal power is registered in a different location with respect to its auto-correlation. Later, the pitch frequency was calculated, which is the minimum frequency with better signal power, and this frequency is unique in every voice signal, and it is used to identify a speaker.

The above process applies to both the signal in the database and the signal recorded during the test; so, both the results are available for comparing by the matching filter. The filter

configurations are set as per the database signal statistics, and the test single is then entered as second input to the matching filter. The pitch frequencies from both the inputs are compared, so when a minimum pitch frequency from the database is similar to the same of a microphone (the test signal), only then, the program will display “access is allowed;” otherwise, if the pitch frequencies aren't equal, the program will show “access denied.” Figure 4.7 is depicting the autocorrelation and the matching process.

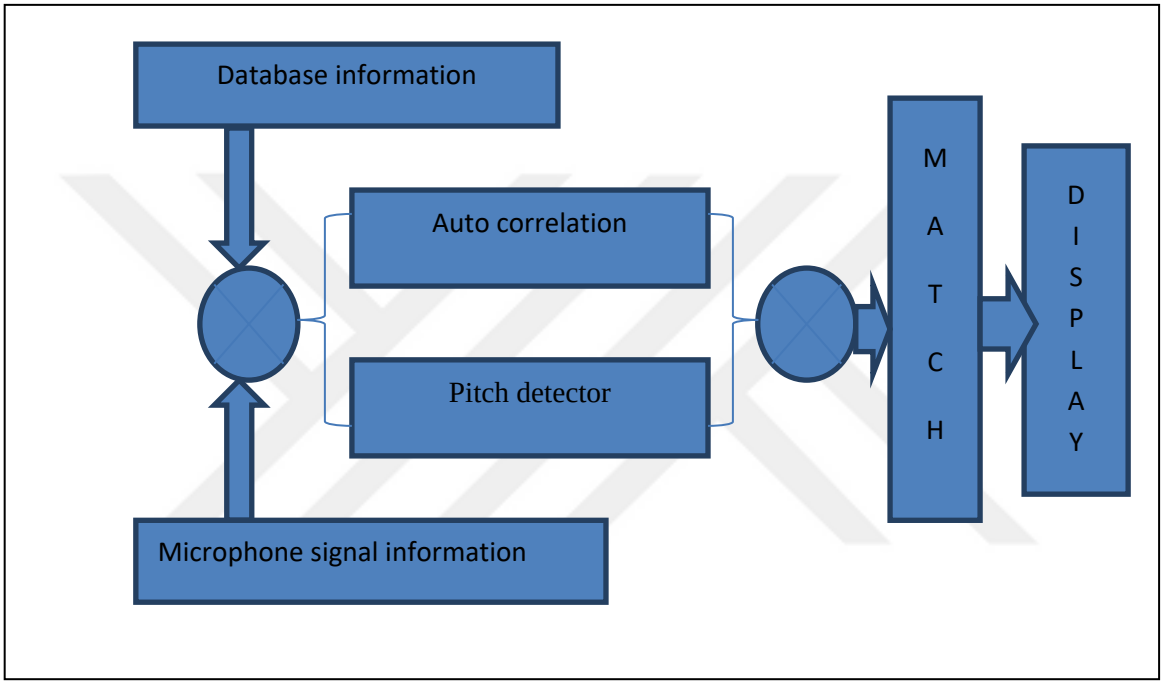

Figure 4-7: Auto correlation and matching filter of voice signal.

4.5 System strength/comparative concept

By designing this paradigm, a speaker will have all the features to identify himself and get the proper feed-back but that depends on the authentication matching process. This is not only looking after classical recognition of vice signal, efforts are made to approach the signal processing concepts according to the real-world requirements. As per market demand, such program may be utilized in security systems, which exist in many applications including web user identifiers (online identification process), and it can be offline in the local servers of security systems. The following points summarize the project outcomes as compared to the literature survey.

- By observing the literature survey that was accomplished to understand relevant approaches of similar nature, this project has multiple approaches pertaining to the digital signal processing unlike the most previous studies.
- In voice recording, two channels with deeper voices were recorded to get the best audio from the input device.
- Voice pre-processor was part of our system to eliminate the unwanted interferences; that was done by removing the noise and unvoiced components at signal boundaries. It was stored in the database as a reference signal so that only the required voice information will be available for the recognition scheme.
- The features were extracted in various stages for robust voice information recognition.
- Frequency property is used as underlying approach, and this idea was taken from the theory of voice, which states that every person has a unique voice pitch so there are no two persons in the world with similar voice. It should be noted that the pitch frequency is the main source of identification because it belongs to only a single person.
- Another speech recognition property is auto correlation, which assures that the input signal is typically similar to the reference signal, for example, if the user recoding was about three alphabets (A B C), and the input was like (A B F) so, the result will appear as: "Access denied due to unmatched information." It should be noted that the autocorrelation is used to recognize the voice information (speech content).
- In our project, two important security attributes are ensured for robust speech recognition. The first attribute is that the same person should access the system, which means even if other user had the password, he/she will not have any access. Secondly, the same person should provide the same speech information (same password). Therefore, it is a double-checking security system.

5 Results and Discussion

5.1 Outline

After setting up the system and describing the entire process with principle keywords, now we are going to illustrate the results of the process and discuss the outcomes. The project was executed using Matlab while the voice recognition was done at minimum pitch frequency obtained from the autocorrelation and advanced signal processing. The following sections demonstrate the recording algorithms, which were used for advanced processes. First, the speech signal was recorded, then the noise was removed, and filtering process was accomplished before the voice recognition process. In the end, all the signals were plotted during every stage involved in the Voice Recognition Scheme (VRS).

5.2 Validation

A speaker may start entering his information using a microphone and physical sound card linked with the computer. After doing so, the system saves the audio in the Matlab directory. The database is involving voice signal representation as samples while each sample carries the voice information according to the time domain. In other words, the sound converts into numerical data by breaking the mentioned speech signal in dots/points of time.

A speaker may record his input for database, and the program allows recording for five seconds. The reason behind five second recording is to simplify the computational complexity as the capacity of computer RAM and allowing the processor to process it very quickly.

The first input signal was entered into Matlab as a data file having two columns as well as multiple rows, the first column acts as a placeholder for time points/sample time. The second column acts as a placeholder for voice data within that particular time. Figure 5.1 depicts the signal of first entry. The signal is highlighted in red color, and it acquires the shape of a database signal. The signal amplitude was taken along the x-axis while y-axis shows the recording time.

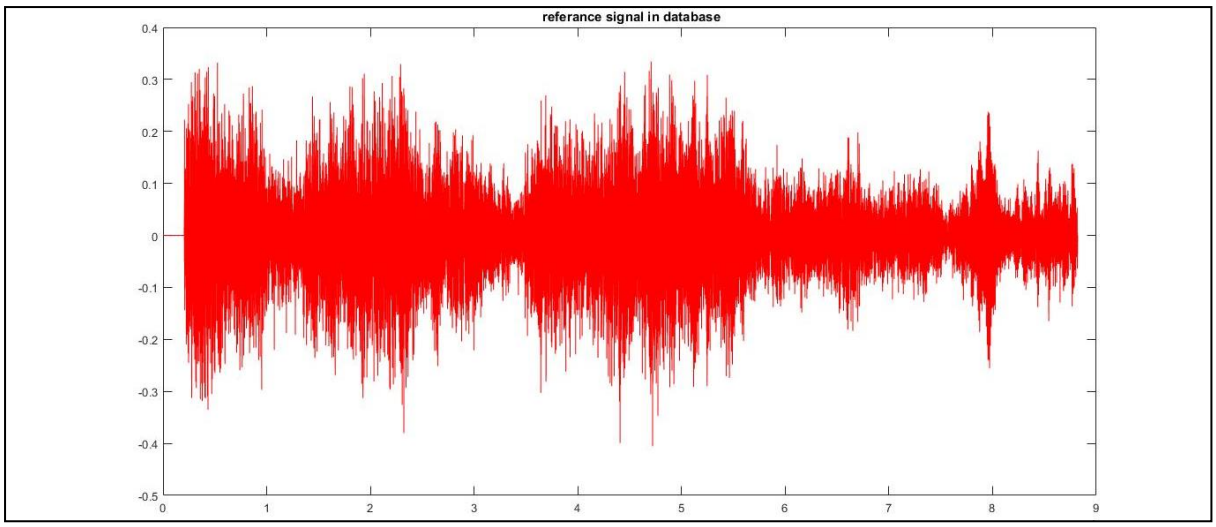

Figure 5-1: Recorder signal as database input

5.3 Signal representation

A speech signal is recorded at voice hearing frequency 44,100Hz stereo quality with double channel and 32bit resolution. Later, the speech signal was processed with the silence removing program and coded for producing smaller signal size to simplify the computational complexity and increase the response time. Normally, this kind of signal processing takes a long time due to dependence on recorded signal duration, size, and the machine properties including the random-access memory (RAM) and the processor speed. Hence, by removing the silent segments of that signal, only the important information will remain available for the coming process, and the rest of the signal parts will be neglected. By the end of this stage, the signal shown in Figure 5.2 will be obtained.

With same time limits and same amplitude, the signal has only the voiced information. The starting part of this signal is illustrated in Figure 5.2, which is showing no information at zero amplitude/signal level as it implies that every initial slot consists of silence. Such process is handled by applying a concept called as windowing such as all speech signals are broken into similar size of windows where DSP can easily work.

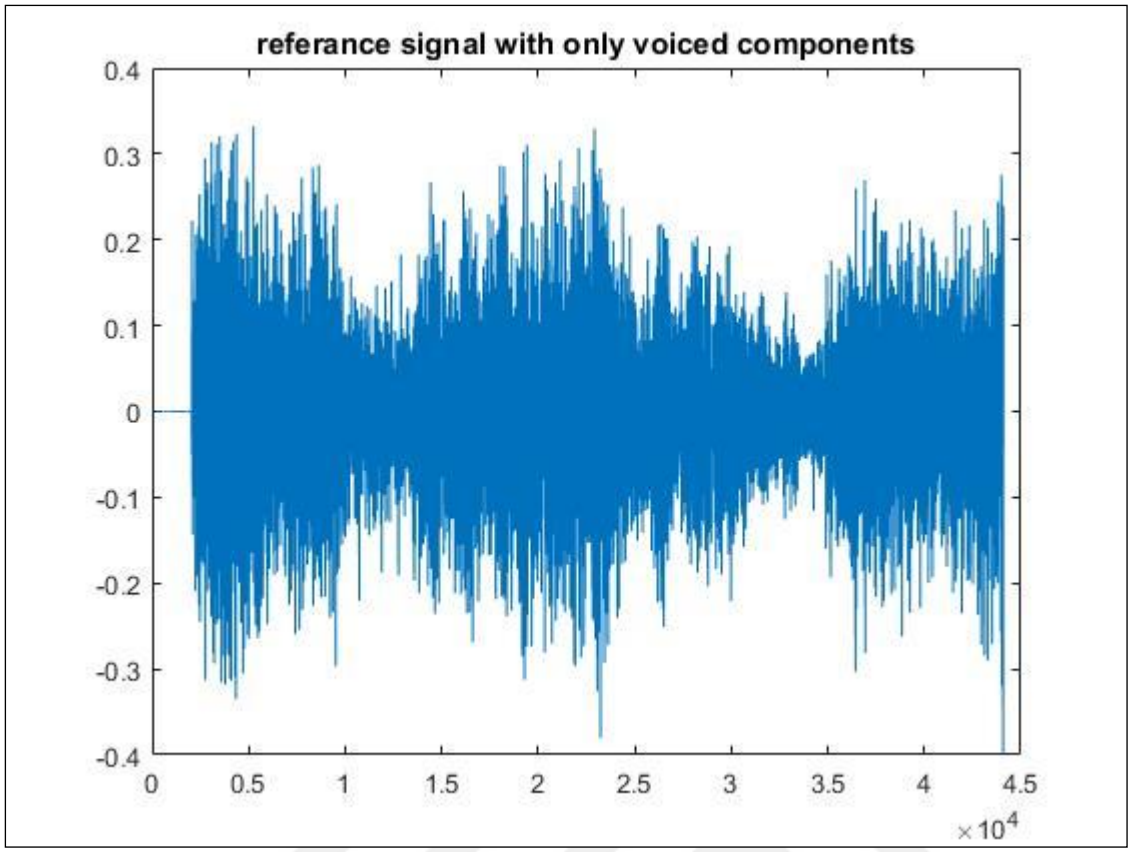

Figure 5-2: Signal after recording with no silent parts

After that, the signal is entered in a frequency domain using Fast Fourier transform (FFT), so that the information of speech signal can be identified according to frequency. Now a signal is presented as power/amplitude that distributes among the frequencies as Figure 5.3 depicts.

Double side frequency signal is plotted, and the plotted values are obtained by FFT. The signal's real part has been plotted while the complex part is neglected. Furthermore, single-sided FFT is possible to plot as a normal output shown in Figure 5.3, which involves symmetrical two components with same information; however, the samples' number is halved when it is divided by two and only that is plotted. The same analysis may apply to single-sided FFT results, which reveal exactly the similar outcomes.

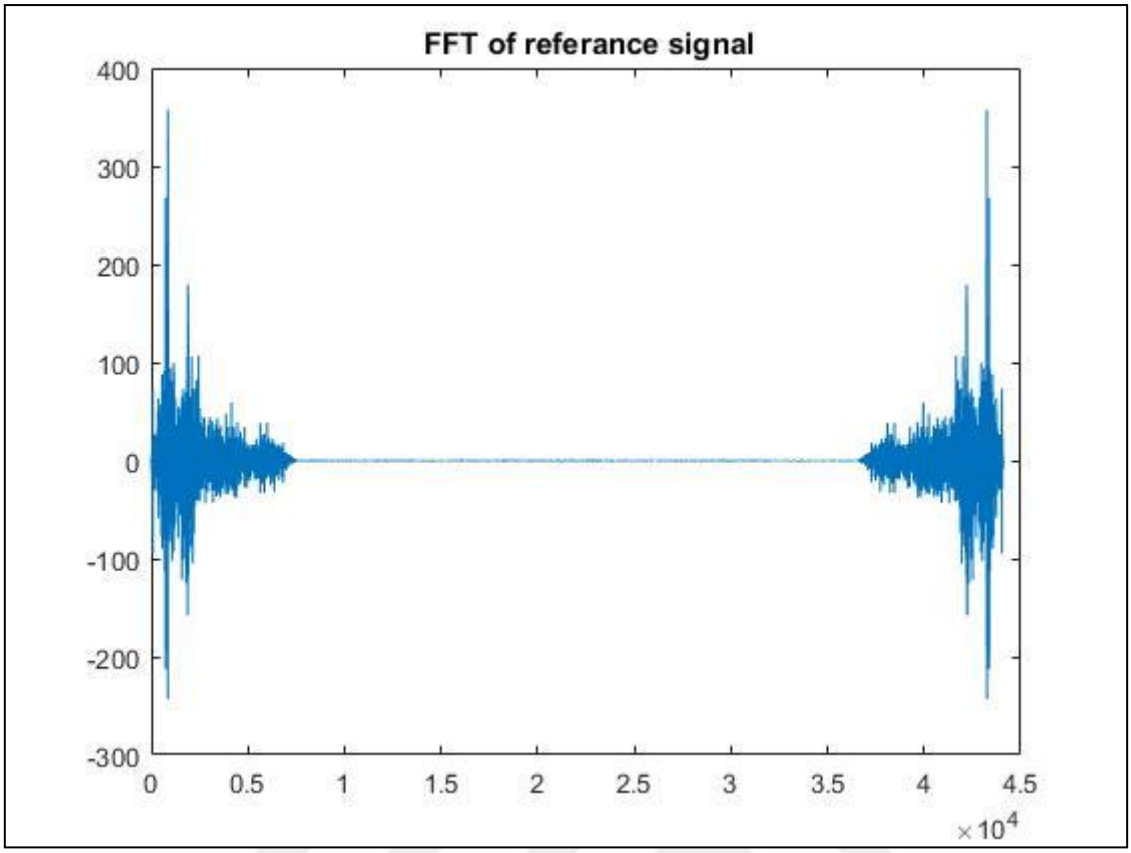

Figure 5-3: Two-sided Fast Fourier Transform

5.4 External test input

After preparing the system with first input that is stored in Matlab directory, the second input is entered as external data for the testing purpose will be directed to the Matlab workspace by letting the speaker who requests access to the system to record a new signal. This signal acts as security code or password of voice, while a speaker must be the same person, who entered his voice in the algorithm earlier. Moreover, he/she has to repeat the sentence that he/she had already entered in the system. After understanding these criteria, a user/speaker may now give the second input to the program by saying the same sentence that already exists in the database. Figure 5.4 is depicting the second/external/test input.

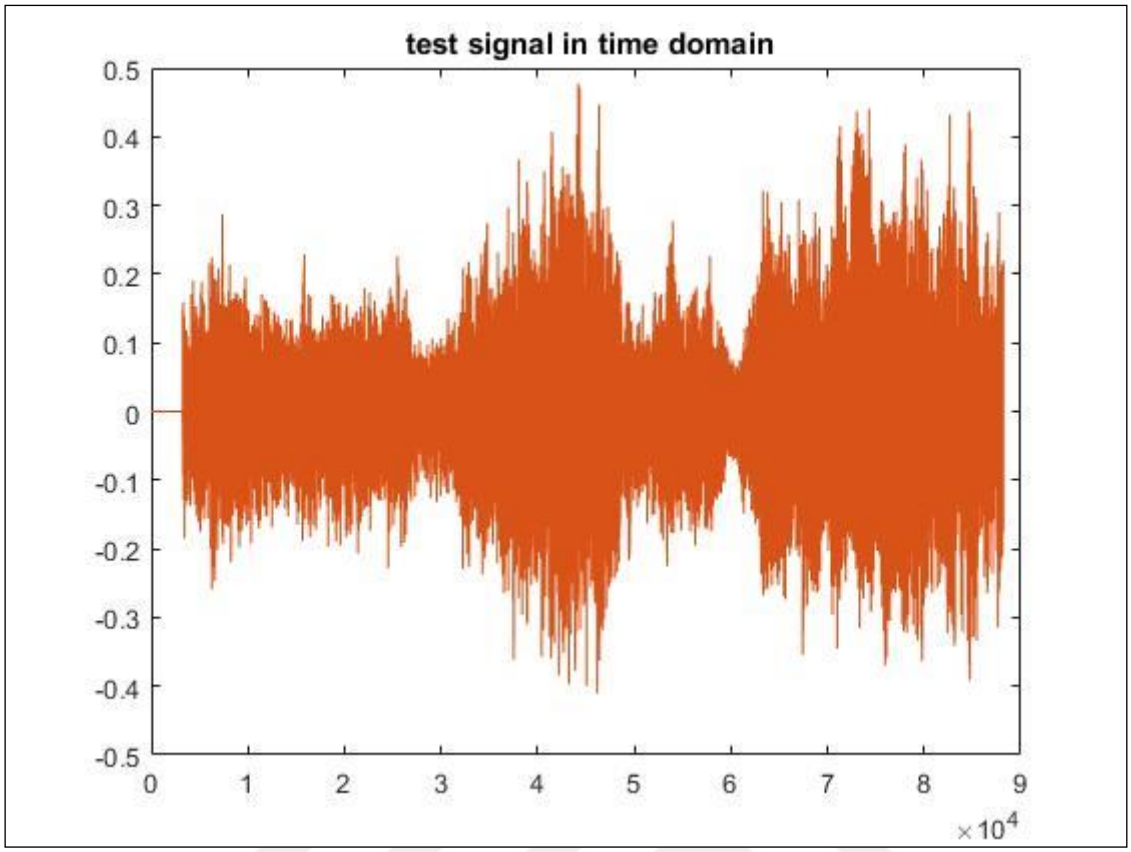

Figure 5-4: The external second input of speech.

5.5 Recognition procedure

Both the data lines show that the input in database and input of test undergo specific types of digital signal processing. The DSP technologies are applied for extracting a signal's pitch frequencies and the time of pitch occurrence. Auto correlation is also used to monitor the signal similarity, and the location of maximum output power in a signal, which is stored for the further process. Pitch frequencies from the reference input and the test input are matched in the last stage in a voice recognition system (VRS). During the matching process, the system finds out whether the pitch frequencies are identical or not, and accordingly, it reveals the message in the MATLAB command window. Figure 5.5 is depicting the auto correlation pertaining to the first input signal.

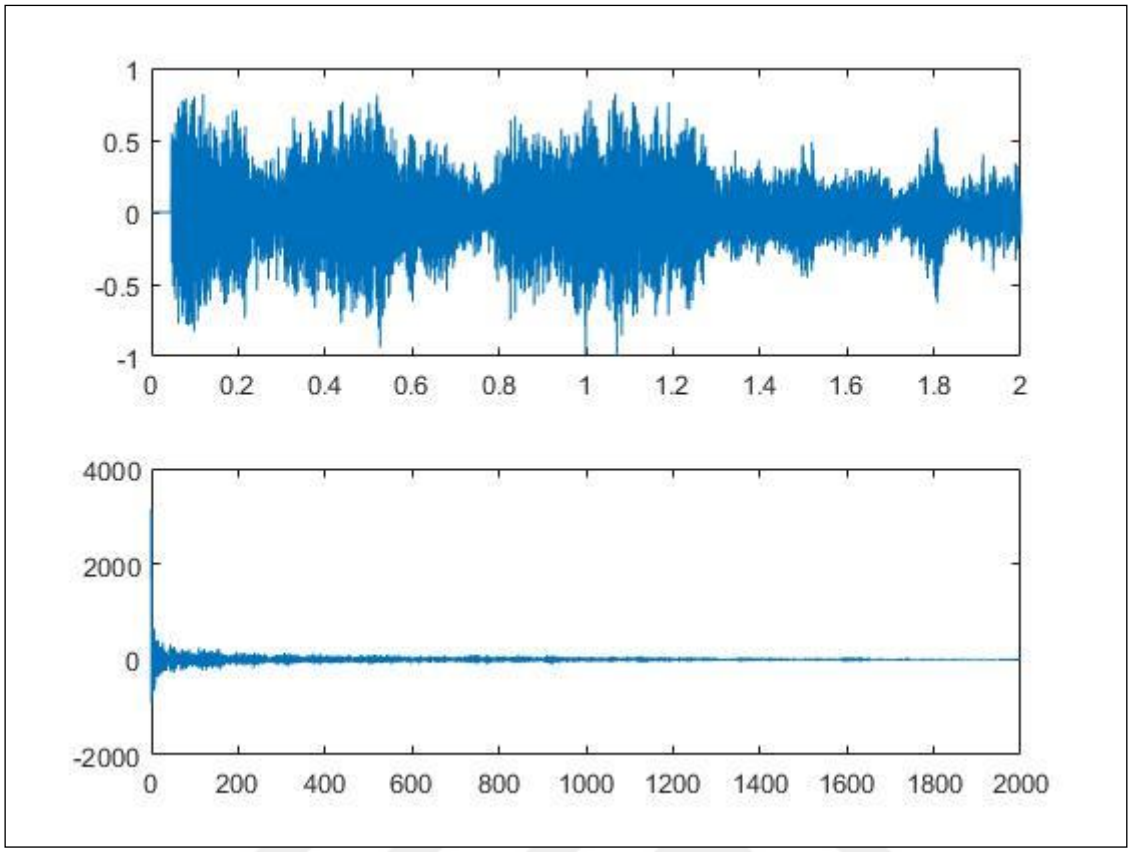

Figure 5-5: The first reference input with auto-correlation results

6 Conclusion

As far as the early signal processing stage is concerned, many theories have been created to analyze the analog signals such as signals of heat and light sensors. These approaches are used for analog-digital conversion (ADC) for making analog signals appropriate for digital environments specifically because the digital world is providing great facilities for quick and less expensive solutions. A speech is recorded in the form of analog signal, which emerges out of vocal cord vibrations. Nowadays, speech is used for the purpose of security and personal identity. It is a unique attribute in humans that every person has different speech frequency and because of that, different people have different voices.

So far, a lot has been accomplished in terms of making and using a speech recognition system having capability to distinguish the voices of different of people; however, the program made in Matlab environment provides voice recognition solution irrespective of language.

Advanced digital processing of signals was used in this project to achieve best recognition results. The utilized system showed attributes such as less time complexity and quicker response to the user. For this kind of experiment, the data is always entered in Matlab into two parts, the reference input and the external input. The database is loaded with reference speech signal when the DSP is applied to extract its pitch frequency. At that time, the program demands another input for recognition. Normally, such system has a great contribution to the data or system security. The user should be the same because the voice signal is required to be the same so that the condition “same person and same password” is validated. In this way, nobody other than the intended user will be able to access the system and feed the data in the database.

Such work should be done with applications, which require high level of security allowing only the authorized person to access the data. For future research, the mentioned system should be tested for larger databases allowing large number of users while all of them will be able to enter a password and access the system.

7 References

[1]. Aseem Saxena, Amit Kumar Sinha, Shashank Chakrawarti, Surabhi Charu," Speech Recognition Using Matlab", International Journal of Advances in Computer Science and Cloud Computing, ISSN: 2321-4058 Volume- 1, Issue- 2, Nov-2013.

[2]. Atheer Tahseen Hussein, "Analysis of Voice Recognition Algorithms using MATLAB", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS080082 www.ijert.org (This work is licensed under a Creative Commons Attribution 4.0 International License.) Vol. 4 Issue 08, August-2015.

[3]. Hardik Chhatbar, Janak Trivedi, Rahul Chauhan, Darshan Bhatt, "SECURE SPEECH CONTROLLED ROBOT USING MATLAB AND ARDUINO", International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, Issue 04, [April – 2015] ISSN(Online):2349–9745; ISSN (Print):2393-8161.

[4]. Lavin Jalan, Rahul Masrani, Roshan Jadhav, Tejaswini Palav, "SPEECH RECOGNITION BASED LEARNING SYSTEM", International Journal of Engineering Trends and Technology- Volume4Issue2- 2013.

[5]. Er. Abhishek Thakur, Assistant Prof. Neeru Singla, "Design of Matlab-Based Automatic Speaker Recognition and Control System", Er. Abhishek Thakur* et al. / (IJAEST) INTERNATIONAL JOURNAL OF ADVANCED ENGINEERING SCIENCES AND TECHNOLOGIES Vol No. 8, Issue No. 1, 100 – 106.

[6]. P. Yasotha1 N. Rajasekaran2 M. Nathiya3, "Automatic Spectral Analysis for Speech to Text Conversion with Embedded Execution", IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 09, 2014 | ISSN (online): 2321-0613.

[7]. Siddhant C. Joshi1, Dr. A.N. Cheeran2, "MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 7, July 2014.

[8]. Shiv Kumar, Member, IACSIT, IAENG Aditya Shastri and R.K. Singh, "An Approach for Automatic Voice Signal Detection (AVSD) Using Matlab", International Journal of Computer Theory and Engineering, Vol. 3, No. 2, April 2011 ISSN: 1793-8201.

[9]. Neha Sharma, Shipra Sardana, “Designing a Real Time Speech Recognition System using MATLAB”, International Journal of Computer Applications (0975 – 8887) National Conference on Latest Initiatives& Innovations in Communication and Electronics (IICE 2016).

[10]. Angelo A. Beltran Jr.1, Ericson D. Dimaunahan2, Donde A. Deveras, “Speaker Dependent Voice Recognition Using Discrete Wavelet Transform”, International Journal of Scientific Engineering and Technology ISSN: 2277 – 1581 Volume No. 4, Issue No. 8, pp: 443 – 446. 01 Issue 7, July 2014.

[11]. Luqman Gbadamosi, “VOICE RECOGNITION SYSTEM USING TEMPLATE MATCHING”, International Journal of Research in Computer Science eISSN 2249-8265 Volume 3 Issue 5 (2013) pp. 13-17 www.ijorcs.org, A Unit of White Globe Publications doi: 10.7815/ijorcs. 35.2013.070.

[12]. V.Naveen Kumar1, Y Padma Sai2, C Om Prakash, “Design and Implementation of Silent Pause Stuttered Speech Recognition System”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 3, March 2015.

[13]. Abhishek Thakur1, Rajesh Kumar2, Amandeep Bath3, Jitender Sharma4, “Automatic Control of Instruments Using Efficient Speech Recognition Algorithm”, IJEEE, Vol. 1, Spl. Issue 1 (March 2014) e-ISSN: 1694-2310 | p-ISSN: 1694-2426.

[14]. Amrutha S1, Aravind S2, Ansu Mathew3, Swathy Sugathan4, Rajasree R5, iyalakshmi,

“Voice Controlled Smart Home”, International Journal of Emerging Technology and dvanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015).

[15]. Abdelmajid Hassan Mansour, Gafar Zen Alabdeen Salh*2, Hozayfa Hayder Zeen Alabdeen, “Voice recognition Using back propagation algorithm in neural networks”, International Journal of Computer Trends and Technology (IJCTT) – volume 23 Number 3– May 2015.

[16]. Hairol Nizam Mohd. Shah*, Mohd. Zamzuri Ab Rashid, Mohd. Fairus Abdollah, Muhammad Nizam Kamarudin, Chow Kok Lin and Zalina Kamis, “Biometric Voice

Recognition in Security System", Indian Journal of Science and Technology, Vol 7(2), 104–112, February 2014.

[17]. Sankirna D. Joge1, Prof. A. S. Shirsat, "Different Language Recognition Model", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 7, July 2016.

[18]. Yogesh Watile1, Pratik Ghotkar 2, Bhushan Rohankar, "COMPUTER CONTROL WITHVOICE COMMAND USING MATLAB", INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING Vol. 3, Issue 6, June 2015.

[19]. Siddhant C. Joshi, Dr. A.N.Cheeran, "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 6, June 2014.

[20]. Manan Vyas, "A GAUSSIAN MIXTURE MODEL BASED SPEECH RECOGNITION

SYSTEM USING MATLAB", Signal & Image Processing: An International Journal (SIPIJ) Vol.4, No.4, August 2013.

[21]. Silvy Achankunju, Chiranjeevi Mondikathi, "Voice & Speech Based Security System Using MATLAB", International Journal of Emerging Trends in Electrical and Electronics (IJETEE – ISSN: 2320-9569) Vol. 11, Issue. 2, June 2015.

[22]. S. R. Suralkar1, Amol C. Wani2, Prabhakar V. Mhadse, "Speech Recognized Automation System Using Speaker Identification through Wireless Communication", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834, p-ISSN: 2278-8735. Volume 6, Issue 1 (May. - Jun. 2013), PP 11-18 www.iosrjournals.org.