Speaker Identification System Using LabVIEW

A thesis
submitted towards the partial fulfillment for
the requirements of the degree of

Master of Engineering
In
Electronic Instrumentation and Control Engineering

Submitted By

Manish Sharma
Roll No-80751012

Under the esteemed guidance of

Sunil Kumar Singla
Sr.Lecturer,EIED

DEPARTMENT OF ELECTRICAL AND INSTRUMENTATION ENGINEERING
THAPAR UNIVERSITY
PATIALA –147004

June - 2009

DEDICATED

TO

MY PARENTS

CERTIFICATE

This is to certify that my work presented in this thesis entitled ***"Speaker Identification System Using LabVIEW"*** *submitted in partial fulfillment of the requirement for the award of the degree of* ***Master of Engineering in Electronic Instrumentation and Control Engineering*** *at* ***Thapar University, Patiala****, is an original record under supervision and guidance of* ***Mr. Sunil Kumar Singla.***

The matter embodied in this report has not been submitted anywhere for the award of any degree.

Date: 10-7-2009

(Manish Sharma)

Roll No - 80751012

It is certified that the above statement made by the student is correct to the best of our knowledge and belief.

(Sunil Kumar Singla)
Sr. Lecturer, EIED
(Supervisor)
Thapar University, Patiala

Countersigned By:

S. Ghosh 14/7/09

(Dr. Smarajit Ghosh)
Professor & Head, EIED
Thapar University, Patiala

16/7

(Dr. R. K. Sharma)
Dean of Academic Affairs
Thapar University, Patiala

ACKNOWLEDGEMENT

The real spirit of achieving a goal is through the way of excellence and austerous discipline. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various personalities.

First of all, I render my gratitude to the ALMIGHTY who bestowed self-confidence, ability and strength in me to complete this work. Without his grace this would never come to be today's reality.

With deep sense of gratitude I express my sincere thanks to my esteemed and worthy Supervisor **Mr. Sunil Kumar Singla** in the Department of Electrical and Instrumentation Engineering for his valuable guidance in carrying out this work under his effective supervision, encouragement, enlightenment and cooperation. Most of the novel ideas and solutions found in this thesis are the result of our numerous stimulating discussions. His feedback and editorial comments were also invaluable for writing of this thesis.

I shall be failing in my duties if I do not express my deep sense of gratitude towards **Dr. Smarajit Ghosh,** Professor and Head of Electrical and Instrumentation Department who has been a constant source of inspiration for me throughout this work.

I am greatful to **Dr. R.K. Sharma**, Dean of Academic Affair for his constant encouragement that was of great importance in the completion of the thesis.

I extend my thanks to **Dr. K.K. Raina**, Deputy Director, **Dr. Abhijit Mukherjee**, Director Thapar University, for their valuable support that made me consistent performer.

I am also thankful to all the staff members of the Department for their full cooperation and help. My greatest thanks are to all who wished me success especially my parents, my friends whose support and care makes me stay on earth.

Place: Thapar University, Patiala

Manish Sharma
Roll No. 80751012

ABSTRACT

Now days, Biometrics is being used extensively for the purpose of security. Biometrics deals with identifying individuals with their physiological such as fingerprint DNA, ECG etc or behavioral traits i.e. rhythm, gait, voice etc. Voice is a most natural way of communication and non-intrusive as a biometric, Voice biometric has characteristic of acceptability, cost, easy to implement as no special equipment is required. Also Voice based biometric system can be easily combined with other biometric systems to enhance the reliability and security of the system.

In the present work a speaker identification system has been developed. The developed system uses the LabVIEW (Laboratory Virtual Instrument Engineering Workbench) 8.5 platform. Speaker Identification involves features extraction, preprocessing, pattern matching, decision-making. Silence removing of voice signal is key factor to improve the identification. In feature extraction stage, Mel frequency cepstrum coefficients (MFCC) have been calculated which provides a better measure of Speaker Identification than the other features. Speaker identification can be done by various methods but in this thesis vector quantization based recognition system using LabVIEW has been developed and tested. The developed system is user friendly and provides the results in real time. A database of 20 person having 5 samples per person including male and female has been created. The experiments conducted on the above database suggest that an accuracy of 90% has been achieved with the developed system.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

In the modern world, there is an ever-increasing need to authenticate and identify individuals automatically. Securing personal privacy and deterring identity theft are national priorities. Biometrics, the physical traits and behavioral characteristics that make each of us unique, are a natural choice for identity verification. It is an emerging technology that promises an effective solution to our security needs. It can accurately identify or verify individuals based upon their unique physical or behavioral characteristics. It is a key that can be customized to an individual's access needs opening doors for one while keeping others out. We can use a biometric to access our home, our account, or to invoke a customized setting for any secure area or application. In this chapter we explore the various types of biometric authentication techniques and their deployment potential. We take a look into the emerging technologies in this field and note their potential applications and future prospects.

1.1 Biometrics

The term biometrics is now widely known as "the science of measuring physical characteristics, to verify a person's identity and derived from the Greek words bio (life) and metric (to measure) which includes voice recognition, iris and face scans, and fingerprint recognition." This definition represents a recently created application used in the industrial-tech world.

Since biometrics is a system for measuring unique biological traits for the purpose of identification; it includes utilization of time clocks, the "easy way" to track and to report employees authentication to increase security, and the enhancement of access with the convenience of hand readers or finger prints; so, there is no further need for ID badges or time cards and the biometric system also eliminates the "buddy punching" of time cards or employees clocking each other in. When some recognition systems verify the identities of individuals by the size and shape of the hand, they do so without the fingerprints or palm prints being utilized.

Biometric characteristics can be divided in two main classes:

- **Physiological** are related to the shape of the body. The oldest traits that have been used for more than 100 years are fingerprints. Other examples are face recognition, hand geometry and iris recognition.
- **Behavioral** are related to the behavior of a person. The first characteristic to be used, still widely used today, is the signature. More modern approaches are the study of keystroke dynamics and of voice.

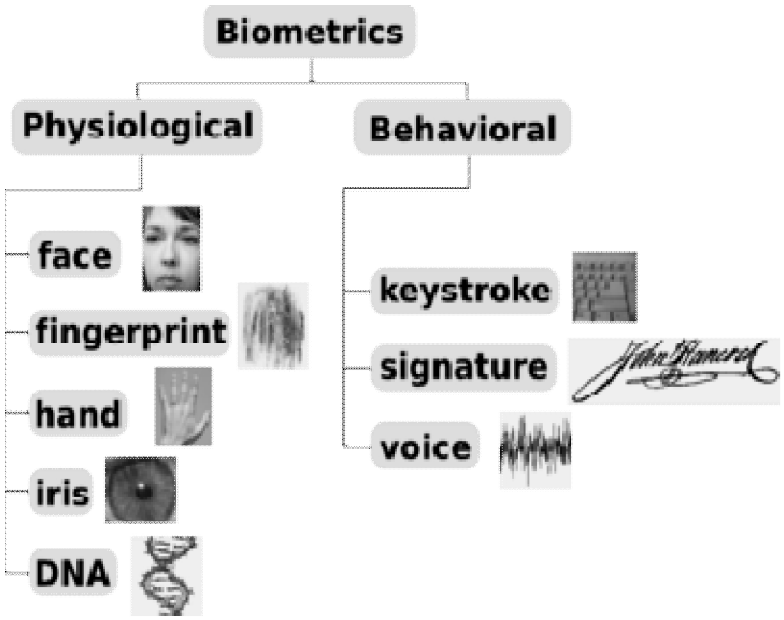


Figure: 1.1 Biometrics Characteristics.

The physiological and/or behavioral characteristic can be used as a biometric characteristic as long as it satisfies the following requirements [3]:

1.2 History of Biometrics

Among the first known examples of practiced biometrics was a form of member printing used in China in the fourteenth century, as reported by the Portuguese historian Joao de Barros. The Chinese merchants were stamping children's palm and footprints on paper with ink to distinguish the babies from one another.

In the 1890s, an anthropologist and police desk clerk in Paris named Alphonse Bertillon sought to fix the problem of identifying convicted criminals and turned biometrics into a distinct field of study. He developed a method of multiple body measurements that was named after him (the Bertillonage technique—measuring body lengths). Police throughout the world used this system until it proved to be exceedingly prone to error as many people shared the same measurements. After this failure, the police started using fingerprinting—developed by Richard Edward Henry of Scotland Yard—after the methods used by the Chinese centuries before.

- European explorer Joao de Barros recorded the first known example of fingerprinting, which is a form of biometrics, in China during the 14th century. Chinese merchants used ink to take children's fingerprints for identification purposes
- In 1890, Alphonse Bertillon, a Parisian police desk studied body mechanics and measurements to help identify criminals. The police used his method, the Bertillonage method, until it falsely identified some subjects. The Bertillonage method was quickly abandoned in favor of fingerprinting, brought back into use by Richard Edward Henry of Scotland Yard.
- Karl Pearson, an applied mathematician studied biometric research early in the 20th century at University College of London. He made important discoveries in the field of biometrics through studying statistical history and correlation, which he applied to animal evolution. His historical work included the method of moments, the Pearson system of curves, correlation and the chi-squared test.
- In the 1960s and '70s, signature biometric authentication procedures were developed, but the biometric field remained fixed until the military and security agencies researched and developed biometric technology beyond fingerprinting.
- 2001 Super Bowl in Tampa, Florida -- each facial image of the 100,000 fans passing through the stadium was recorded via video security cameras and checked electronically against mug shots from the Tampa police. No felons were identified and the video surveillance led many civil liberties advocates to denounce biometric identifying technologies.

- Post 9/11 -- after the attacks, authorities installed biometric technologies in airports to ID suspected terrorists, but some airports, like Palm Beach International, never reached full installation status due to the costs of the surveillance system.
- July 7th, 2005 London, England -- British law enforcement is using biometric face recognition technologies and 360-degree "fish-eye" video cameras to ID terrorists after four bombings on subways and on a double-decker bus. In fact, London has over 200,000 security cameras and surveillance cameras that have been in use since the 1960s.

1.3 Biometric Security Systems

It has been quite a long time since the first Biometric System (BS) was introduced. However, until now, they have not become widely used. This is usual in case of a new, not well-tested, and unverified technology, among which could be the biometric systems included. Skepticism, of course, has its place in our live and is necessary to improve our work. Such skepticism is doubled in case of the security systems. Imperfections in the security systems are not tolerated and may not be. A perfect research work is necessary in course of science thank the increased sensitivity to the quality of the security systems. This is a reason for absence of the biometric security systems in the real-world applications. However, even when the popularity of the biometric systems is not high yet, they are supposed to become very important soon.

1.3.1 Basic Definitions

To avoid possible misunderstanding, some common term definitions follow.

Definition 1: ***Security System*** = *a system that provides some sort of security or security services.*

Security services can be an authentication, a public/secret key providing (for the purposes of an encryption and a decryption), or some other services that strongly depend on a perfect verification and/or identification of an individual. This definition suits perfectly many applications. The obligatory login service is the first of them. Nowadays, the login is based upon a login name selection and a proper password insertion. However, this is a very weak protection of critical information or private

resources. The password based protection could be cracked and the secret information could be misused. This, of course, is not a desirable attribute.

Definition 2: ***Biometric*** *= based upon a physiological or a behavioral feature of a human.*

Physiological features are possible to acquire directly from the human body, i.e. a fingerprint, a retina scan, or a face image. Behavioral features are acquired indirectly. The behavioral features depend on behaviour. Many times, they can be influenced knowingly. These features are i.e. a typing, a signature, or speech. Overall, the human body is able to provide us various features, but not all of them are possible to be applicable in a security system. Features applied to a security system *must be unique*.

Definition 3: ***Biometric Security System (BSS)*** *= a security system, protection of which is based on the biometric features.*

1.3.2 Design of Biometric Security System (BSS)

Design of Biometric Security System (BSS) is the last goal of the dissertation. The speaker recognition can be integrated in many ways. The BSS will be formed from the following security elements: common password, fingerprint authentication, and a voice-based authentication. The designed BSS should be able to accept the basic security elements mentioned above. Besides, it should provide tools for generating a cryptographic key applicable to the cryptographic services – an encryption and a decryption.

1.4 Advantages and Disadvantages of the BSS

The biometric security systems have their advantages and disadvantages. The advantages are *universality*, *uniqueness*, *low circumvention, scalability* and, in some cases, *permanence*. The disadvantages are *exactingness*, *difficult implementation*, *cooperation unwillingness*, and, in some cases, *inconstancy*, which is a counterpart to the permanence of some of them.

1.4.1 Advantages

1.4.1.1 Universality

The biometric features are relatively universal. Most people have their fingers, they can speak, and their cells contain the DNA. Of course, there are some exceptions. People, who lost their finger or arm, cannot use fingerprint-based systems etc. However, all of us have the DNA. Mass use of the biometric security systems is not breaking by a difficult implementation, cooperation unwillingness, an inconstancy, or exactingness only. Another obstacle is inability of some people to pass an enrolment. Still, this could be solved exactly by a multi-biometric security system.

1.4.1.2 Uniqueness

The biometric security systems are supposed to be very reliable, because the biometric features used in the authentication process of the biometric security systems are unique. The features are extracted from the various parts of the human body. They can be extracted from the DNA, a fingerprint, and many others. Many of them will not be used until near future [11]. The essential is the uniqueness – there is (or should be) just one individual having some feature set. It is not valid in every case, but it holds mostly true.

1.4.1.3 Low Circumvention

The biometric security systems are very safe. It is not possible to deceive the protected devices and services thanks to their biometric protection. Most biometric features are possible to acquire by the authorized man only, so that his presence at the point, where the authorization device is placed, is necessary. However, it could be possible to fool the security system, i.e. you can cut off a finger and try to persuade the system to accept it. Still, the security system designers are able to prevent the system from accepting such a sham. Nowadays it is possible to use a fingerprint scanner able to measure flow of blood in veins in the scanned finger. This scanner proves aliveness of the human, who is being identified.

1.4.1.4 Scalability

The individual biometrical technologies can be stacked and can be built in a multilevel authentication system. This multilevel system can include a standard login, a voice login, a fingerprint login, and many others. Stacking of the separate technologies in one complex unit increases the security of the whole system. If one level of the multilevel system is broken or cheated, the others could decrease a break-through possibility. The break-through possibility decreases with the count and strength of each of the login levels.

1.4.1.5 Permanence

Counterpart to the inconstancy of some biometric features is their permanency. An example of the permanent feature set is a set of features acquired from the DNA. DNA is permanent and does not change. The same could be said of the fingerprints. Nevertheless, in case of the fingerprints, a man could cut himself to his finger, which would change the fingerprint structure. Next relatively permanent biometric feature is a retina scan. The other biometric features are not permanent – they are short-term (relatively to the length of a human's life).

1.4.2 Disadvantages

1.4.2.1 Exactingness

Some of the biometric features can be very exacting to acquire. These features can be very exact and unique. Among them could be included e.g. the DNA. Though the DNA is unique for all of us (apart from the twins), it is very difficult to extract and to analyze it quickly enough. Besides, the price of the analysis is not low. This excludes the DNA from the real time applications – it is not worth the advantages it can bring these days.

1.4.2.2 Difficult Implementation

It is very difficult to implement a reliable biometric security system. Nowadays, teams of developers and researchers are working on a safe and reliable implementation of the biometric security system. Some parts of the system have

already been finished and proved them to be well enough – fingerprint recognition, face recognition and partially the speaker recognition. Nevertheless, these systems are standalone and the security of them is not as high as it could be.

1.4.2.3 Cooperation Unwillingness

Some humans are not happy with acquiring their biometrical features. Results of the public inquiry being gone ahead recently show that the most people dislike scanning their retinas. Many of them dislike scanning their fingerprints and faces and the least of them dislike recording of their voices. It is clear then that it would be very useful to develop a reliable technology based upon the speech processing. The most of us are ready to let the machine analyze our voices rather than anything other.

1.4.2.4 Inconstancy

Some biometric features are permanent, but some of them are not. The human voice is not constant during the whole life. Mostly is this effect obvious in the changes of teenagers' voices. Aside should not stay the influence of illnesses or psychological. This can be the greatest difficulty. The human voice is not the only one instable, even many other biometric features change during the human's life.

1.5 Biometric Modules

There are four modules by which a biometric system works. As figure 1.2 shows there are basically four main modules but depend upon different application it can be varied [3].

1.5.1 Sensor Module

It captures the biometric data of an individual. An example is a fingerprint sensor that images the ridge and valley structure of a user's finger.

1.5.2 Feature Extraction Module

In which the acquired biometric data is processed to extract a set of salient or discriminatory features. For example, the position and orientation of minutiae points (local ridge and valley singularities) in a fingerprint image are extracted in the feature extraction module of a fingerprint based biometric system.

1.5.3 Matcher Module

In which the features during recognition are compared against the stored templates to generate matching scores. For example, in the matching module of a fingerprint based biometric system, the number of matching minutiae between the input and the template fingerprint images is determined and a matching score is reported. The matcher module also encapsulates a decision-making module, in which a user's claimed identity is confirmed (verification) or a user's identity is established (identification) based on the database.

1.5.4 System Database Module

It is used by the biometric system to store the biometric templates of the enrolled user. The enrolment module is responsible for enrolling individual into the biometric system database. During the enrolment phase, the biometric characteristic of an individual is first scanned by a biometric reader to produce a digital representation (feature values) of the characteristic. The data capture during the enrolment process may or may not be supervised by a human depending on the application.

A quality check is generally performed to ensure that the acquired sample can be reliably processed by successive stages. In order to facilitate matching, a feature extractor to generate a compact but expressive representation, called a template, further processes the input digital representation.

Depending on the application, the template may be stored in the central database of the biometric system or be recorded on a smart card issued to the individual [33]. Usually, multiple templates of an individual are stored to account for variation observed in the biometric trait and the templates in the database may be updated over time.

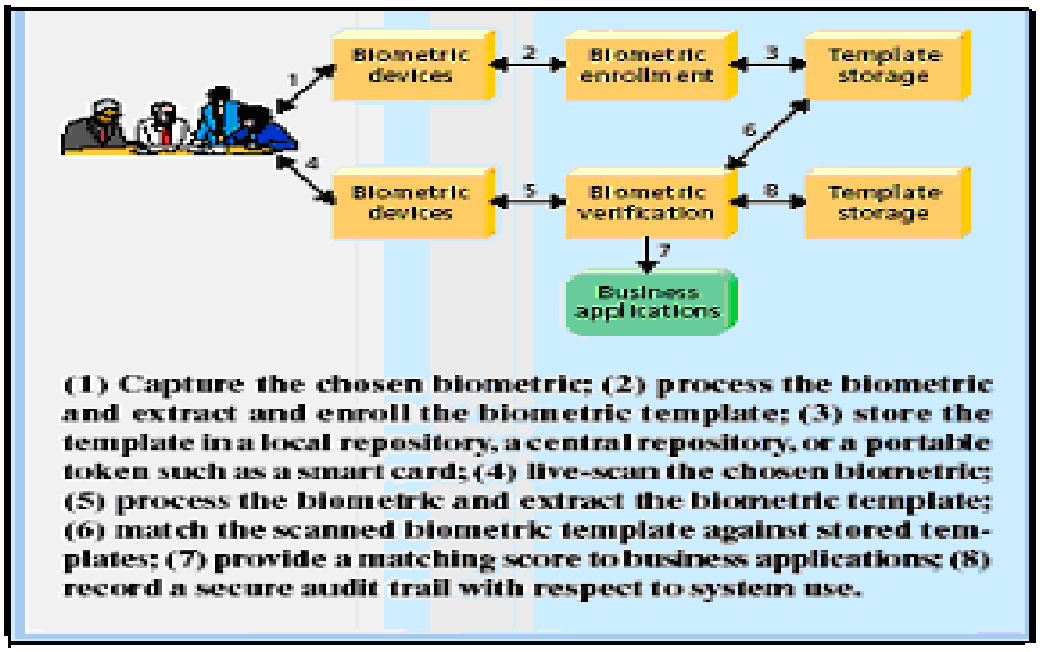


Figure: 1.2 Biometric Modules

1.6 Commonly Used Biometrics

There are number of biometric methods in use. Some of them are:

1.6.1 Facial Recognition

Facial recognition records the spatial geometry of distinguishing features of the face. Different vendors use different methods of facial recognition, however, all focus on measures of key features of the face. Because a person's face can be captured by a camera from some distance away, facial recognition has a clandestine or covert capability (*i.e.* the subject does not necessarily know he has been observed). For this reason, facial recognition has been used in projects to identify card counters or other undesirables in casinos, shoplifters in stores, criminals and terrorists in urban areas.

Facial disguise is of concern in unattended authentication applications. It is very challenging to develop face recognition techniques which can tolerate the effects of aging, facial expressions, slight variations in the imaging environment and variations in the pose of face with respect to camera (2D & 3D rotations) [30].

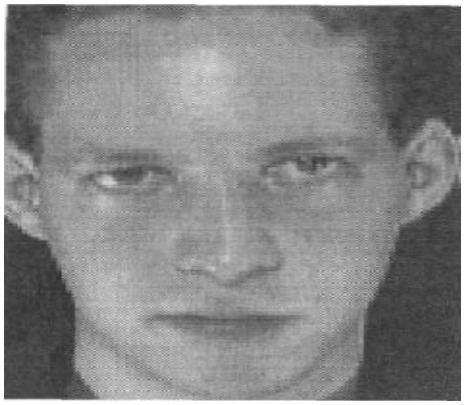


Figure: 1.3 Recognition based on face.

1.6.2 Fingerprint

A fingerprint is a pattern of ridges and furrows located on the tip of each finger. Fingerprints were used for personal identification for many centuries and the matching accuracy was very high [14]. The fingerprint biometric is an automated digital version of the old ink-and- paper method used for more than a century for identification, primarily by law enforcement agencies. The biometric device involves users placing their finger on a platen for the print to be electronically read. The minutiae are then extracted by the vendor's algorithm, which also makes a fingerprint pattern analysis. Fingerprint biometrics currently has three main application arenas: large-scale Automated Finger Imaging Systems (AFIS) generally used for law enforcement purposes, fraud prevention in entitlement programs, and physical and computer access.

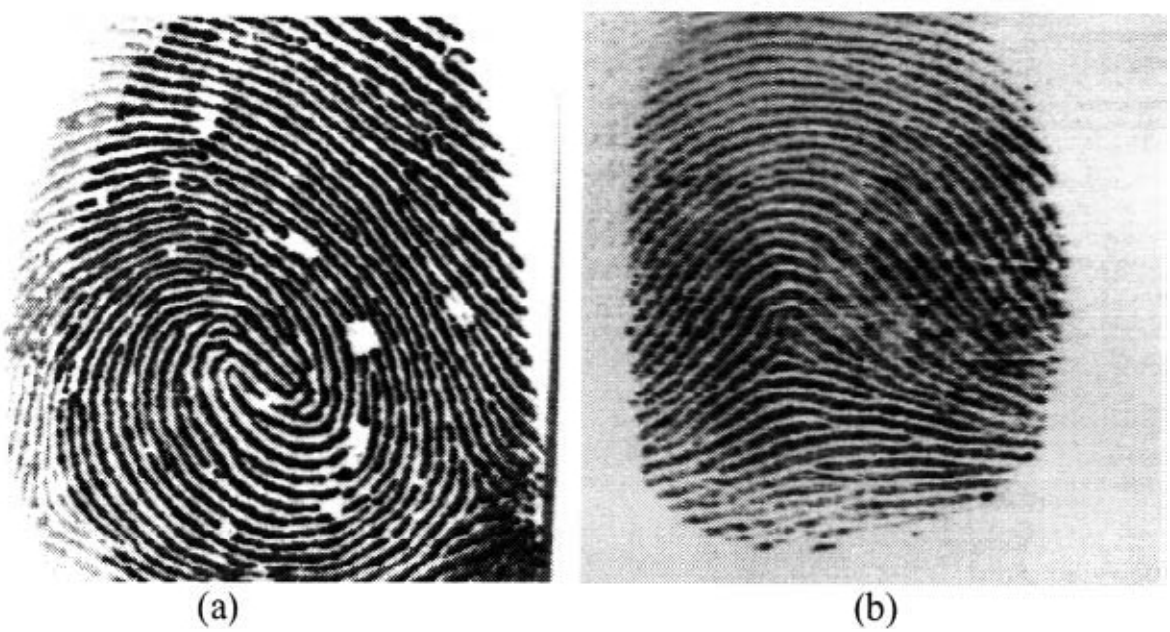


Figure: 1.4 A fingerprint image could be captured from the inked impression of a finger or directly imaging a finger using frustrated total internal reflection tae chnology. The former is called an inked fingerprint (a) and the latter is called live-scan fingerprint (b).

1.6.3 Hand/Finger Geometry

In recent years, hand geometry (Figure 1.5) has become a very popular access control biometrics which has captured almost half of the physical access control market [30]. Hand or finger geometry is an automated measurement of many dimensions of the hand and fingers. Neither of these methods takes actual prints of the palm or fingers. Spatial geometry is examined as the user puts his hand on the sensor's surface and uses guiding poles between the fingers to properly place the hand and initiate the reading. Finger geometry usually measures two or three fingers. Hand geometry is a well-developed technology that has been thoroughly field-tested and is easily accepted by users. Because hand and finger geometry have a low degree of distinctiveness, the technology is not well-suited for identification applications.

Figure 1.5 Recognition based on hand geometry.

1.6.4 Iris Scan

The iris begins to form in the third month of gestation and the structures creating its pattern are largely complete by the eight month. Its complex pattern can contain many distinctive features such as arching ligaments, furrows, ridges, crypts, rings, corona, freckles and a zigzag collarette [18]. Responses of the iris to changes in light can provide an important secondary verification that the iris presented belongs to a live subject. Irises of identical twins are different, which is another advantage. A careful balance of light, focus, resolution and contrast is necessary to extract a feature vector from localized image. While the iris seems to be consistent throughout adulthood, it varies somewhat up to adolescence.

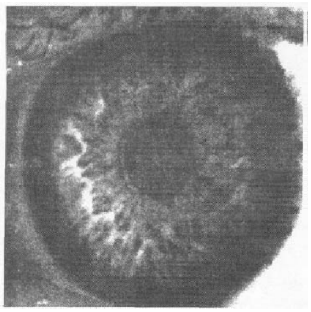


Figure: 1.6 Recognition Based on Iris.

1.6.5 Keystroke Dynamics

Keystroke dynamics is an automated method of examining an individual's keystrokes on a keyboard [31]. This technology examines such dynamics as speed and pressure, the total time taken to type particular words, and the time elapsed between hitting certain keys. This technology's algorithms are still being developed to improve robustness and distinctiveness. One potentially useful application that may emerge is computer access, where this biometric could be used to verify the computer user's identity continuously.

1.6.6 Dynamic Signature Verification

Dynamic signature verification is an automated method of measuring an individual's signature [33]. This technology examines such dynamics as speed, direction, and pressure of writing; the time that the stylus is in and out of contact with the "paper," the total time taken to make the signature; and where the stylus is raised from and lowered onto the "paper."

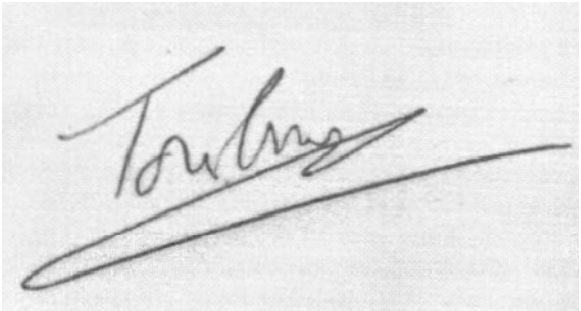


Figure: 1.7 Recognition based on signature.

1.6.7 Speaker / Voice Recognition

Voice or speaker recognition uses vocal characteristics to identify individuals using a pass-phrase. The matching strategy may typically employ approaches based on hidden Markov model, vector quantization, or dynamic time warping [32]. A telephone or microphone can serve as a sensor, which makes it a relatively cheap and easily deployable technology. However, voice recognition can be affected by environmental factors such as background noise. This technology has been the focus of considerable efforts on the part of the telecommunications industry and the U.S. government's intelligence community, which continue to work on improving reliability.

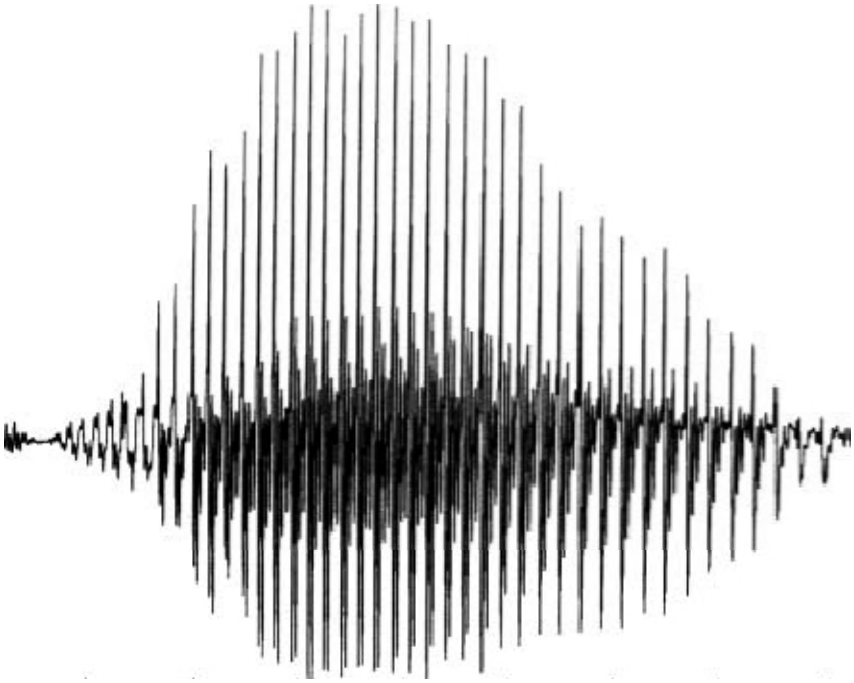


Figure: 1.8 Voice signal representing an utterance of the word "seven".

1.6.8 Infrared Facial Thermograms

Human body radiates heat and the pattern of heat radiation is a characteristic of each individual body [14]. An infrared sensor could acquire an image indicating the heat

emanating from different parts of the body (Figure 1.9). These images are called thermograms. The method of acquisition of the thermal image unobtrusively is akin to the capture of a regular (visible spectrum) photograph of the person. Any part of the body could be used for identification. The absolute values of the heat radiation are dependent upon many extraneous factors and are not completely invariant to the identity of an individual; the raw measurements of heat radiation need to be normalized, e.g., with respect to heat radiating from a landmark feature of the body.

Figure: 1.9 Identification based on facial thermograms

1.6.9 Ear as Biometrics

It is known that the shape of the ear and the structure of the cartilegenous tissue of the pinna are distinctive3. The features of an ear are not expected to be unique to each individual. The ear recognition approaches are based on matching vectors of distances of salient points on the pinna from a landmark location (Figure 1.10) on the ear [26]. No commercial systems are available yet and authentication of individual identity based on ear recognition is still a research topic.

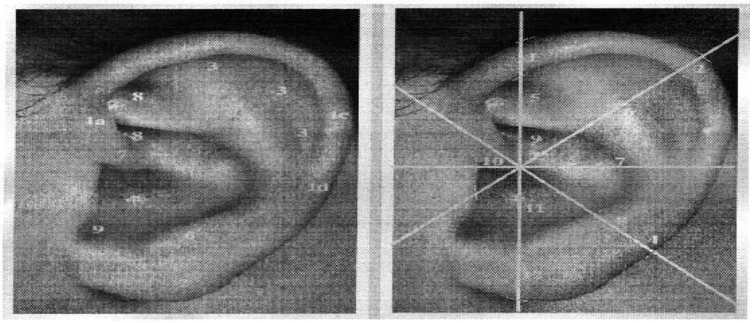


Figure: 1.10 An image of an ear and the features used for ear-based recognition.

1.6.10 DNA

DNA (Deoxyribonucleic Acid) is the one-dimensional ultimate unique code for one's individuality - except for the fact that identical twins have the identical DNA pattern. It is, however, currently used mostly in the context of forensic applications for identification [16]. Three issues limit the utility of this biometrics for other applications: (i) contamination and sensitivity: it is easy to steal a piece of DNA from an unsuspecting subject to be subsequently abused for an ulterior purpose; (ii) automatic real-time identification issues: the present technology for genetic matching is not geared for online unobtrusive identifications.

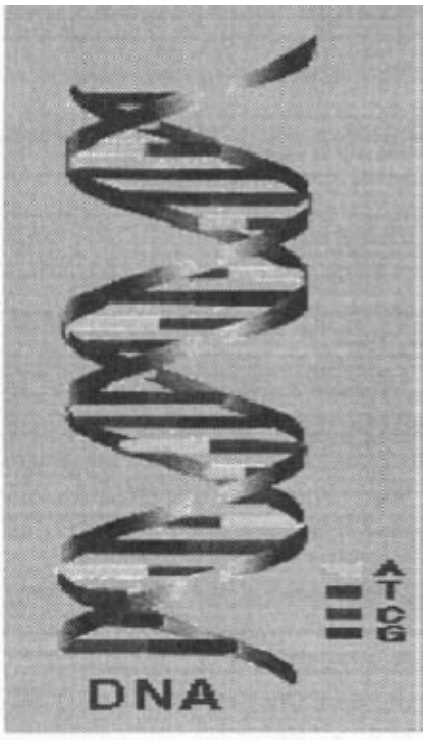


Figure: 1.11 DNA is double helix structure made of four bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)

1.7 Application of Biometric Systems

The applications of biometrics can be divided into the following three main groups:

Commercial applications such as computer network login, electronic data security, e-commerce, Internet access, ATM, credit card, physical access control, cellular phone, PDA, medical records management, distance learning, etc.

Government applications such as national ID card, correctional facility, driver's license, social security, welfare-disbursement, border control, passport control, etc.

Forensic applications such as corpse identification, criminal investigation, terrorist Identification, parenthood determination, missing children, etc.. Traditionally, commercial applications have used knowledge based systems (e.g., PINs and passwords), government applications have used token based systems (e.g., ID cards and badges), and forensic applications have relied on human experts to match biometric features. Biometric systems are increasingly deployed in large scale civilian applications. The Schiphol Premium scheme at the Amsterdam airport, for example, employs iris scan cards to speed up the passport and visa control producers. The passengers enrolled in this scheme insert their card at the gate and look into the camera; the camera acquires the image of the traveller's eye and processes it to locate the iris, and compute the iris code the computed iris code is compared with the data residing in the card to complete user verification. A similar scheme is also being used to verify the identity of Schiphol airport employees working in high-security areas. Thus, biometric systems can be used to enhance user convenience while improving security.

1.8 Problem Formulation

From the above discussion it has been conclude that no single biometric is expected to effectively meet the requirements of authentication. The match between a specific biometric is determined depending upon the operational mode and the properties of the biometric characteristic. Voice is a most natural way of communication and non-intrusive as a biometric, Voice biometric has characteristic of acceptability, cost, easy to implement, no special equipment required. So, we have chosen to develop a Speaker Identification System.

i). Study the different features of the voice signal

ii). Study the different methods/ techniques of Speaker Identification system.

iii). Develop a Speaker Identification system using LabVIEW.

iv). Test the developed system for its accuracy.

Chapter 2 Literature Survey

2.1 Introduction

Speech and hearing, man's most used means of communication, have been the objects of intense study for more than 150 years—from the time of von Kempelen's speaking machine to the present day. With the advent of the telephone and the explosive growth of its dissemination and use, the engineering and design of evermore bandwidth-efficient and higher-quality transmission systems has been the objective and providence of both engineers and scientists for more than seventy years. This work and investigations have been largely driven by these real-world applications which now have broadened to include not only speech synthesizers but also automatic speech recognition systems, speaker verification systems, speech enhancement systems, efficient speech coding systems, and speech and voice modification systems. The objectives of the engineers have been to design and build real workable and economically affordable systems that can be used over the broad range of existing and newly installed communication channels.

2.2 Production and Classification of Speech Sounds

A simplified view of speech production is given in Figure 2.1, where the speech organs are divided into three main groups: the lungs, larynx, and vocal tract. The lungs act as a power supply and provide airflow to the larynx stage of the speech production mechanism. The larynx modulates airflow from the lungs and provides either a periodic puff-like or a noisy airflow source to the third organ group, the vocal tract. The vocal tract consists of oral, nasal, and pharynx cavities, giving the modulated airflow its "color" by spectrally shaping the source. Sound sources can also be generated by constrictions and boundaries, not shown in Figure 2.1, that are made within the vocal tract itself, yielding in addition to noisy and periodic sources, an impulsive airflow source. We have here idealized the sources in the sense that the anatomy and physiology of the speech production mechanism does not generate a perfect periodic, impulsive, or noise source.[1] Following the spectral coloring of

the source by the vocal tract, the variation of air pressure at the lips results in a traveling sound wave that the listener perceives as speech.

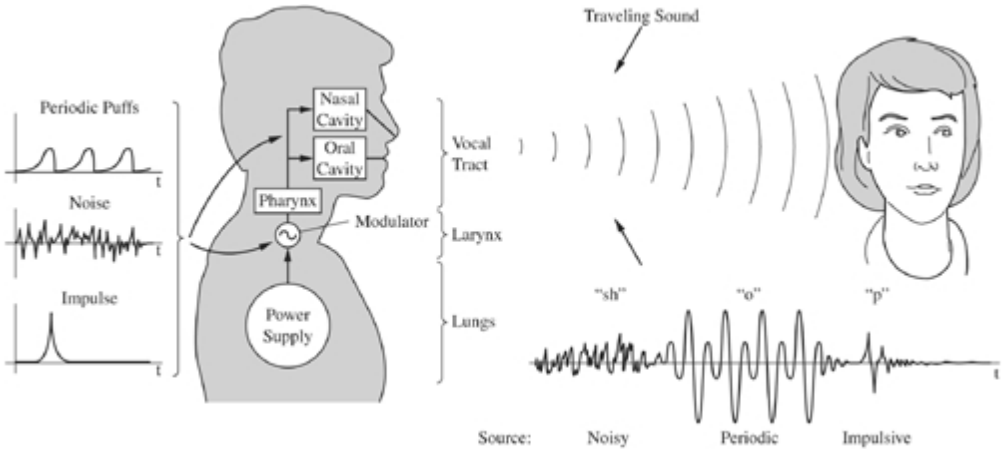


Figure: 2.1 Simple view of speech production system.

2.2.1 Models for Speech Production

A schematic longitudinal cross-sectional drawing of the human vocal tract mechanism is given in Figure 2.2 [19]. This diagram highlights the essential physical features of human anatomy that enter into the final stages of the speech production process. It shows the vocal tract as a tube of non uniform cross-sectional area that is bounded at one end by the vocal cords and at the other by the mouth opening. This tube serves as an acoustic transmission system for sounds generated inside the vocal tract. For creating nasal sounds like /M/, /N/, or /NG/, a side-branch tube, called the nasal tract, is connected to the main acoustic branch by the trapdoor action of the velum. This branch path radiates sound at the nostrils. The shape (variation of cross-section along the axis) of the vocal tract varies with time due to motions of the lips, jaw, tongue, and velum. Although the actual human vocal tract is not laid out along a straight line as in Figure 2.2, this type of model is a reasonable approximation for wavelengths of the sounds in speech.

The sounds of speech are generated in the system of Figure 2.2 in several ways. Voiced sounds (vowels, liquids, glides, nasals) are produced when the vocal tract tube is excited by pulses of air pressure resulting from quasi-periodic opening and closing of the glottal orifice (opening between the vocal cords). The vowels /UH/, /IY/, and /EY/, and the liquid consonant /W/. Unvoiced sounds are produced by creating a constriction somewhere in the vocal tract tube and forcing air through that constriction, thereby creating turbulent air flow, which acts as a random noise excitation of the vocal tract tube. Examples are the unvoiced fricative sounds such as /SH/ and /S/. A third sound production mechanism is when the vocal tract is partially closed off causing turbulent flow due to the constriction, at the same time allowing quasi-periodic flow due to vocal cord vibrations. Sounds produced in this manner include the voiced fricatives /V/, /DH/, /Z/, and /ZH/. Finally, plosive sounds such as /P/, /T/, and /K/ and affricates such as /CH/ are formed by momentarily closing off air flow, allowing pressure to build up behind the closure, and then abruptly releasing the pressure. All these excitation sources create a wide-band excitation signal to the vocal tract tube, which acts as an acoustic transmission line with certain vocal tract shape-dependent resonances that tend to emphasize some frequencies of the excitation relative to others.

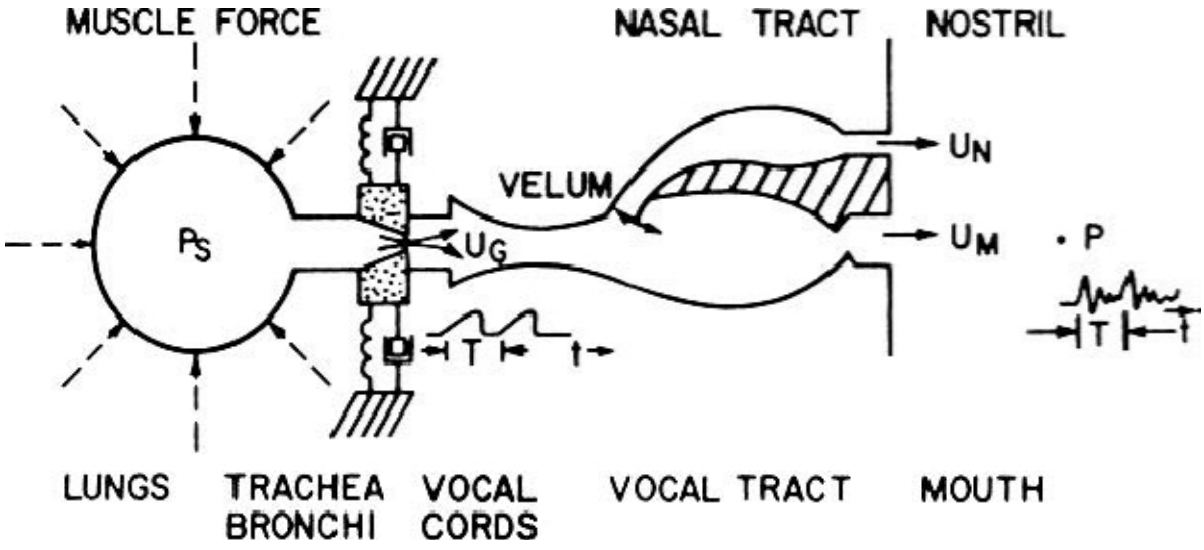


Figure: 2.2 Schematic model of the vocal tract system.

2.3 Hearing and Auditory Perception

In this topic the properties of human sound perception that can be employed to create digital representations of the speech signal that are perceptually robust has been discussed.

2.3.1 The Human Ear

Figure 2.3 shows a schematic view of the human ear showing the three distinct sound processing sections, namely: the outer ear consisting of the pinna, which gathers sound and conducts it through the external canal to the middle ear; the middle ear beginning at the tympanic membrane, or eardrum, and including three small bones, the malleus (also called the hammer), the incus (also called the anvil) and the stapes (also called the stirrup), which perform a transduction from acoustic waves to mechanical pressure waves; and finally, the inner ear, which consists of the cochlea and the set of neural connections to the auditory nerve, which conducts the neural signals to the brain.

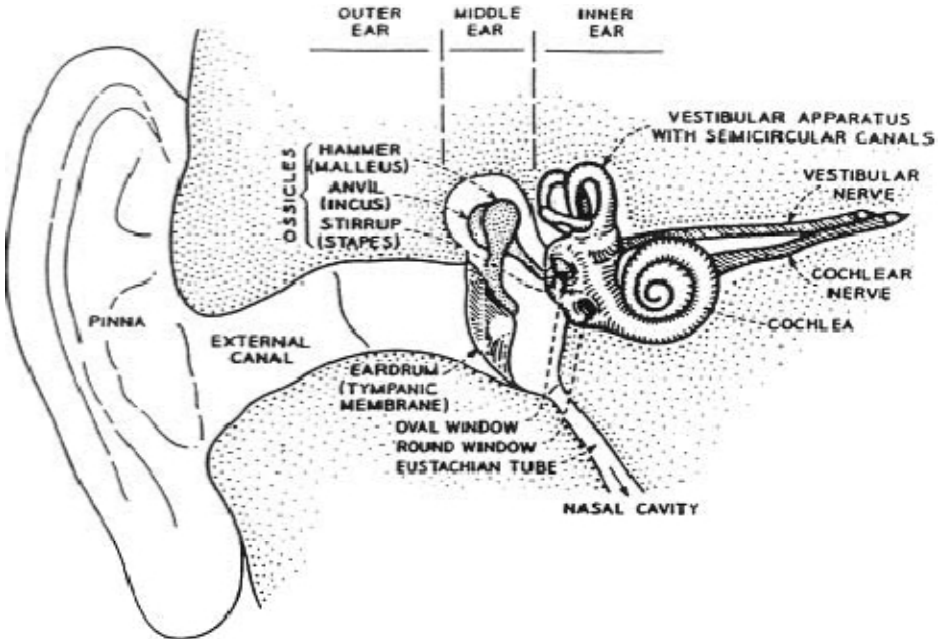


Figure: 2.3 Schematic view of the human ear (inner and middle structures enlarged).

Figure 2.4 [25] depicts a block diagram abstraction of the auditory system. The acoustic wave is transmitted from the outer ear to the inner ear where the ear drum

and bone structures convert the sound wave to mechanical vibrations which ultimately are transferred to the basilar membrane inside the cochlea. The basilar membrane vibrates in a frequency-selective manner along its extent and thereby performs a rough (non-uniform) spectral analysis of the sound. Distributed along the basilar membrane are a set of inner hair cells that serve to convert motion along the basilar membrane to neural activity. This produces an auditory nerve representation in both time and frequency. The processing at higher levels in the brain, shown in Figure 2.4 as a sequence of central processing with multiple representations followed by some type of pattern recognition, is not well understood and we can only postulate the mechanisms used by the human brain to perceive sound or speech. Even so, a wealth of knowledge about how sounds are perceived has been discovered by careful experiments that use tones and noise signals to stimulate the auditory system of human observers in very specific and controlled ways. These experiments have yielded much valuable knowledge about the sensitivity of the human auditory system to acoustic properties such as intensity and frequency.

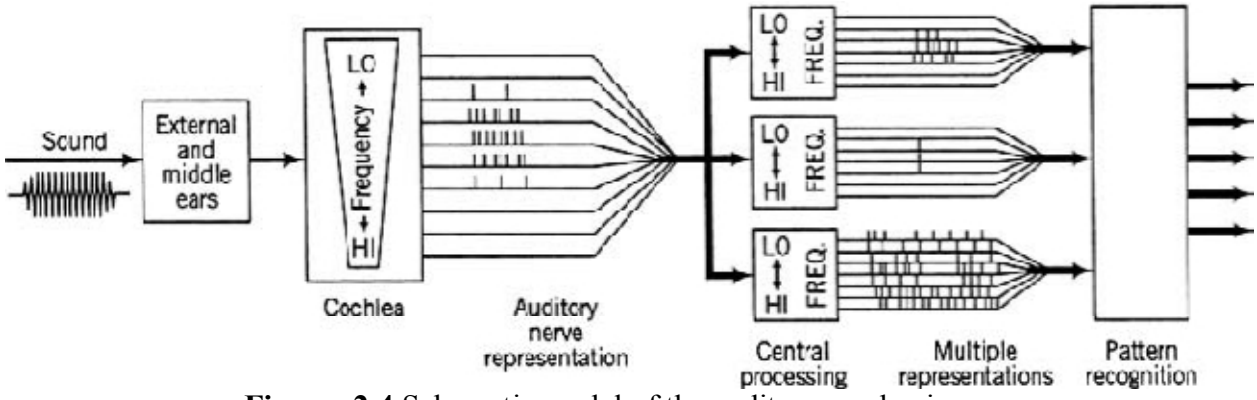


Figure: 2.4 Schematic model of the auditory mechanism

2.3.2 Perception and Loudness

A key factor in the perception of speech and other sounds is loudness. Loudness is a perceptual quality that is related to the physical property of sound pressure level. Loudness is quantified by relating the actual sound pressure level of a pure tone (in dB relative to a standard reference level) to the perceived loudness of the same tone

(in a unit called phons) over the range of human hearing (20 Hz–20 kHz). This relationship is shown in Figure 2.5 [7][17]. These loudness curves show that the perception of loudness is frequency-dependent. Specifically, the dotted curve at the bottom of the figure labeled "threshold of audibility" shows the sound pressure level that is required for a sound of a given frequency to be just audible (by a person with normal hearing). It can be seen that low frequencies must be significantly more intense than frequencies in the mid-range in order that they be perceived at all.

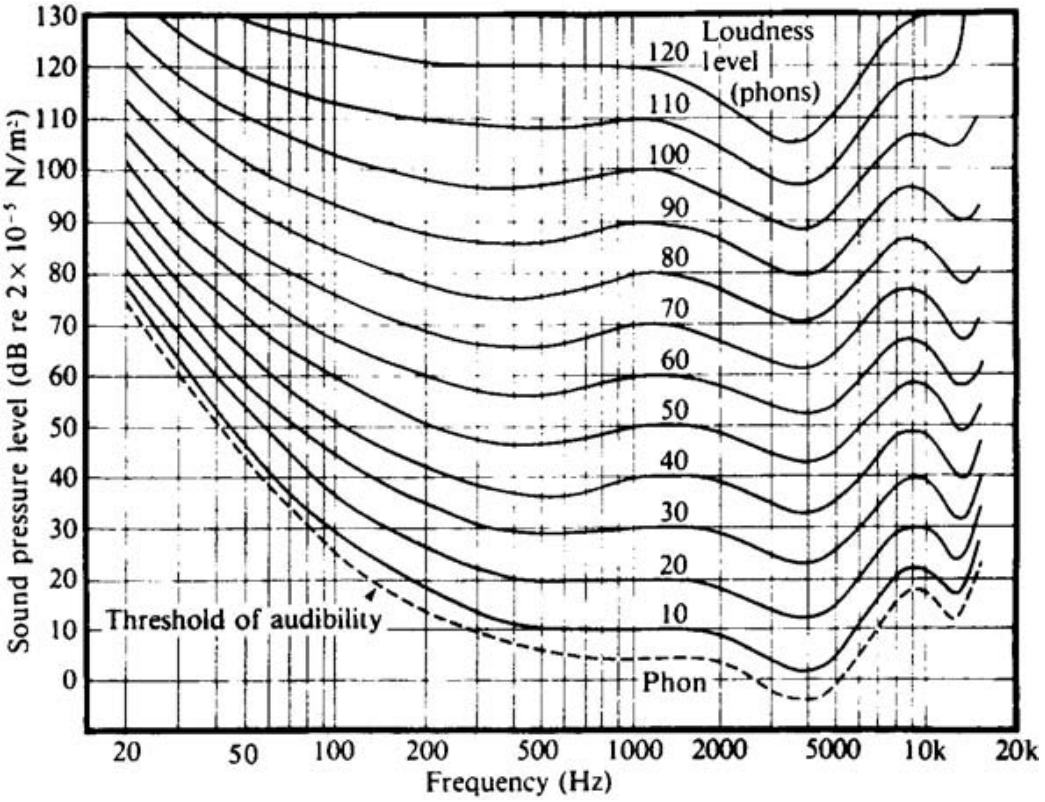


Figure: 2.5 Loudness level for human hearing

The solid curves are equal-loudness-level contours measured by comparing sounds at various frequencies with a pure tone of frequency 1000Hz and known sound pressure level. For example, the point at frequency 100 Hz on the curve labeled 50 (phons) is obtained by adjusting the power of the 100 Hz tone until it sounds as loud as a 1000 Hz tone having a sound pressure level of 50 dB. Careful measurements of this kind show that a 100 Hz tone must have a sound pressure level of about 60 dB in order to be perceived to be equal in loudness to the 1000 Hz tone

of sound pressure level 50 dB. By convention, both the 50 dB 1000 Hz tone and the 60 dB 100 Hz tone are said to have a loudness level of 50 phons (pronounced as /F OW N Z/). The equal-loudness-level curves show that the auditory system is most sensitive for frequencies ranging from about 100 Hz up to about 6 kHz with the greatest sensitivity at around 3 to 4 kHz. This is almost precisely the range of frequencies occupied by most of the sounds of speech.

2.3.3 Pitch Perception

Most musical sounds as well as voiced speech sounds have a periodic structure when viewed over short time intervals, and such sounds are perceived by the auditory system as having a quality known as pitch. Like loudness, pitch is a subjective attribute of sound that is related to the fundamental frequency of the sound, which is a physical attribute of the acoustic waveform [34]. The relationship between pitch (measured on a nonlinear frequency scale called the mel-scale) and frequency of a pure tone is approximated by the equation [34]:

$$\text{Pitch in Mel} = 1127\ log_e\ (1 + F/700) \tag{2.1}$$

which is plotted in Figure 2.5. This expression is calibrated so that a frequency of 1000 Hz corresponds to a pitch of 1000 mels. This empirical scale describes the results of experiments where subjects were asked to adjust the pitch of a measurement tone to half the pitch of a reference tone. To calibrate the scale, a tone of frequency 1000 Hz is given a pitch of 1000 mels. Below 1000 Hz, the relationship between pitch and frequency is nearly proportional. For higher frequencies, however, the relationship is nonlinear. For example, (2.1) shows that a frequency of f = 5000 Hz corresponds to a pitch of 2364 mels. The psychophysical phenomenon of pitch, as quantified by the melscale, can be related to the concept of critical bands [10]. It turns out that more or less independently of the center frequency of the band, one critical bandwidth corresponds to about 100 mels on the pitch scale. This is shown in Figure 3.5, where a critical band of width _fc = 160 Hz centered on fc = 1000 Hz maps into a band of width 106 mels and a critical band of width 100 Hz centered on 350 Hz maps into a band of width 107 mels. Thus, what we know about pitch

perception reinforces the notion that the auditory system performs a frequency analysis that can be simulated with a bank of bandpass filters whose bandwidths increase as center frequency increases.

2.4 State of Art

At present, there are many various applications based on the speech technology. The speech technology can be divided into three main parts: the speech recognition, the speaker recognition, and other recognition. However, emphasis is laid on the speech recognition more than on the other types. The most widely used techniques in the speech technology are the hidden Markov models, or the neural networks experiencing their renaissance. Figure 2.6 shows the classification of speech processing technology.

2.4.1 Speech Recognition

In the speech recognition, there is a machine trying to answer the question "What was said by the speaker?". Thus, the content of the speech is recognized. Result of this process should be a transcription of the text said by the speaker, accomplishment of a speech command, or another action. The speech recognition can be roughly divided into the following groups:

- Continuous speech recognition is recognition of the phones, words, or sentences continuously spoken by the speaker.
- Recognition of the isolated speech units, i.e. the units (phones, words, or others) are said separately and the task is to find them in a signal and to recognize them.
- Recognition and understanding of the meaning, i.e. the meaning of the speech should be recognized, which can be reached by context recognition.

There are many ways of usage of the speech recognition technology [12] [9]. Before all, phone banking, device controlling, or just simple type writing (speech-to-text transcription) can be mentioned.

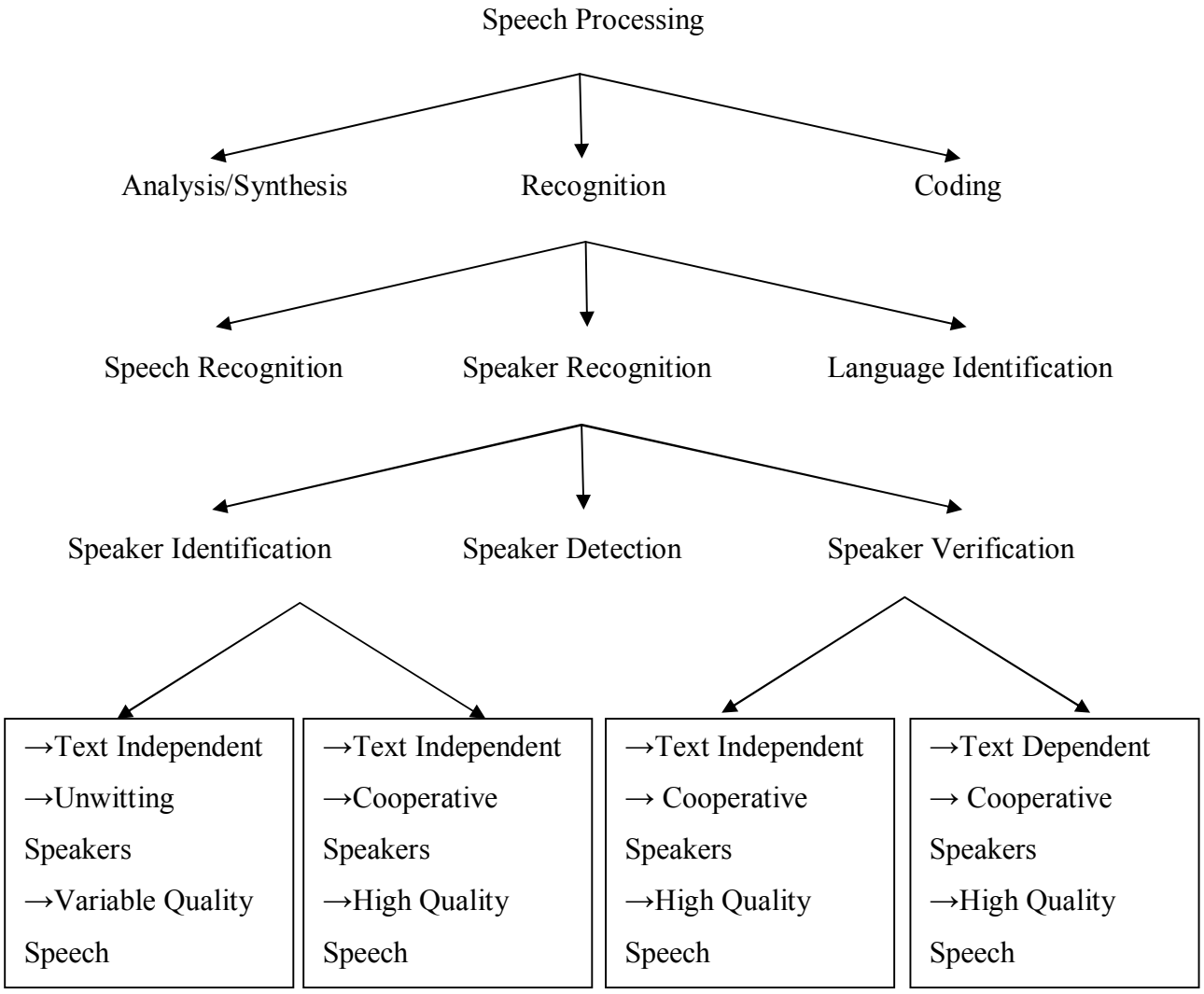


Figure: 2.6 Classification of speech Processing

2.4.2 Speaker Recognition

The speaker recognition can be divided into two main groups: speaker identification and speaker verification. The process of the speaker identification answers a question "Who is speaking?" On the other hand, the speaker verification answers a question "Is the one, who is speaking, really the one, who he is claiming to be?" i.e. in case of the identification, we are trying to identify an unknown voice among other voices, and in case of the verification, we would like to determine a similarity of two voices, one of which is known and the other one is unknown.

2.4.3 Other Recognition

The other recognition includes everything that is not a speech or speaker recognition. Among the other recognition can be included stress recognition or alcohol detection. These sorts of recognition are based on a fact that some features of the voice change under the influence of stress. Some research on this topic was done at the BUT FEEC, where some individuals (students in this case) were reading a text in a common situation and then they were asked to read the same text before the final state exam. Likewise, the alcohol changes voice characteristics even when there are only few thousandths per mile in blood. There are many other sorts of the recognition using voice, among them can be named mental condition recognition, gender recognition, or age recognition.

2.5 Speech in the Biometry

Voice (or vocalization) is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. Voice is not always produced as speech, however. Infants babble and coo; animals bark, moo, whinny, growl, and meow; and adult humans laugh, sing, and cry. Voice is generated by airflow from the lungs as the vocal folds are brought close together. When air is pushed past the vocal folds with sufficient pressure, the vocal folds vibrate. If the vocal folds in the larynx did not vibrate normally, speech could only be produced as a whisper. Your voice is as unique as your fingerprint. It helps define your personality, mood, and health.

The range of use of the speech technology in the biometry is wide. The most common one is the user authentication using his/her voice. In some cases, another speech technology is applicable, e.g. in case of the phone banking can be the system fully automatic, which is realized by combination of the speech recognition, speaker recognition, and speech synthesis.

The speaker recognition is becoming very important in the biometry. Nowadays, only the fingerprint technology is fully accepted so that many real applications based on this technology have been created since now. The speech technology is in the frame to be as

successful as the fingerprint technology, but there are some limiting factors, which make this technology not as popular as it could be.

2.6 Speaker Recognition

Speaker Recognition is a system that can recognize a person based on his/her voice. This is achieved by implementing complex signal processing algorithms that run on a digital computer or a processor. This Application is analogous to the fingerprint recognition system or other biometrics recognition systems that are based on certain characteristics of a person.

There are several occasions when we want to identify a person from a given group of people even when the person is not present for physical examination. For example, when a person converses on a telephone, all we have is the person's voice for analysis. It then makes sense to develop a recognition system based on voice.

According to the dependency of text, there are two kinds of input voice that is Text Dependent and Text Independent.

2.6.1 Text-dependent

The content of speech is known to the system. It is based on the assumption that the speaker is cooperative. This is the one that easier to implement, and with higher accuracy.

2.6.2 Text-independent

In this condition, the content of speech is unknown. In other words, there is no prior knowledge in the system. This is the one that harder to implement, but with higher flexibility.

2.7 Approaches to the Speaker Recognition

There are two different approaches to the speaker recognition. The first one is based on the speaker verification and the second one is based on the speaker identification. When verifying a speaker, there are two possible answers – 'yes' or 'no', i.e. the speaker is ('yes')

or is not ('no') the one, who he/she claims to be. When identifying a speaker, there are other two possible results. If the unknown speaker were found in the voice database, the answer would be an identification number of the speaker. Hence, the answer would be 'yes', since the user was found in the database. If the unknown speaker were not found in the voice database, the answer would be unknown (unauthorized) user, i.e. the answer would be 'no'[12].

Main difference between the two approaches is that in case of the speaker verification the user has to tell the system, who he/she claims to be. Then, the system compares the stored pattern of the claimed user with the new pattern of the unknown user and answers 'yes' or 'no'. Hence, a one-to-one comparison is performed. In case of the speaker identification, the system itself determines the user. Thus, a one-to-many comparison is performed. If the user's voice were not found in the database, the answer would be 'no', otherwise the answer would be 'yes'. In other words – you can see two unwanted situations.

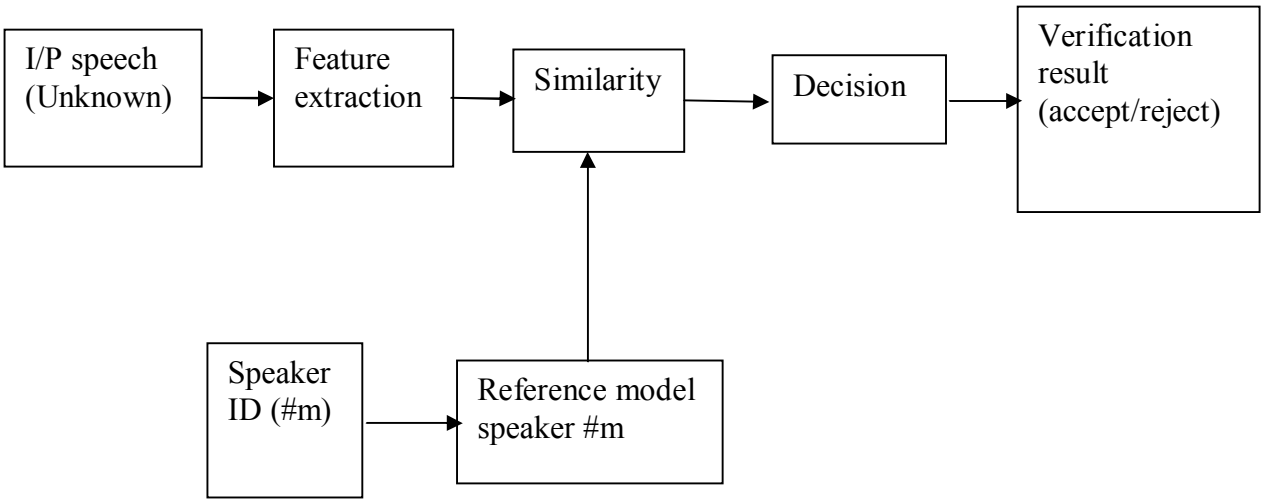


Figure:2.7 The speaker recognition - Verification approach.

First, the system accepts a user though he/she is not authorized to access to the system resources (this corresponds to the FAR). In case of the speaker verification, the algorithm decided the unknown voice to be similar enough to the claimed voice to let in the user. In case of the speaker identification, the algorithm decided the unknown voice to be similar to a voice stored in the voice database though this was not true.

Second, the system rejects the user though he/she is a valid user. In case of the speaker verification, the algorithm did not find similarity between the unknown voice and the claimed voice. In case of the speaker identification, there was no stored voice similar enough to the unknown voice to let in the user. The speaker verification and identification approaches are described by the Figure: 2.7, Figure 2.8 [24].

Result of the speaker recognition need not to be the answer 'yes' or 'no'. It can be a probability of the unknown user being the claimed user (in case of the verification) or a probability of the unknown user is one of the users stored in a database (in case of the identification). Such output should lie in an interval of (0,1). It is usable in a multi-biometric security system, where a fusion of the outputs of the individual biometric subsystems is done. Now, we know enough to design a speaker recognition application. This application can use one of the approaches mentioned above to accept or reject a user.

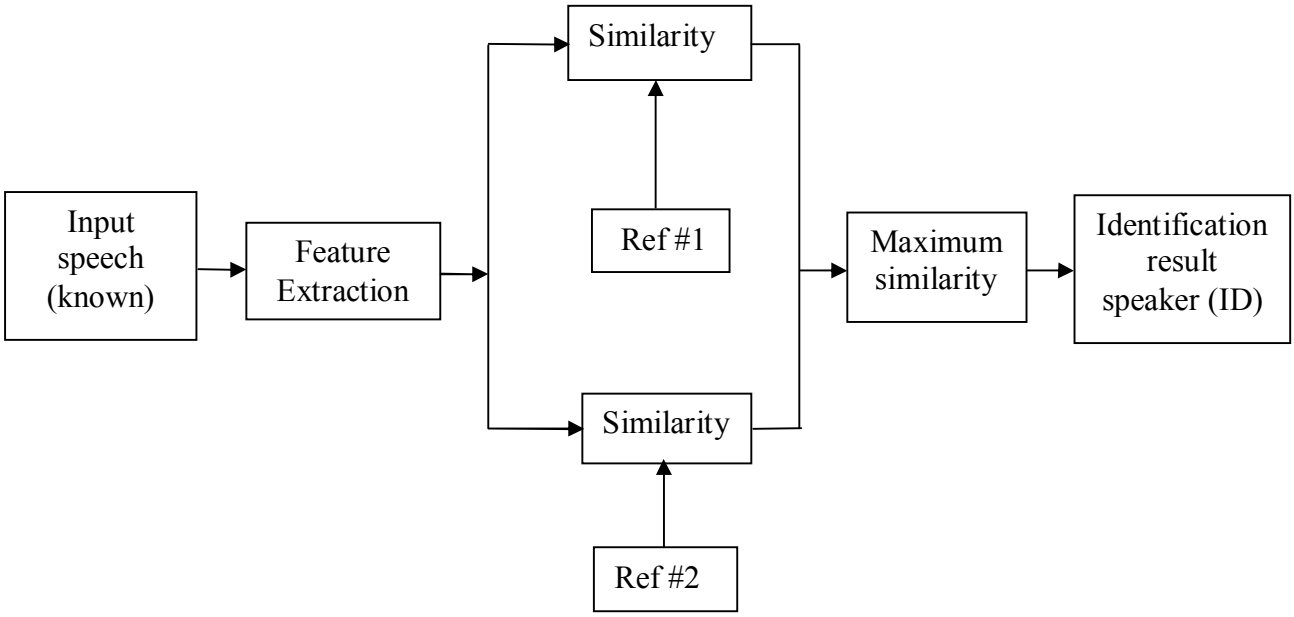


Figure:2.8 The speaker recognition - Identification approach.

At the highest level, all speaker recognition systems contain two main modules feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing

extracted features from his/her voice input with the ones from a set of known speakers [8]. We will discuss each module in detail in later sections.

All speaker recognition systems have to serve two distinguished phases. The first one is referred to the enrolment or training phase, while the second one is referred to as the operational or testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples. In the testing phase, the input speech is matched with stored reference model and a recognition decision is made.

Speaker recognition is a difficult task. Automatic speaker recognition works based on the premise that a person's speech exhibits characteristics that are unique to the speaker. However this task has been challenged by the highly variant of input speech signals. The principle source of variance is the speaker himself/herself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health conditions (e.g. the speaker has a cold), speaking rates, and so on. There are also other factors, beyond speaker variability, that present a challenge to speaker recognition technology [24]. Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets).

For a BSS, the speaker verification is a reasonable choice, but the speaker identification can be used too in the way the verification is used. There are many ways of accomplishing the recognition, but the most popular is the method based on the Gaussian Mixtures Model. Usually, the process of the speaker recognition consists of the following steps:

- Speech Processing
- Feature extraction
- Pattern matching
- Decision making

2.8 Speech Processing

The first phases of the speech signal processing are the recording, the digitizing, and the pre-processing.

2.8.1 Recording and Digitizing

Recording of a signal is the first stage of the whole speaker recognition process. Recording and digitizing is very important and can influence the speaker recognition very much. Although it is so important, it will not be described in detail. Relevant information can be found in the references [7][15]. The analogue speech signal is recorded using a microphone. Quality of the microphone can influence final quality of the recognition. However, not every time is it possible to have a first-rate microphone to use. At this point, the algorithms can show their strengths. Subsequently to the recording, the analogue signal is sampled and quantized.

Speech signals are usually represented as functions of continuous variable t, which denotes time. The analogue speech signal $S_a\{t)$ can be defined as a function varying continuously in time. The processed signals are sampled with a sampling period Ts. Then, we can define a sample of a discrete-time signal as

$$S(n) = Sa(nTs) \tag{2.2}$$

Which means $t = nT_s$. The signal $s\{n)$ is called *digital signal.* According to the sampling period can be defined the sampling frequency as $F_s = T_s^{-1}$. Usually, the sampling frequency of the speech signals lies in the range $8000 < F_s < 22050$. Due to the limitations of the human organs for the speech production and the auditory system, the typical speech communication is limited to a bandwidth of 8 kHz, which results in the sampling frequency of 16 kHz, which ordinarily satisfy our requirements.

Another problem of the digitizing is quantization. The amplitude of the analogue signal must be quantized due to the limitations of the computer technology [20]. Problems like the quantization error are not taken into consideration. The signals

nowadays are quantized in common to 16 bits, which means that the amplitudes of the analogue signal are limited to the values from an interval of (-32768,32767).

We work implicitly only with the digital signal *s (n),* which is sampled with the sampling frequency F_s. The recorded digital signal is of a finite length, which is referred to as N_{total}. The signal is quantized to 16 bits and normalized to the range from -1 to 1.

2.8.2 Framing

Framing is next important step in the signal processing process. The recorded discrete signal *s(n)* has always a finite length N_{total}, but is usually not processed whole. The signal is framed - cut to pieces. Length of one frame is $N<<N_{total}$ samples. Given total length of the signal N_{total} total count of all frames is:

$$j = \text{int}\left(\frac{N_{total}}{N}\right) \tag{2.3}$$

Where the function int(x) returns the integer part of x. The length N of the frames in the real application is usually based on the physiological characteristic of a vocal tract[22].

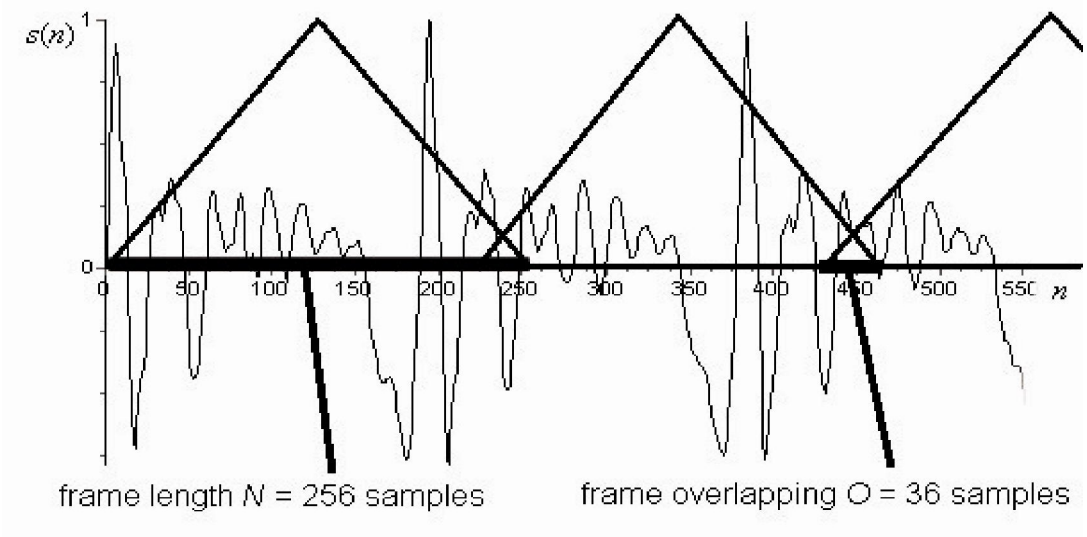


Figure: 2.9 Framing of a speech signal – overlapping frames and triangular windows.

The vocal tract is not able to change its shape faster than fifty times per second, which gives us a period of 20 milliseconds. When the sampling frequency is F_s =16000 Hz, then we get the length of the Frame $N = F_s.T = 16.10^3.20.10^{-3}$=320 samples[20].

Usually, an overlapping of the individual frames is defined. The overlapping is used to increase precision of the recognition process. The length of the frame is increased ordinarily to the power of two, i.e. in our case it would be 512 samples. The power of two is chosen because of the Fast Fourier Transformation (FFT), which is performed fastest in case the input length of the signal equals the power of two. Then the overlapping is chosen so that it is O = 512-320 = 192 samples long. When the overlapping is defined, changes the total count to

$$j = \text{int}\left(\frac{N_{total}}{N - O}\right) \tag{2.4}$$

In the Figure 2.1 you can see an illustration of speech signal framing by overlapping frames. The frames are represented by the thick black triangles. The appropriate values are marked.

After we chose proper length and overlapping of the frames, we can process them further. When we work with the whole signal, we speak of a long-term (LT) processing. When we work with the individual frames, we speak of a short-term (ST) processing. The individual frames themselves are defined as

$$S(j.n) = S(j * (N - O) + n) \tag{2.5}$$

Where *s{n)* is the original signal, *s{j,n)* is the 7-th frame, *N* is length of one frame, and *O* is the length of the overlapping.

2.8.3 Windowing

Before further processing, the individual frames are windowed. The frame itself is implicitly windowed by a rectangular window. However, spectral characteristic of the rectangular window is unsuitable. This is why other windows are applied. Windowed signal is defined as

$$S_w(n) = S(n).W(n) \tag{2.6}$$

Where $s_w(n)$ is the windowed signal, $s(n)$ is the original signal N samples long, and $w(n)$ is the window itself. The window $w(n)$ should be of the same length as the original signal $s(n)$, which is N in this case[4]. There are many types of windows using for this purpose. The rectangular window is defined as

$$w(n) = \begin{cases} 1, 1 \le n \le N \\ 0, otherwise \end{cases} \tag{2.7}$$

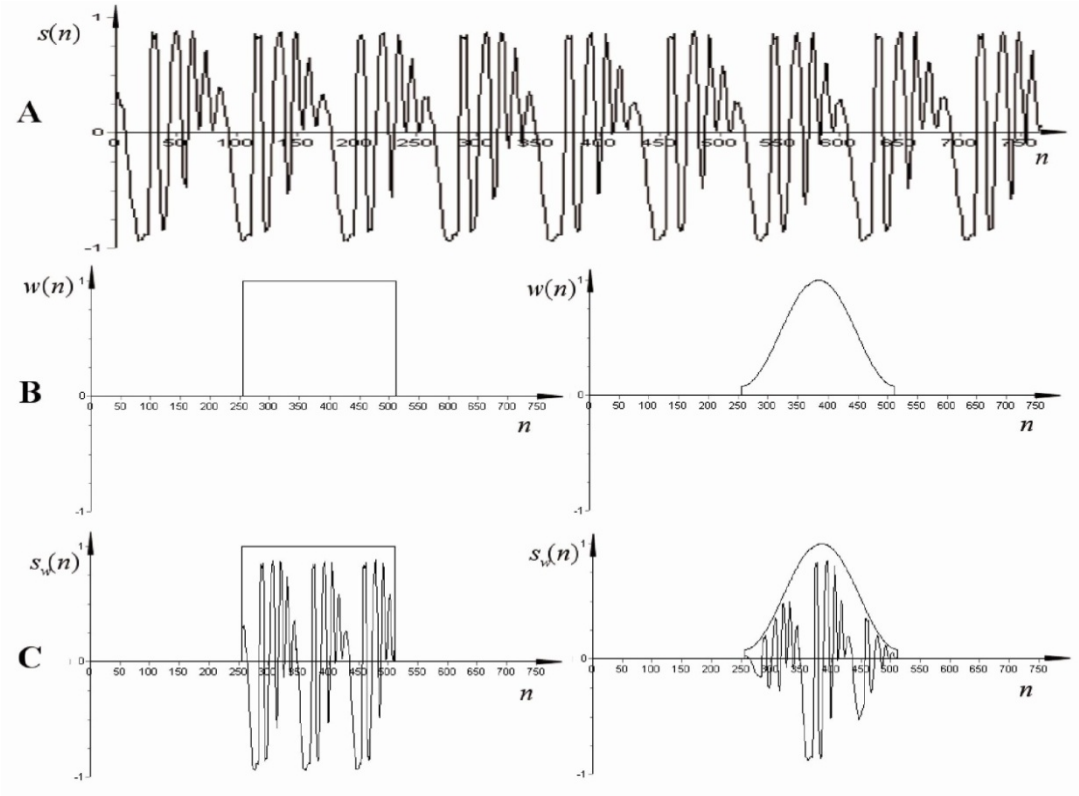


Figure: 2.10 An example of windowing. A) original signal, B) rectangular window (left) and Hamming window (right), and C) influence of a rectangular window (left) and a Hamming window (right).

The most frequently used window type is a Hamming window[13] defined as

$$w(n) = \begin{cases} 0{,}54 - 0{,}46 \cdot \cos\left(2\pi \dfrac{n}{N}\right), n \notin \langle 1, N \rangle \\ 0, n \notin \langle 1, N \rangle \end{cases}, \tag{2.8}$$

Where N is length of the window. In the Figure 2.10 you can see influence of a rectangular and a Hamming window on the original signal. In this case, the length of the original signal is different from the length of the window, i.e. the window is shifted relatively to the beginning of the signal.

2.8.4 Pre-emphasis

Pre-emphasis [9] is processing of the input signal by a low order digital FIR filter. This filter is usually the first order FIR filter defined as

$$S_p(n) = S(n) - \lambda s(n-1) \qquad (2.9)$$

Where **λ** is a pre-emphasis coefficient lying usually in an interval of (0.9-1), *s{n)* is the original signal, and s_p{n)* is a pre-emphasized signal. The pre-emphasis is used to flatten spectrally the input signal in favor of vocal tract parameters. It makes the signal less susceptible to later finite precision effects [22].

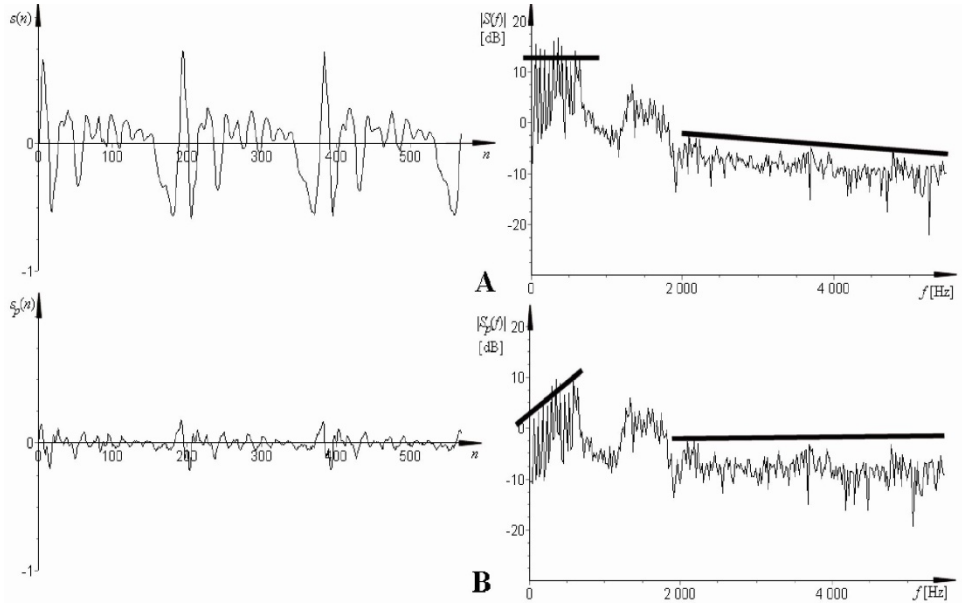


Figure: 2.11 Influence of the pre-emphasis. A) an original signal, B) the same signal after the pre-emphasis.

Effect of the pre-emphasis is obvious in the Figure 2.11, where a waveform of a vowel /ah/ as in the word *cut* is drawn with the corresponding frequency spectrum (A) and a waveform of the same vowel after the pre-emphasis with the

corresponding frequency spectrum (B). Changes in the frequency spectrum are emphasized by the thick black lines.

2.9 Features Extraction

Feature extraction is a crucial phase of the speaker verification process. Well-chosen feature set can result in quality recognition as well as wrongly chosen feature set can result in a poor recognition. Many features described in the following are originally defined for an infinite continuous signal. This is not useful for the real applications. Most of the real applications work with a finite discrete signal. All the equations in this work are defined to be valid for the finite discrete signal. You can see the original definitions in the quoted references [28].

2.9.1 Energy (E)

Energy of a signal expresses strength of the signal. Its value is usable to voice activity detection, because the energy of the voice is higher than the energy of the noise (which is not valid in case of consonants – they are more similar to the noise). Definition of the energy of a finite discrete signal is

$$E = \sum_{n=1}^{Ntotal} S^2(j,n) \tag{2.10}$$

Where N_{total} is length of the whole finite discrete signal $s(n)$. Energy calculated this way is energy of the whole signal so that it is called Long-term Energy (LTE). Given the signal divided into J frames N samples long we can define energy of one frame as

$$E(j) = \sum_{n=1}^{N} S^2(j,n) \tag{2.11}$$

where $s(j,n)$ is the j-th frame of the original signal $s(n)$ and $j \in \{1,2,...,J\}$. Energy of one frame is called Short-Time Energy (STE).

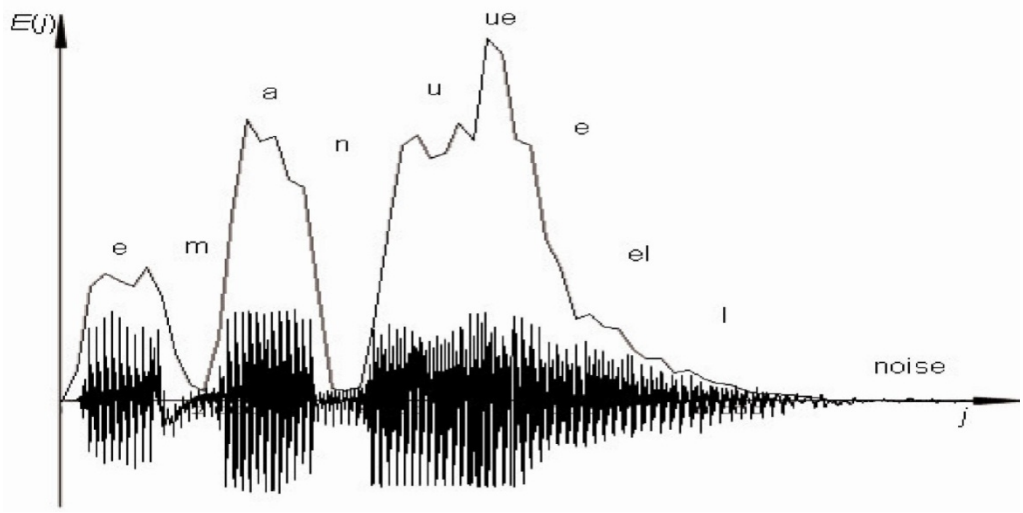


Figure: 2.12 Shape of the short-time energy of the frames of a Czech word 'Emanuel'.

Figure 2.12 shows a shape of the short-time energy of all the frames as they go one after another in a Czech word 'Emanuel'. You can distinguish quite obviously the background noise and the consonants from the vowels. Energy of the background noise depends on the quality of the microphone used to record the signal [9]. The lower the signal-to-noise ratio of the microphone, the better can we distinguish the background noise from the vowels and even from the consonants.

2.9.2 Zero-Crossing Rate (ZCR)

Zero-Crossing Rate (ZCR) expresses how many times crosses the signal zero. Value of the zero-crossing rate suits best to distinguish noise from the vowels or the consonants. The zero-crossing rate of the noise is usually much higher than the zero-crossing rate of the vowels. However, the zero-crossing rate of some consonants is high as well, which makes the recognition of consonants more difficult. The zero-crossing rate is formulate as

$$Z(j) = \frac{1}{2}\sum_{n=1}^{N-1} | \, sign(S((j,n)) - sign(S((j,n+1)) \tag{2.12}$$

Where N is length of the j-th frame of a signal s(n), j∈ {1,2,...,J}. The function sign(s(n)) is defined as

$$sign(s(n)) = \begin{cases} +1, when s(n) > 0 or s(n) = 0 and s(n-1) > 0 \\ -1, when s(n) < 0 or s(n) = 0 and s(n-1) < 0 \end{cases} \quad (2.13)$$

Figure 2.13 shows an example of a shape of the short-time zero-crossing rate of all frames as they go one after another in a German word 'Wiesbaden'. You can clearly see high values of the zero-crossing rate of the noise and lower values of some phones.

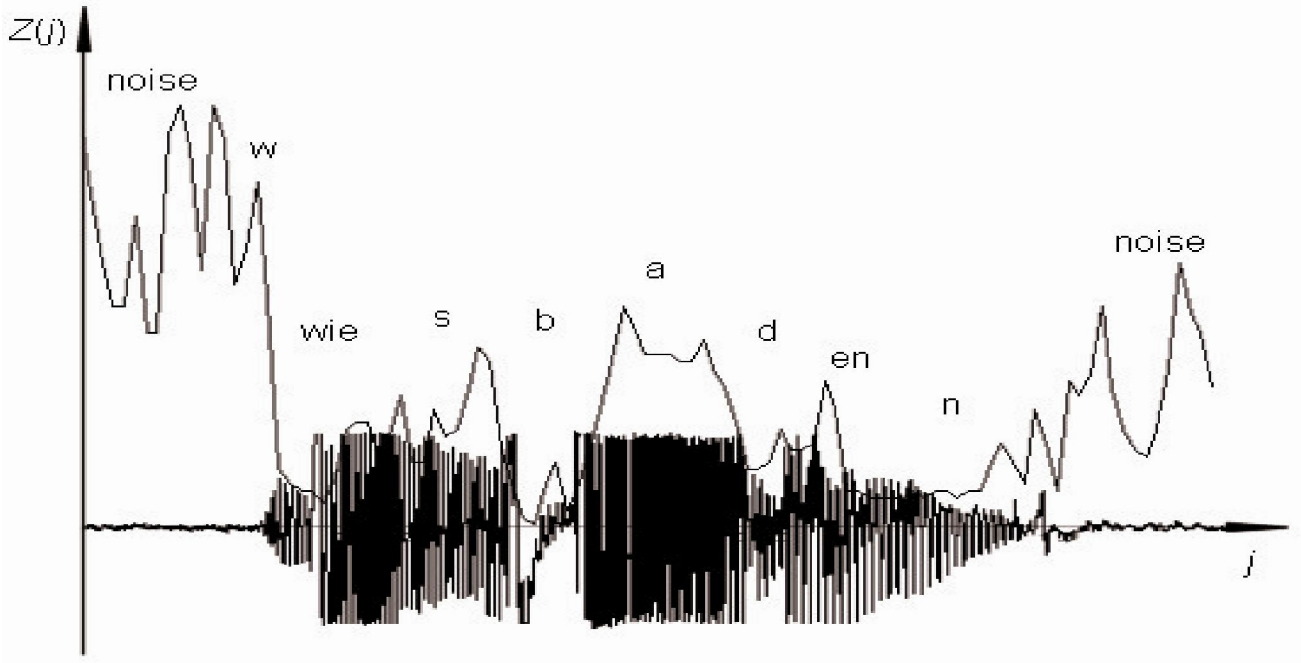


Figure: 2.13 Shape of the short-time zero-crossing rate of the frames as in a German word 'Wiesbaden'**.**

2.9.3 Autocorrelation

Autocorrelation is a mathematical tool used frequently in the signal processing for analyzing functions or series of values (time domain signals). It is the cross-

correlation of a signal with itself. Autocorrelation is useful for finding repeating patterns in a signal, such as determining the presence of a periodic signal, which has been buried under noise, or identifying the fundamental frequency of a signal, which does not actually contain that frequency component, but implies it with many harmonic frequencies [28]. The continuous autocorrelation $R_c(\tau)$ is the continuous cross-correlation of the function f(t) with itself, at lag τ, and is defined as

$$R_c(\tau) = f^*(-\tau) * f(\tau) = \int_{-\infty}^{\infty} f(t+\tau) f^*(t) dt \qquad (2.14)$$

Where f*(t) represents the complex conjugate. For a real function, f(t) = f(t). Usually, we work with a sampled signal, i.e. with a finite discrete signal. Hence, the discrete autocorrelation R(k) of the discrete finite signal s{n) at t he lag k is formally defined as

$$R(k) = \sum_{n=1}^{N-k} (s(n) - \mu)(s(n+k) - \mu) \qquad (2.15)$$

Where μ is the mean value of the discrete finite signal s{n), and N is length of the signal. In the speech processing the first 12-32 autocorrelation coefficients are used. Usually, the order of the autocorrelation is approximately M =Fs +4 , where Fs is the sampling frequency given in kHz.

2.9.4 Linear Predictive Coding (LPC)

The coefficients of the Linear Predictive Coding (LPC) play an important role in the speech signal processing. An approximation S_{pred} (n) of a speech signal s(n) can be calculated by a linear combination of the LP coefficients and M previous samples of the original signal s(n) (autoregressive model). This approximation is called linear prediction of the original signal[28]. The linear prediction is defined as

$$S_{pred}(n) = -\sum_{m=1}^{M} a(m) s(n-m) \qquad (2.16)$$

where a(m), m =1,2……M , are the LPC coefficients, M is number of the LPC coefficients called order of the linear prediction, s(n) is the original signal.

2.9.5 Discrete Fourier Transform (DFT)

The Fourier Transform is very useful one, because it uses complex exponentials as its basis functions. A digital system *T* is a system that, given an input signal *s{n),* generates an output signal.

$$y\{n) = T\{s\{n)\} \quad (2.17)$$

The system T is shift-invariant (or, in terms of the continuous signal, time-invariant), then

$$y\{n-n_0) = T\{s\{n-n_0)\} \quad (2.18)$$

Linear digital systems, so-called *linear shift-invariant* (LSI) systems, are described as

$$y(n) = \sum_{-\infty}^{\infty} s\{n)h\{n-k) \; = \; s\{n) * h\{n) \quad (2.19)$$

Where * denotes the convolution operator. The convolution operator is commutative, associative, and distributive. The LSI systems are characterized by the impulse response *h{n)* of the system.

Given a Fourier transform *F* and its inverse F^{-1}, we can define cepstrum coefficients. The cepstrum coefficients are given as the inverse Fourier transform of the log-energy of the Fourier transform of a signal *s(n)*:

$$c(\nu) = F\text{-}1\{\log |F\{s(w)\}|2 \} \quad (2.20)$$

The cepstrum coefficients can be used for some speaker recognition specific purposes. It is useful for pitch estimation, since at the position of the pitch and at the positions of its harmonics there are peaks. Actually, the coefficients of the cepstrum can be used to separate the vocal tract impulse response and the generating signal. The vocal tract response could be used for the speaker recognition, since they are speaker dependent. However, there is a question asking how many of the cepstrum coefficients should be used for this purpose. These coefficients were tested together with the other ones[5].

2.9.6 Discrete Cosine Transform (DCT)

The Discrete Cosine Transform (DCT) is widely used for the speech processing. It is so often used because of its energy compaction, which results in its coefficients being more concentrated at low indices than the coefficients of the DFT. This allows us to approximate a signal using fewer coefficients. There are several definitions of the DCT. Coefficients *C(k)* of one of them are defined as

$$C(k) = \sum_{n=0}^{N-1} s(n) \cos\left(\pi k \frac{n=1}{2N} \right), \quad k = 0,1.....N-1 \tag{2.21}$$

Where *s(n)* is a real signal. Its inverse is given by

$$s(n) = \frac{1}{N}\left[C(0) + 2\sum_{k=1}^{N-1} C(k) \cos\left(\pi k \frac{n+1}{2N} \right) \right], n = 0,1.....N-1 \tag{2.22}$$

The DCT-II can be derived from the DFT when assuming *s(n)* is a real periodic signal with a period of *2N* and with an even symmetry *s(n)* = *s(2N-1-n)*.The DCT is used for computing of the Mel-Frequency Cepstrum Coefficients or the Speaker Dependent Cepstrum Coefficients.

2.9.7 Mel-Frequency Cepstrum Coefficients (MFCC)

Some psychoacoustic experiments were undertaken to derive scales attempting to model the natural response of the human perceptual system, since the cochlea of the inner ear acts as a spectrum analyzer. Fletcher's work pointed to the existence of critical bands in the cochlear response. One class of critical band scales is called Bark frequency scale. It ranges from 1 to 24 Barks, corresponding to 24 critical bands of hearing. Another scale like the Bark frequency scale is the Mel-frequency scale, which is linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above 1 kHz [28]. The Mel-scale is based on experiments with simple sinusoidal tones. It can be approximated by

$$B(f) = 1125 \cdot \ln\left(1 + \frac{f}{700} \right) \tag{2.23}$$

Then its inverse is given by

$$B^{-1}(b) = 700 \cdot \left(e^{\frac{b}{1125}} - 1 \right) \tag{2.24}$$

The Mel-Frequency Cepstrum Coefficients (MFCC) are defined as the real cepstrum of a windowed short-term signal derived from the FFT of that signal. It differs from the real cepstrum in the nonlinear frequency scale used for approximation of the behaviour of the auditory system.

Given the DFT of an input signal defined, we define a filter bank with I filters. The i_{th} filter, i =1, 2, K, I, is defined as

$$H(i,f)\begin{cases} 0, & f < f(i-1) \\ 2\dfrac{f - f(i-1)}{(f(i+1) - f(i-1)(f(i) - f(i-1))}, & f(i-1) \le f \le f(i) \\ 2\dfrac{f(i+1) - k}{(f(i+1) - f(i-1)(f(i+1) - f(i))}, & f(i) < f \le f(i+1) \\ 0, & f > f(i+1) \end{cases} \tag{2.25}$$

These filters compute the average spectrum around each central frequency with increasing bandwidths. The boundary frequencies f (i), $0 \le i \le I + 1$, are uniformly spaced in the Mel-scale:

$$F(i) = \left(\frac{N}{F_s} \right) B^{-1} \left(B(F_{low)} \cdot i \frac{B(F_{high}) - B(F_{low})}{I+1} \right) \tag{2.26}$$

Where N is the size of the FFT, i.e. it is the length of one frame, I is the count of the filters, F_s is the sampling frequency in Hz, F_{low} and F_{high} are the lowest and the highest frequencies of the filter bank in Hz, B is given, and its inverse B^{-1} is given then, we can express the log-energy at the output of each filter as

$$C(i) = \ln\left(\sum H(i,k) \cdot |S(k)|^2 \right), \quad i = 1,2....., I \tag{2.27}$$

Where S(k) is the magnitude of the FFT of the signal s(n), which is N samples long. The Mel-frequency Cepstrum is the discrete cosine transform of the I filter outputs

$$C(j) = \sum_{i=0}^{I-1} C(i).\cos\left(\pi n \frac{i-1}{2I}\right), j = 0,1....., I \qquad (2.28)$$

The count I of the filters varies from 24 to 40 in different applications. However, for the speech recognition only the first 13 coefficients are used.

A short-term frame of 25 ms multiplied by the Hamming window is typically used to calculate the MFCC and the delta coefficients. The frame overlapping is usually 10 ms.

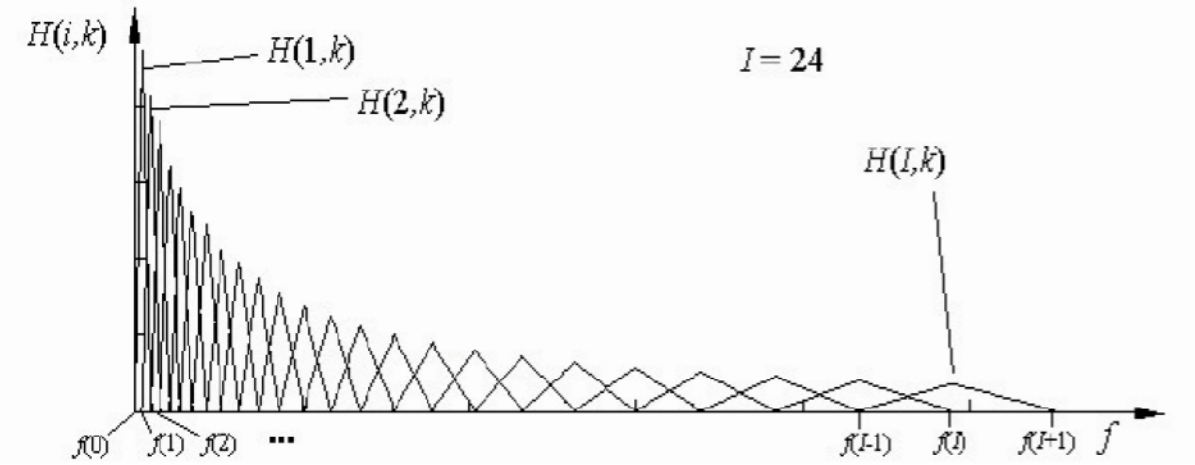


Figure: 2.14 Example of a filter bank used for computation of the MFCC. In this case there are I = 24 triangular filters obtained.

2.9.8 Average Long-Term LPC Spectrum

Long-term statistics of many various features are used often to recognize speech. However, not only the area of the speech recognition is a domain of the long-term statistics. The average long-term LPC spectrum is applicable to the speaker verification as well.

2.10 Pattern Matching

After the input signal (a speech signal) was processed and proper features were extracted, it is possible to come to the next important phase of the speaker recognition process – the pattern matching. The pattern matching can be accomplished in many ways. There are many techniques – nonparametric techniques (nearest-neighbour classification, fuzzy

classification, reduced coulomb energy networks etc.), methods based on maximum-likelihood and Bayesian parameter, linear Discriminant functions, neural networks, or stochastic methods. For this task a method based on estimation of the maximum-likelihood was chosen – the Gaussian Mixtures Models (GMM).

2.10.1 Nearest-Neighbor Modeling

Nearest-neighbor models have been popular in nonparametric classification [21]. This approach is often thought of as estimating the local density of each class by a Parzen estimate and assigning the test vector to the class with the maximum local density. The local density of a class (speaker) with enrollment data X at a test vector y is defined as

$$P_{nn}(y, X) = \frac{1}{V[d_{nn}(y,X)]} \quad (2.29)$$

Where d_{nn}(y, X) = minxj∈X _y−xj_ is the nearest neighbor distance and V(r) is the volume of a sphere of radius r in the D-dimensional feature space.

The log-likelihood score of the test utterances Y with respect to a speaker specified by enrollment X is given by

$$S_{nn}(Y; X) \approx -\sum \ln [d_{nn}(y, X)] \quad (2.30)$$

And the speaker with the greatest s(Y; X) is identified.

2.10.2 Vector Quantization Modeling

Vector quantization constructs a set of representative samples of the target speaker's enrollment utterances by clustering the feature vectors. Although a variety of clustering techniques exist, the most commonly used isk-means clustering [36]. This approach partitions N feature vectors into K disjoint subsets Sj to minimize an overall distance such as

$$D = \sum_{j=1}^{J} \sum(x_i - \mu_j) \quad (2.31)$$

Where μj = (1/Nj)_xi∈Sj xi is the centroid of the Nj samples in the j-th cluster.

The algorithm proceeds in two steps:

1. Compute the centroid of each cluster using an initial assignment of the feature vectors to the clusters.

2. Reassign xi to that cluster whose centroid is closest to it. These steps are iterated until successive steps do not reassign samples.

Once the VQ models are established for a target speaker, scoring consists of evaluating D in (36.11) for feature vectors in the test utterance. This approach is general and can be used for text-dependent and text independent speaker recognition, and has been shown to be quite effective [15]. Vector quantization models can also be constructed on sequences of feature vectors, which are effective at modeling the temporal structure of speech.

2.10.3 Gaussian Mixture Models

In the case of text-independent speaker recognition where the system has no prior knowledge of the text of the speaker's utterance, Gaussian mixture models (GMMs) have proven to be very effective. This can be thought of as a refinement of the VQ model. Feature vectors of the enrollment utterances X are assumed to be drawn from a probability density function that is a mixture of Gaussians given by

$$P(x|\lambda) = \sum_{k=1}^{K} w_k P_k(x|\lambda_k) \tag{2.32}$$

$$P(x \mid \lambda_k) = \frac{1}{(2\pi)^{D/2} \mid \sum_k \mid^{1/2}} e^{-(1/2)x-\mu k)' \sum_k^{-1} (x-\mu k)} \tag{2.33}$$

λ represents the parameters (μ_i, Σ_i ,w_i) of the distribution. Since the size of the training data is often small, it is difficult to estimate full covariance matrices reliably. In practice, $\{\Sigma_k\}$ are assumed to be diagonal. Given the enrollment data X, the maximum likelihood estimates of the λ can be obtained using the expectation-maximization (EM) algorithm [23]. The K-means algorithm can be used to initialize the parameters of the component densities. The posterior probability that x_t is drawn from the component $p_m(x_t|\lambda_m)$ can be written

$$P_m(m|x_t, \lambda) = \frac{w_m P_m(x_t|\lambda_m)}{P(x_t|\lambda)} \tag{2.34}$$

The two steps of the EM algorithm consist of computing $P(_m|x_t, \lambda)$ given the current model, and updating the model using the equations above. These two steps are iterated until a convergence criterion is satisfied. Test utterance scores are obtained as the average log-likelihood given by

$$S(Y|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log [P(y_t|\lambda)] \quad (2.35)$$

Speaker identification is often based on maximum similarity means maximum loglikehood with the database is the result.

2.10.4 Hidden Markov Models

Hidden Markov models (HMMs) [23] for phones, words, or phrases, have been shown to be very effective [2]. Passwords consisting of word sequences drawn from specialized vocabularies such as digits are commonly used. Each word can be characterized by an HMM with a small number of states, in which each state is represented by a Gaussian mixture density. The maximum-likelihood estimates of the parameters of the model can be obtained using a generalization of the EM algorithm [23].

The Maximum Likelihood (ML) training aims to approximate the underlying distribution of the enrollment data for a speaker. The estimates deviate from the true distribution due to lack of sufficient training data and incorrect modeling assumptions. This leads to a suboptimal classifier design. Some limitations of ML training can be overcome using discriminative training of speaker models in which an attempt is made to minimize an overall cost function that depends on misclassification or detection errors. Discriminative training approaches require examples from competing speakers in addition to examples from the target speaker. In the case of closed-set speaker identification, it is possible to construct a misclassification measure to evaluate how likely a test sample, spoken by a target speaker, is misclassified as any of the others.

2.10.5 Support Vector Modeling

Traditional discriminative training approaches such as those based on MCE have a tendency to over train on the training set. The complexity and generalization ability

of the models are usually controlled by testing on a held-out development set. Support vector machines (SVMs) [27] provide a way for training classifiers using discriminative criteria and in which the model complexity that provides good generalization to test data is determined automatically from the training data. SVMs have been found to be useful in many classification tasks including speaker identification. The original formulation of SVMs was for two-class problems. This seems appropriate for speaker verification in which the positive samples consist of the enrollment data from a target user and the negative samples are drawn from a large set of imposter speakers. Many extensions of SVMs to multi class classification have also been developed and are appropriate for speaker identification. There are many issues with SVM modeling for speaker recognition, including the appropriate choice of features and the kernel.

2.10.6 Other Approaches

Most state-of-the-art speaker recognition systems use some combination of the modeling methods described in the previous sections. Many other interesting models have been proposed and have been shown to be useful in limited scenarios. Eigen voice modeling is an approach in which the speaker models are confined to a low dimensional linear subspace obtained using independent training data from a large set of speakers. This method has been shown to be effective for speaker modeling and speaker adaptation when the enrollment data is too limited for the effective use of other text-independent approaches such as GMMs [29]. Artificial neural networks have also been shown to be useful in some situations, perhaps in combination with GMMs. When sufficient enrollment data is available, a method for speaker detection that involves comparing the test segment directly to similar segments in enrollment data has been shown to be effective [6].

2.11 Decision-Making

Decision-making is the final step of the speaker recognition. From the previous steps, we have the result of the pattern matching. Now, we have to decide, whether the result of the matching is positive or negative.

The decision-making can be done in many ways, but every-time it finishes by a simple threshold based (linear) classification. You can have complex pattern matching systems, sophisticated tools, but the recognition must end with a final answer - ‘yes’ (for the positive recognition - acceptance) or ‘no’ (for the negative recognition - rejection). There are only two possible answers. Hence, there is no other way than a threshold classification.

Chapter 3 Speaker Identification System

3.1 Introduction

Speaker Identification is defined as deciding if a speaker a specific person or is among a group of persons. This is different than the speaker verification problem, which is deciding if a speaker is who he claims to be. Automatic text independent speaker identification technology is intended for automatic identification of a speech signal of unknown voice by paired comparing with 'speaker', existing in the database of system. Comparison is conducted by calculation of 'true' and 'false' spots (spots of correspondences) and with the further determination of probability of Accuracy. Each speaker card besides information about current speaker (first, last name, birthday, gender, and so on) is characterized by examples of audio files with the speaker voice.

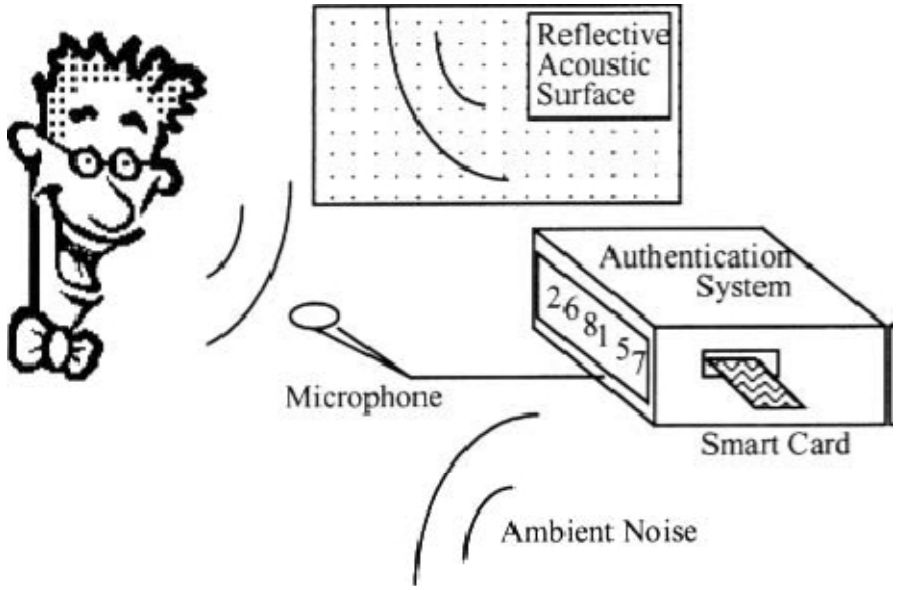


Figure: 3.1 A typical Speaker Identification System.

A typical Automatic Speaker Identification ASI setup is shown in Figure 3.1. The speaker, who has previously enrolled in the system, presents an encrypted smart card containing his identification information. He then attempts to be authenticated by speaking a

prompted phrase(s) into the microphone. In addition to his voice, ambient room noise and delayed versions of his voice enter the microphone via reflective acoustic surfaces. Prior to a identification session, users must enroll in the system (typically under supervised conditions). During this enrollment, voice models are generated and stored (possibly on a smart card) for use in later identification sessions.

This signal is analyzed by an identification system that makes the decision if a speaker a specific person or is among a group of persons or possibly to report insufficient confidence and request additional input before making the decision.

3.2 Hardware Requirements

The system comprises of mostly software portion but had some hardware involved too. The hardware that has been used was:

- Microphone
- Sound Card
- P.C.

3.2.1 Microphones

A quality microphone is a key, when utilizing Automatic Speaker Recognition (ASR). In most cases, a desktop microphone just won't do the job. They tend to pick up more ambient noise that gives ASR programs a hard time. So choice of microphone is critical for the success of ASR system.

Hand held microphones are also not the best choice as they can be cumbersome to pick up all the time. While they do limit the amount of ambient noise, they are most useful in applications that require changing speakers often, or when speaking to the recognizer isn't done frequently (when wearing a headset isn't an option).

The best choice and by far the most common is the headset style. It allows the ambient noise to be minimized, while allowing having the microphone at the tip of one’s tongue all the time. Headsets are available without earphones and with earphones (mono or stereo).

3.2.2 Sound Cards

Because speech requires a relatively low bandwidth, just about any medium-high quality 16-bit sound card will do the job. Sound must be enabled in kernel, and must have correct drivers installed. Sound card quality often starts a heated discussion about their impact on accuracy and noise.

Sound cards with the 'cleanest' A/D (analog to digital) conversions are recommended, but most often the clarity of the digital sample is more dependent on the microphone quality and even more dependent on the environmental noise. Electrical "noise" from monitors, PCI slots, hard-drives, etc. is usually nothing compared to audible noise from the computer fans, squeaking chairs, or heavy breathing.

Some ASR software packages may require a specific sound card. It's usually a good idea to stay away from specific hardware requirements, because it limits many of possible future options and decisions. One will have to weigh the benefits and costs if it considering packages that require specific hardware to function properly.

A sound card creates a sound file in wav format from the data input through the microphone. The process of converting that data into a file to be recorded to the hard disk is:

- The sound card receives a continuous, analog-waveform-input signal from the microphone jack. The analog signals received vary in both amplitude and frequency.
- Software in the computer selects which input(s) will be used.
- The mixed, analog waveform signal is processed in real-time by an analog-to-digital converter (ADC) circuit chip, creating a binary digital output of 1s and 0s.
- The digital output from the ADC flows into the DSP. The DSP is programmed by a set of instructions stored on another chip on the sound card. One of the functions of the DSP is to compress the now-digital data in order

to save space. The DSP also allows the computer's processor to perform other tasks while this is taking place.

- The output from the DSP is fed to the computer's data bus by way of connections on the sound card (or traces on the motherboard to and from the sound chipset).
- The digital data is processed by the computer's processor and routed to the hard-disk controller. It is then sent on to the hard-disk drive as a recorded 'wav' file.

3.2.3 Computers/Processors

ASR applications can be heavily dependent on processing speed. This is because a large amount of digital filtering and signal processing can take place in ASR. It's possible to do some Speaker Recognition with 100MHz and 16M RAM, but for fast processing (large dictionaries, complex recognition schemes, or high sample rates), it should shoot for a minimum of a 400MHz and 128M RAM. Because of the processing required, most software packages list their minimum requirements. The processor with 1.6 GHz and 1.5GB RAM has been used in the system implementation.

3.3 Software Platform

The software platform used by us was LabVIEW (Laboratory Virtual Instrument Engineering Workbench).

3.3.1 LabVIEW

LabVIEW is a programming environment in which you create programs using a graphical notation (connecting functional nodes via wires through which data flows); in this regard, it differs from traditional programming languages like C, C++, or Java, in which you program with text. However, LabVIEW is much more than a programming language. It is an interactive program development and execution

system designed for people, like scientists and engineers, who need to program as part of their jobs. The LabVIEW development environment works on computers running Windows, Mac OS X, or Linux. LabVIEW can create programs that run on those platforms, as well as Microsoft Pocket PC, Microsoft Windows CE, Palm OS, and a variety of embedded platforms, including Field Programmable Gate Arrays (FPGAs), Digital Signal Processors (DSPs), and microprocessors.

Using the very powerful graphical programming language that many LabVIEW users affectionately call "G" (for graphical), LabVIEW can increase your productivity by orders of magnitude. Programs that take weeks or months to write using conventional programming languages can be completed in hours using LabVIEW because it is specifically designed to take measurements, analyze data, and present results to the user.

LabVIEW offers more flexibility than standard laboratory instruments because it is software-based. You, not the instrument manufacturer, define instrument functionality. Your computer, plug-in hardware, and LabVIEW comprise a completely configurable virtual instrument to accomplish your tasks. Using LabVIEW, you can create exactly the type of virtual instrument you need, when you need it, at a fraction of the cost of traditional instruments. When your needs change, you can modify your virtual instrument in moments.

LabVIEW tries to make your life as hassle-free as possible. It has extensive libraries of functions and subroutines to help you with most programming tasks, without the fuss of pointers, memory allocation, and other arcane programming problems found in conventional programming languages. LabVIEW also contains application-specific libraries of code for data acquisition (DAQ), General Purpose Interface Bus (GPIB), and serial instrument control, data analysis, data presentation, data storage, and communication over the Internet. The Analysis Library contains a multitude of useful functions, including signal generation, signal processing, filters, windows, statistics, regression, linear algebra, and array arithmetic.

3.3.2 How Does LabVIEW Work?

LabVIEW uses terminology, icons, and ideas familiar to scientists and engineers. It relies on graphical symbols rather than textual language to define a program's actions. Its execution is based on the principle of **dataflow**, in which functions execute only after receiving the necessary data. Because of these features, you can learn LabVIEW even if you have little or no programming experience.

A LabVIEW program consists of one or more **virtual instruments (VIs)**. Virtual instruments are called such because their appearance and operation often imitate actual physical instruments. However, behind the scenes, they are analogous to main programs, functions, and subroutines from popular programming languages like C or Basic. Hereafter, LabVIEW program is referred as a "VI". Also, be aware that a LabVIEW program is always called a VI, whether its appearance or function relates to an actual instrument or not.

A VI has two main parts: a **front panel**, a **block diagram.**

- The *front panel* is the interactive user interface of a VI, so named because it simulates the front panel of a physical instrument (see Figure 3.2). The front panel can contain knobs, push buttons, graphs, and many other controls (which are user inputs) and indicators (which are program outputs). You can input data using a mouse and keyboard, and then view the results produced by your program on the screen.
- The *block diagram* is the VI's source code, constructed in LabVIEW's graphical programming language, G (see Figure 3.3). The block diagram is the actual executable program. The components of a block diagram are lower-level VIs, built-in functions, constants, and program execution control structures. You draw wires to connect the appropriate objects together to define the flow of data between them. Front panel objects have corresponding terminals on the block diagram so data can pass from the user to the program and back to the user.

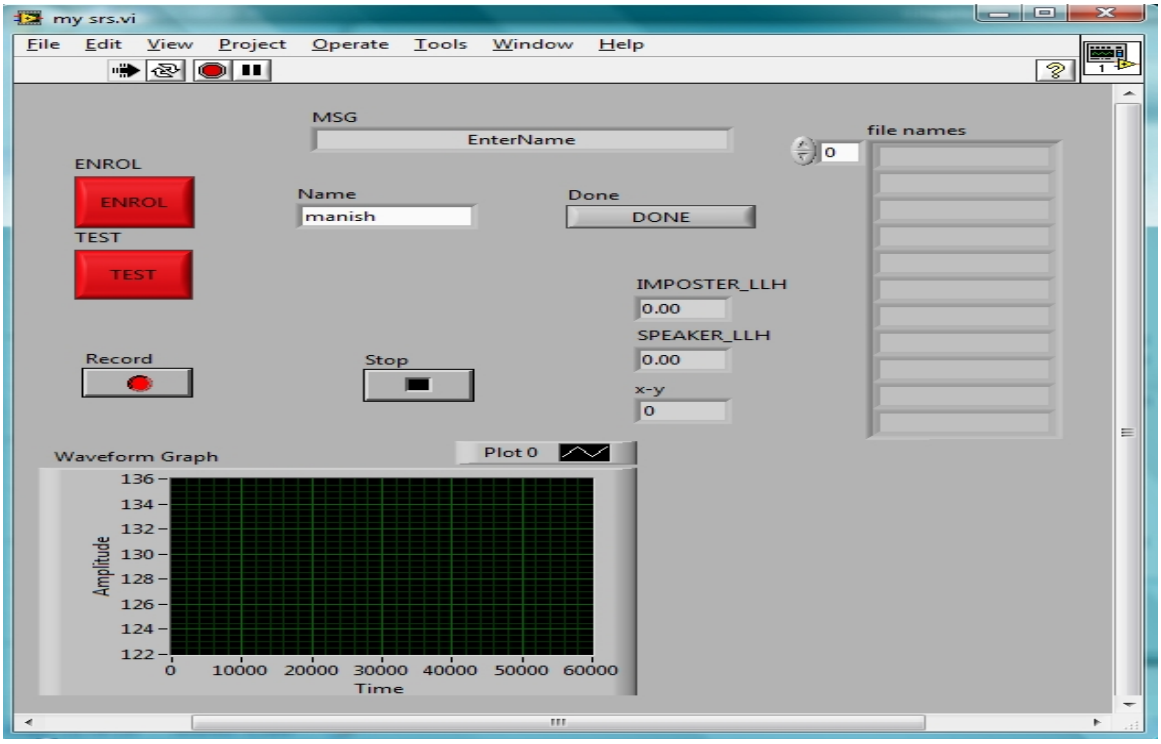


Figure: 3.2 Front Panel of Speaker Identification of System.

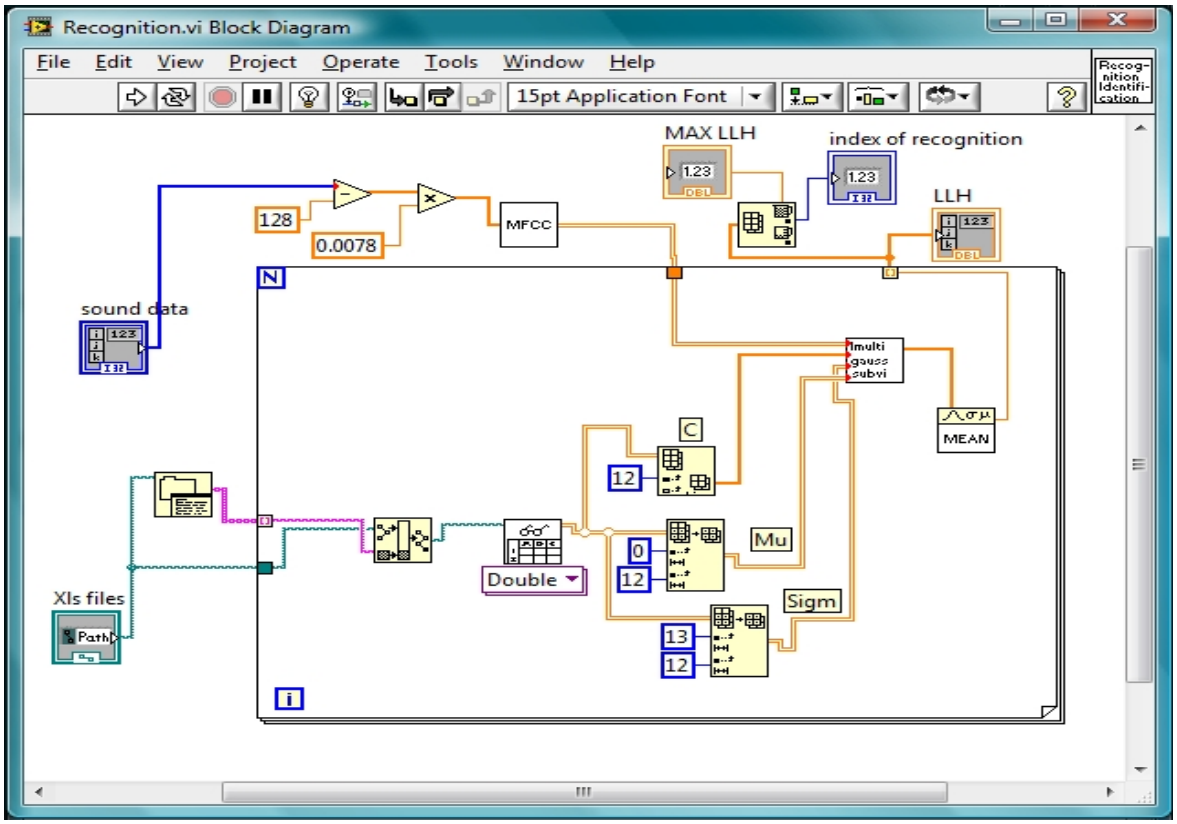


Figure: 3.3 Block Diagram of Speaker Identification of System.

3.4 Database

The Database is made in LabVIEW, which includes both male and female. Twenty university students have been chosen and 5 sample of each were taken. In this way a total of 100 voice sample were used for training and testing. In which Recording and Stop button is used to start and stop recording. Figure 3.4 and 3.5 show the Block diagram and front panel of the VI sound recording.

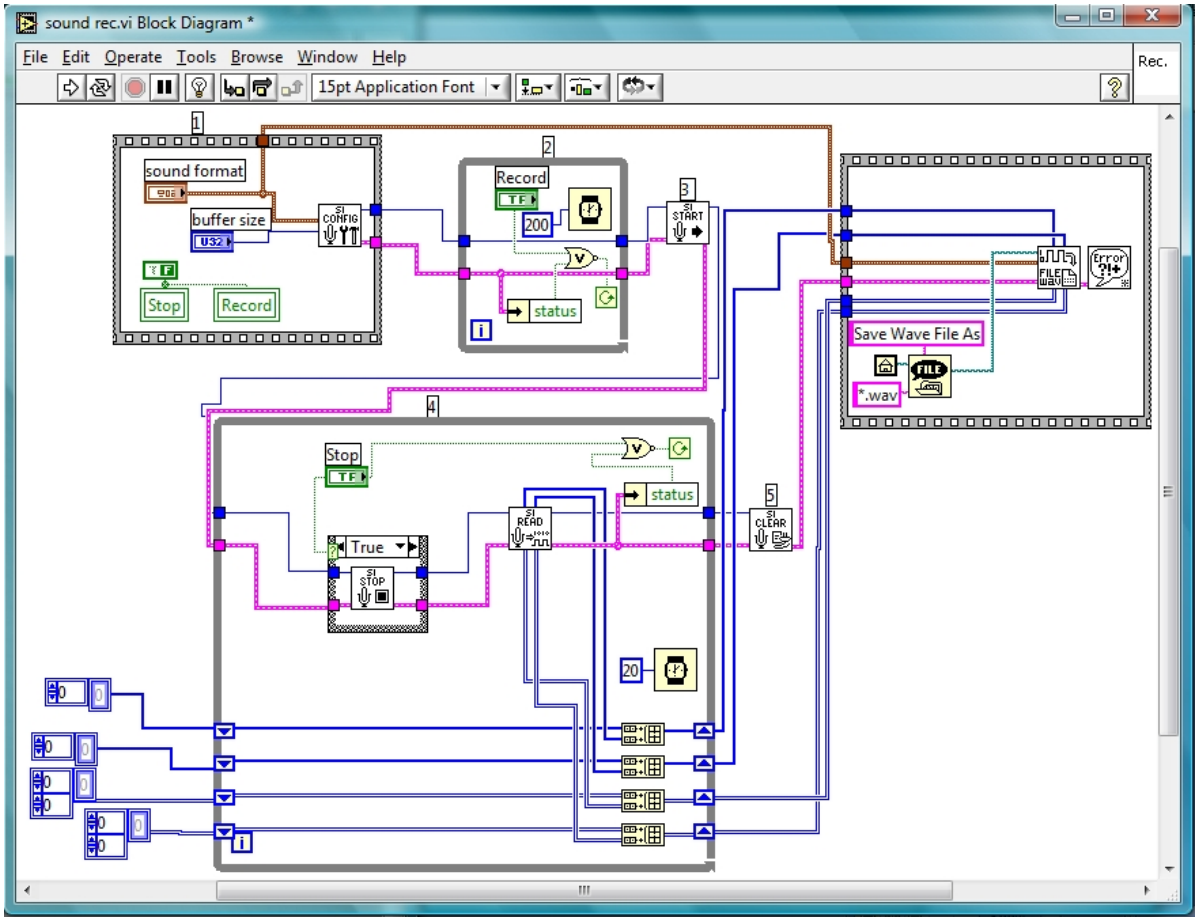


Figure: 3.4 Block Diagram of VI sound recording.

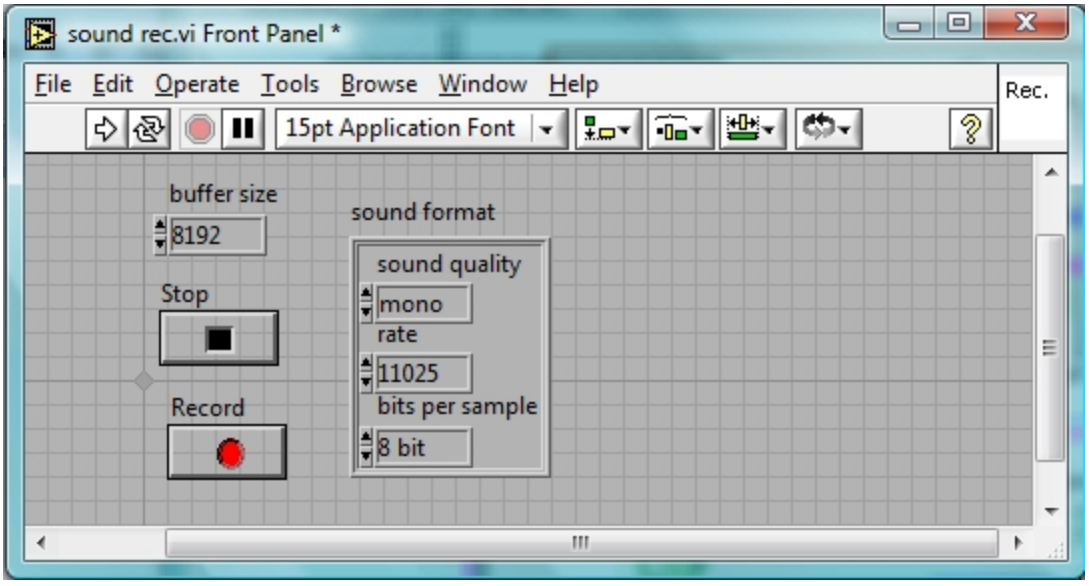


Figure: 3.5 Front Panel Diagram of VI sound recording.

The various steps of recording the speech signal are:

i) Configure the Input channel -:

Sound Format: Sound format specifies

a) How the sound quality is set up (Mono or Stereo)

b) Sets its Recording rate (speed-11025, 22050 or 44100) ,and

c) Sets up the sound as 8 or 16-bit sound.

Buffer size: Buffer size is the size of the internal buffer that Lab VIEW uses to transfer data from a device. The default value is 8192 bytes.

ii) Initialize sound recording

iii) Start sound recording

iv) Store data in 20ms. Frame until stop is pressed. Shift registers are provided to accumulate data in 20ms frames.

v) Returns resources used during recording to the system.

vi) Store sound data in wave file.

Configuration used in the system is:

i) Sound Quality : Mono

ii) Recording rate: 11025

iii) Bit per sample : 8-bit

Flow chart for capturing the speech signal is shown in figure 3.6

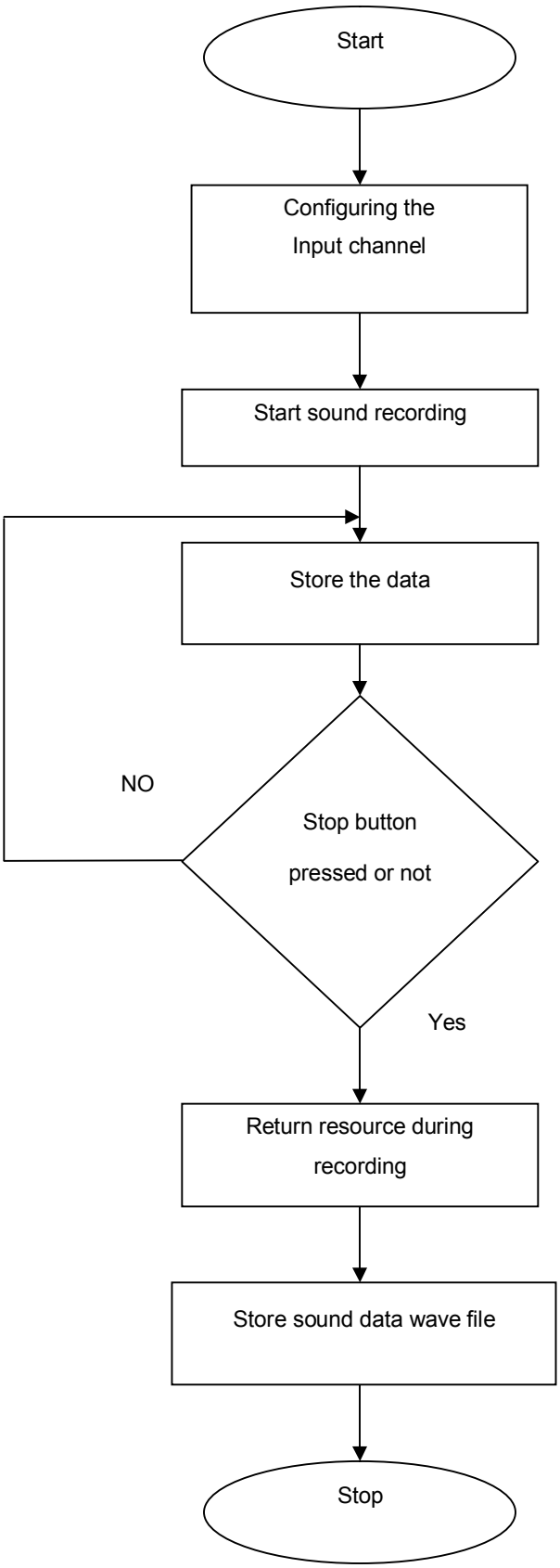


Figure: 3.6 Flow Chart of Capturing the Sound Signal.

3.5 Software Implementation

LabVIEW software of Speaker Identification includes two steps:

1) Registration of the user
2) Testing for Authentication

3.5.1 Registration

Registration is the necessary step in the automatic speaker identification system. In the registration process new user are added to the database. Figure 3.7 shows the complete flow chart of registration process. The various steps for registration are:

3.5.1.1 Name of the user

The process of registration starts by entering the name of the user. After the person enters his name the system checks if the name previously exists in the database. If a match is found then the system prompts to re-enter the name. This process is continued until the name has no match in the database. Figure 3.4 shows the flow chart of Enrollment process.

3.5.1.2 Biometric Signature of the user i.e. capturing the speech signal

The system then asks for voice print, of the user to be passed on to the system. This is done using a microphone. Capturing the speech signal has already been explained in this chapter.

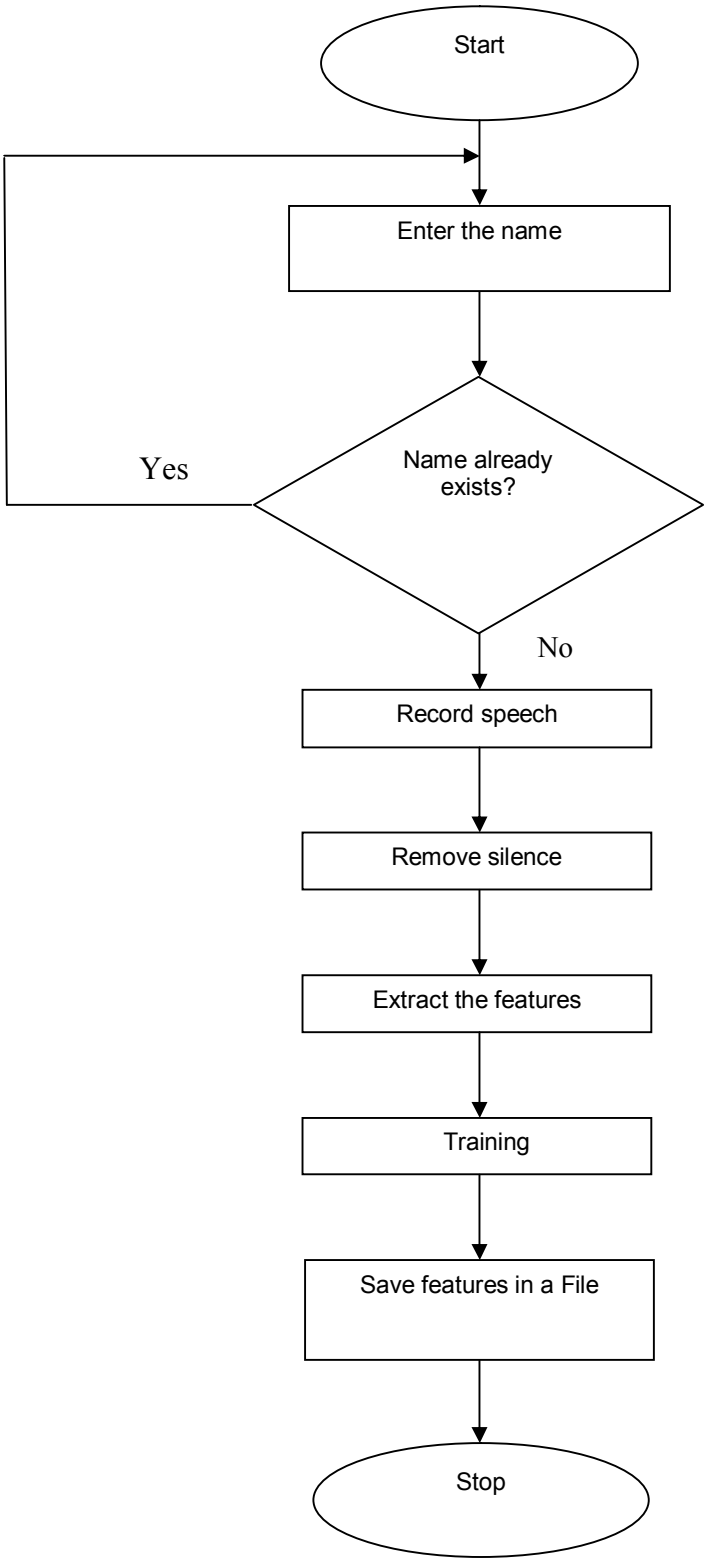


Figure: 3.7 Flow Chart of Registration of a Person.

3.5.1.3 Silence Removing

This function removes the silence intervals from the input speech based on an envelope threshold. The input signal is up-sampled, segmented to remove samples that fall below a threshold, and then re-sampled back to the original sampling rate, and filtered to smooth out the discontinuities where pauses in active speech occurred. The threshold used is a scaled function of the median of the envelope. The default threshold is one-fourth of the median of the envelope. The Block diagram of VI Remove silence is shown in figure 3.8

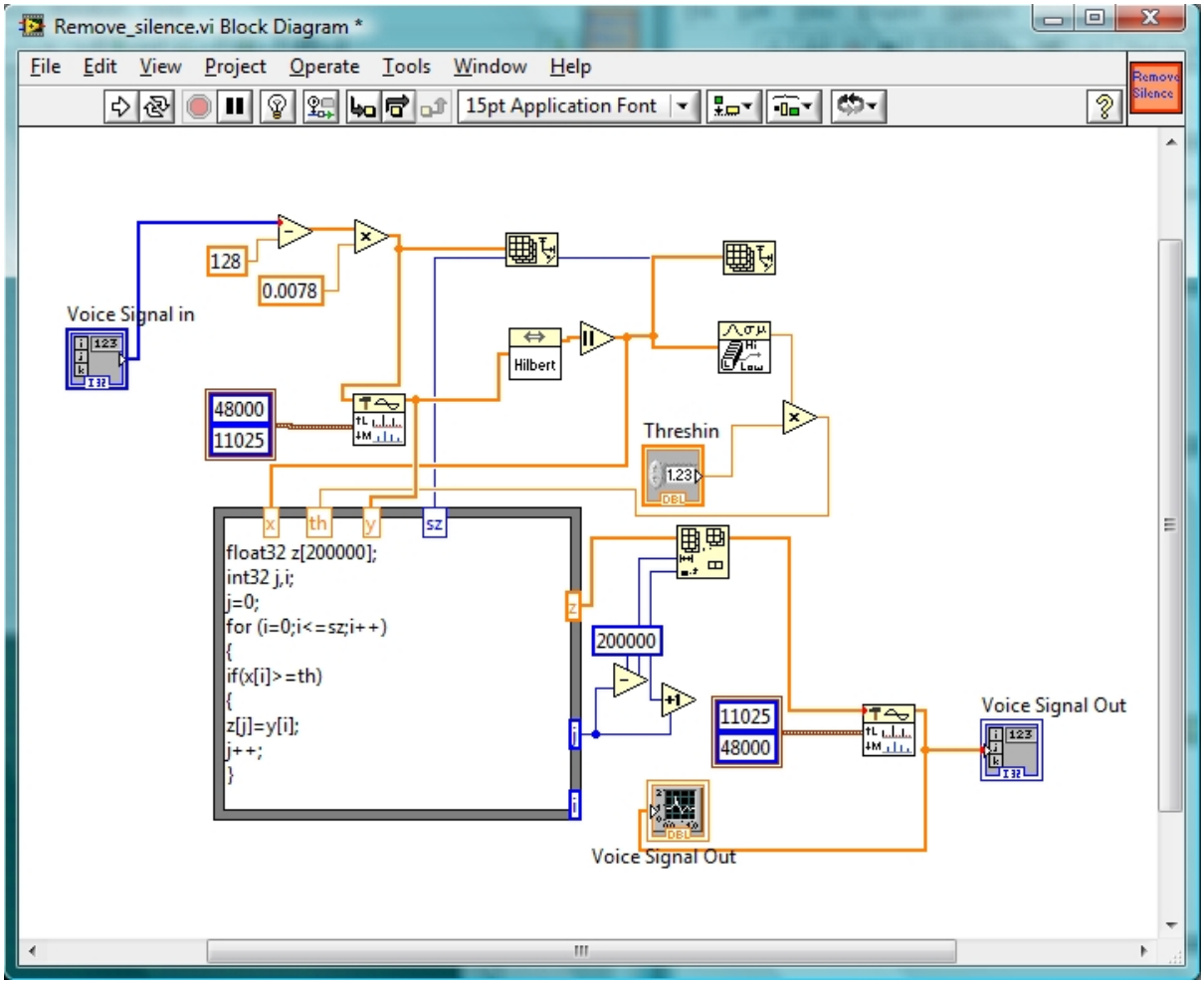
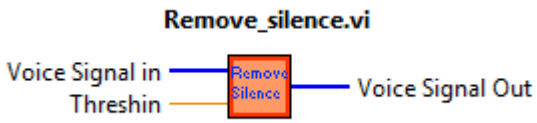


Figure: 3.8 Block diagram of VI remove silence.



The Various subvi's used in silence removing function are:

- **Voice Signal in:** Vectorized Speech signal input. If sig has multiple columns the first signal is considered the reference signal in which the pause segments are detected and these are applied to all other columns in the matrix.
- **Threshin:** Scale on threshold applied to the envelope for detecting silence periods. The actual threshold is computed by multiplying Threshin by the median value of the envelope samples. The default is (.25), which is 25% of the median envelope value over input signal.
- **Voice Signal Out:** Vectorized Signal output with silence interval corresponding to the reference signal removed.

Flow chart of removing silence is shown in figure 3.8.

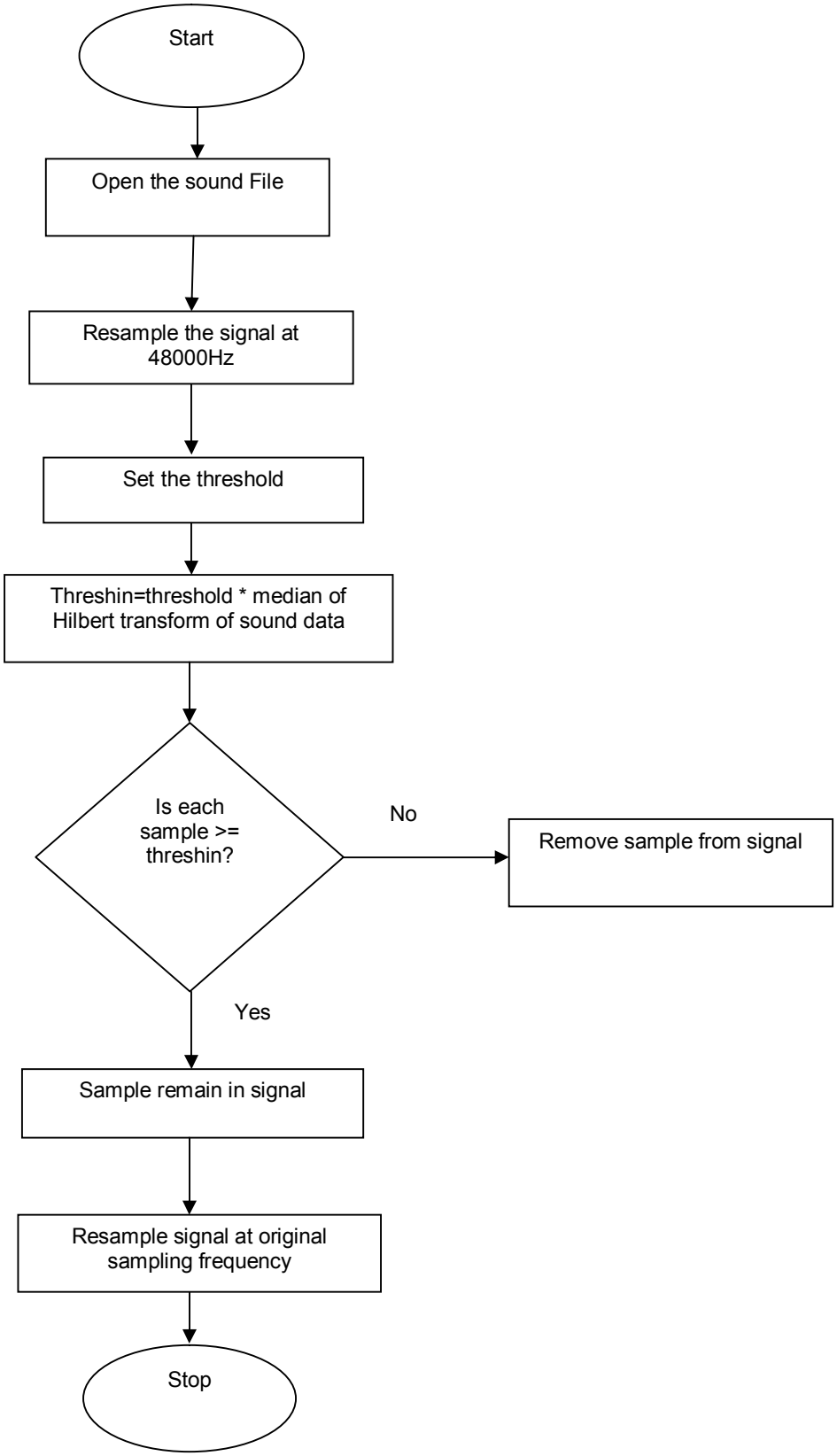


Figure: 3.9 Flow Chart of Removing Silence

3.5.1.4 Extracting the features i.e. Mel-cepstrum of the speech signal

Some psychoacoustic experiments were undertaken to derive scales attempting to model the natural response of the human perceptual system, since the cochlea of the inner ear acts as a spectrum analyzer. Fletcher's work pointed to the existence of critical bands in the cochlear response. One class of critical band scales is called Bark frequency scale. It ranges from 1 to 24 Barks, corresponding to 24 critical bands of hearing. Another scale like the Bark frequency scale is the Mel-frequency scale, which is linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above 1 kHz [28]. The Mel-scale is based on experiments with simple sinusoidal tones. Figure 3.10 shows the block diagram to calculate the MFCC.

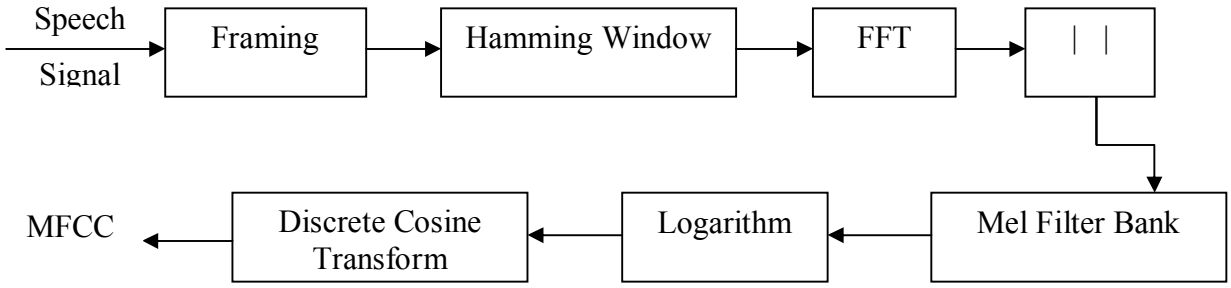


Figure: 3.10 Block diagram to calculate MFCC [37].

The speech waveform is first windowed with the analysis window w[n] &the discrete STFT X (n,ω) is computed as

$$X(n,\omega_k) = \sum_{m=-\infty}^{m=+\infty} x(m) * w(n-m)\ \exp(-j\omega_k m) \qquad (3.1)$$

Where $\omega = \frac{2\pi k}{N}$ with N as the DFT length. The magnitude is then weighted by a series of filter frequency responses whose center & edge frequencies and bandwidths match with those of auditory critical band filters. This has been called as a Mel-Scale filter bank.

Let us now move on with our discussion on Mel-Cepstrum. By weighing the magnitude of STFT with filter bank and each frame is multiplied with entire

filter bank. The next step is to compute energy in STFT weighted by each Mel-scale filter response. The freq response of l^{th} filter is denoted with $V_l(\omega)$. The resulting energies are given as:

$$E_{mel}(n,l) = \frac{1}{A_l} \sum_{k=L_l}^{U_l} \left| V_l(\omega_k) X(n,\omega_k) \right|^2 \qquad (3.2)$$

Where L_l & U_l denote the upper and lower cutoff frequencies of the l^{th} filter and where

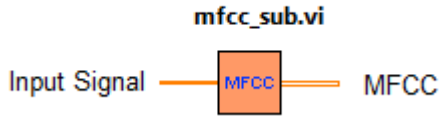
$$A_l = \sum_{k=L_l}^{U_l} \left| V_l(\omega_k) \right|^2 \qquad (3.3)$$

Which normalizes the filters according to their varying bandwidths so to give equal energies for flat input spectrum?

The real cepstrum associated with E_{mel} is known as Mel-cepstrum and is computed as:

$$C_{mel}(n,m) = \frac{1}{R} \sum_{l=0}^{R-1} \log\{Emel(n,l)\} \cos\left(\frac{2\pi}{R} lm\right) \qquad (3.4)$$

Where m is the no. of cepstral coefficients and R = no. of filters



- **Input Signal:** This Input Signal is Vectorized Voice Signal.
- **MFCC:** The Output of this block is computed MFCC.

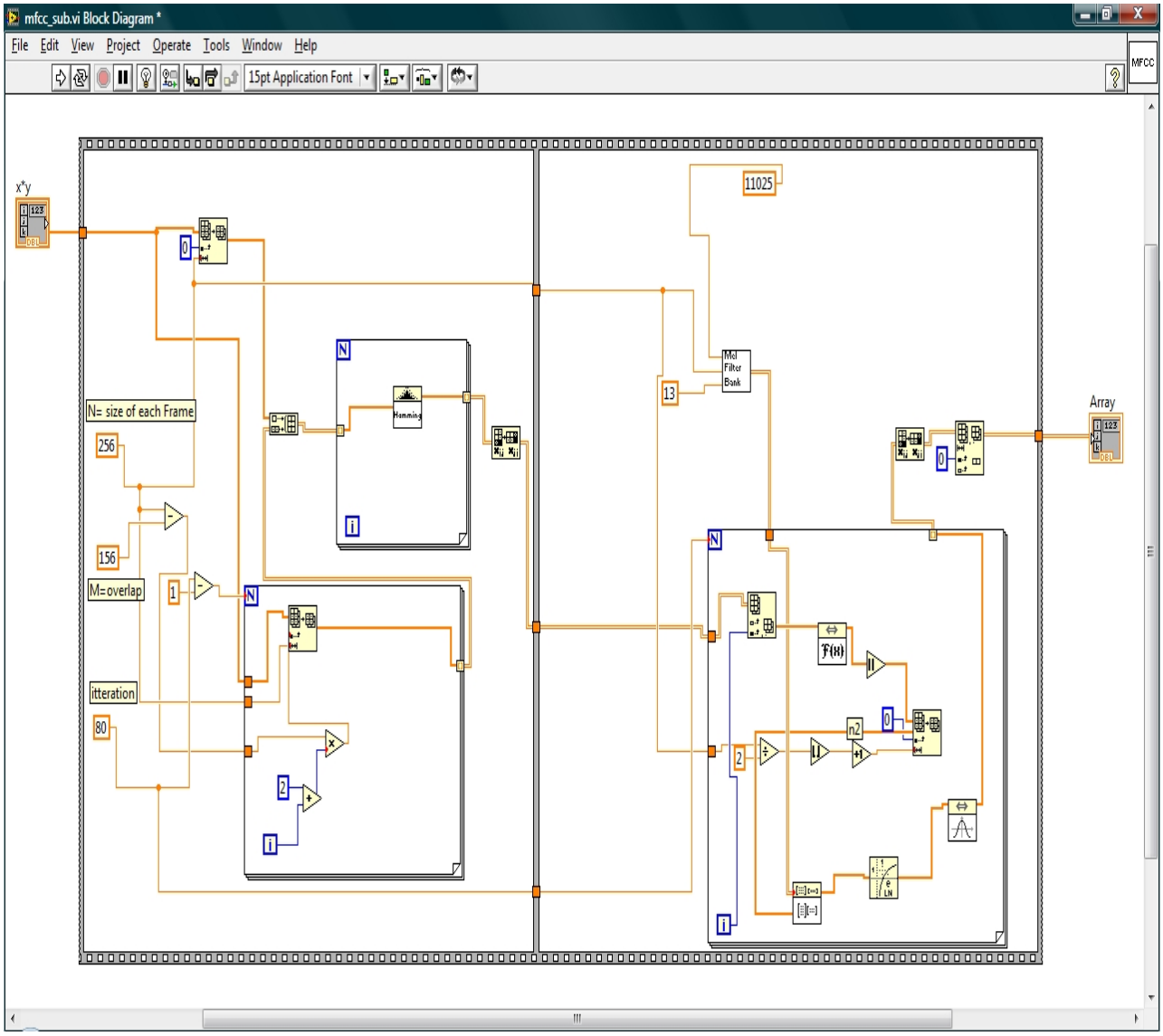


Figure: 3.11 Block diagram of extracting MFCC.

3.5.1.5 Vector Quantization

Vector quantization constructs a set of representative samples of the target speaker's enrollment utterances by clustering the feature vectors. Although a variety of clustering techniques exist, the most commonly used isk-means clustering [36]. This approach partitions N feature vectors into K disjoint subsets Sj to minimize an overall distance such as

$$D = \sum_{j=1}^{J} \sum (x_i - \mu_j) \qquad (3.5)$$

Where μj = (1/Nj)_xi∈Sj xi is the centroid of the Nj samples in the j-th cluster.

The algorithm proceeds in two steps:

1. Compute the centroid of each cluster using an initial assignment of the feature vectors to the clusters.
2. Reassign xi to that cluster whose centroid is closest to it. These steps are iterated until successive steps do not reassign samples.

Once the VQ models are established for a target speaker, scoring consists of evaluating D in (3.5) for feature vectors in the test utterance.

Suppose we know *a priori* the speech segments corresponding to the sound classes in both the training and test data. We then form averages for the *i* th class as

$$\bar{C}_i^{ts}[n] = \frac{1}{M}\sum_{m=1}^{M} C_i^{ts}[mL, n] \quad (3.6)$$

And

$$\bar{C}_i^{tr}[n] = \frac{1}{M}\sum_{m=1}^{M} C_i^{tr}[mL, n] \quad (3.7)$$

Where for convenience we have removed the "Mel" notation. We the compute a Euclidean distance with respect to each class as

$$D(i) = \frac{1}{R-1}\sum_{n=1}^{R-1}(\bar{C}_i^{ts}[n] - \bar{C}_i^{tr}[n])^2 \quad (3.8)$$

Finally, we average over all classes as

$$D(I) = \frac{1}{I}\sum_{i=1}^{I} D_i \quad (3.9)$$

Where *I* is the number of classes. To form this distance measure, we must identify class distinctions for the training and test data.

vector-quantization (VQ) method, using the *k*-nearest neighbor clustering algorithm each centroid in the clustering is derived from training data and represents an acoustic class, but without identification or labeling. The distance measured used in the clustering is given by the Euclidean distance between a

feature vector and a centroid. In recognition, i.e., in testing, we pick a class for each feature vector by finding the minimum distance with respect to the various centroids from the training stage. We then compute the average of the minimum distances over all test feature vectors. In speaker identification, for example, we do this for each known speaker and then pick a speaker with the smallest average minimum distance.

3.5.2 Identification

During testing the system, the system checks the identity of the person. The various steps for testing are:

1) Enter the biometric signature i.e. speech signal.
2) Silence removing.
3) Extract the features i.e. Mel Frequency Cepstrum Coefficient.
4) Pattern Matching
5) Decision – Who is identified?

Figure: 3.12 shows the Flow chart of the testing of a user that the user is identified or not.

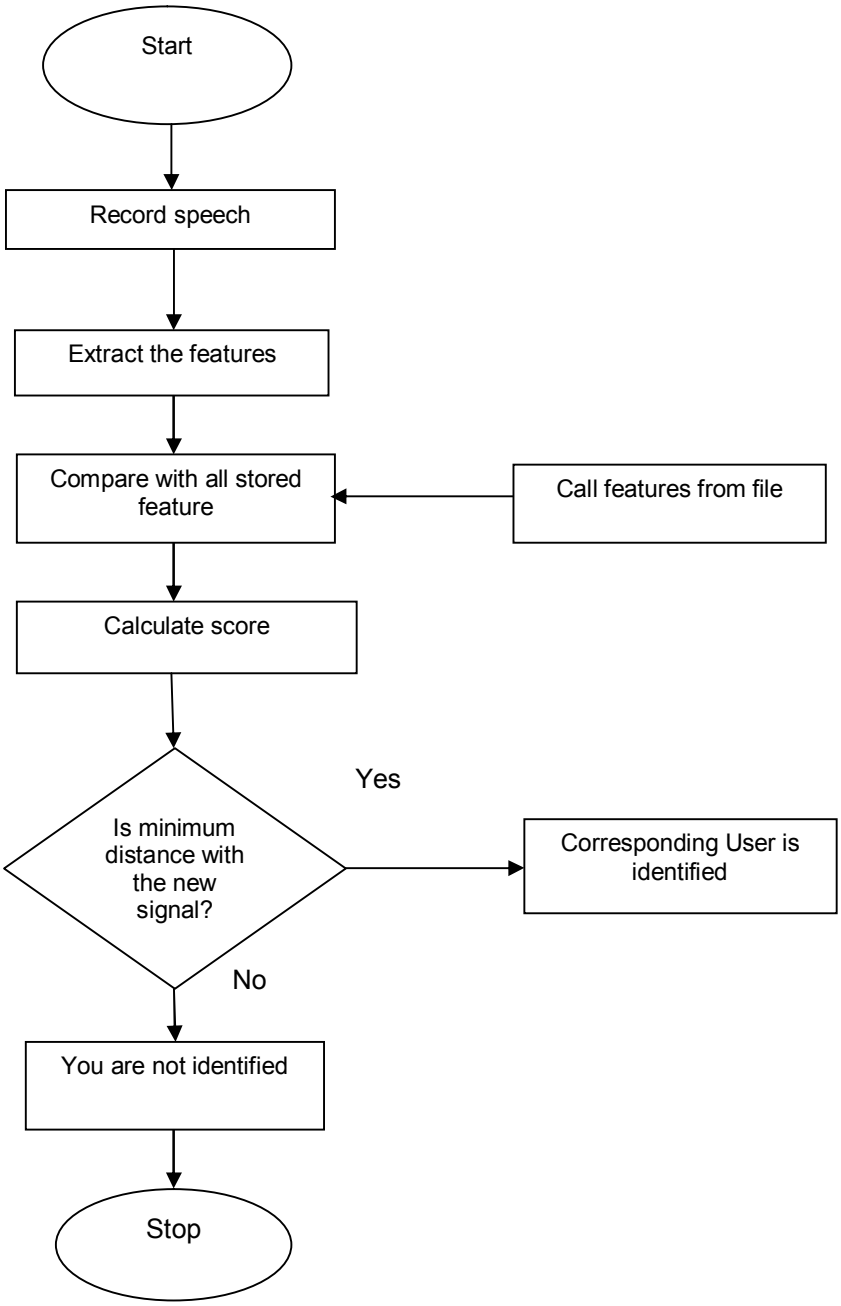


Figure: 3.12 Flow Chart of the Identification of Person.

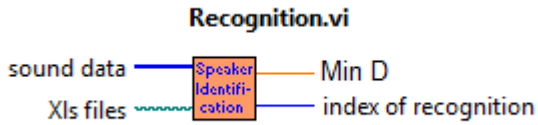
3.5.2.1 Enter the biometric signature i.e. speech signal

The system then prompts the user to enter his voice through microphone. The user presses the record button and record the speech for required time.

3.5.2.2 Extract the features i.e. Mel Frequency Cepstrum Coefficient

After this the done the system then extracts the spectral feature which has been chosen as the Mel-cepstrum (discussed earlier).

3.5.2.3 Pattern Matching By Comparing Loglikehood and Decision Making



The Recognition.vi gives the output as the maximum loglikehood and the index. While the input is vectorized sound data and the feature data file of the speaker for the user is claimed. In this VI the sound data of speaker is matched with all speaker s in database by calculating the distance D. The smallest average minimum distance with the stored database is the speaker identification. The output of this VI is the index of maximum LLH so that the speaker on that index could be found. The block diagram of recognition (identification) is shown in Fig 3.3.

Chapter 4 Results and Discussion

4.1 Introduction

Experiments had to be conducted to prove validity of the algorithms and to test accuracy of system. The experiments were had been conducted on a voice database created using LabVIEW to determine the validity of the developed algorithm.

4.2 Registration

The first step is Registration the user in the data base. For it registration program is employed which enroll the user by:

- Entering the Name
- Entering the biometric signal so that speech signal of the user as shown in figure 4.1.

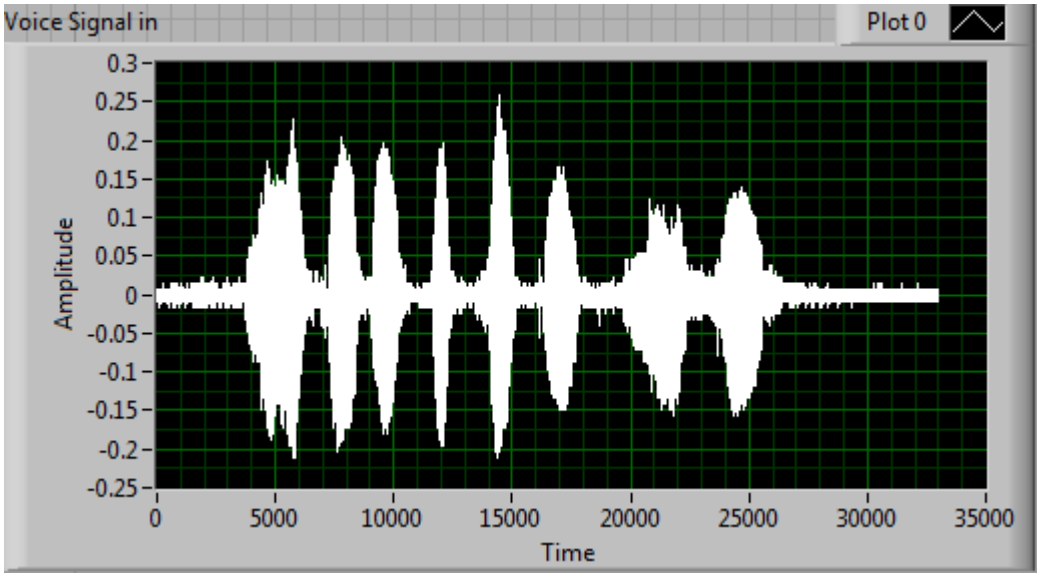
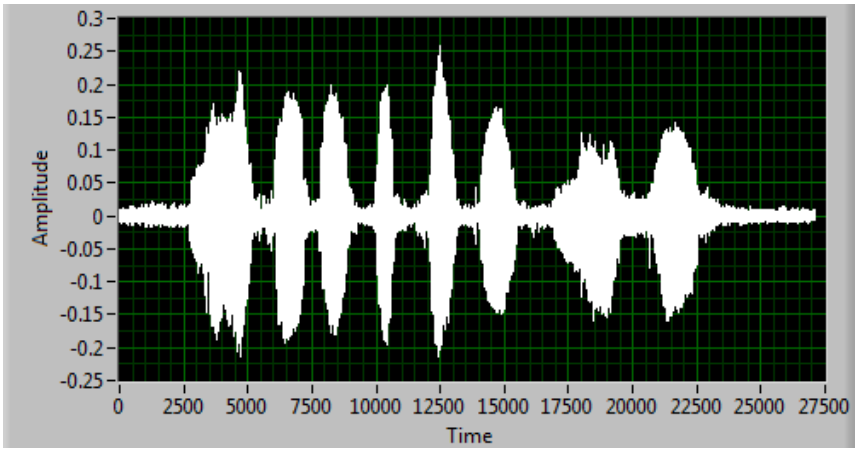


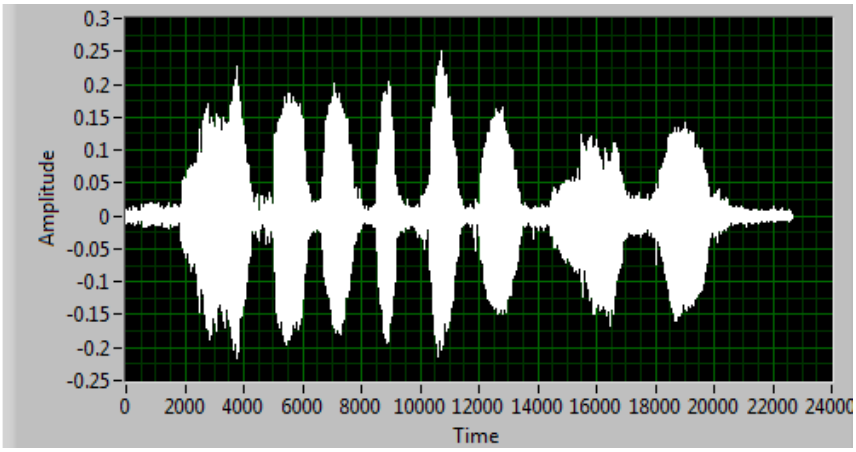
Figure: 4.1 Input voice signals for registration.

4.2.1 Silence Removing

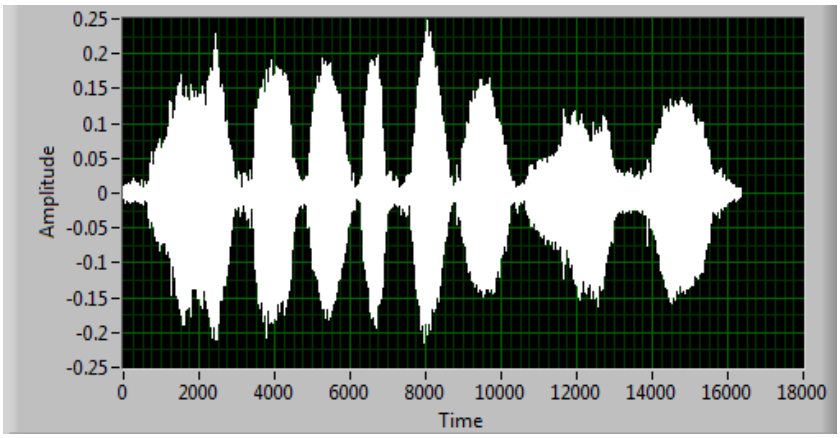
Experiments had been conducted on the data base to find out the optimum (best) value of the threshold for silence removal. The speech signal of figure 4.1 with various value of threshold has been shown in figure 4.2.



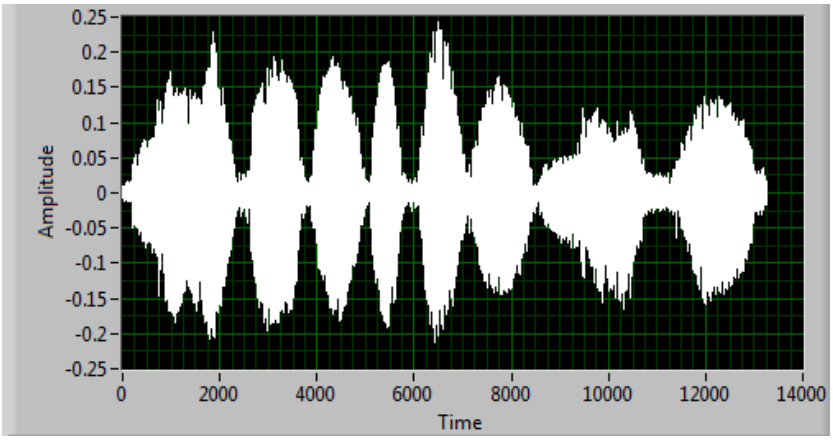
Threshold=0.25



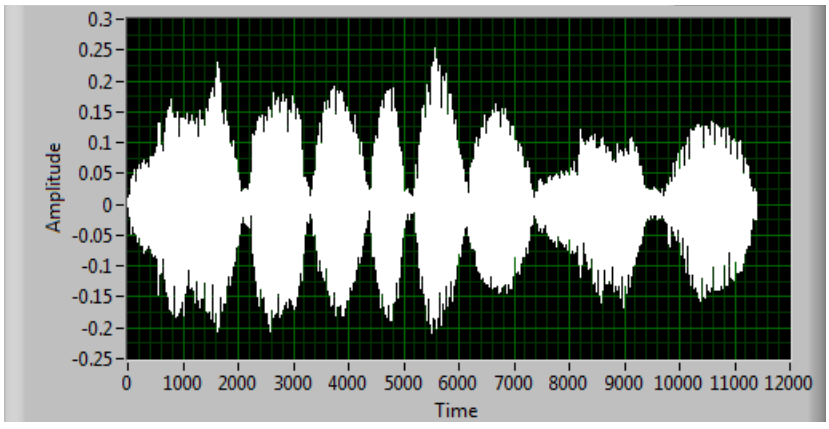
Threshold=0.5



Threshold=1.0



Threshold=1.5



Threshold=2.0

Figure: 4.2 Input voice signals after removing silence at different thresholds.

It has been observed that the signal with a threshold of 1.0 gives the best result.

4.2.2 Feature Extraction (MFCC of Speech Signal)

Mel frequency Cepstrum Coefficient (MFCC) has been extracted from the speech signal. The MFCC of a speech signal of figure 4.1 with a threshold of 1.0 for silence removing has been listed in table 4.1.

Figure: 4.1 Features (MFCC) of speech signal

1.043	0.472	1.305	1.462	1.051	1.099	1.206	0.976	0.875	0.439	0.097	-0.001
2.611	2.29	1.799	1.073	0.631	0.19	0.119	0.303	0.094	0.209	-0.091	-0.083
2.6	2.962	1.895	0.955	-0.125	0.408	0.336	0.466	-0.206	-0.133	0.13	0.072
2.534	2.814	1.958	1.043	0.416	0.412	-0.034	0.193	-0.045	-0.117	-0.357	-0.073
2.456	2.792	1.887	0.991	-0.014	0.287	0.019	0.29	-0.125	-0.207	-0.028	0.151
2.255	2.807	2.207	0.824	0.29	0.379	0.066	0.141	-0.313	-0.253	0.013	0.189
2.855	2.629	1.698	0.817	0.265	-0.067	0.151	0.409	-0.011	-0.383	-0.163	-0.049
3.895	1.302	0.328	0.475	0.051	0.171	0.457	0.457	-0.186	-0.167	-0.046	-0.209
3.105	0.334	0.414	0.868	0.64	0.315	0.845	0.215	-0.338	-0.293	0.059	-0.154
2.732	-0.144	0.392	1.208	0.43	0.294	0.887	0.146	-0.262	-0.298	0.376	-0.067
2.415	-0.372	0.754	0.861	0.538	0.462	0.951	0.073	-0.374	-0.144	0.363	-0.007
2.069	-0.123	0.917	0.816	0.457	0.604	0.956	0.018	-0.425	-0.145	0.424	-0.177
1.933	-0.168	1.101	0.638	0.53	0.5	0.958	-0.003	-0.352	-0.094	0.37	-0.224
1.858	0.069	1.088	0.405	0.53	0.866	0.825	-0.086	-0.331	-0.129	0.278	0.012
1.921	0.109	1.221	0.037	0.664	0.89	0.852	-0.122	-0.208	-0.105	0.119	0.122
1.799	0.101	1.296	-0.235	0.622	0.977	0.608	-0.162	-0.097	0.137	0.176	0.019
1.466	0.642	1.263	-0.505	0.687	0.965	0.544	-0.053	-0.006	0.061	0.131	0.041
1.474	1.057	1.395	-0.822	0.709	0.934	0.533	0.181	-0.051	0.019	0.166	0.044
1.798	1.253	1.646	-1.367	0.692	0.852	0.528	0.302	-0.073	-0.165	0.164	0.17
1.762	1.686	1.739	-1.647	0.426	0.783	0.304	0.408	0.34	-0.086	0.073	0.06
1.906	2.233	1.768	-1.814	0.015	0.814	0.183	0.579	0.532	-0.084	-0.032	-0.203
2.137	2.408	1.917	-1.814	-0.11	0.945	-0.015	0.641	0.53	-0.284	-0.135	-0.358
2.235	2.168	1.961	-1.116	0.105	0.925	-0.162	0.548	0.227	-0.44	-0.169	-0.284
2.583	2.424	1.961	-0.979	0.108	0.996	-0.048	0.131	0.259	-0.271	-0.26	-0.215
2.489	2.852	1.915	-0.3	-0.023	0.765	0.219	0.266	0.099	-0.286	-0.197	-0.161
2.355	3.001	1.727	0.351	0.524	0.395	0.25	0.343	0.111	-0.162	-0.062	-0.152
2.394	2.383	1.005	0.618	0.717	0.579	0.464	0.661	0.562	0.51	0.081	0.086
1.877	2.549	1.102	0.879	0.575	0.944	0.833	0.472	0.522	0.303	0.304	0.255
1.767	2.073	1.161	1.248	0.963	0.719	0.406	0.53	0.737	0.436	0.614	0.345
1.569	1.164	1.41	-0.018	1.088	1.178	0.961	0.668	0.646	0.707	0.156	0.047

1.571	2.103	1.91	0.423	0.555	0.878	0.497	0.304	0.234	0.489	0.219	0.222
2.114	2.681	2.271	-0.281	-0.338	0.755	-0.231	0.095	0.334	-0.304	0.171	0.057
2.159	2.617	2.091	-0.964	0.183	0.844	-0.421	0.169	0.186	-0.397	0.007	-0.255
2.121	2.679	1.805	-1.207	0.312	0.923	-0.457	0.215	0.383	-0.435	-0.256	-0.061
1.967	2.279	2.162	-1.379	0.28	0.906	-0.355	0.572	0.138	-0.482	-0.267	-0.196
2.05	2.167	2.228	-1.386	0.292	0.843	-0.307	0.705	-0.015	-0.478	-0.164	-0.174
1.882	2.363	2.105	-1.503	0.164	0.957	-0.234	0.605	0.163	-0.534	-0.171	-0.14
2.077	2.077	1.876	-1.423	0.268	0.858	-0.233	0.578	0.343	-0.542	-0.217	-0.178
2.247	1.91	1.697	-1.242	0.056	0.968	-0.172	0.485	0.336	-0.46	-0.096	-0.189
2.571	1.782	1.746	-1.095	-0.423	0.879	-0.016	0.614	0.375	-0.342	0.014	-0.199
2.793	1.863	1.653	-1.074	-0.525	0.886	0.452	0.185	0.274	0.174	-0.038	-0.409
2.789	1.86	1.632	-0.517	-0.572	0.557	0.665	0.355	0.316	0.141	0.105	-0.338
2.378	2.339	1.401	0.14	0.075	0.523	0.81	0.405	0.187	0.095	0.057	-0.164
2.783	1.842	0.717	1.068	0.683	0.518	0.29	0.297	0.402	0.299	0.439	0.068
2.4	1.496	0.628	1.008	0.684	1.03	0.596	0.761	0.741	0.389	0.654	0.178
2.284	2.172	1.197	0.77	0.739	0.504	0.494	0.537	0.287	0.265	0.157	0.063
1.789	2.239	2.461	-0.023	0.154	0.14	0.639	0.5	0.099	0.178	-0.317	-0.135
2.235	1.849	2.452	-0.774	-0.245	0.826	0.414	0.27	0.184	-0.003	-0.19	-0.313
2.457	1.829	2.066	-1.277	-0.129	0.981	0.353	0.239	0.028	0.1	-0.166	-0.314
2.944	1.498	2.03	-1.378	-0.158	0.835	0.236	0.559	-0.119	0.236	-0.275	-0.204
2.517	2.018	2.145	-1.314	-0.429	0.469	0.342	0.714	0.184	0.328	-0.371	-0.206
2.706	1.783	1.902	-1.038	-0.604	0.698	0.29	0.652	0.21	0.142	-0.218	-0.274
2.76	1.607	1.676	-1.346	-0.161	0.941	0.258	0.305	0.098	0.289	-0.419	-0.339
2.722	1.578	1.372	-0.906	0.159	0.829	0.187	0.393	-0.122	0.236	-0.532	-0.247
2.385	1.835	1.894	-0.9	0.054	0.549	0.32	0.438	0.011	0.019	-0.301	-0.212
2.312	2.319	2.246	-0.823	-0.101	0.771	0.312	0.36	0.035	-0.124	-0.197	-0.209
1.961	2.395	2.495	-0.27	-0.075	0.859	0.311	0.193	0.229	-0.293	-0.151	-0.205
0.844	3.092	2.07	0.44	0.521	0.132	0.513	0.337	0.202	0.097	0.005	-0.02
0.244	2.554	0.853	0.695	1.302	0.473	0.733	0.468	0.712	0.575	0.298	0.285
0.877	2.03	1.288	0.769	0.951	0.473	0.63	0.562	0.51	0.456	0.464	0.029
2.664	2.389	1.363	0.042	-0.054	0.42	0.331	0.311	0.084	-0.004	-0.004	-0.085
2.929	1.949	1.3	-0.459	-0.127	0.398	0.611	0.648	-0.241	-0.086	-0.235	-0.28
2.951	1.963	1.097	-0.674	-0.215	0.316	0.614	0.617	-0.125	0.242	-0.262	-0.513
3.194	1.821	1.036	-0.57	-0.437	0.356	0.765	0.559	-0.081	0.197	-0.176	-0.486
3.032	2.153	0.999	-0.784	-0.09	0.211	0.549	0.666	-0.024	0.148	-0.12	-0.522
2.953	2.162	1.058	-0.628	0.125	0.395	0.459	0.448	-0.132	0.142	-0.212	-0.423
2.598	2.443	1.474	-0.098	0.261	0.2	0.615	0.522	0.024	-0.083	-0.054	-0.251
2.308	3.086	1.733	0.624	0.576	0.339	0.215	0.633	0.318	0.269	0.086	0.056
2.248	3.113	1.388	0.869	0.792	0.347	0.353	0.453	0.365	0.219	0.187	-0.034
2.324	1.821	0.943	0.586	0.851	0.622	0.792	0.622	0.683	0.553	0.552	0.152
0.9	1.724	0.55	1.039	1.244	0.846	0.732	0.583	0.814	0.606	0.44	0.173

-0.76	2.559	0.851	1.602	1.255	0.961	0.998	0.706	0.614	0.407	0.47	0.22
-0.261	2.843	1.042	1.351	0.809	0.638	0.713	0.502	0.491	0.319	0.538	0.204
1.985	2.94	1.05	0.066	0.45	0.072	0.218	0.685	0.211	0.089	-0.016	-0.083
2.453	2.558	1.255	-1.06	0.423	0.459	0.178	0.49	0.214	0.005	-0.421	-0.058
2.733	2.386	1.244	-1.451	-0.014	0.54	0.111	0.61	0.507	0.288	-0.414	0.018
2.703	2.408	1.494	-1.506	-0.227	0.245	0.344	0.803	0.679	0.172	-0.336	-0.095
2.702	2.255	1.437	-1.597	-0.356	0.448	0.494	0.575	0.739	0.286	-0.188	-0.059
2.756	2.486	1.343	-1.682	-0.206	0.498	0.372	0.592	0.771	0.279	-0.109	0.039
2.732	2.597	1.525	-1.643	-0.31	0.707	0.131	0.711	0.726	0.207	-0.064	0.107

4.3 Identification

Identification has been done by calculating the Loglikehood of the new input speech signal from a Speaker with the estimated μ, $\sum$, C of the speakers that already registered.

- Enter Voice Print (Speech Signal)
- Remove Silence
- Extract MFCC
- Open file of stored μ, $\sum$, C.
- Calculate the distance D and compare.
- Result is the name of the person who is identified.

Table: 4.2 MFCC of new speaker for testing.

2.658	2.653	1.689	1.334	0.471	0.295	0.056	0.025	-0.008	-0.384	0.069	0.293
2.682	2.823	1.423	0.871	0.586	0.024	0.038	0.364	0.005	-0.314	-0.131	-0.157
3.027	2.256	1.426	0.813	0.527	0.074	0.13	0.341	-0.08	-0.374	-0.194	-0.101
3.968	0.626	0.1	0.298	0.226	-0.045	0.433	0.534	-0.14	-0.191	0.124	-0.167
3.575	-0.027	0.036	0.539	0.481	-0.142	0.766	0.596	-0.271	0.001	0.195	-0.165
3.219	-0.189	0.11	0.627	0.447	0.179	0.747	0.257	-0.287	-0.086	0.346	-0.13
2.917	-0.129	0.181	0.344	0.581	0.536	0.668	0.028	-0.282	0.047	0.306	-0.172
2.655	-0.309	0.606	-0.109	0.712	0.551	0.693	-0.004	-0.293	0.156	0.287	-0.26
2.638	-0.283	0.581	-0.056	0.58	0.556	0.869	-0.212	-0.174	0.167	0.255	-0.227
2.655	-0.154	0.723	-0.226	0.435	0.756	0.702	-0.146	-0.181	0.172	0.19	-0.129
2.568	0.16	0.757	-0.463	0.383	0.892	0.561	-0.043	-0.088	0.221	0.151	-0.183
2.767	0.364	0.947	-0.767	0.297	0.993	0.453	0.123	0.067	0.064	0.147	-0.153
2.618	0.757	1.132	-1.208	0.078	1.11	0.367	0.28	0.166	0.125	-0.02	-0.09
2.788	1.014	1.142	-1.401	-0.168	0.95	0.337	0.577	0.322	0.022	-0.084	-0.13

3.068	1.499	1.275	-1.55	-0.621	0.771	0.554	0.871	0.189	-0.116	-0.075	-0.202
3.277	1.924	1.183	-1.229	-0.697	0.815	0.629	0.591	0.37	-0.172	-0.252	-0.363
3.35	1.502	0.894	-0.238	-0.218	0.675	0.555	0.3	0.088	-0.153	-0.386	-0.2
2.885	2.634	1.365	0.138	-0.188	0.629	0.666	0.54	-0.154	-0.284	0.005	-0.329
2.265	2.463	1.754	0.763	0.222	0.532	0.497	0.447	0.26	-0.213	-0.103	0
1.589	1.794	1.307	0.952	0.649	0.466	0.142	0.436	0.629	0.494	0.424	-0.036
2.7	1.887	1.701	-0.106	-0.363	0.593	0.023	0.162	0.501	-0.027	-0.063	-0.223
2.322	2.317	1.739	-0.782	-0.224	0.966	0.31	0.084	0.352	-0.215	-0.277	-0.294
1.783	2.001	1.693	-0.661	-0.05	0.815	0.219	0.334	0.492	-0.508	-0.189	-0.298
1.949	1.926	1.603	-0.817	0.189	1.028	-0.1	0.483	0.419	-0.414	-0.183	-0.264
1.852	1.962	1.592	-0.895	0.12	0.998	-0.157	0.425	0.561	-0.352	-0.245	-0.08
1.951	2.03	1.603	-1.05	-0.016	1.085	0.022	0.318	0.691	-0.304	-0.106	-0.149
2.141	1.753	1.788	-1.117	-0.131	1.209	-0.012	0.358	0.655	-0.334	0.024	-0.116
2.371	1.724	1.766	-1.03	-0.211	1.054	0.238	0.325	0.451	0.045	-0.177	-0.025
2.54	1.715	1.347	-0.797	0.073	0.756	0.339	0.266	0.398	0.044	-0.172	0.006
2.849	2.175	1.155	-0.565	-0.373	0.686	0.637	0.233	0.498	0.031	0.161	-0.329
3.174	1.843	0.69	0.089	-0.047	0.709	0.662	0.329	0.046	0.106	0.082	-0.432
2.683	2.134	0.964	0.684	0.541	0.358	0.462	0.579	0.428	0.419	0.304	0.042
2.024	1.344	1.048	0.772	0.67	0.63	0.83	0.663	0.636	0.887	0.297	0.27
2.733	2.176	1.103	0.593	0.51	0.603	0.482	0.363	0.363	0.464	0.301	-0.011
2.937	1.509	1.273	0.002	-0.036	0.199	0.58	0.595	0.205	0.172	0.064	-0.166
2.89	1.545	1.393	-0.39	-0.523	0.247	0.666	0.538	0.305	0.207	0.103	-0.332
2.65	1.676	1.528	-0.781	-0.409	0.33	0.735	0.628	0.129	0.129	-0.004	-0.36
2.68	1.514	1.752	-1.025	-0.441	0.503	0.574	0.664	0.251	0.061	-0.101	-0.35
2.594	1.525	1.659	-1.172	-0.297	0.724	0.428	0.569	0.273	0.071	-0.135	-0.437
2.548	1.168	1.474	-1.094	0.024	0.704	0.349	0.412	0.179	0.091	-0.194	-0.402
2.397	1.144	1.453	-1.006	0.211	0.491	0.201	0.594	0.173	0.069	-0.39	-0.255
2.241	1.166	1.373	-0.847	0.303	0.405	0.284	0.593	0.047	-0.159	-0.28	-0.271
2.078	1.796	1.66	-1.048	0.198	0.494	0.425	0.401	-0.004	-0.185	-0.391	-0.257
2.439	2.055	1.701	-1.032	0.164	0.525	0.193	0.448	-0.105	-0.078	-0.469	-0.149
2.579	2.271	1.458	-0.589	0.059	0.402	0.204	0.378	-0.34	-0.033	-0.295	-0.202
2.478	1.93	1.561	-0.131	-0.205	0.751	-0.025	0.081	-0.109	0.344	-0.278	-0.12
2.707	2.914	1.308	0.311	-0.09	0.327	0.459	0.371	0.075	0.06	-0.216	-0.24
2.596	2.454	1.296	0.336	0.237	0.398	0.318	0.361	0.331	0.514	0.245	-0.028
3.034	1.934	1.489	-0.046	-0.627	0.314	0.686	0.326	-0.132	0.212	-0.121	-0.189
2.919	1.473	1.094	-0.142	-0.491	0.608	0.79	0.251	-0.048	-0.001	-0.232	-0.274
2.763	1.58	1.018	-0.188	-0.589	0.717	0.829	0.458	-0.112	0.168	-0.368	-0.397
2.875	1.584	0.825	-0.118	-0.798	0.652	0.863	0.582	0.053	0.106	-0.268	-0.411
2.756	1.473	1.029	-0.272	-0.628	0.625	0.825	0.365	0.08	0.195	-0.31	-0.408
2.595	1.539	1.036	-0.053	-0.17	0.503	0.696	0.258	0.14	-0.107	-0.173	-0.461
2.66	2.206	0.914	0.575	0.244	0.245	0.508	0.607	0.117	-0.084	-0.152	-0.351

2.372	3.116	1.607	1.008	0.666	0.478	0.239	0.458	0.055	0.011	0.241	0.06
1.729	3.257	2.218	1.544	0.559	0.282	0.244	0.029	0.058	0.38	0.225	-0.098
2.008	2.371	0.828	1.224	0.35	0.43	0.496	0.451	0.559	0.245	0.256	0.121
0.579	2.437	0.957	1.017	0.541	0.639	0.492	0.691	0.49	0.074	0.194	-0.002
2.34	2.431	0.979	-0.395	0.515	0.483	0.206	0.47	0.265	-0.033	-0.316	0.109
2.203	2.354	1.12	-0.915	0.034	0.235	0.452	0.553	0.239	0.258	0.016	-0.146
2.043	2.623	1.377	-1.066	-0.225	0.085	0.489	0.681	0.764	0.26	-0.053	-0.1
2.175	2.617	1.338	-1.133	-0.302	0.08	0.283	0.94	0.949	0.269	-0.036	0.032
2.342	2.846	1.368	-1.162	-0.396	0.046	0.245	1.079	0.962	0.321	-0.002	0.023
2.354	2.624	1.433	-0.922	-0.4	0.101	0.252	0.968	0.821	0.355	0.06	-0.013
2.459	2.611	1.447	-0.945	-0.353	0.599	0.178	0.589	0.577	0.209	0.162	-0.107
2.647	2.464	1.307	-0.174	0.029	0.626	0.022	0.554	0.542	-0.337	-0.048	-0.095
1.618	2.317	1.465	0.918	0.636	-0.083	0.21	0.441	0.268	0.301	0.111	0.02
1.182	2.116	1.51	0.742	0.946	0.317	0.251	0.6	0.321	0.568	0.376	0.273
2.791	2.398	1.556	-0.453	-0.097	0.48	0.01	0.445	0.353	0.107	-0.078	-0.016
2.464	1.729	1.517	-0.371	-0.011	0.722	0.147	0.601	0.194	-0.139	-0.166	-0.297
2.376	1.384	1.477	-0.586	0.039	0.825	0.095	0.605	0.115	0.001	-0.239	-0.227
2.354	2.116	1.514	-0.739	-0.362	0.794	0.235	0.631	0.375	0.124	-0.25	-0.32
2.368	1.899	1.103	-0.374	-0.299	0.864	0.223	0.573	0.372	-0.086	-0.394	-0.198
2.368	1.951	0.946	-0.291	-0.158	0.957	0.239	0.492	0.193	-0.214	-0.311	-0.247
2.718	2.124	0.976	-0.195	-0.086	0.654	0.436	0.337	0.128	-0.121	-0.379	-0.246
2.694	2.638	1.183	0.005	-0.222	0.511	0.543	0.302	0.003	0.035	-0.394	-0.336
2.684	2.779	1.522	0.39	-0.461	0.509	0.81	0.148	-0.398	0.121	-0.333	-0.135
3.018	2.716	1.251	0.597	0.119	0.536	0.589	0.043	-0.353	-0.063	0.058	-0.167
2.846	2.996	1.273	0.878	0.276	0.45	0.529	0.184	-0.05	-0.027	-0.041	-0.162

4.4 Result of Identification

During testing the system, the system checks the identity of the person. The experiments had been conducted on the database stored in the lab and following result had been obtained as shown in table 4.3.

Table: 4.3 Results

Stored Database / Test data	A	B	C	D	E	F	G	H	I	J
A	4	0	0	0	0	0	0	1	0	0
B	0	5	0	0	0	0	0	0	2	0
C	0	0	4	0	0	0	1	0	0	0
D	0	0	0	3	0	0	0	3	1	0
E	0	1	0	0	5	0	0	0	0	0
F	2	0	0	1	0	4	0	0	0	0
G	0	0	0	1	0	0	5	0	0	0
H	0	2	0	0	0	0	0	5	0	0
I	0	0	2	2	0	0	0	0	4	0
J	0	1	0	0	0	2	0	0	0	5
K	1	0	1	0	0	1	0	1	0	0
L	0	0	0	1	0	0	1	0	0	1
M	0	0	0	0	0	0	0	0	0	0
N	1	0	1	0	1	0	0	1	0	0
O	0	1	0	0	0	0	0	0	0	0
P	0	0	0	0	0	2	2	0	0	0
Q	0	1	0	0	0	0	0	0	0	2
R	0	0	0	1	0	0	1	0	2	0
S	1	0	0	0	0	0	0	0	0	0
T	0	0	0	0	2	0	0	2	0	0
	K	L	M	N	O	P	Q	R	S	T
A	0	0	1	0	0	1	0	1	0	0
B	3	0	0	1	0	0	1	0	0	1
C	0	0	0	0	0	0	0	0	0	0
D	1	0	1	0	2	0	0	1	0	0
E	0	0	0	0	0	0	0	0	0	0
F	0	2	0	0	0	0	1	0	2	0
G	0	0	0	0	0	0	0	0	0	2
H	0	0	0	1	0	0	1	0	2	0
I	1	0	0	0	0	0	0	2	0	0
J	0	0	0	0	0	0	0	0	0	0
K	5	0	1	0	0	1	0	0	1	0
L	0	3	0	0	0	0	0	0	0	0
M	0	0	4	0	3	0	1	0	0	0
N	0	0	0	5	0	0	0	0	0	0
O	0	1	0	0	5	0	0	0	2	0

P	0	1	0	0	0	5	0	1	0	2
Q	0	0	0	0	0	0	5	0	0	0
R	0	1	3	1	0	3	0	5	1	0
S	0	0	0	0	0	0	0	0	4	0
T	1	0	0	1	0	0	2	0	0	3

The A, B, C…….T represents the different subjects in the above table. From the above table it can be seen that out of 100 true tests 88 have been found accurate so the true rejection rate is 12%.

Chapter 5 Conclusion and Future Scope

5.1 Conclusion

This work describes Speaker Recognition systems as a part of the Biometric Security System. Mainly, this work aims at the speaker identification and the research in this field. The Speaker Identification system using Gaussian Mixture Model is implemented on LabVIEW 8.5 platform. There is two session of system first is *Registration* and second is *Testing*. In Registration session firstly silence is removed from the voice print to improve the precision of the recognition. The feature have been extracted and stored in a file to be compared with the query . In testing session voice print of unknown speaker is taken then silence removed and correspondence have been found between the stored features and the features obtained from the query signal. Experiments have been conducted on the database stored in the lab and it has been observed that the system is accurate to a value of 88%.

5.2 Future Scope

Automatic speaker recognition system using LabVIEW is an efficient program giving almost 90% of accuracy still there are chances to improve it.

i) The main problem to the system is from the external noise. By using some another noise elimination methods, the performance of the system can be improved.
ii) Different method of the Silence Remove can be used to improve it.
iii) High quality microphone can be used to improve the system accuracy.
iv) The system can be tested on larger database.

References

1 A.Barney, C.H Shadle, and P.O.A.L. Davies, “Fluid Flow in a Dynamical Machanical Model of the Vocal Folds and Tract. I: Measurements and Theory,” J.Acoustical society of America, Vol. 105, no, 1, pp 444-445, jan. 1999.

2 A.E. Rosenberg, S. Parthasarathy, “Speaker background models for connected digit password speaker verification”, Proc. ICASSP pp. 81– 84, 1996.

3 A.K. Jain, A. Ross and S. Prabhakar, “An Introduction to Biometric Recognition,” IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image and Video Based Biometrics, vol.14, no.1, pp.4-20, Jan.2004.

4 A.V Oppenheim, R.W Schafer, Buck, J.R.: Discrete-Time Signal Processing, 2nd ed, Upper Saddle River, NJ, Prentice Hall, 1999.

5 B. Gold, N Morgan, “Speech and Audio Signal Processing”, New York, USA, John Wiley & sons, inc, 2000.

6 D. Gillick, S. Stafford, B. Peskin, “Speaker detection without models”, Proc. ICASSP 2005.

7 D. W. Robinson and R. S. Dadson, “A re-determination of the equal-loudness contours for pure tones,” British Journal of Applied Physics, vol. 7, pp. 166– 181, 1956.

8 D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech and Audio Process., pp. 72–83, Jan. 1995.

9 D.R Rodman, “Computer Speech Technology, Boston”, Mass.: Artech House, 1999.

10 E. Zwicker and H. Fastl, “Psycho-acoustics”. Springer-Verlag, 2nd Edition, 1990.

11 F Orság, “Vision für die Zukunft, Biometrie, Kreutztal”, DE, b-Quadrat, , pp. 131-145, ISBN 3-933609-02-X, 2004.

12 F Orság, “Some Basic Techniques of the Speech Recognition”, In: Proceedings of 8th Conference STUDENT EEICT 2002, Brno, CZ, FEKT VUT, pp. 5, ISBN 80-214-2116, 2002.

13 F. J Harris, “On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform”, Proceedings of the IEEE. Vol. 66, pp. 66-67,1978.

14 F. J. Prokoski, R. B. Riedel, and J. S. Coffin, “Identification of individuals by means of facial thermography,” in Proceedings of The IEEE 1992 International Carnahan Conference on Security Technology: Crime Countermeasures, Atlanta, GA, USA 14-16 Oct., pp. 120-125, IEEE, 1992.

15 F.K. Soong, A.E. Rosenberg, L.R. Rabiner, B.H. Juang: A vector quantization approach to speaker recognition, Proc. IEEE ICASSP pp. 387–390, 1985.

16 Federal Bureau of Investigation Educational Internet Publication, ‘‘DNA testing,” http://www.Jbi.gov/kids/dna/dna.htm, 1997.

17 H. Fletcher and W. J. Munson, “Loudness, its definition, measurement and calculation,” Journal of Acoustical Society of America, vol. 5, no. 2, pp. 82– 108, October 1933.

18 J. Daugman, “How Iris Recognition Works,” IEEE Transaction on Circuits and Systems for Video Technology, Vol.14, no.1, pp.21-30, Jan.2004.

19 J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, “Synthetic voices for computers,” IEEE Spectrum, vol. 7, pp. 22–45, October 1970.

20 J.R Deller, J.H.L Hansen, Proakis, J.G. Discrete-Time Processing of Speech Signals, New York, USA, IEEE Press, ISBN 0-7803-5386-2, 2000.

21 K. Fukunaga, “Introduction to Statistical Pattern Recognition”, 2nd edn, Elsevier, New York, 1990.

22 L Hui-Ling, “Toward a high-quality singing synthesizer with vocal texture control”, PhD thesis, 2002.

23 L.R. Rabiner, B.-H. Juang, “Fundamentals of Speech Recognition” (Prentice-Hall, Englewood Cliffs, 1993 .

24 M Drahanský, F Orság, “Fingerprints and Speech Recognition as parts of the biometry”, In: Proceedings of 36th International Conference MOSIS '02, Ostrava,

CZ, MARQ, pp. 177-183, ISBN 80-85988-71-2, 2002

25 M. B. Sachs, C. C. Blackburn, and E. D. Young, “Rate-place and temporalplace representations of vowels in the auditory nerve and anteroventral cochlear nucleus,” Journal of Phonetics, vol. 16, pp. 37–53, 1988.

26 M. Burge and W. Burger, ‘‘Ear biometrics for machine vision,” in 21Workshop of Austrian Association for Pattern Recognition, http://www.cast.uni-linz.ac.at/st/vision/

27 M. Przybocki, Martin, “An overview, Digital Signal Process” , The NIST 1999 speaker recognition evaluation pp. 1–18 ,2000.

28 M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, “Robust prosodic features for speaker identification,” Proc. ICSLP’96, vol.3, pp.1800–1804, Oct. 1996.

29 O. Thyes, R. Kuhn, P. Nguyen, J.C. Junqua, “Speaker identification and verification using eigenvoices”, Proc. ICASSP pp. 242–245, 2002.

30 P. J. Phillips, P. J. Rauss, and S. Z. Der, ‘‘The FERET (Face Recognition Technology) evaluation methodology”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 97, June 1997.

31 S. A. Bleha and M. Obaidat, ‘‘Computer users verification using the perceptron algorithm”, IEEE Trans. Systems Man Cybernetics, Vol. 23, pp. 900-902, 1993.

32 S. Furui, ‘‘Recent advances in speaker recognition”, in Lecture Notes in Computer Science 1206, Proceedings of Audio- and Video Biometric Person Authentication AVBPA '97, First International Conference, Crans-Montana, Switzerland, March 12-14, pp. 237-252, Springer-Verlag, Berlin, 1997.

33 S. Liu and M. Silverman, “A Practical Guide to Biometric Security Technology”, IEEE Journal on IT Professional, vol.3, no.1, pp.27-32, Jan.-Feb. 2001.

34 S. S. Stevens and J. Volkman, “The relation of pitch to frequency”, American Journal of Psychology, vol. 53, p. 329, 1940.

35 V. Nalwa, ‘‘Automatic on-line signature verification,” Proceedings of the IEEE, Vol. 85, pp. 2 13-239, February 1997.

36 X. Huang, A. Acero, H.W. Hon “Spoken Language Processing: A Guide to Theory, Algorithm and System Development”, Prentice-Hall, Englewood Cliffs,

2001.

37 Z. Tychtl, J. Psutka, “Speech Production Based on the Mel-Frequency Cepstral Coefficients”.