Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Anbar
College of Computer Science and Information Technology
Department of Computer Science

Deep Learning Algorithms Based Voiceprint Recognition System in Noisy Environment

A Thesis

Submitted to the Department of Computer Science - College of Computer Science and Information Technology, University of Anbar in Partial Fulfillment of the Requirements for master's degree of Science in Computer Science.

By

Hajer Yass Khdier

Bachelor of Computer Science – 2013

Supervised By

Dr. Salah Awad Salman

Dr. Wesam Mohammed Jasim

1441 A.H. 2020 A.D.

بسم الله الرحمن الرحيم

(إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ وَالْفُلْكِ الَّتِي تَجْرِي فِي الْبَحْرِ بِمَا يَنفَعُ النَّاسَ وَمَا أَنزَلَ اللَّهُ مِنَ السَّمَاءِ مِن مَّاءٍ فَأَحْيَا بِهِ الْأَرْضَ بَعْدَ مَوْتِهَا وَبَثَّ فِيهَا مِن كُلِّ دَابَّةٍ وَتَصْرِيفِ الرِّيَاحِ وَالسَّحَابِ الْمُسَخَّرِ بَيْنَ السَّمَاءِ وَالْأَرْضِ لَآيَاتٍ لِّقَوْمٍ يَعْقِلُونَ)

صدق الله العظيم

الآية(١٦٤) سورة البقرة

اسم الطالـــــبه: هاجرياس خضير
الكليـــــــــــة: علوم الحاسوب وتكنولوجيا المعلومات - قسم علوم الحاسبات
عنوان الرسالة: خوارزميات التعلم العميق لأنظمة تمييز البصمة الصوتية في البيئات الصاخبة

Supervisors' Certification

*We certify that we read this thesis entitled "**Deep Learning Algorithms Based Voiceprint Recognition System in Noisy Environment**" that is carried out under our supervision at the Department of Computer Science of the University of Anbar, by the student "**Hajer Yass Khdier**" and that in our opinion, it meets the standard of a thesis for the degree of Master of Science in Computer Science.*

Signature :

Name : ***Dr. Salah Awad Salman***

Date : / /2020

Signature :

Name : ***Dr. Wesam Mohammed Jasim***

Date : / /2020

Certification of the Examination Committee

*We the examination committee certify that we have read this thesis entitled " **Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment** " and have examined the student " **Hajer Yass Khdier** ", in its contents and what is related to it, and that in our option it is adequate to fulfill the requirements for the degree of **Master of Computer Science**.*

Signature:
Name: ***(Chairman)***
Date: / / 2020

Signature:
*Name**:*** ***(Member)***
Date: / / 2020

Signature:
Name: ***(Member)***
Date: / / 2020

Signature:
Name: ***(Supervisor)***
Date: / / 2020

Signature:
Name: ***(Supervisor)***
Date: / / 2020

Approved by the Dean of the College of Computer Science and Information Technology, University of Anbar.

Signature:
Name: ***Assist. Prof. Dr. Salah Awad Salman***
Title: ***Dean of the College***
Date: / / 2020

Student name: Hajer Yass Khdier
Thesis title: Deep Learning Algorithms Based Voiceprint Recognition System in Noisy Environment

Acknowledgments

*First of all, I would like to express my thanks and gratitude to **ALLAH the Almighty**, who granted me all graces.*

This thesis would not have been possible without the support and encouragement of many people. foremost to my supervisors Dr. Salah Awad and Dr. Wesam M. Jasim, thank you for all your support and encouragement throughout the thesis. It has been both an honor and a pleasure to work with them and learn from them.

To my dear parents, especially my mother, who always supported me throughout the dissertation period, thank you from the bottom of my heart for my dear mother, and my dear brother and sisters, thank you for always believing in me and encouraging me to follow my dreams. I could not have achieved any of this without the support and encouragement that you have always given me.

To my husband Ali, you have come into my life and have turned everything in my life beautiful. You have always helped me and supported me. Thank you for your love and your support for me. This is something I will always love.

Finally, I express my deep gratitude for my colleagues and friends at University of Anbar, I thank you all for the encouragement and support you have given me during this period, my love and best wishes are with all of you.

Dedication

This thesis is dedicated:

To my soul and my heart pulse "my Father and Mother".

Hajer Yass Khdier
2020

Abstract

Voiceprint Recognition (VPR) is the mechanism by which a user's so-called identity is determined using characteristics taken from their voice, where this technique is one of the world's most useful and common biometric recognition techniques particularly the fields- relevant to security. These can be used for authentication, monitoring, forensic identification of speakers, and a variety of related activities.

The aim of this thesis is to design a deep learning strategy, which will provide a way to implicitly learn the voiceprint recognition in noisy environments.Two approaches were used for VPRS and were used the same structure of convolution neural network for two approaches and trying to increase the accuracy of the system by deal with a huge dataset adding to its background a random noise to prove the efficiency of the system in noisy conditions.

In this thesis, Attempt is applied to create a system that recognizes human speaker identity using Convolutional Neural Network (CNN). Used CNN for both feature extraction and deep learning algorithm,thus will enhance the ability of the system to be much accurate and be more efficient. The CNN architecture is designed to work with both MFCC-CNN and RW-CNN. In both cases, the proposed CNN inputs are images, i.e. the network dealt with images, where the same CNN architecture is used for both methods .The obtained findings show that both methods are similar in their accuracy 0.96 and mean square error 3.2000e-08 results but differents in performance where the time results show that RW-CNN is better than MFCC-CNN whether with or without noise. In other words RW-CNN is more efficient in clean and noisy environments from MFCC-CNN.

Keywords: Convolutional Neural Network, Deep Learning,Voiceprint Recognition System,

List of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations	
AI	***Artificial Intelligence***
ANN	***Artificial Neural Network***
AVPS	***Automatic Voiceprint Systems***
CNN	***Convolution Neural Network***
DCT	***Discrete Cosine Transform***
DFT	***Discrete Fourier Transform***
DL	***Deep Learning***
DWT	***Discrete Wavelet Transform***
FFT	***Fast Fourier Transform***
GMM	***Gaussian Mixture Model***
HMM	***Hidden Markov Models***
LPC	***Linear Predictive Coefficients***
LPCC	***Linear Prediction Cepstral Coefficient***
MFCC	***Mel Frequency Cepstral Coefficients***
ML	***Machine Learning***
MSE	***Mean Square Error***
PLP	***Perceptual Linear Prediction***

RBM	***Restricted Boltzman Machine***
RELU	***Rectified Linear Unit***
RNN	***Recurrent Neural Network***
RW	***Raw Waveform***
VPI	***Voiceprint Identification***
VP	***Voiceprint***
VPRS	***Voiceprint Recognition System***
VQ	***Vector Quantization***
VPV	***Voiceprint Verification***
WPD	***Wavelet Packet Decomposition***

Chapter One

General Introduction

Chapter One

General Introduction

1.1 Introduction

Voiceprint Recognition system (VPRs) is the method of distinguishing automatically who is speaking based on individual information found in a human's voice. No two individuals are identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary, and so on [1,2].

Voice is the fundamental, normal, and productive type of specialized technique for individuals to interface with each other. The voice signal contains numerous degrees of information. On the one hand, the message is passed on through expressed words and this message contains data about the individual, for example, language, feeling, sexual, and age. This data is the issue that decides the personality of the individual. The programmed acknowledgment of voiceprint and discourse acknowledgment are firmly related. While discourse acknowledgment defines its objectives at perceiving the expressed words in discourse. The point of programmed voiceprint acknowledgment is to personalize the speaker by extraction, characterization, and acknowledgment of the data contained in the voice signal [3,4].

This innovation makes it conceivable to utilize the speaker's voice to check their character and afterward empower control of access to administrations. For example, voice dialing and phone message, telebanking, phone shopping, database get to related administrations, data administrations, security control for classified data regions, and remote access to PCs without the genuine need to convey a charge, Mastercard or recall the financial balance secret phrase or some other security locks. Voiceprint acknowledgment innovation is required to make a large group of new administrations that will make our day by day lives progressively advantageous [5].

1.2 Biometrics

The biometric term gets from the Greek words bios (life) and metric. Biometrics alludes to advancements that quantify and analyze human characteristics, uses the remarkable examples of physical or social qualities of clients for validation or ID in other expression Biometrics is the study of examining physical or conduct attributes explicit to every person to have the option to verify their personality. It is utilized as a type of ID and access control. These attributes are quantifiable and particular properties that are utilized to recognize and portray individuals using biometrics. In which, a man can be distinguished in perspective on "who she/he is" rather than "what she/he has" (a card, token, scratch) or "what she/he knows" (secret key, PIN), Where these traditional methods were replaced by modern techniques, which are that use biometrics technique, where Passwords have some conspicuous disadvantages that could be taken, lost, or overlooked. Interestingly, biometrics offer an elective answer for the assignment of individual validation or recognizable proof dependent on biometric attributes. To be overlooked or lost is inconceivable, and dissimilar to passwords, they are difficult to manufacture. There are some the biometric characteristics that can be characterized for a person; for instance, fingerprint, finger-vein, iris, voice, face, and so on [6,7] .

1.3 Biometric Types

Biometric alludes to exceptional qualities of a distinct individual which does not change with time [8]. There are two kinds of biometric strategies ,where Each biometric characteristic has its strengths and weaknesses. Below some properties for Biometric:(1) universality, which means that each entity should have a characteristic; (2) uniqueness, which indicates that no two persons should have the same characteristic; (3) permanence, which means that the characteristic should be invariant over time; and (4) collectability, which indicates that the characteristic can be quantitatively determined. There are some other significant requirements in practice, (5) efficiency, referring to achievable identification accuracy, resource requirements for achieving reasonable identification accuracy and work or environmental factors affecting the accuracy of identification, (6) acceptability, indicating to what degree individuals are willing to accept the biometric system, and (7) circumvention, referring to how easy it is to trick the system[10,11].

• Physical biometrics

It is regarding the model of the body. The physical features of an individual such as fingerprints, hand geometry, iris, face, and DNA are known as physical biometrics [9].

• Behavioral biometrics

It is regarding with conduct of the individual. Behavior techniques are for recognizable proof focus on the activities of an individual, allowing the client a chance to control his activities. Biometrics dependent on these techniques contemplates the elevated level of internal variations (mood, health condition, and so on), that is the reason such strategies are helpful just inconsistent use. It incorporates keystroke, signature, and voice [9].

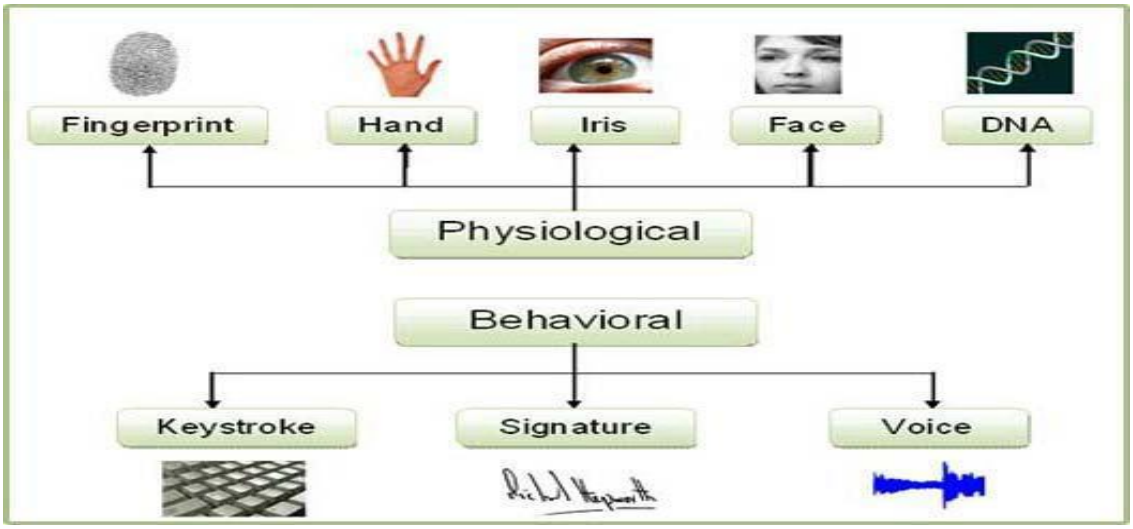
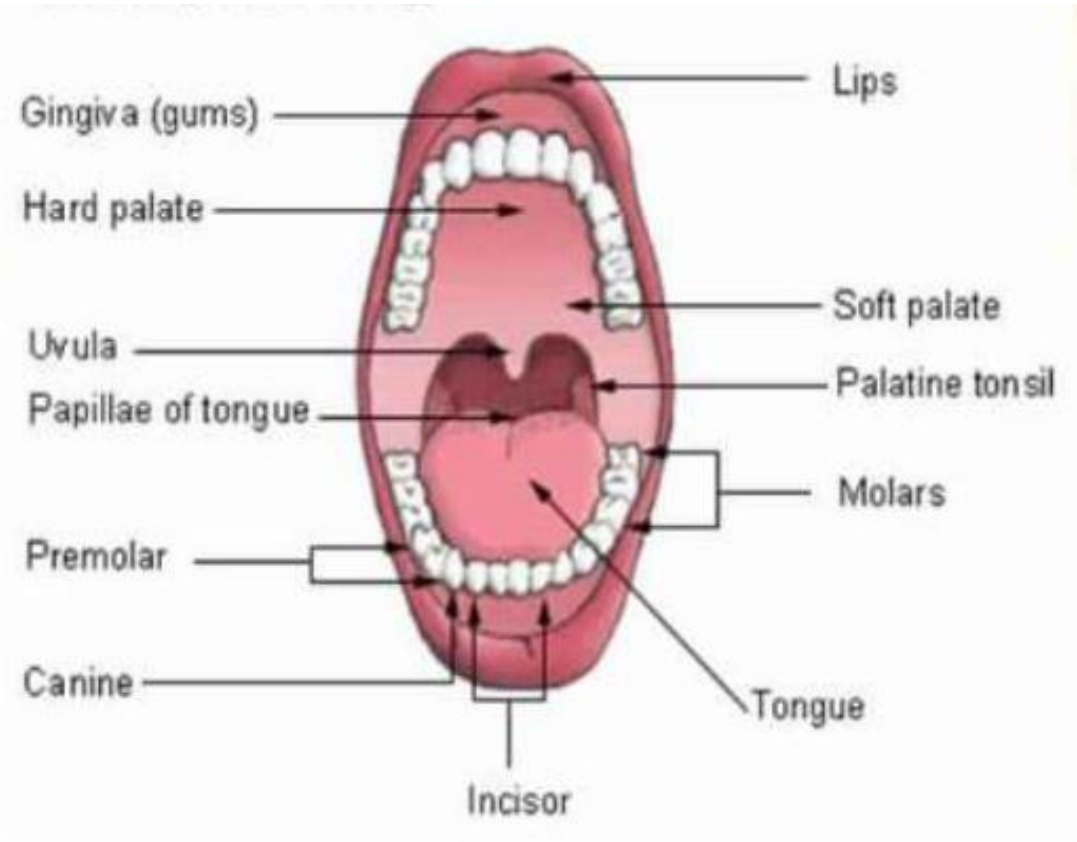


Figure 1.1: Types of Biometrics [9].

Voiceprint recognition (VPR) is the distinguishing proof of an individual dependent on an interesting trademark on their voice. Voice trademark is the mix among physical and behavioral biometric. For the physical piece of view, voice is consistent because it relies upon the size or state of the mouth, lips, vocal tracts, and nasal pits, and so on. For the conducting part, voice isn't steady. It may be changed depending on a person's feelings, affliction, or age [12]. Figure 1.2 shows the physical and behavioral biometric of voiceprint recognition where (a) the physical part and (b) the behavioral part.



(a)

(b)

Figure 1.2: physical and Behavioral biometric of VPR ,where (a) Physical part and (b) Behavioral part [8,6]

1.4 Biometric Recognition System

Biometric system, an ID system dependent on individuals' physiological or behavioral properties (e.g., iris, fingerprint, face, voice, and so on.), which attract sharp interest because of its accuracy, agility, and convenience with distinguishing identification and verification functions. Biometric systems are Inspired from conventional secret word based systems by their utilization of the client's conduct and physiological components as biometric keys. In contrast to conventional security systems, which commonly have physical keys, clients require direct cooperation with sensors in the biometric device [13]. In general, a model biometric system incorporates stages, these are; feature extraction stage, template dataset (data collecting stage), and matching stage,where the input of a biometric voice, A set of features is extracted from the required biometric voice in the feature extraction stage, The extracted features are saved in the dataset as model data. Finally, the matching stage is responsible for comparing the unknown and model data to access an accept or reject that person. A biometric system performs authentication in two stages the enrollment stage and verification stage as illustrated in figure 1.3. In the enrollment stage, a user presents the voice to the voiceprint sensor and the voiceprint is required by the sensor module. The required features of the voiceprint are extracted and saved in the model dataset for comparison in the verification stage. In the verification stage, the voiceprint of a query is collected by the sensor module. The feature representations of the query voiceprint go through the same process as in the enrollment stage to obtain query data. The query data are then compared with the model dataset so that a matching outcome is attained [7].

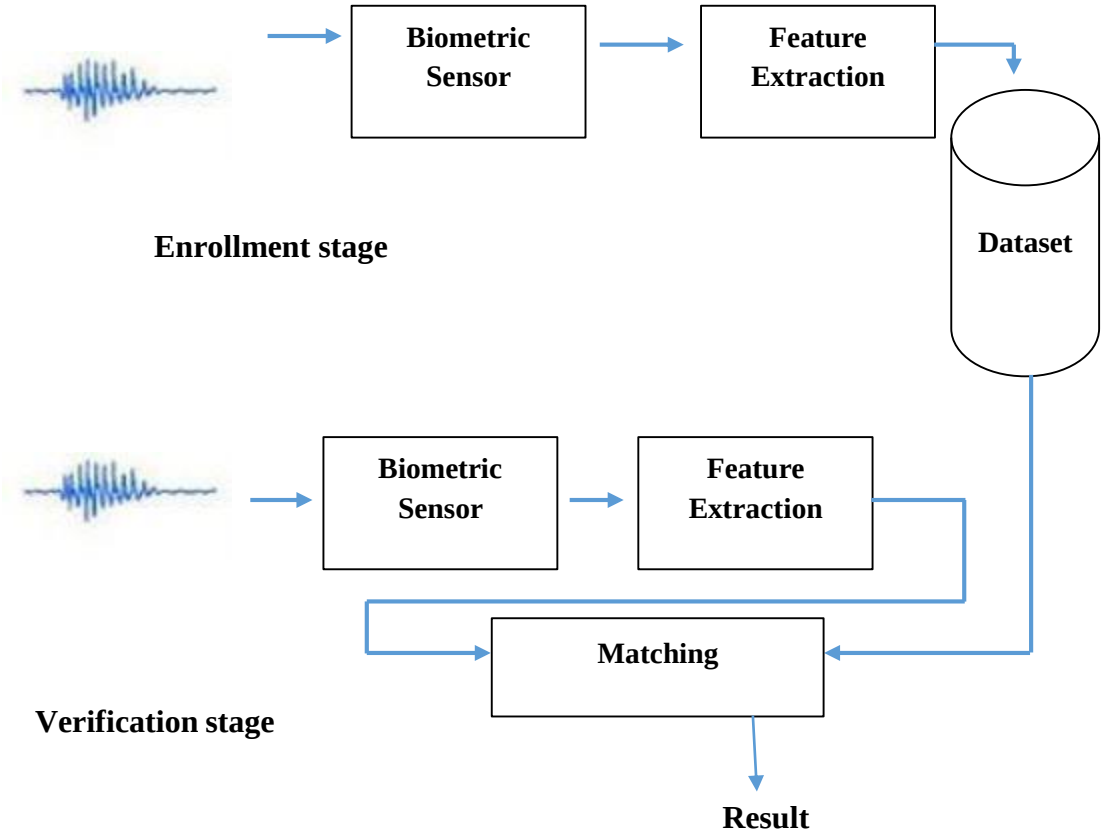


Figure 1.3:General structure of biometric system[9].

1.5 Deep Learning (DL)

A portion of machine learning is Deep Learning. Deep Learning is influenced by the structure and potential of the Artificial Neural Network, a human neuron. Artificial Neural Networks are systems that, without an explicitly specific program, learn to take actions based on examples. The architecture of ANN consists of three layers, namely input , output and one hidden layer. The foundation of deep learning is ANNs [52]. Since the 1950s, a little subset of Artificial Intelligence (AI), frequently called Machine Learning (ML) has changed a few fields over the most recent couple of decades.Artificial Neural Networks (ANN) is a subfield of ML and it was this subfield that produced -DL. Since its beginning, DL has been making ever bigger disturbances and demonstrating extraordinary accomplishment in every application area. DL utilizes either profound models of learning or progressive learning draws near. It is a class of ML that grew to a great extent from 2006 forward [14].

1.6 Literature Review

This section introduces Literature Review on voiceprint recognition which becames impotant in the last years.

T.Kinunen et al, In 2006 focus on decreasing the computational load of identification while attempting to keep the recognition accuracy reasonably high.The authors use MFCC as feature extraction and the quantization distortion with Euclidean distance as feature matching. They used two types of the dataset, the TIMIT, and the NIST dataset.The results of the proposed method were faster and more accurate than the common methods. The methods were applied using the languages C / C++ [15].

A.Revathi et al, In 2009 proposed perceptual features and iterative clustering approach for performing both isolated and continuous speech and dependent and independent speaker recognition,where used PLP(perceptual linear predictive cepstrum) and MF-PLP(Mel frequency perceptual linear predictive cepstrum) as feature extraction technique for both speech and speaker recognition, used VQ(Vector Quantization) developed by K-means clustering algorithm as classification technique named VQ codebook models,where using TI Digits_1, TI Digits_2 and TIMIT databases, where speech signal is pre-emphasized using a difference operator and passing through feature extraction (PLP and MF-PLP) and the passing through classification process (VQ codebook),where recognition accuracy for MF-PLP feature extraction isolated digit recognition system in Speaker independent better than PLP is 91%,while in Speaker dependent is equal is 99%, Recognition accuracy for MF-PLP feature extraction continuous speech recognition for both is better than PLP is 99.5% It is found that MF-PLP performs better than PLP for both speaker-independent isolated digits recognition and continuous speech recognition [16].

L.Muda et al, In 2010 presented the findings of the voice recognition study using the MFCC and DTW techniques using MATLAB. The non-parametric method for modeling the human auditory perception system, Mel Frequency Cepstral Coefficients (MFCCs) were utilized as extraction techniques. The result of the input voice signals of two different speakers confirmed that the input test voice was matched optimally with the

reference template which was stored in the database. Comparing the template with incoming voice was achieved via a pairwise comparison of the feature vectors in each. Through this study, the optimal warping path was achieved where the test input matched with the reference template. These findings were consistent with the theory of dynamic time warping [17].

V.Kulkarni et al, In 2012 presented Different techniques for feature extraction such as DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform), DST (Discrete Sine Transform), Hartley, Walsh, Haar, and Kekre transforms for comparing between their performance and for feature matching will used minimum Euclidean distance as a measure. The results obtained by DFT, DCT, DST, and Hartley transform give comparatively similar results (Above 96%). Used Dataset consists of 107 speakers where each speaker has 20 samples. The results obtained by using DFT, DCT, DST, and Hartley transform give comparatively similar results (Above 96%). The results obtained by using Haar and Kekre transform are very poor. The best results are obtained by using DFT (97.19% for a feature vector of size 40) [18].

A.Raji et al, In 2015 proposed an unconstrained text-independent recognition system using the Gaussian Mixture Model (GMM) as feature matching and MFCC as feature extraction technique. The result presented for 8 people-was fairly good but cannot be considered of being perfect authentication means [19].

J.Lee et al, In 2017 applied two types of sample-level deep convolutional neural networks that take raw waveforms as input and uses filters with small granularity. The first one is a basic model that consists of convolution and pooling layers. The second one is an improved model that additionally has residual connections, squeeze-and-excitation modules, and multi-level concatenation. It shows that the sample-level models reach state-of-the-art performance levels for the three different categories of sound. The authors also visualized the filters along with layers and compared the characteristics of learned filters. The results show the possibility that they can be applied to different audio domains as a true end-to-end model [20].

M.Ravanelli et al, In 2019 were proposed a SincNet, as a novel CNN for processing raw audio samples that encourages the first layer to discover meaningful filters by exploiting parametrized sinc functions, which implement rectangular bandpass filters. In contrast to standard CNNs, which learn all the elements of each filter, only low and high cutoff frequencies of band-pass filters are directly learned from data. This inductive bias offers a very compact way to derive a customized front-end, that only depends on some parameters with a clear physical meaning. The experiments, conducted on both speaker and speech recognition, show that the proposed architecture converges faster, performs better, and more computationally efficient than standard CNNs. The SincNet outperforms other systems on both TIMIT (462 speakers) and Librispeech (2484 speakers) datasets[21].

T.Kim et al, In 2019 the authors work study on convolutional neural networks (CNNs) on audio classification, where they did comparison between CNN based wavelet and CNN based spectrogram on audio classification and they proposed sample CNN based wavelet audio, where they used three different audio domains: music, speech, and acoustic scene sound .The language used in this work TensorFlow and Keras. The best performance was obtained by SampleCNNs when the SampleCNNs have the smallest filter and stride sizes [22].

D.Salvati et al, In 2019 were present a raw waveform (RW) end-to-end computational scheme for speaker identification based on CNN's with noise and reverberation data augmentation (DA). The proposed CNN consists of 5 one-dimensional convolutional layers, 3 fully connected layers, and a classification layer with softmax function.The CNN is designed for a frame-to-frame analysis to handle variable-length signals.They analyze the identification performance with simulated experiments in noisy and reverberation conditions comparing the proposed RW-CNN with the melmel-frequency cepstral coefficients (MFCCs) features. The source speech signals were taken from the TSP speech database to produce noisy and reverberant expression. The TSP speech database consists of 1378 pronouncements from 23 speakers (12 females, 11 males).The results show that the method offers robustness to adverse conditions. The RW-CNN

outperforms the MFCC-CNN in noise conditions, and they have similar performance in reverberant environments [23].

S.Bunrit et al, In 2019 In this report, the CNN-based approach for text-independent speaker recognition is suggested. Each signal wave sample is converted into a spectrogram. Compared to the classic Mel-frequency cepstral coefficients (MFCCs) based featured extraction method classified by support vector machine (SVM), the suggested CNN that the spectrogram image use as an input also compares to a case where raw signal wave image is used to the CNN model.It shows that the proposed CNN-based method trains are the best compared to the other two methods on spectrogram voice picture. The average result of the testing classification developed by the proposed method is 95.83 per-cent accuracy. The system based on MFCC is 91.26 pe-rcent and the raw signal wave picture trained for CNN is just 49.77 per-cent. For text-independent methods , the proposed method is very effective [24].

1.7 Problem Statements

When attempting to build a voiceprint recognition device, some problems arise. Most of these issues are due to the fact that it is almost difficult to pronounce a word exactly the same way on two separate occasions. How easily the word is spoken, stressing various parts of the word, background noise, and so on. are several variables that continually alter the human speech signal. For that purpose, propose a deep learning strategy that will provide a way to learn the recognition of voiceprint implicitly.

1.8 Aim of the Thesis

The main objective of this thesis is to design a deep learning strategy, which will provide a way to implicitly learn the voiceprint recognition in noisy environments and trying to increase the accuracy of the system by using special structures of convolution neural networks have been used for the classification and the system ability to deal with a huge dataset with adding random noise to prove the efficiency of the system in noisy conditions. Where, This work passes in two phases,the first phase have been using the Mel-Frequency Cepstral coefficients (MFCC) were applied to the raw waveform to extract the features (cepstral coefficients). Then, an appropriate architecture of the DL

algorithm which is in particular convolution neural network algorithm was applied. (MFCC-CNN). The second phase have been using convolution neural network directly on raw waveform(RW-CNN).The main obtain results from the MFCC approaches were compared with that obtained from applying the CNN directly on the raw waveform with and without adding noise. Comparing these results was illustrated to prove the efficiency of DL on recognition of a person from his/her voice in noisy environments.

1.9 Outline of The Thesis

The thesis consists of five chapters (including this chapter), which include a general introduction to voiceprint recognition, an outline of the entire thesis, and a literature review of previous studies in the field of voiceprint recognition.

Chapter Two Provides the background of voiceprint recognition and deep learning algorithms that have been applied in this field and how these technologies have been grown in recent years.

Chapter Three Explains the proposed system and all its details including system requirements and tools that are used to implement this system.

Chapter Four Contains the simulation of the proposed system and the discussion and analysis of the results that are obtained through this work.

Finally, **Chapter Five** Includes the most important work conclusions and recommendations for future work.

Chapter Two

Theoretical Background

Chapter Two

Theoretical Background

2.1 Introduction

Adoption of biometric technologies was witnessed widespread in the last decade, such as fingerprint scanners on laptops, cameras with built in face recognition capabilities at airport terminals and stadiums, and voice based authentication technologies for mobile account access. Among biometric authentication technologies, voice-based authentication plays a pivotal role due to the exponential growth in the smartphone user base, and then it provides unprecedented simplicity. However, the human voice can easily be recorded over large distances simply via a standard phone line without needing any special reader system. In addition, Voiceprint recognition is a biometric system that uses the characteristic features extracted from their speech samples to perform the computing task of validating the claimed identity of a user.Biometric recognition and authentication are now favored over traditional methods such as pin and passwords. The biometric character based identification approach is easier, unique to each individual, and more reliable .One such program that uses biometrics is a speech recognition system. Voiceprint recognition is the identification of an individual from the characteristics of his/her speech (voice biometrics). It uses the biometrics of voice and the acoustic characteristics of speech which are different in different people. The acoustic patterns display anatomy such as the size and shape of the vocal tract, and the learned behavioral patterns such as voice pitch and speech style [25].

Like any other biometric device, the voiceprint recognition system (VPRS) also has two parts: the registration part and the verification part. The registration part;is the training process, in which voice sequences are obtained from different speakers and the appropriate features are extracted to form a specific model (voiceprint model) that identifies the speaker uniquely. In the verification part, also called the testing phase (also known as the recognition phase), the voice sequence is taken in real time. The features of voice are extracted and compared to previously created model. [2,25].

In this function, multiple feature extraction algorithms are used, such as: Linear Predictive Coefficients (LPC), Mel Frequency Cepstral Coefficients (MFCC), Linear Cepstral Coefficient (LPCC), Discrete Wavelet Transform (DWT), Wavelet Packet Decomposition (WPD) and Perceptual Linear Prediction (PLP),and so on. The general audio-classification approach includes extracting special features [26].

Figure 2.1 displays a generic block diagram for VPRS. The input for both the training and the test process is the human voice, which is a slowly time-varying signal. The feature extraction block extracts the features for each person which is his/her featuresare stored in the template or compared based on the training or testing phase. [2].

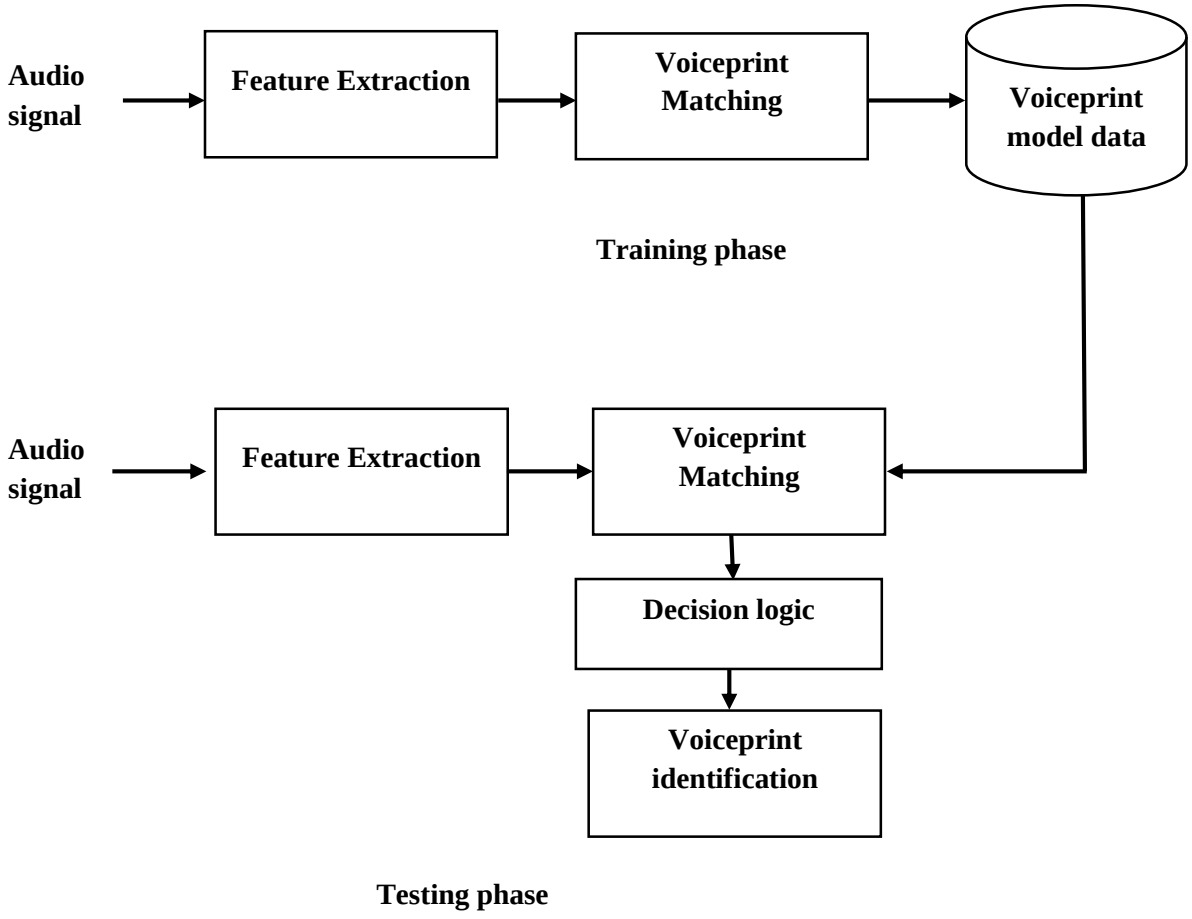


Figure 2.1: Generic Block Diagram of a VPRS [34].

2.2 Classification of VPRS

VPRS can be broken down into several categories. Figure 2.2 shows the different classifications of voiceprint recognition system:

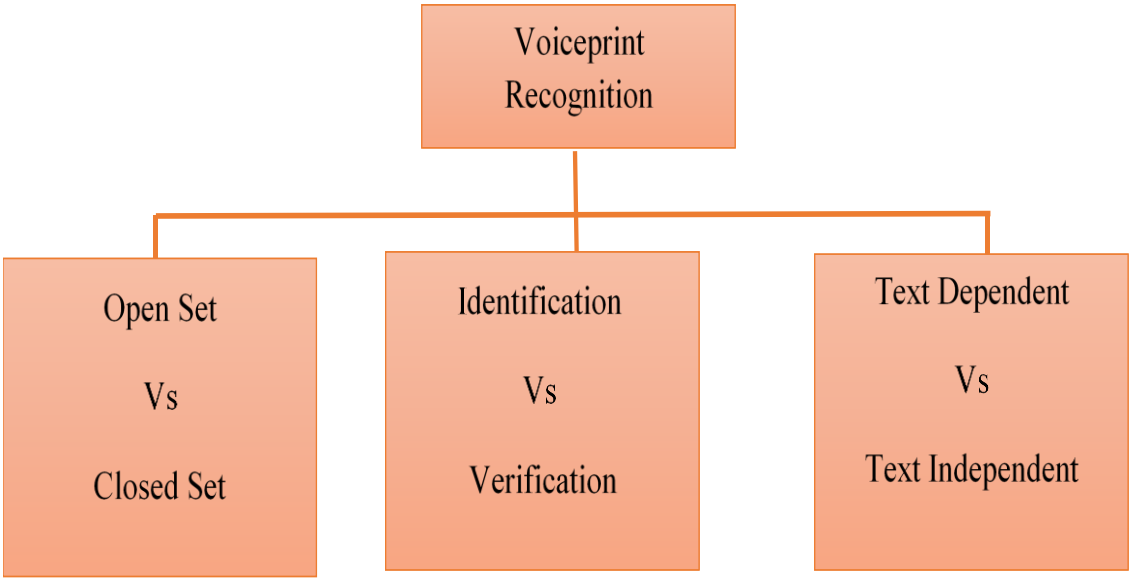


Figure 2.2: VPRS classification [27].

2.2.1 Open Set vs. Closed Set

Voiceprint Recognition can be divided into an open set and closed set voiceprint recognition. The classification category is based on the collection of qualified speakers available in a program. Let's get them in depth.

1. Open Set: there can be a variety of qualified speakers in an open set system. In which, there are an open set of speakers and the number of speakers can be more than one.

2. Closed Set: a closed set system only has a defined (fixed) number of registered users [27] on the system.

In this thesis, a closed group of qualified speakers has been employed.

2.2.2 Identification vs. Verification

This category of classification is the most important among the lot. Automatic voiceprint identification and verification are often considered to be the most natural and economical methods for avoiding unauthorized access to physical locations or computer systems. Let us discuss them in detail.

1. Voiceprint Identification (VPI): It is the method of deciding which registered speaker provides a given utterance. In the identification process, the voice of each speaker is gathered and used to create the corresponding model for that speaker. The compilation of voice models for all speakers is called the speaker dataset [27,28].

2. Voiceprint Verification (VPV): This is the method of approving or denying a speaker's identity claim. The basic differences between voiceprint identification and verification systems are outlined in Figure 2.3 and Figure 2.4 [29,27].

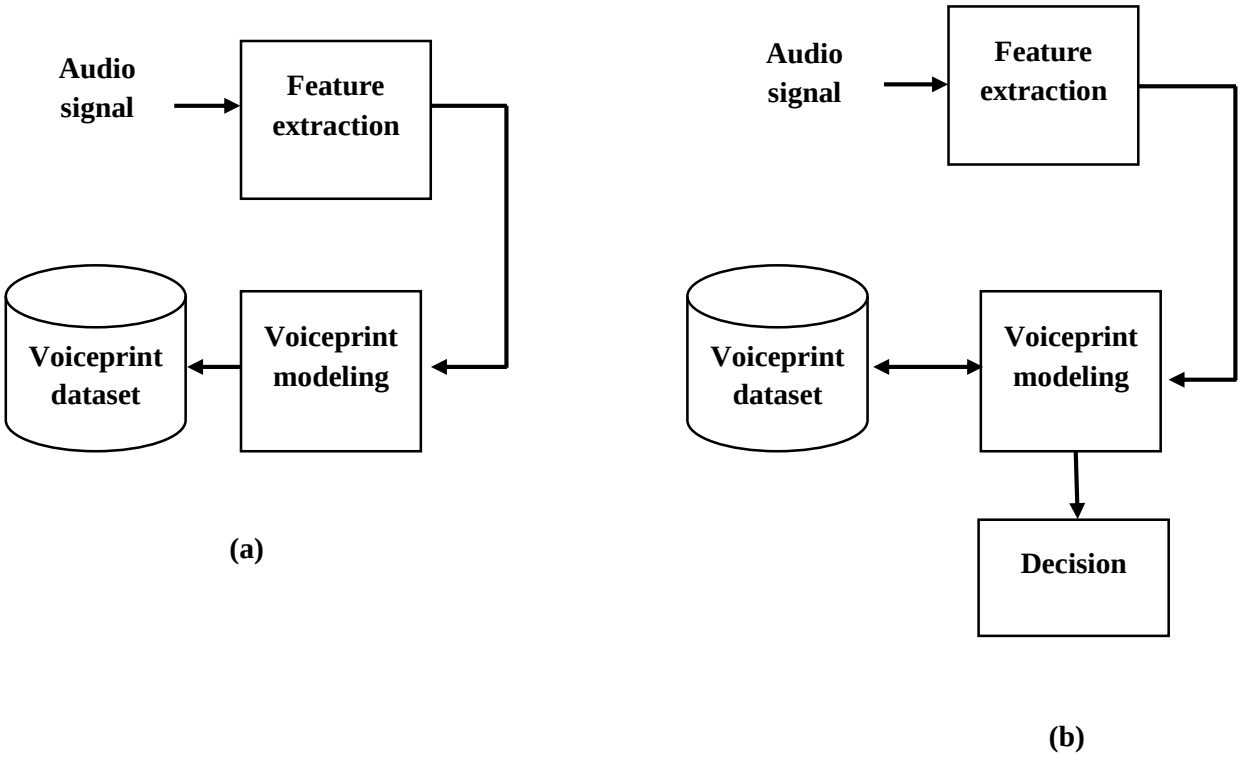


Figure 2.3: VPRS: (a) The VPI phase; (b) The VPV phase [28].

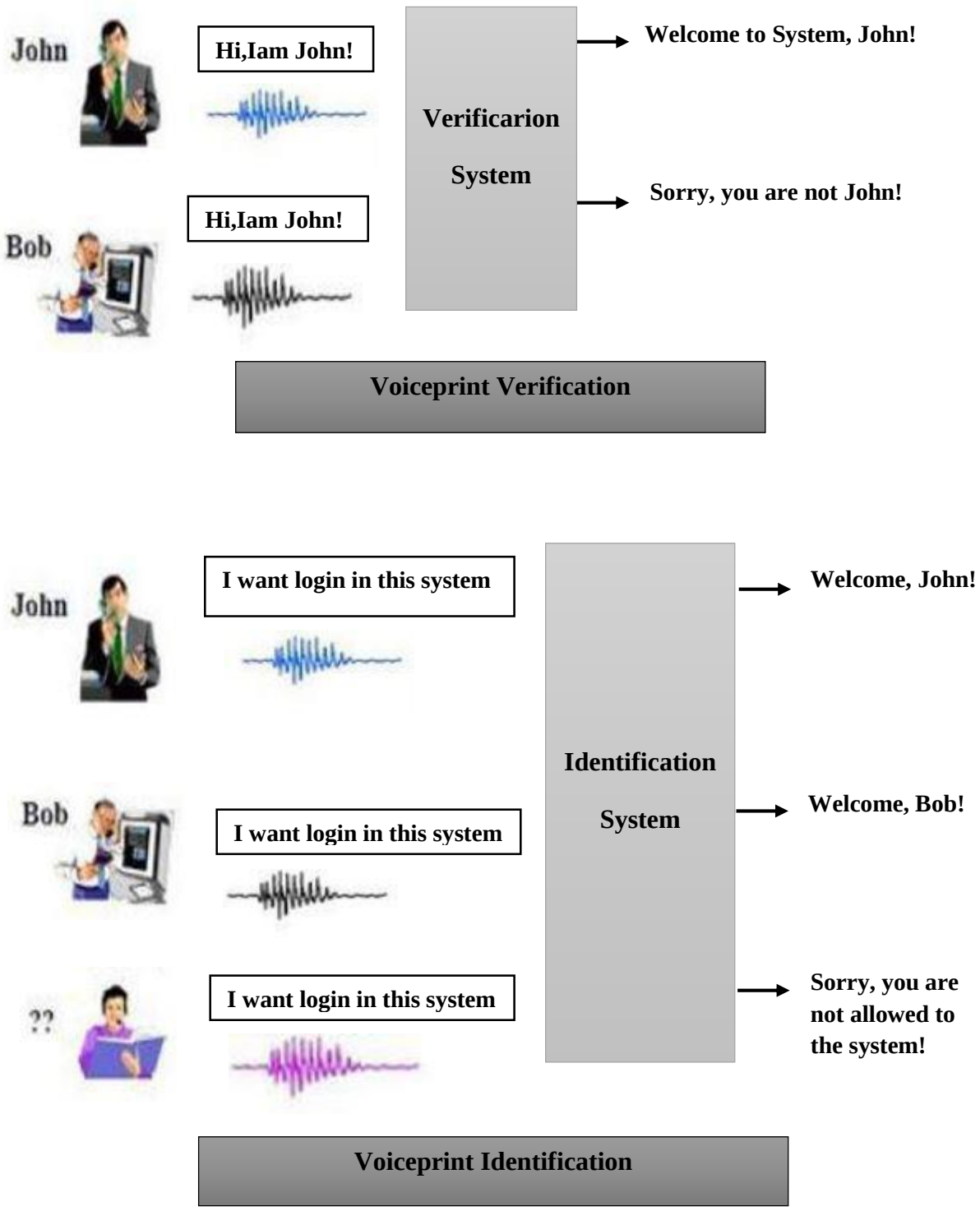


Figure 2.4: An example about VPI and VPV [27].

2.2.3 Text-Dependent vs. Text-Independent

This is another classification category for voice recognition systems. Each category is based on the text that the speaker spoke during the process of identification. Let us talk about each in detail[27,29].

1. Text-Dependent: in this case, the test utterance is similar to the text used during the training phase. The test speaker has previous system experience.

2. Text-Independent: In this case, the test speaker has "no previous knowledge of the training process contents and can say anything.

2.3 Voiceprint Benefits and Disadvantages

Before disclosing the basic properties of the voiceprints and their potential utility it is important to understand the power of biometric keys. How many times you use your cell phone or credit card to forget the PIN. The number may then fall into the wrong hands and build an incredibly awkward situation. Additionally, this form of protection is based on encryption algorithms that can be broken and proven to be ineffective on their many times. Biometric keys come to solve some of these problems since they are "printed" in the human body and cannot be lost as such. Replication of these keys includes the direct participation of the proprietor, and this replication is sometimes difficult depending on the type of biometric key. In conclusion, biometric keys are special, hard to replicate or forget, and it always is with their registered user [30,31]. A good biometric key must follow certain specifications. The extracting, measuring, saving, and comparing are need to be easy. Voiceprints satisfy all these requirements because very expensive hardware is not required to perform any of these operations. The only equipment needed is a microphone and a laptop. Most modern banking apps depend on the use of telephone lines. Voiceprints are the only appropriate biometric technology capable of working without the need for external hardware in this area. To sum up, the front-end infrastructure needed for speaker recognition is very simple and is in many cases already in place, so generally making a VPR run requires a low investment. There is a possible feature in the voice, which may be that the voiceprint does not require the presence of the person as it can be recorded anywhere and sent via

the phone where almost every person owns a microphone or a phone device compared to fingerprints and iris and face scanning where it requires a scanner. The key drawbacks of voiceprint recognition VPR are that the voice depends on the speaker 's safety, that the voice is inconsistent during life, and that the microphone or channel used to transmit the audio has a significant impact on the audio signal. A cough or flue may result in a handicap to the production of voice and ultimately alter a person's natural speech, in this case, an AVP may not be able to recognize an approved speaker. It is also obvious that in his childhood, a person's voice is very different from that in the adult stage, not to mention the eldership. Some algorithms can handle this problem by re-adapting models of speakers over time. Finally, the microphone used during the training phase may be different from the one used in the test. It can lead to a malfunction of the models with the new microphone recorded voice. Some algorithms can reduce the effects of the microphone and the channel of transmission over the speech signal [30].

2.4 Simple VPRS

Voice is probably the most important mode of contact with humans. Human voice or speech is an information-rich signal that transmits a wide range of information such as language material, emotions of the speaker, tone of speech, and so on. The VPR seeks to separate, identify, and recognize a speaker based on speech characteristics. Several methods can simplify the process of speaker recognition. Such systems typically involve two phases of extraction the features and matching or classification of the features, where the element of classification has two components: the pattern matching and the decision. Figure 2.5 depicts a generic voiceprint recognition system. The feature extraction module estimates a collection of speech signal features that reflect some speaker-specific information, where the voice(s) of each speaker is collected and used to construct the corresponding speaker model. The compilation of voice models for all speakers is called the voice dataset [32,28].

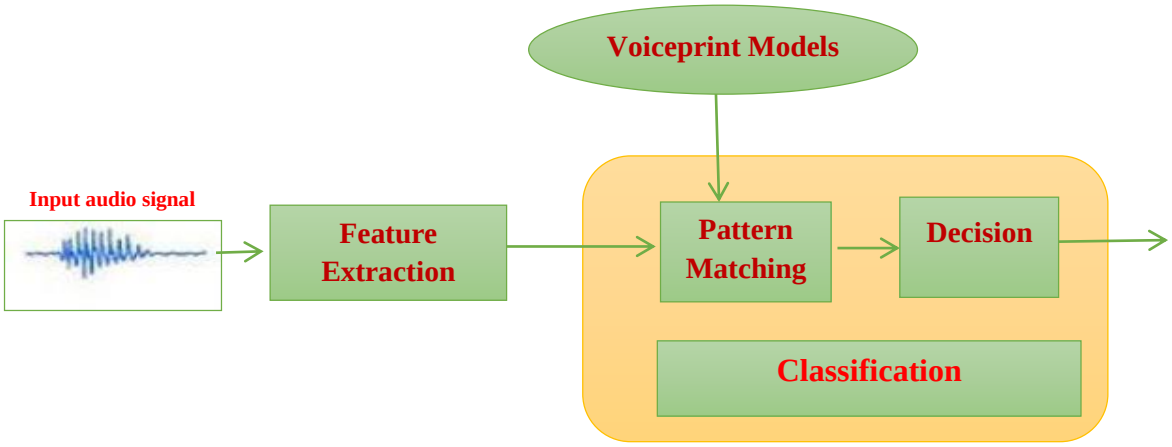


Figure 2.5: Generic method for recognizing VPRS [28].

Feature matching is responsible for matching the approximate features of the speaker versions. There are many types of pattern matching methods used in speaker recognition and the corresponding models [13]. Some of these methods include Hidden Markov Models (HMM), Dynamic Time Warping (DTW), and Vector Quantization (VQ), Artificial Neural Networks (ANNs), and Deep Learning Algorithms [28].

2.5 The Traditional VPRS

The speech signal is defined as there is information in a message that is transmitted by a speaking person. This information is spoken language, the speaker's feeling, gender, and identity [33]. VPR is a process of distinguishing a person from his / her speech signal based on particular information obtained. As explained in the previous chapter. Two forms of VPR exist; VPI and VPV. VPI is the process of determining the identity of the person from among peoples who speaks. VPV is the method of accepting or denying the identity of a speaker who has previously been determined [34].

Figure 2.6 [38] displays a general block diagram of typical VPRS. VPRS involves the process of converting a voice waveform into features [36]. Audio recognition works based on the analysis of speech signal characteristics that differ between individuals. Each has in their speech a distinctive property [37], where these properties were used as input to a classifier (ANN), Which eventually gives us the identification or verification

decision. As explained in previous chapter VPR methods classified into text-dependent and text-independent depending on the text, which is said. In this case of the system of text-dependent voiceprint recognition, it requires the person to speak the same sentences in both training and testing phases, while text-independent voiceprint recognition the sentences may not be the same in both training and testing phases [34]. The feature extraction and classification are very important in every VPRS, where the MFCC algorithm was used for features extract phase.

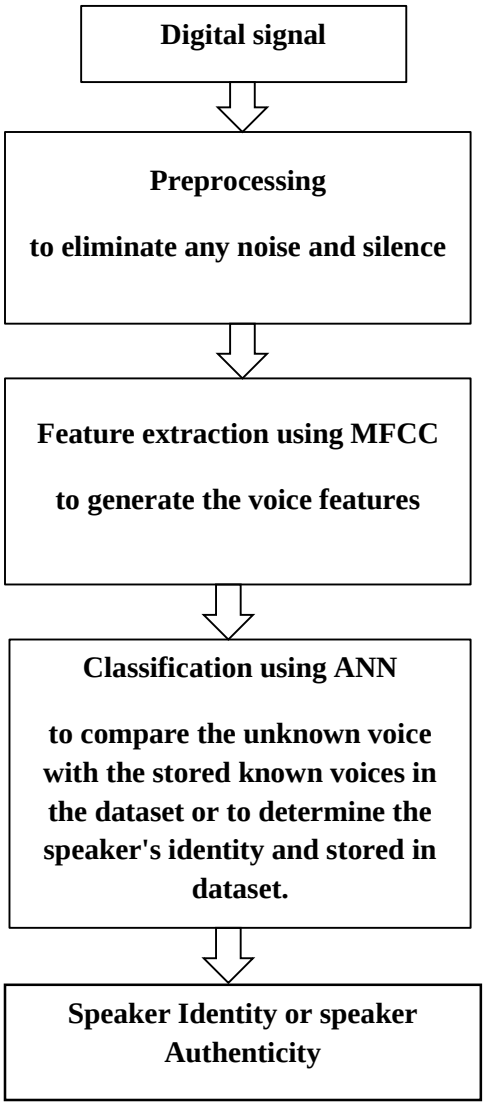


Figure 2.6: Traditional VPRS [38]

In Figure 2.6, the human audio is a digital signal type to obtain the digital data for each generated signal. Then, to eliminate any noise and silence in the speech signal, it goes through the preprocessing step, to generate the voice features it goes via MFCC. The voice coefficients may then move through ANN in order to compare the unknown voice with the stored known voices in the dataset or to determine the speaker's identity and stored in dataset. This segment discusses the speech results and the analysis of the VPR using conventional MFCC and ANN techniques.

2.6 Voice Feature Extraction

Speech signal encompasses a broad variety of information about speakers, including 'high-level' information including language, context, spoken language, mood,and so on. High-level features include more comprehensive speaker-dependent details; and are very hard to find. Instead, information at low levels, such as pitch, speed, tempo, band, audio spectrum,and so on. Can be easily extracted and is considered to be highly efficient for the implementation of automated VPRS. The MFCC algorithm proposes an efficient way to extract enough information to distinguish between one person and another [32]. The purpose of this module is to transform the speech waveform into a collection of features, or rather feature vectors, used for further analysis.

2.6.1 Mel Frequency Cepstral Coefficient (MFCC)

MFCC is based on evidence that low-frequency part information is more relevant in phonetic terms than high-frequency sounds [32]. Figure 2.7 demonstrates the step by step algorithm used to extract MFCC functions.

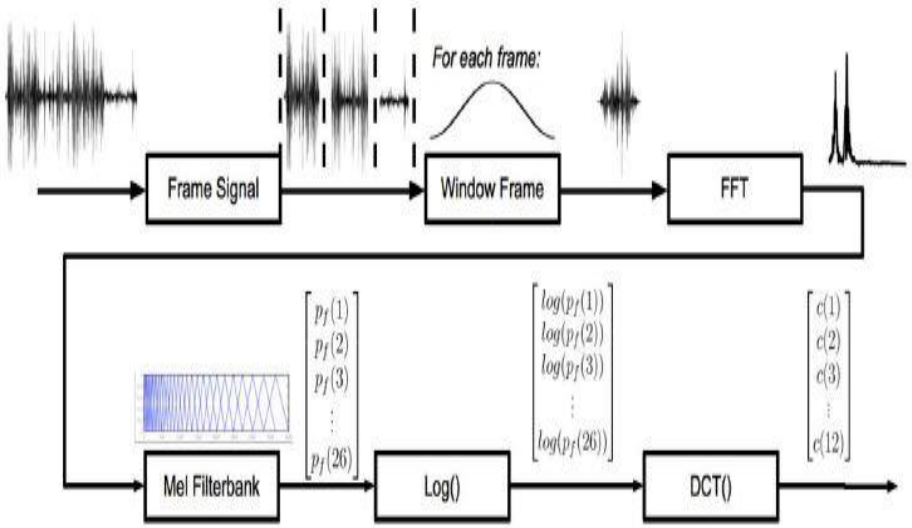


Figure 2.7: MFCC Processor [39]

1 - Frame Blocking

In this stage, the continuous speech signal is divided into N sample frames, where by M samples separate the adjacent frames with the value M less than N. The first frame integrates N's first measurements. The second frame begins from M samples and overlaps it with N-M samples, and so on. The cycle continues until one or more frames are used for the entire speech.

In this work, the M and N values were selected to be N= 256, and M= 128.

The N value is chosen as 256 because it is presumed that the voice signal is intermittent over the duration. As a power of 2, the length frame 256 can also be used to quickly apply the Discrete Fourier Transform (DFT), also known as the FFT (Fast Fourier Transform) [27,40]. Figure 2.8 shows audio signal in framing stage.

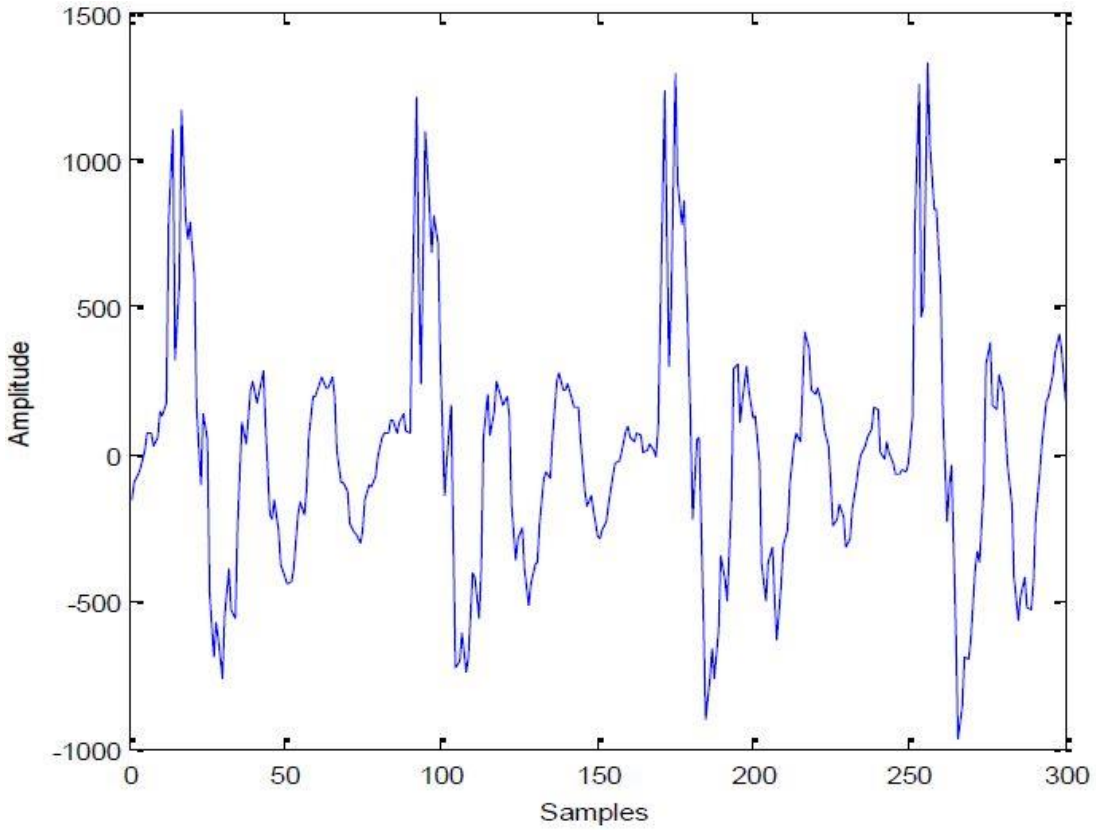


Figure 2.8: Framing step for Audio Signal [41].

2 - Windowing

The next stage is to window each frame so that the signal discontinuities at the beginning and end of each frame are minimized. The principle applied here is to eliminate the spectral distortion by using the window at the beginning and end of each frame to taper the signal to nil. If the window is defined as w(n), 0 ≤n≤N-1, where N is the frame length, then the result of windowing is the signal of equation (2.1) [27,40].

$y(n) = x(n)w(n)$, 0 ≤ n ≤ N-1 (2.1)

In this thesis, the used Hamming window has the form:

$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$, 0 ≤ n ≤ N-1 (2.2)

Where N represents the number of samples in each frame.Figure 2.9 shows windowing stage for audio signal.

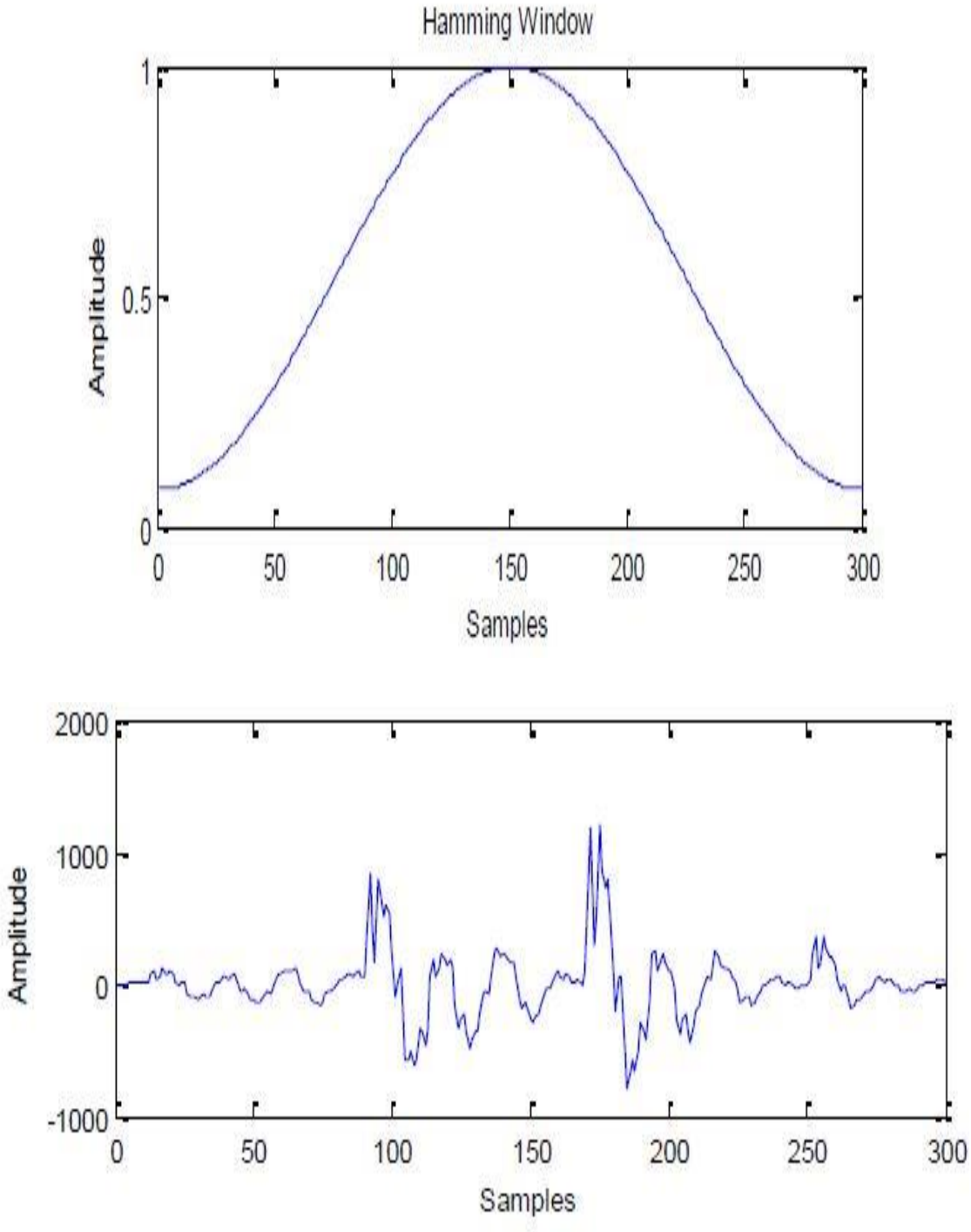


Figure 2.9: Hamming Window Form [41].

3 - Fast Fourier Transform (FFT)

The third stage is the Fast Fourier Transform (FFT) application which converts every frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm for implementing the Discrete Fourier Transform (DFT), which is defined in the N sample set { xn } as follows:

$Xk = \sum_{n=0}^{N-1} xne^{-\frac{j2\pi kn}{N}}$, K=0,1,2,…,N-1 (2.3)

Xk represents a series of audio signals after transformed it into a frequency domain [27,40]. Figure 2.10 illustrate the fft stage for audio signal.

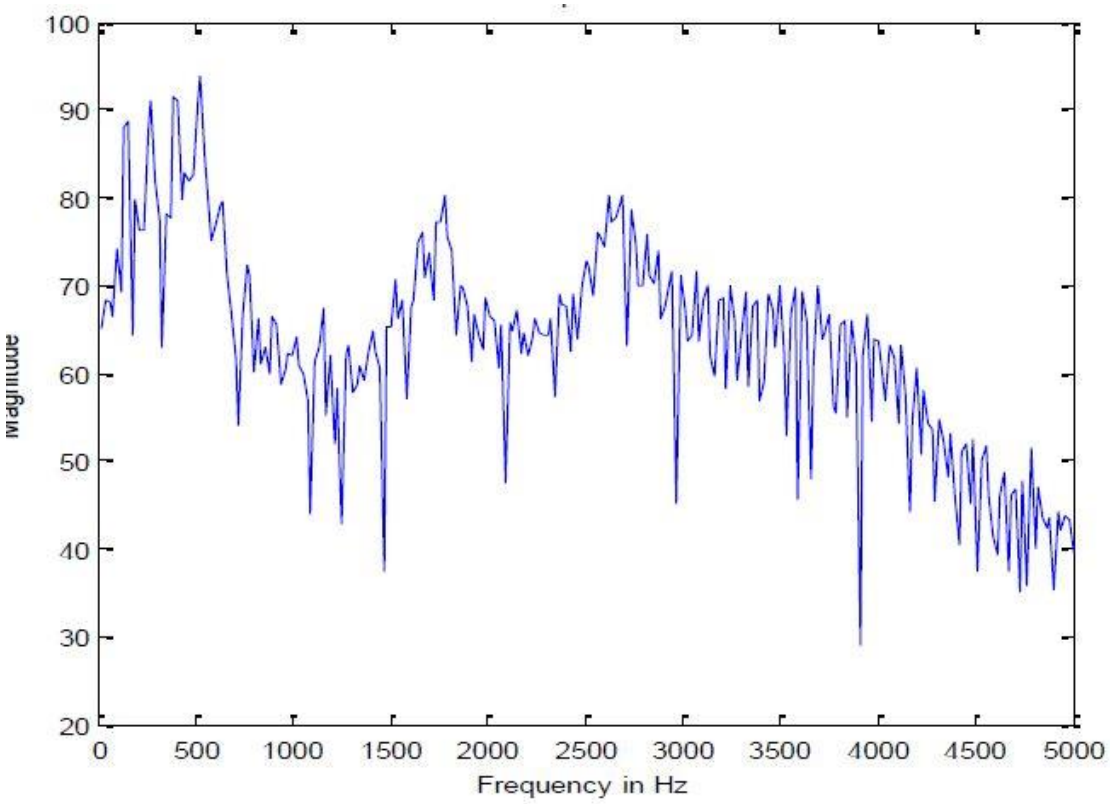


Figure 2.10: FFT for Audio signal [41].

4 - Mel Filter Bank

A filter bank consists of a series of filters used to extract information from input signal studies, demonstrating that human interpretation of sound frequency content for speech signals does not obey a linear scale, where of frequency measured in Hertz, while sound measurement unit on a scale called the " Mel" scale. The frequency scale for mel has a linear spacing of frequencies below 1000 Hz and a logarithmic spacing

above 1000 Hz. Can, therefore, use the following approximate formula for calculating the mels in Hz for a given frequency f [20].

$$mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (2.4)$$

Where f is the sampling frequency (sampling frequency or sampling rate, which is the number of bits used per sample), 8kHz, 16kHz, 11, kHz and 22kHz are typically the same.Figure 2.11 shows Mel Filter Bank stage fo audio signal.

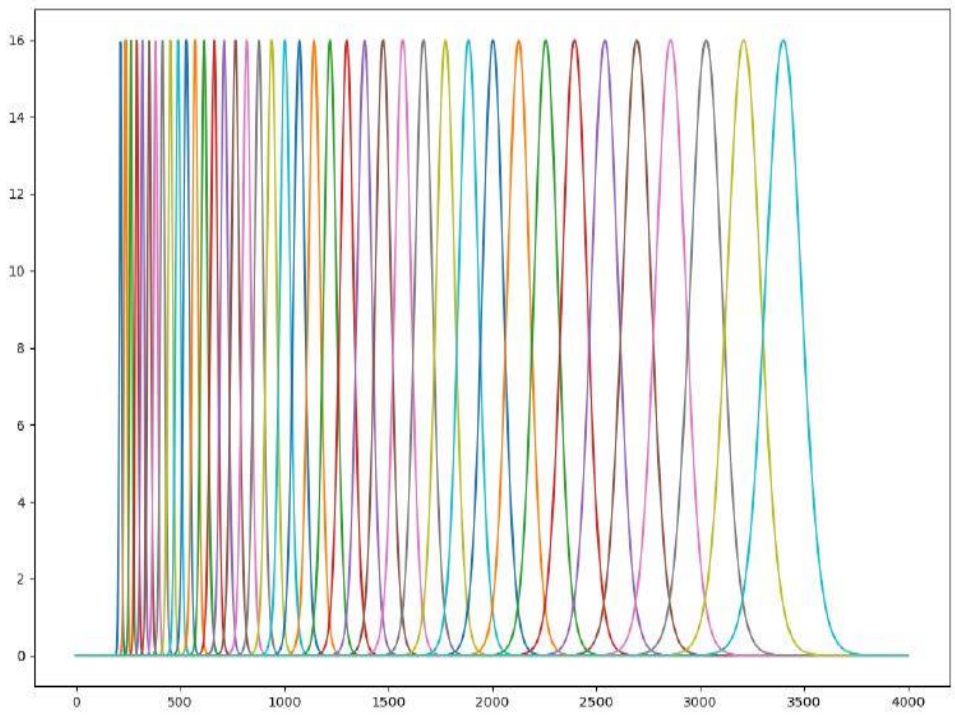


Figure 2.11: Mel-Filter Bank [42].

5 - Cepstral Coefficients using Discrete Cosine Transform (DCT)

For extract attribute as MFCCs, the term "cepstrum" is the same word of the term "spectrum." [39]. Using DCT to transform the Mel log spectrum into a time domain. The result of that conversion is the cepstrum coefficient of Mel frequency. In other

words, this step results in coefficients that are known as features. The Cepstral Coefficients were determined using the formula (2.5) [27,40].

$$Ceps = dct(\log(FFT(ywindowed))) \quad (2.5)$$

Implementing the MFCC For VPRS and it is results will be shown in Chapter 4. Figure 2.12 shows Cepstral Coefficients using DCT for audio signal.

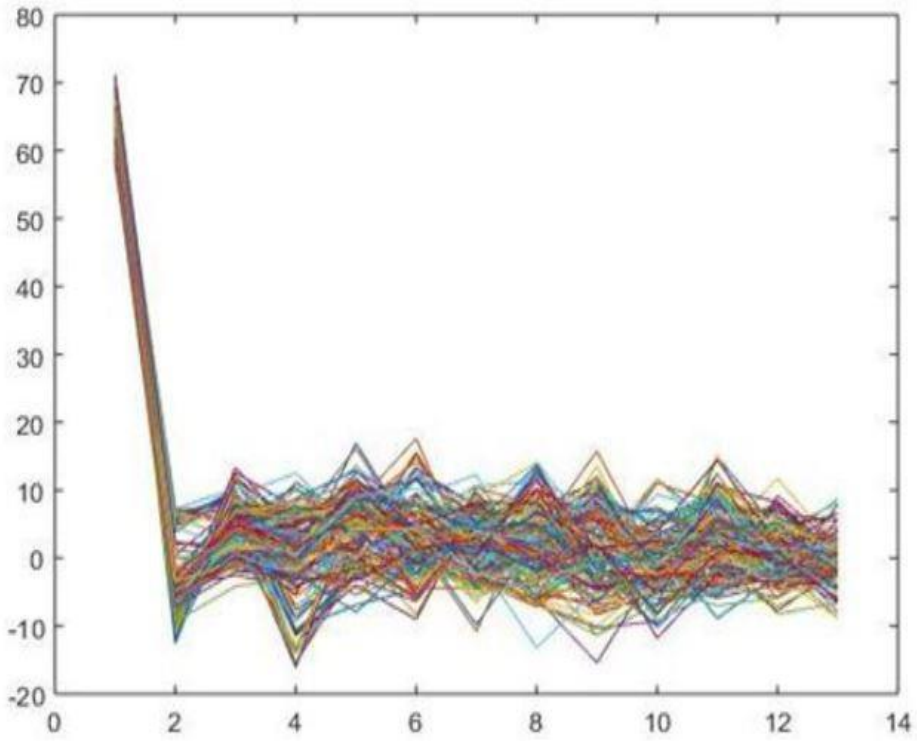


Figure 2.12: Cepstral Coefficients [43].

2.7 Artificial Neural Network (ANN)

ANN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the new structure of the information processing system. It is composed of a huge number of highly interconnected processing elements (neurons) working together to solve specific problems. ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. Learning

in biological systems adds adjustments to the synaptic connections that exist between the neurons [44].

An ANN consists of three layers: an input layer, an output layer and hidden layers as shown in figure 2.13. Pattern recognition in NN is one of the important steps in Image Processing and speech processing. The first step in pattern recognition is to select a set of features or attributes from the speech sample that will be used to classify the pattern. Next, the original pattern must be transformed into a representation that can be easily manipulated programmatically [45].

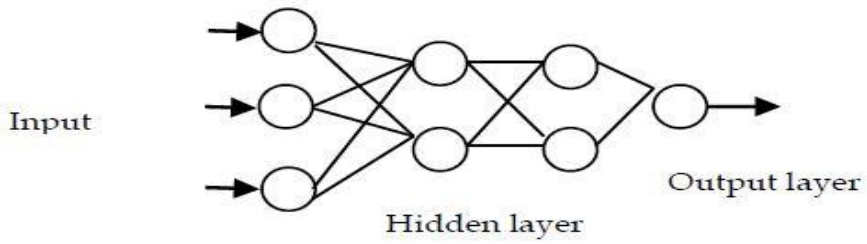


Figure 2.13: ANN Structure [45].

For each layer, the nodes or units are connected to the nodes for neighboring layers. Every relation has a value in weight. The inputs are multiplied by the respective weights and each unit is summed. The sum then undergoes a transformation based on the activation function, which is often a sigmoid function, tan hyperbolic, or Rectified Linear Unit (ReLU). Such functions are used because they have a mathematically favorable derivative which makes it easier to measure partial error delta derivatives concerning individual weights. Sigmoid and tanh both squeeze the input into or alternative to a small output set, i.e. 0/1 and -1/+1 respectively. Implemented saturated nonlinearity as the plateaus or saturates the respective thresholds before / after the output. In contrast, ReLu exhibits both saturating and unsaturating behaviors with f(x)=max(0,x). The function 's output is then fed into the next layer as input to the subsequent unit. The output layer end result is used as the solution to the problem [46].

Implementation of Artificial Neural Networks is composed of.

1. Acquire data collection for training and testing

2. Train the network

3. Predict with Test Data

Artificial Neural Networks can be classified into the following different types.

1. Feedforward Neural Network

2. Recurrent Neural Network (RNN)

3. Radial Basis Function Neural Network

4. Kohonen Self Organizing Neural Network

5. Modular Neural Network

2.8 Deep Learning (DL)

ML technology supports the modern society in many ways, where machine learning has become more and more popular in research and has been incorporated in a large number of applications, including multimedia concept retrieval, image classification, video recommendation, social network analysis, text mining, and so on. Where this technology has been used in many applications on a wide range such as cameras and smartphones and also used in visual data processing, speech and audio processing, and many other applications. DL is part of a wider family of machine learning methods focused on representations of the learning data,as opposed to task-specific algorithms. DL has enabled many practical applications of machine learning and by extension the overall field of Artificial Intelligence. Compared to shallow learning deep learning has the advantage of building deep architectures to learn more abstract information. The most important property of deep learning methods is that it can automatically learn feature representations thus avoiding a lot of time-consuming engineering. Better chip processing abilities, considerable advances in the machine learning algorithms, and affordable cost of computing hardware are primarily crucial reasons for the booming

of deep learning. Traditional ML relies on shallow networks which are composed of one input and one output layer, and no more than one hidden layer between input and output layers. DL is qualified when more than three layers exist in a network including input and output layers. Therefore, the more the number of hidden layers is increased, the more the network gets deeper [47,48].

2.9 DL algorithms

Deep learning has been growing very fast, several new networks and new structures appear every few months, currently, some popular deep learning algorithms are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM), and Autoencoders algorithms.;In this thesis, CNN will discuss in detail.

2.9.1 Convolutional Neural Network (CNN) Overview

CNN was firstly introduced by Kunihiko Fukushima. It was later proposed by Yann LeCun. He combined CNN with back-propagation theory to recognize handwritten digits and document recognition. His system was eventually used to read hand-written checks and zip codes. CNN uses convolutional layers and pooling layers. Convolutional layers filter inputs for useful information. They have parameters that are learned so that filters are adjusted automatically to extract the most useful information for a certain task. Multiple convolutional layers are used that filter images for more and more abstract information after each layer. Pooling layers are used for limited translation and rotation invariance. Pooling also reduces memory consumption and thus allows for the usage of more convolutional layers [48,49].

2.9.1.1 Convolution Operation

For each stage in CNN / layer of convolution. Convolution is a mathematical operation is used to convolute with the incoming data, where input data and a convolution kernel are subjected to a particular mathematical operation to generate a feature map. Convolution is often interpreted as a filter, where the kernel filter is applied once at a time to convolute with the data, where the resulting of convolution operation is the stack of N feature maps as shown in figure 2.14, when N filters are applied. Convolution is described formally as follows [48,24].

$$h(t) = \int_{-\infty}^{\infty} f(T)g(t-T)dT \qquad (2.6)$$

where;f(T) represents the input,g(t-T) represent kernel filter as shown in the following example:

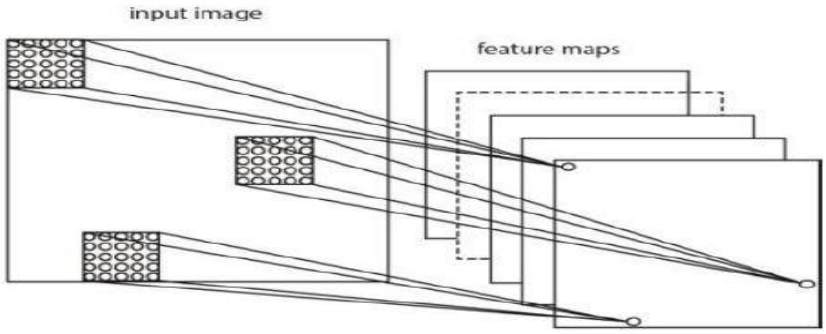


Figure 2.14: Convolution operation in CNN [24].

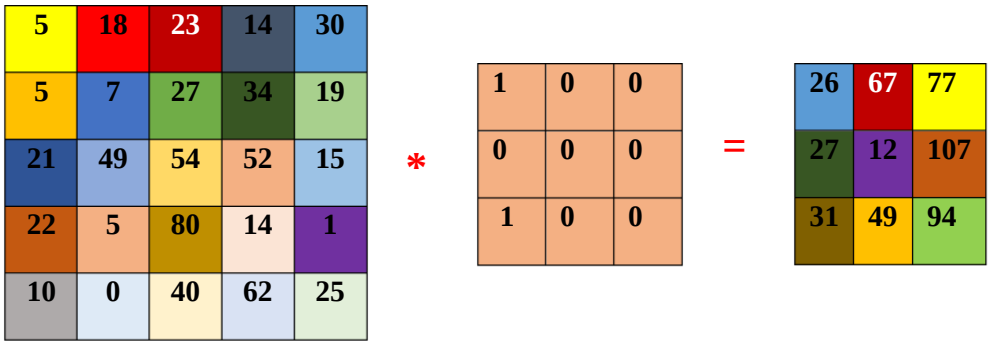


Figure 2.15: An Example on Convolution Operation.

CNN typically works with two dimensional convolution operation as summarized in figure 2.15 The leftmost array is input data. The array in the middle is convolution kernel and the rightmost array is a feature map. The feature map is calculated by sliding convolution kernel over the entire input array. The convolution process is an element wise operation followed by a sum. For example, when the right upper 3×3 array is convoluted with a convolution kernel, the result is 77 .The convolution operation is usually known as kernels. By different choices of kernels, different operations of the images can be obtained. Operations are typically including edge detection, blurring, sharpening,and so on. By introducing random matrices as a convolution operator, some interesting properties might be discovered. As a result of convolution in neural networks, the image is split into perceptrons, creating local receptive fields and finally compressing the perceptrons in feature maps. All in all, learning a meaningful convolutional kernel is one of the central tasks in CNN when applied to computer vision tasks [48].

2.9.1.2 Convolution Layers

A typical CNN architecture consists of convolutional and pooling (or subsampling or downsampling) layers. convolution layer is the first layer in CNN that performs convolution operation which explained in previous paragraph. Coming after the Convolution layer the non-linearity function named Rectified Linear Unit (ReLU). There are other non-linear functions such as Hyperbolic Tangent or Sigmoid that can also be used instead of ReLU, however, ReLU has been found to better perform. ReLU is a special implementation that combines non-linearity and rectification in layers of CNN [24,48]. ReLU Formula defined as follows:

$$F(x)=Max(0,x) \quad (2.7)$$

Where;x is the input (feature maps).

Feature maps resulting from the convolution stage will pass through a ReLU function which it's work replaces all the negative values in the feature map with zero where keeps the others as the original to remove the features' non-linear property as shown in figure 2.16.

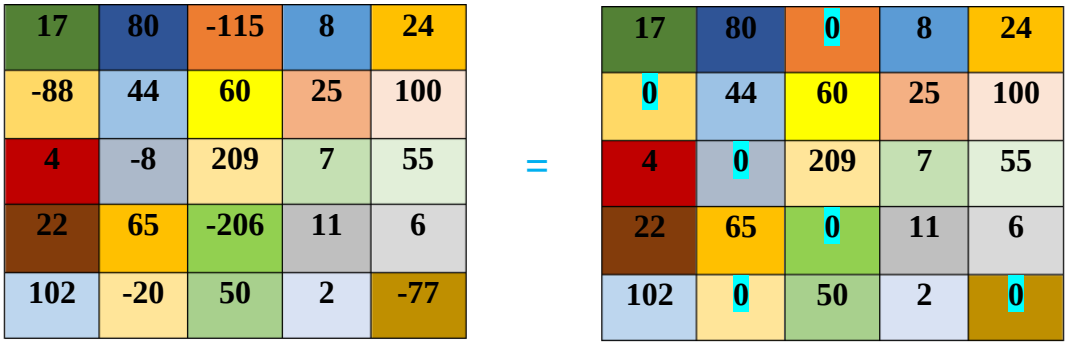


Figure 2.16: An Example on ReLU Function Operation.

The result from RELU function passes through the pooling layer which is responsible for reducing the size of the activation feature maps. Although it reduces the dimensionality of each feature map, it retains the most important information. There are different strategies of the pooling which are max-pooling, average-pooling, and mean- pooling. Max-pooling takes the maximum of the input data. Average-pooling takes the average value of the input data. mean pooling takes a minimum value of the input data as shown in figure 2.17. Pooling makes the input representations or feature dimensions smaller and more manageable [24,48].

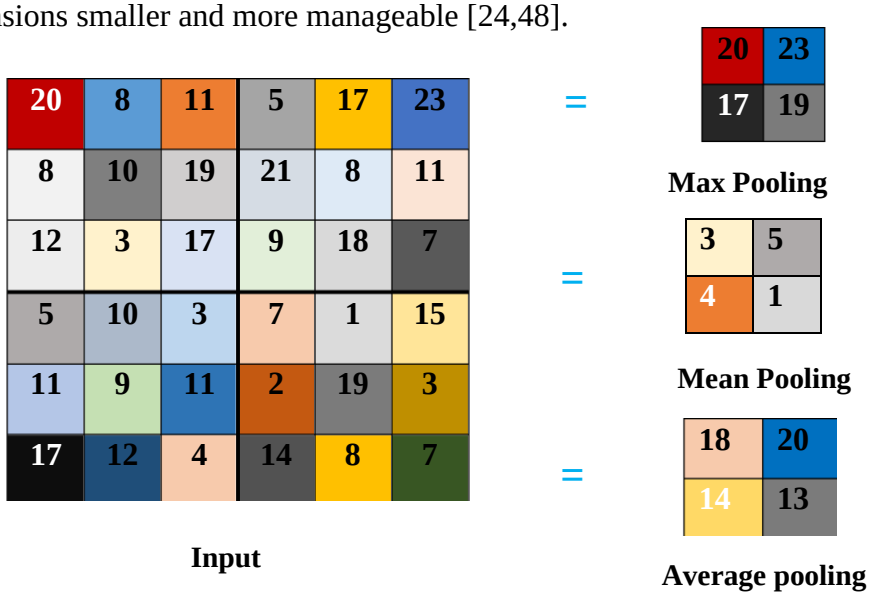


Figure 2.17: Pooling Operation.

After many convolutions, nonlinearity by ReLU and pooling layers, are arranged consecutively to create a network in the process of feature learning. Some other CNNs could be included such as the standardization, Batchnormalization and dropout layer,

and others. The extracted features learned from the feature learning process will then be organized in a vector form and it's considered as input to Fully Connected layers (FC) to perform classification process same as a traditional neural network but the difference is the layer before the output contains softmax function whichtakes a vector of arbitrary real valued scores and squashes it to a vector of values between zero and one. Note there are can be no convolutional layers after an FC [24].

Softmax Function formula defined as follows [24].

$$SoftMax(zj) = e^{zj} / \sum_{k=1}^{K} e^{zk} \quad ,j=1,2,\ldots K \qquad (2.8)$$

In other words, Softmax Function formula defined as follows:

Softmax(x_i)=exp(x_i)/sum(exp(x_i)) , i=1,2,3….

Where; x_i is the input (feature maps).

2.10 Learning Types for DL algorithms

1 - Unsupervised learning, which is intended to capture the high-order correlation of the observed or visible data for pattern analysis or synthesis purposes when no information about target class labels is available. Unsupervised feature or representation learning in the literature refers to this category of deep networks [50].

2 - Supervised learning, which is intended to provide direct discriminative power for pattern classification purposes, often by characterizing subsequent class distributions that are conditioned on the observable data. The target label data for such supervised learning are also available in direct or indirect forms. They are also called deep-seated discriminative networks [50].

3 - Hybrid Learning, where the target is discrimination which is assisted, often in a significant way, with the outcomes of unsupervised networks. This can be accomplished by better optimization or/and regularization of the- networks in category (2). The goal can also be reached when discriminative criteria for supervised learning are used to estimate the parameters in each of the category (1) above unsupervised networks [50].

Chapter Three

The Proposed System

Chapter Three

The proposed system

3.1 Introduction

In this chapter, the proposed system will be explained in deep with each step details, and how CNN was implemented to recognize the human identity from their voices. As it was explained in the previous chapters, the CNN was used on a raw waveform. In this way, CNN will extract features from the audio stream and start the learning at the same time. In which, four Hundred (400) human voices datasets were used. Each class contains 10 voices; 8 voices were for training and the other 2 were for testing. The proposed system contains a process for recording voices from the PC microphone and use it in both training and testing.

In the proposed system, the noise removal function used to remove noise from an audio signal, the noise_adds function used to add random noise to an audio signal, the preprocessing function, which removes the silence from audio files and converted them into a format that can be used in the training process. Then all these formed data will be passed to the training process which CNN as image takes these data and extract its features through several layers. Then, it was trained these features for future recognition.

The system shows high accuracy and minimum mean square error for both methods. The audio files and the recording by microphone files were used to test the system and the system in most cases were positive.Details of each step of the proposed system will be illustrated in this chapter. The work of CNN and the used parameters for best functionality and performance is illustrated in the next sections as well.

In this thesis, CNN was implemented without the need to MFCC as feature extraction. It simply uses raw audio data and convolutionally filtering the audio to extract features from voices. Also, implemented the system with the need to MFCC to extract features from voices. Also, implemented the traditional system by using MFCC and ANN and

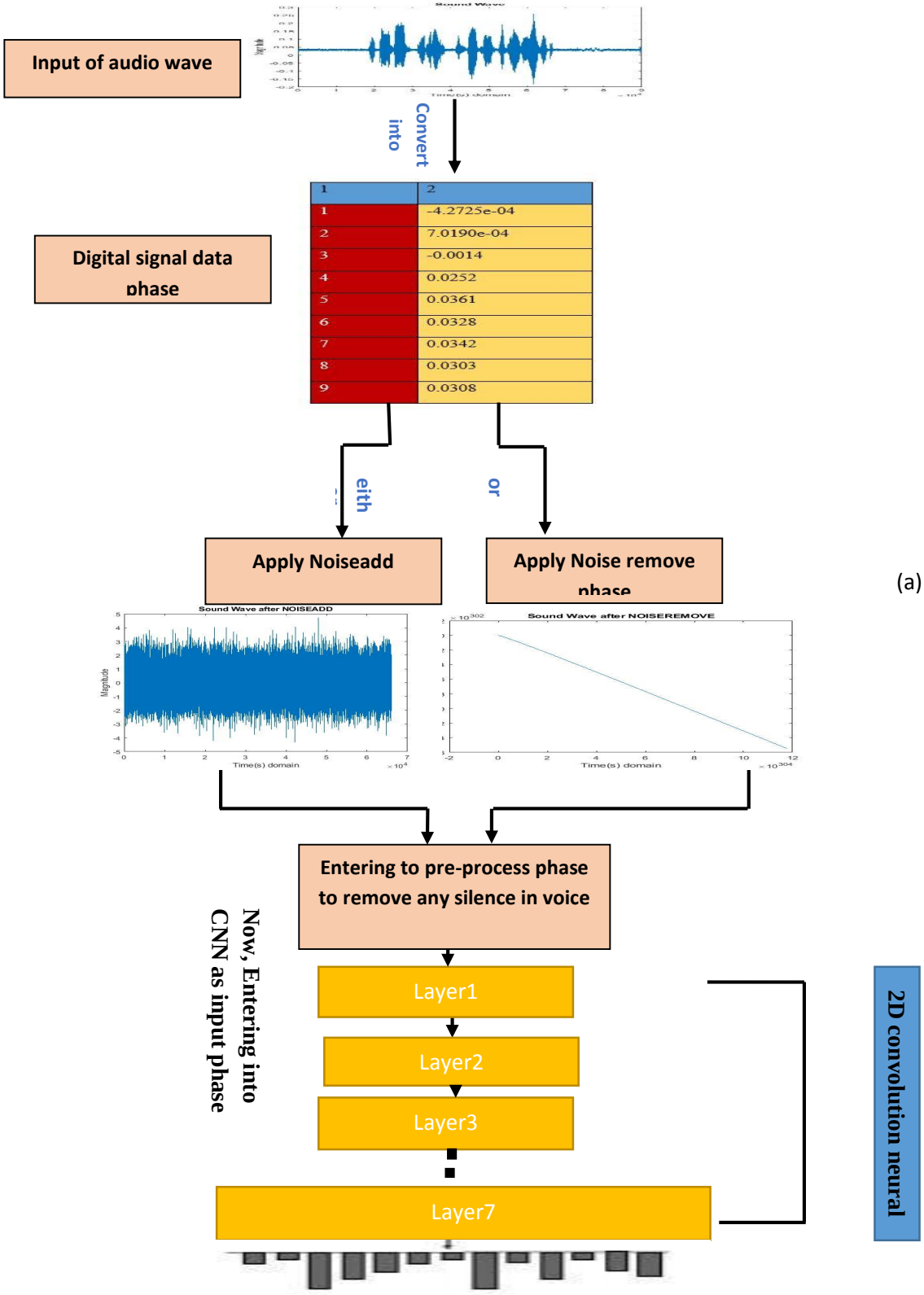
their results were compared. Algorithm 3.1 shows the used steps or algorithms to implement traditional VPRS.

Algorithm (3.1): describes implementation of traditional VPRS
Input: Audio Dataset //wavesignals
Output: network, contains trained speakers

Step1: read wave signals // Read audio files
Step2: Filter and amplification the wave signal //perform pre-processing stage
To remove any noisy & silence
That exist in wave_signals
//The input to MFCC algorithm is_ a vector of pre-processed wave_signals
Step3: N=256 // Determine the frame size
Step4:M=128 // Determine the overlapping
Step5: Segment the wave signals into some frames // The process of segmentin the audio signal
Step6: Apply hamming window for each frame by using equation (2.2)
Step7: Execute Fast Fourier Transform for each window // To transform signal from time domain to frequency domain
Step8: Applying Mel-filter bank by using equation (2.4)
Step9: Calculate Cepstral coefficients (DCT) by using equation (2.5)
//The output of these steps is MFCC features are input to a feed-forward neural network
Step10: Determine number of hidden layers
Step11: Determine train ratio, test ratio // Where trainRatio=0.7 and testRatio=0.3
Step12: Determine the number of epochs // Number of training epochs
Step13: Determine the goal // goal = 1e-10

Step14: Determine the learning rate value // lr = 0.1 , lr=0.5
Step15: Execute the training process
End.

In the traditional method, VPRS relies on MFCC to extract the characteristics from the signal for the extraction of the feature process. The speaker audio is recorded using a microphone, which transforms the speech of humans into an electrical signal. Then, with several frames, it transforms into a digital signal. Each frame contains 256 samples which overlap the frame size by 50 per-cent. Then, the hamming window is added to each frame to decrease the discontinuities that occur at the end and at the frame start. In addition, feed-forward neural network for classification purposes is implemented with one input layer, one output layer, and one hidden layer.


Input of audio wave
Convert into
Digital signal data phase
1 2
1 -4.2725e-04
2 7.0190e-04
3 -0.0014
4 0.0252
5 0.0361
6 0.0328
7 0.0342
8 0.0303
9 0.0308
eith
or
Apply Noiseadd
Apply Noise remove phase
Sound Wave after NOISEADD
Sound Wave after NOISEREMOVE
Time(s) domain
Entering to pre-process phase to remove any silence in voice
Now, Entering into CNN as input phase
Layer1
Layer2
Layer3
Layer7
2D convolution neural

(a)

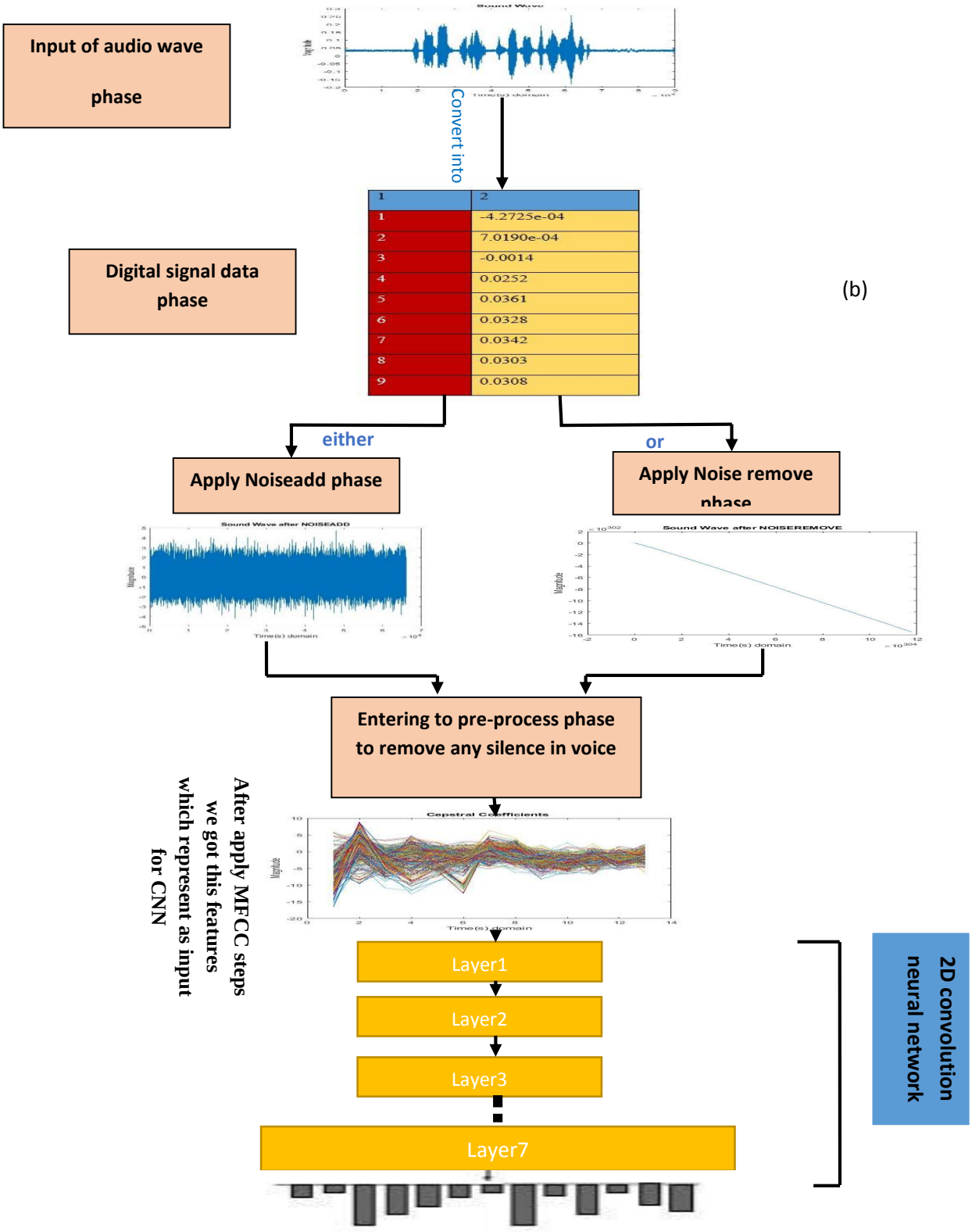


Figure 3.1 :The proposed system where (a) RW-CNN steps (b) MFCC-CNN steps

Figure 3.1 shows the proposed sytem for two methods MFCC-CNN and RW-CNN. In figure (a) The audio signal is converted into digital data is a first phase for the system called reading audio files phase,and the second phase apply noise remove or noise add function where noise remove to remove any noise that exist in audio signal and noise add to add random noise on audio signal To distort the audio signal,the third phase applying preprocessing phase to remove any silence that exist in the audio signal, and finally the resulted data taken from the previous phases was collected and passed through all CNN layers . In figure (b) the same stages in figure (a) but after the third phase which is applying preprocessing phase ,the fourth phase applying MFCC steps to extract features and using it as input for network,where the same structure of the network used for both methods.

3.2 Proposed CNN Structure

CNN based training is utilized from scratch in this proposal. Therefore, the appropriated CNN model for the studied dataset is generated using an empirical experiment. The structure of a CNN model proposed for Dataset shows in Table I. The model trains on 100000 epochs with a learning rate of 1e-1. The stochastic gradient descent with momentum (SGDM) optimization algorithm is employed. The momentum term is set to 0.95.

The architectures of the proposed CNN demonstrate in Table3.1, consists of an input layer with size 28*28, one convolution layer which contains 2 filters with size 3*3 and uses relu function in this layer, one pooling layer which contains 2*2 max-pooling, After convolution and pooling, the multi-dimension "outputs" usually are converted to a vector to be used as the inputs of the fully connected non-linear layers. And stack2line layer is to indicate this converting. The last layer is a fully connected layer which contains nonlinear type as softmax and the number of classifications is 10.

Table 3.1: Structure of The Proposed CNN Model

Layer No.	Layer Name	Details
1	Input	28×28×1
2	Convolution	3×3, 2 filters, pad 1
3	Relu	F(x)=Max(0,x)
4	Maxpooling	2×2
5	Stack2line	
6	Relu	F(x)=Max(0,x)
7	Relu	F(x)=Max(0,x)
8	FullyConnectedLayer	10
9	SoftMax	F(x)=exp(x)/sum(exp(x))
10	ClassificationLayer	10

3.3 The Proposed System Stages

This section, explains the stages used in the proposed system as follows.

3.3.1 Read Audio Files

The first step in such a system is to read the audio files and extract their samples in a form 1-D and saved as a 2-Dimensional array that can be used. Reading audio operation is implemented to read the samples in one vector as shown in figure 3.2. Reading the audio file is in the form of digital data as shown in table 3.2. The type of audio files is ".wav", which is called Wave Audio Files. Algorithm 3.2 shows how wave files can be read.

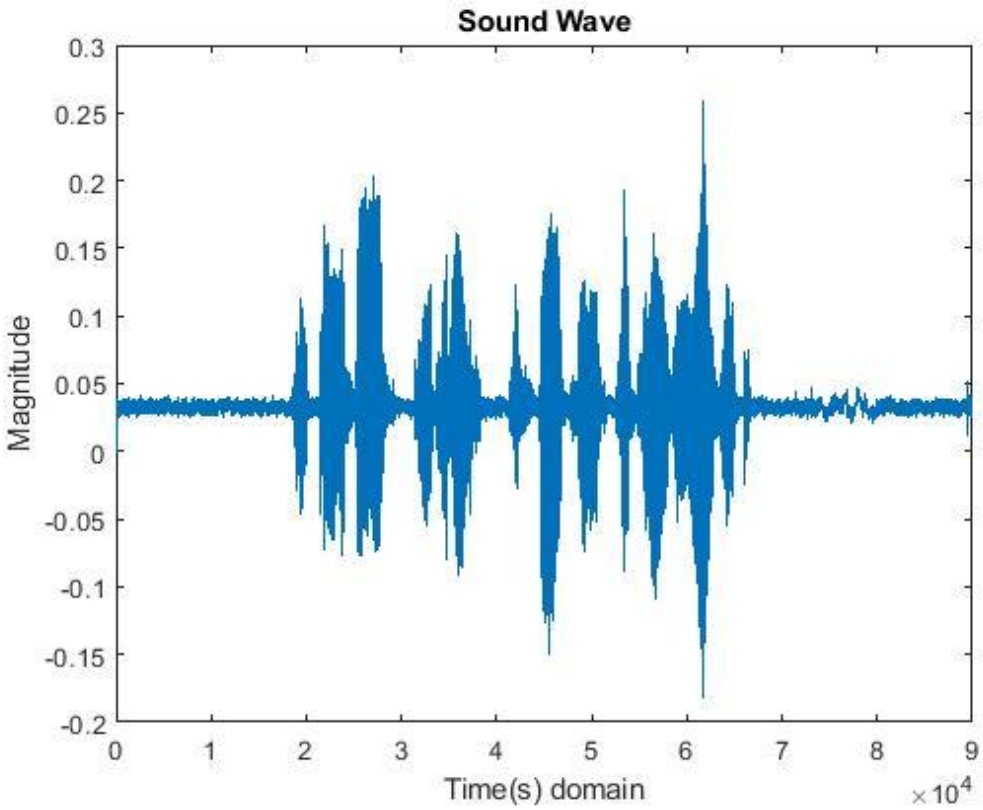


Figure 3.2: Audio Signal Form.

Table 3.2: An example of reading an audio file.

1	2
1	-4.2725e-04
2	7.0190e-04
3	-0.0014
4	0.0252
5	0.0361
6	0.0328
7	0.0342
8	0.0303
9	0.0308
10	0.0291
11	0.0297
12	0.0278
13	0.0285

Algorithm (3.2): describes Read Audio Files phase
Input: filename //Path of filename, FS //Sampling Rate
Output: Samples //Vector of audio data
Step1: Create a Channel without reading audio tags Step2: Create audio options with the following parameters (FS, Channel) Step3: Open file with these options Step4: Samples = FileRead() Step5: End

After reading the audio files, the system is entered either into the noise removal stage using the noise removal () function, remove any noise present in the sound. Or into noise add where; randomly add noise to the sound by using the noise_add () function.

3.3.2 Noise Remove

The removing noise process is to remove any unwanted samples in the audio files. The audio signal after removing the noise illustrated in figure 3.3.

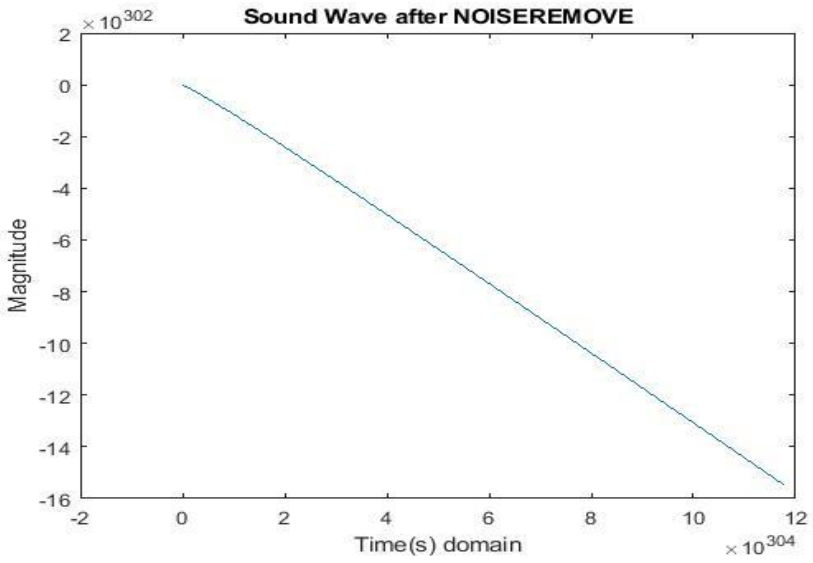


Figure 3.3: Audio signal After Applying Removing Noise

3.3.3 NoiseAdd

The adding noise process is to add random noise on the audio signal. Figure 3.4 shows the adding noise process.

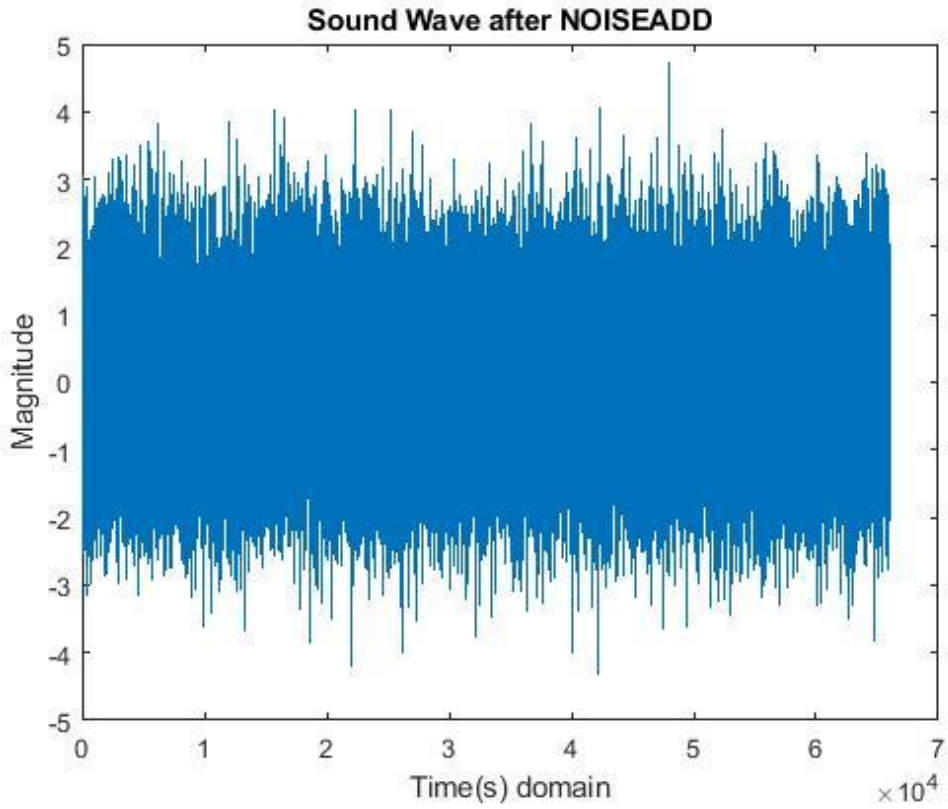


Figure 3.4: Audio signal After Applying Noiseadd

3.3.4 Preprocessing

The data samples extracted from the audio file were simply taken and then applying several processes on it. In the first process from this phase, it must make sure that the samples extracted from the file are 1-D vector Then, the silence is trimmed and removed. The silence in the Audio should be detected and then removed from the audio. Next, the amplification is applied to this signal by calculating the mean of the signal and multiplying it by 2. In the final process, make dimention size for samples uniform. Algorithm 3.3 shows the Pre-processing phase.

Algorithm (3.3): Pre-processing phase
Input: Samples, FS, Interval // Recording time interval in seconds where intval = 3
Output: New_Samples //A set of 1-D removed audio data from areas of silence.
Procedure: Step1: Convert Samples to 1D vector Step2: Define i = 1 to take each sample from samples of an array Step3: While Absolute value of Samples[i] < 0.002 And i < 48000 //The silence is trimmed and removed Step4: i = i + 1 Step5: Next Step6: NewSamples = Samples – the mean value of all Samples Step7: NewSamples = NewSamples * 2 // Step6 and step7 To make an amplification of the audio signal Step8: NewSamples = uniform(Samples) //Make dimention size for samples uniform. Step9: End

Then the audio files are stored in 2_D array.

3.3.5 Post Normalization

Normalizing data is very important to maintain its data balance and removing redundancy data that has been occurred.

3.4 CNN Algorithm

In this section, the resulted data taken from the previous phases was collected and passed through all CNN layers for both the feature extraction and the training. Implementing the CNN in the proposed system was done by initializing it is parameters with seven layers. The following algorithm 3.4 shows the initialization of the CNN.

Algorithm(3.4): describes Initialization of the CNN
Input: (Dataset)
Output: cnnConfig // A structure contains all information about CNN

Procedure:
Step1: initialize cnnConfig structure // cnnConfig is an object to create the network
Step2: Set the first layer as an input layer for CNN
cnnConfig.layer[1].type = 'input' // Determining the first layer as input layer
Step3: Set the dimension of the input layer // Determining the dimension of the input layer 28*28
cnnConfig.layer[1].dimension = [28 28 1]
Step4: Set the second layer to convolution filter // Determining the second layer as convolution filter
cnnConfig.layer[2].type = 'conv'
Step5: Set the dimension of convolution filter
cnnConfig.layer[2].filterDim = [3 3] // Determining the dimension of convolution filter 3*3
Step6: determine the number of filters to the second layer
cnnConfig.layer[2].numFilters = 2 // Determining the number of filters to the second layer is 2
Step7: The second layer will be set is relu
cnnConfig.layer[2].nonLinearType = 'relu' // Determining the second layer function as relu

Step8: The convolution matrix will be one row array of 20 elements

cnnConfig.layer[2].conMatrix = ones(1,20) // Determining the dimension of convolution matrix

Step9: Set the third layer to pool

cnnConfig.layer[3].type = 'pool' // Determining the third layer as pool

Step10: Set the dimension of pool to 2x2 // Determining the dimension of pool 2*2

cnnConfig.layer[3].poolDim = [2 2]

Step11: And the type of this pool is maxpool // Set the type of pool is maxpool

cnnConfig.layer[3].poolType = 'maxpool'

Step12: Set the fourth layer to stack to line

cnnConfig.layer[4].type = 'stack2line' // Determining the fourth layer as stack2line

Step13: Set the fifth layer to Relu (0,max(Convolved Features from previous laysers)))

cnnConfig.layer[5].type = 'relu' //Set the fifth layer to Relu

Step14: Set the dimension of the fifth layer to 360

cnnConfig.layer[5].dimension = 360

Step15: Also set the sixth layer to relu with dimension 60

cnnConfig.layer[6].type = 'relu' ////Set the sixth layer to Relu

Step16: cnnConfig.layer[6].dimension = 60 // Set the dimension of relu equal to 60

Step17: Set the last layer to softmax (exp(Weights * input)) with dimension 10

cnnConfig.layer[7].type = 'softmax' // Determining the seventh layer using softmax as function

Step18: cnnConfig.layer[7].dimension = 10 // Determining the dimension of last layer is 10

Initializing parameters of the CNN was the first step in the implementation of the proposed system. Next, the data samples was loade from the dataset. As it mentianed earlier in this chapter, the dataset contains 400 human voices each one contains 10

audio files. Using the following algorithm 3.5 , all these dataset were uploaded in one massive 2-D matrix.

Algorithm(3.5): Prepaird Data phase
Input: (dataset)
Output: cdb, labels

Procedure:
Step1: set number of training samples. We have 10 audios for each class, we will have 8 for training and 2 for testing
Define n = 8 // number of training samples for each person
Step2: Define FS = 16000 // Sampling rate
Step3: Define intval = 3 // Recording time interval in seconds
Step4: Define len=fs*intval // Total length of voice sample
Step5: Define stime=0.015 // Time duration in which voice is stationary
Step6: Calculate frame size [fsize,osize,nwin]=calsize(stime,intval,len) by the following formulas:
fsize = len * stime / intval // Calculate frame size
osize = fsize / 2
nwin = len – 1 / osize
Step7: Set no. of filters or channels in mel filter bank
Define noc=40 // number of mel filter bank in mfcc
Step8: Define dirs = get list of directories in dataset (the dataset consists of list of folders representing classes each one has 10 audios)
Step9: Set number of classes to process
Define p = 400 // determine number of speakers in dataset
Step10: for i = 1 to p // loop on number of speakers
Step11: Define d = dirs[i] //Take one of the directories according to i as index

Step12: Define files = get list of files in d

Step13: for j = 1 to n

Step14: voicein=audioread(namefile) // Process of read audio file

Step15:Applying noiseremove or noiseadd function on original signal

voiceout = noiseremove(voicein) //Apply noiseremove function

voiceout = noiseadd(voicein) //Apply noiseadd function

Step16: Apply preprocessing function for converting the signal to one vector and remove the silence

voiceout = preprocess(voiceout,fs,intval) //Apply preprocess function

Step17: append voiceout to cdb //To save audio data

Step18: append filename to labels

Step19: next

Step20: end

It can be seen that looping each folder to get audio files (8 files only for training) was used; and preprocessed. Then it was appended to one massive matrix that contains all the data samples which are called cdb. After loading the data, CNN was used to train these data in one long massive process. For training the network, the following algorithm 3.6 shown training process.

Algorithm(3.6): Training the network phase
Input: (dataset)
Output: A group of trained speakers

Procedure:
Step1: Initialize CNN configuration // To prepare (call) network
Net.cnnConfig = CNNInitialization()
Step2: Prepare data
cdb, labels = PrepareData() //To prepare data to input to network
Step3: Set options
options.epochs = 100000; //Number of training epochs
Step4: options.minibatch = 128 //Determine minibatch
Step5: options.alpha = 1e-1 //Determine learning rate
Step6: options.momentum = .95 //Determine momentum
Start training
Step8: End

Two methods are used in this proposal which are MFCC-CNN and RW-CNN. Direct entry to CNN. Where the same CNN structure was used in both methods.The first method is called the standard method and includes the following: reading the audio files and then these audio files pass in a phase of removing or adding random noise to the audio files and then the system goes through the pre-processing stage to remove any silence that exists in the sound then passes the audio files in the MFCC which It has already been explained in detail, to use the features in the audio and before entering the voice for recognition stage using the CNN the features are stored in the 2D matrix because the proposed CNN is 2D so these features are converted and stored in a two-dimensional matrix and then these extracted features will be entered and converted into 2D to CNN to perform a process The training. CNN will take input as a picture and then begin the process of training via the proposed

CNN. In the second method, RW-CNN, the same steps as the first method, but without going through the MFCC phases where the audio files are read and then these audio files pass in a phase of removing or adding random noise to the audio files, and then the system goes through the pre-processing stage to remove any silence in the sound After that, the audio files are stored in the 2D matrix. performance of the network appear in the following chapter.

Chapter Four

Results and Discussion

Chapter Four

Results and Discussion

4.1 Introduction

In this chapter, will discuss the results are get in our hands from the implementation of our proposed system. As we have spoken before, we implement two approaches for VPR. The first one is by applying MFCC as feature extraction for the audio data and then apply CNN for training and recognition. The second one is by applying CNN directly to the raw data. Our CNN is designed as a feature extraction and training method. Adding some noise before applying our proposed method to conclude more results that show if our method is more accurate for different circumstances or not. Now, will explain in detail each case and comparing them to each other.

The used dataset contains VoxForge Speech Audio files for different ages and people of all groups speak the English language in various words. The speaker's Characteristics are: (Gender: Male&female, Age Range: Adult, Language: EN, Pronunciation dialect: New Zealand English).The recording information is: (Microphone make: n/a; Microphone type: Headset mic, Audio card make: unknown, Audio card type: unknown, Audio Recording Software: VoxForge Speech Submission Application O/S. The file information is: (File type: wav, Sampling Rate: 48000, Sample rate format: 16, Number of channels: 1). The dataset size used for this thesis is about 400 classes, each class has 10 voice audio files. The train set taken from each class is eight voices and the rest are for the test set [51].

4.2 Proposed System Implementation

4.2.1 Traditional System

Implemented a conventional framework for VPRS where MFCC was used to extract features and feed-forward neural networks for classification purposes with one input layer, one output layer, and one hidden layer. The used datset comprised of audio recordings of 43 persons (25 male and 18 female) read short sentences and each person

talks 10 different sentences. The weights and bias values were chosen at random. The parameter used in backpropagation for the learning rate is 0.5 and 0.1. The neural network used is a perceptron multilayer with a single hidden layer. The number of output neurons is 43; since 43 people were the number of speakers included in the experimental tests. For the training process, 10 separate samples of speech signals were used for each person. For the training phase, the data set was divided to 70 per-cent and for the test phase to 30 per-cent.

Table 4.1: Results of mean square error across learning rate equal to 0.1.

No. of Cofficients	No. of hidden neuron	No. of epochs	Mean sequare error
13	**30**	**50**	**3.47e-05**
13	**60**	**50**	**3.10e-05**
13	**80**	**50**	**2.49e-06**
13	**100**	**200**	**2.39e-07**
13	**25**	**200**	**1.95e-06**
13	**80**	**200**	**7.49e-07**
13	**30**	**200**	**4.94e-06**

Table 4.2: Results of mean square error -across learning rate equal to 0.5.

No. of Cofficients	No. of hidden neuron	No. of epochs	Mean sequare error
13	**30**	**448**	**3.39e-07**
13	**60**	**448**	**3.65e-08**
13	**80**	**448**	**3.03e-08**
13	**25**	**448**	**6.33e-07**
13	**100**	**164**	**3.33e-07**

13	**80**	**164**	**5.71e-07**
13	**30**	**164**	**2.12e-06**

Tables 4.1 and 4.2 show the mean square error results of test signals generated from constant numbers of DCT coefficients and different numbers of hidden neurons and the learning rate of 0.1and 0.5. The global error does not exceed the minimum target value The best result of a mean square error in table 4.1 is when the epoch number is 200, and the number of hidden neurons is 25. In table 4.2, when the number of an epoch is 164, and the number of hidden neurons is 30.

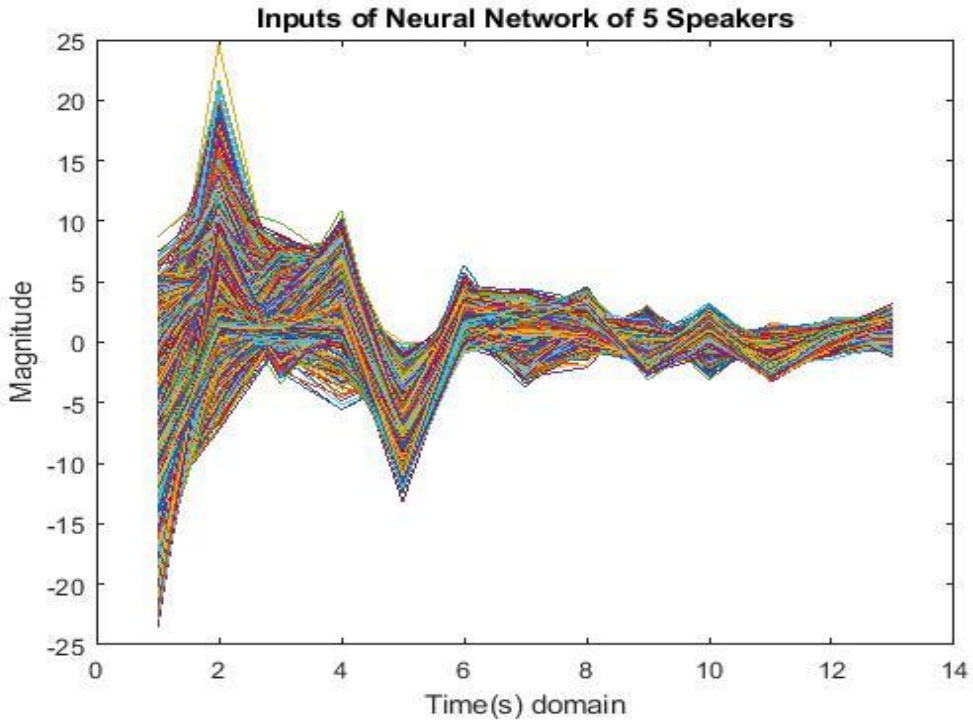


Figure 4.1: Input of neural network of five speakers

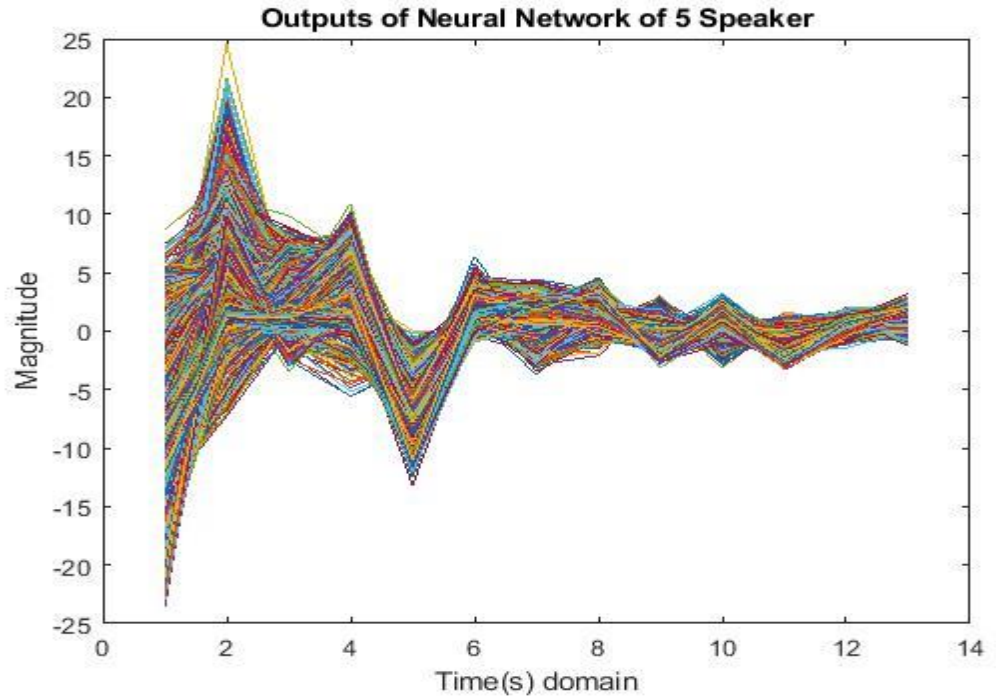


Figure 4.2: Output of neural network of five speaker

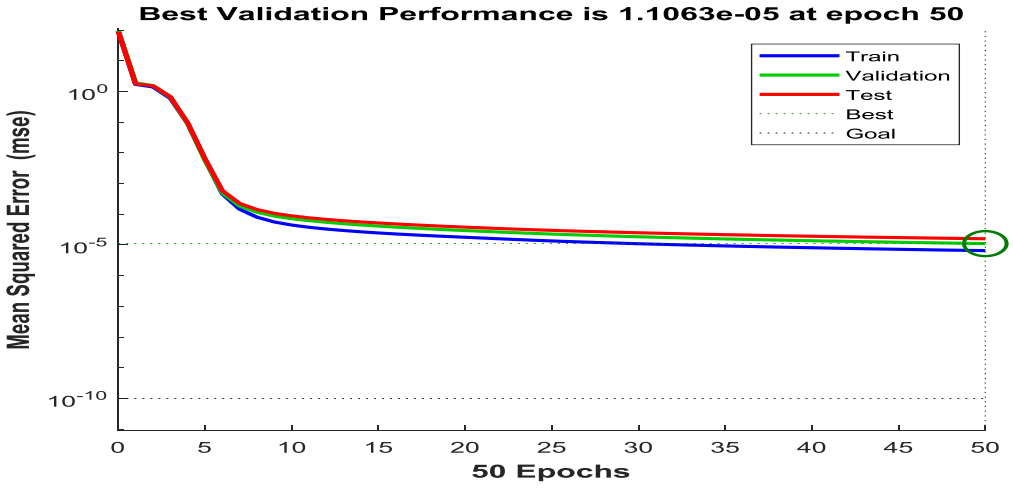


Figure 4.3: Mean square error of five speakers

Five speakers were chosen in figures 4.1 and 4.2 to check the neural network at 50 epochs, the number of hidden neurons is 30 and the number of inputs to the neural network is constant. Eventually, the mean square error as shown in figure 4.3 exceeded 1.1063e-05, where Figure 4.2 almost corresponds to Figure 4.1.

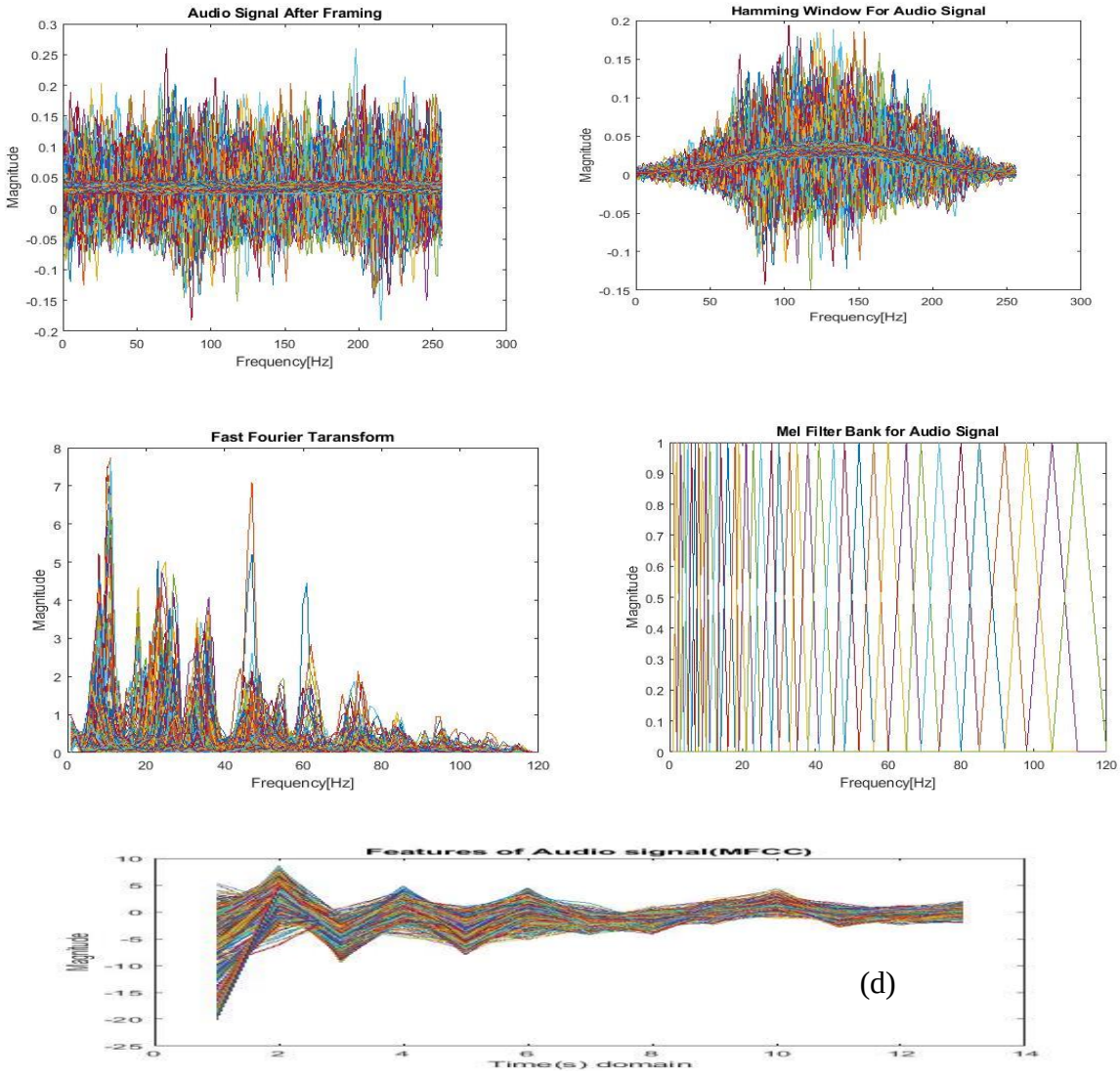
4.2.2 Modern Proposed System

The system was designed in GUI (Graphical User Interface).One main form that has three buttons to launch the options of the system in which, the first button is to add another dataset; when pressing on it button will appear a menu which contains on two options,where the first option is add a dataset and delete the old dataset ,when pressing on it removing old added dataset and add new dataset and appear when press on it, a menu will appear asking to enter the number of the speakers .

After entering the number of speakers such as one and click on OK will appear a menu contains two buttons,The first button is choosing any saved Audio files from PC and the second button is recording audio by the microphone of PC, in other words, can adding dataset from your voice or any person's audio,when choosing recording audio button will appear a menu asking to enter the name of the voice you record by mic of pc and click OK, finally the recorded voice is saved as wave format.The proposed system was implemented using a MATLAB 2018b, with PC of the properties; CPU: Corei7 8generation,8 GB of Ram and Hard SSD.

4.3 Applying MFCC

By applying the Mel-Frequency approach, a feature has been extracted from an audio signal after applying add noise or noise remove and pre-processing phase for CNN as input. The features output of the MFCC stages is illustrated in figure 4.4.



(e)

Figure 4.4:MFCC Results For Audio Signal ,where (a) Audio signal after framing,(b) Hamming window for audio signal,(c) Fast fourier transformation for audio signal,(d) Mel-filter bank for audio signal,(e) Cepstral Coefficients for audio signal

4.4 The Accuracy

The most important part of each ML or DL system is how this system is accurate in his process. As it is known, no system that uses massive data is 100% accurate, since no algorithm is perfect so far, but increased the accuracy as high as possible. Therefore; the first step was to take a look at the accuracy calculation process to understand how the accuracy has been calculated. The meaning of accuracy is the difference between the test set and the predict set for the same samples and it can be calculated as shown in Equation (4.1).

$$Accuracy = \frac{TN}{TP+TN+FP+FN} \quad (4.1)$$

Where:

- True Positive (TP) is the correct identification of a speaker.
- False Positive (FP) is the incorrect identification of a speaker.
- True Negative (TN) is the correct identification of the incorrect speaker.
- False Negative (FN) is the incorrect identification of the incorrect speaker.

For example, if the number of the correct predictions is 90 and the total number of the predictions is 100, then the accuracy will be 90%. It has been considered that the accuracy can be changed from one system to another according to its internal design and circumstances. If some voices have noise in its background, then it should be removed. In this case, some audio data will be corrupted and the accuracy is affected. Moreover, the effect of using the preprocessing of mel-frequency and mel-filter should be considered as it can affect the system accuracy. Using these processes, the proposed system accuracy was evaluated when the audio without any noise and the mel-frequency and mel-filter were considered. The obtained results were explained in table 4.3.

Table 4.3:The accuracy of proposed CNN

Technique	Dataset Size			Number of Traing Data			Accuracy		
MFCC-CNN Without Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**0.96**	**0.96**	**0.96**
MFCC-CNN With Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**0.96**	**0.96**	**0.96**
RW-CNN Without Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**0.96**	**0.96**	**0.96**
RW-CNN With Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**0.96**	**0.96**	**0.96**

The accuracy was shown in table 4.3 for each category of data for both MFCC-CNN and RW-CNN with and without noise. The results showed that the accuracy for both methods with different data size are the same. This because that in both ways the input process for the CNN network is the same. In the MFCC-CNN, the features were used and entered into the CNN network with some extracted features and train the network based on the extracted features. In the second method, the sound is directly entered into the network to extract the features directly. regardless of the increase or decrease in the volume of data. In table 4.3, the first column is the used technique (methods) MFCC-CNN and RW-CNN with and without noise. The second column is the used data size for the proposed system (68,88 and 400). The third column is the number of training data which depends on the second column. In the second column, the used data size 68

is related to the 544 number of training data in the third column. In the first stage, MFCC-CNN was implemented without noise on 68 audio files, means 544 voiceprints, and the result of accuracy was 0.96, as well as it was applied to 88 audio files, 704 voiceprints, and the result of accuracy was equal to 0.96 also applied to 400 audio files, means 3200 voiceprints and accuracy result was 0.96 Also MFCC-CNN was used with noise and on 68 audio files means 544 voiceprints and accuracy result was 0.96 Also it was applied to 88 audio files means 704 voiceprints and accuracy result was 0.96 also It was applied to 400 audio files means 3200 voiceprints and the result of accuracy was 0.96. The second method RW-CNN was applied with without noise on 68 audio files means 544 voiceprints. The accuracy is equal to 0.96, as well as it was applied to 88 audio files, means 704 voiceprints. An audio file, means, on 544 voiceprints, the result of_accuracy was equal to 0.96, as well as it was applied to 88 audio files, means 704 voiceprints, and the result of accuracy was equal to 0.96. Also applied to 400 audio files, means 3200 voiceprints, and the result of accuracy was equal to 0.96.

4.5 Mean Square Error (MSE)

There are another evaluation criterion called MSE (Mean Square Error), which calculates the error that occurred in each epoch and then find its mean. It measures the common of the squares of the errors, that is, the average squared difference among the estimated values and the real value. MSE is a threat function, similar to the predicted fee of the squared mistakes loss. The truth that MSE is almost continually strictly positive (and not zero) because of the randomness or because the estimator does not account for statistics that would produce a more accurate estimation. The MSE can be calculated as in equation 4.2.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (4.2)$$

Where y ia actual value and $\hat{y}$ is predicted value. The MSE was calculated for the same above four cases. In our system, the MSE has reached its minimum by using noise. Table 4.4 shows the MSE of each subset of data in the four circumstances.

Table 4.4: The MSE of Proposed CNN

Technique	Dataset size			Number of TrainingData			MSE		
MFCC-CNN Without Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**3.2000e-08**	**3.2000e-08**	**3.2000e-08**
MFCC-CNN With Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**3.2000e-08**	**3.2000e-08**	**3.2000e-08**
RW-CNN Without Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**3.2000e-08**	**3.2000e-08**	**3.2000e-08**
RW-CNN With Noise	**68**	**88**	**400**	**544**	**704**	**3200**	**3.2000e-08**	**3.2000e-08**	**3.2000e-08**

In Table 4.4, MSE was shows for each category of data for both MFCC-CNN and RW-CNN with or without noise. The results showed that MSE for both methods with different data sizes are the same. It was because that in both ways the input process for the CNN network is the same. In the table, the second column illustrate the used data size for the proposed system (68,88 and 400), while the third column represent the number of training data depending on the second column. The MSE was 3.2000e-08 for all circumstances.

4.6 Performance

One of the most important characteristics in all systems is how much time the system needs to finish it is process, called 'Performance'. The time consuming for training for each case is illustrated in table 4.5. In which, the system performance have been evaluated with and without using noise for both proposed methods with equal number of epochs (100000). It is clear from table 4.5 that the consuming time in training of the RW-CNN method with nand without noise is lower than the consuming time in training the MFCC-CNN with and without noise.

Table 4.5: The Performance of Proposed CNN

Technique	Dataset size			Number of Training Data			Time		
MFCC-CNN Without Noise	68	88	400	544	704	3200	4.0369e+04	8.5479e+04	7.2589e+04
MFCC-CNN With Noise	68	88	400	544	704	3200	2.6688e+04	7.8509e+04	8.7330e+04
RW-CNN Without Noise	68	88	400	544	704	3200	3.5783e+04	1.4750e+04	1.3717e+04
RW-CNN With Noise	68	88	400	544	704	3200	2.6960e+04	1.2887e+04	1.1814e+04

Table 4.6: Compare the Classification Accuracy from each Method

Method	Accuracy(%)		Accuracy of th proposed system
	Max	Average	
MFCC	91.30	91.26	
CNN of signal wave	54.00	49.77	0.96
CNN of Spectrogram	99.00	95.83	0.96

Table 4.6 shows the accuracy when the CNN of Spectrogram method is compared with the MFCC method and CNN of the signal wave method. The experiments conducted on 50 times to the testing set for each method. Maximum and average of the classification accuracy were illustrated for comparison purpose. It reveals the proposed CNN based method trains on the spectrogram image of voice is the best compared to the other two methods. The average classification results of the testing set by the proposed method is 95.83% accuracy. MFCC based method is 91.26% and CNN trained on the image of the raw signal wave is 49.77%. The proposed method is very efficient for a text-independent approach where the only short utterance of voice is needed as an input. Accuracy of the propoed system for both methods is 0.96 in clean and noisy environments; because in both ways the input process for the CNN network is the same.

Table 4.7:The frame identification performance FIA %

	Clean	Noise (SNR=5dB)	Reverb(RT_{60}=0.5 s)	Accuracy of the propsed system (clean and noise)
MFCC-CNN	**45.72**	**16.85**	**37.32**	**0.96**
RW-CNN	**64.01**	**46.24**	**38.51**	**0.96**

Table 4.7 shows the FIA which is the percentage of identification accuracy by considering each frame individually. The capability of the RW-CNN to correctly identify the speaker is widely greater for clean and noisy signals concerning MFCC features. In the reverberation case, the RW and MFCC have similar identification performance. It demonstrated that the RW-CNN is more robust to noise if compared with the MFCC-CNN trained with the same DA dataset, and they have similar performance in reverberant environments. Accuracy of the propoed system for both methods is 0.96 in clean and noisy environments; because in both ways the input process for the CNN network is the same.

Chapter Five

Conclusions and Future Works

Chapter Five

Conclusions and Future Works

5.1 Conclusions

In this thesis, a convolution neural network based Voiceprint Recognition System in Noisy Environment was presented. The CNN architecture is designed to operate for both MFCC-CNN and RW-CNN. The proposed CNN inputs are images in both cases, i.e. the network dealt with images, where the same architecture of CNN used for both methods. The obtained results show that both methods are similar in their accuracy 0.96 and mean square error 3.2000e-08 results but differents in performance where the time results show that RW-CNN is better than MFCC-CNN whether with or without noise. In other words RW-CNN is more efficient in clean and noisy environments from MFCC-CNN.

5.2 Future Works

1 - In our research, some random noise added to the background audio files, and suggestion of the next step use some real noise and try to eliminate them.

2 - Using a very huge dataset of human audio_(millions of Audio).

3 - Connection using between traditional and modern methods for the voiceprint recognition system.

4 - Use CNN directly on digital data of Audio if possible.

References

References

[1] V. Sharma and P. K. Bansal, "***A Review On Speaker Recognition Approaches And Challenges,***" International Journal of Engineering Research and Technology,vol. 2, no. 5, pp. 1581–1588, 2013.

[2] T.B. Mokgonyane,T.J. Sefara,T.I. Modipa,M.M. Mogale,M.J. Manamela and P.J. Manamela "***Automatic Speaker Recognition System based on Machine Learning Algorithms,***" IEEE, January 28-30, 2019.

[3] S.V and G.S.K, "***A Survey on Different Algorithms for Automatic Speaker Recognition Systems,***" International Journal of Engineering Research and General Science,vol. 4, no. 1, pp. 579–587, 2016.

[4] P. Das, K. Acharjee, P. Das, and V. Prasad, "***Voice Recognition System : Speech-To-Text,***" Journal of Applied and Fundamental Sciences ,vol.2,pp.191-195,November 2015.

[5] N. Singh, A. Agrawal and R. A. Khan, "***Principle and Applications of Speaker Recognition Security System,***" International Conference on Artificial Intelligence , June, 2018.

[6] Ch. Kalyani, "***Various Biometric Authentication Techniques : A Review,***" Journal of Biometrics and Biostatistics,vol.8, Issue .5, January, 2017.

[7] W.Yang , S.Wang, J.Hu, G.Zheng and C.Valli, "***Security and Accuracy of Fingerprint-Based Biometrics: A Review,***" Journal of Symmetry , January,2019.

[8] S.S.Harakannanavar,P.Ch.Renukamurthy and K.B. Raja, "***Comprehensive Study of Biometric Authentication Systems, Challenges and Future Trends,***" Int. J. Advanced Networking and Applications,Volume: 10 Issue: 04 Pages: 3958-3968, January, 2019.

[9] S.Shrivastava, "***Biometric: Types and its Applications,***" International Journal of Science and Research, pp. 204–207, April ,2015.

[10] A. Jain, E.Lansing ,R. Bolle and S. Pankanti, "***Introduction To Biometrics,***" *Handbook of biometrics,* no. ii, pp. 19, 2002.

[11] U. Bakshi, R. Singhal and M. Malhotra, "***Biometric Technology: A Look and Survey at Face Recogntion,***" International Journal of Engineering Science Invention, vol. 3, no. 6, pp. 24–30, June,2014.

[12] J. Choudhary, "***Survey of Different Biometrics Techniques,***" International Journal of Modern Engineering Research , Vol. 2, Issue. 5, pp-3150-3155,Sep.-Oct. 2012.

[13] J. Oh , U. Lee, and K. Lee, "***Usability Evaluation Model for Biometric System considering Privacy Concern Based on MCDM Model,***" Hindawi Security and Communication Networksv ,vol. 2019, pp.1-14, 2019.

[14] M. Z. Alom , T. M. Taha , C. Yakopcic , S. Westberg , P. Sidike ,M. S. Nasrin , M. Hasan , B. C. V. Essen , A. A. S. Awwal and V. K. Asari, "***A State-of-the-Art Survey on Deep Learning Theory and Architectures,***" Electronics ,pp. 1–67, 2019.

[15] T.Kinnunen, E. Karpov, and P. Fränti, "***Real-Time Speaker Identification and Verification,***" IEEE Transactions On Audio, Speech, And Language Processing, vol. 14,no. 1,pp.227-288, january, 2006.

[16] A.Revathi, R.Ganapathy and Y.Venkataramani, "T***ext Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach,***" International Journal of Computer science & Information Technology , vol 1, no 2,pp.30-41, November, 2009.

[17] L.Muda, M.Begam and I.Elamvazuthi, "***Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques,***" Journal Of Computing, vol. 2, Issue. 3, pp.138-143,March 2010.

[18] H.B.Kekre and V. Kulkarni, "S***peaker Identification using Frequency Dsitribution in the Transform Domain,***" International Journal of Advanced

Computer Science and Applications,Vol. 3, No.2,pp.73-78, 2012.

[19] U.A.Kamalu, A. Raji and V.I.Nnebedum, "***Identity Authentication Using Voice Biometrics Technique,***" International Journal of Research in Engineering and Technology, vol. 04, Issue: 12 , pp.130-136,December,2015.

[20] J. Lee,T. Kim,J.Park and J. Nam, "***Raw Waveform-based Audio Classification Using Sample-level CNN Architectures,***" 31st Conference on Neural Information Processing Systems arXiv:1712.00866v1 [cs.SD], pp.1-5,4 Dec. ,2017.

[21] M. Ravanelli and Y. Bengio, "***Speech and Speaker Recognition from Raw waveform with Sincnet,***" arXiv:1812.05920v2 [eess.AS] ,pp.1-5,15 Febreuary, 2019.

[22] T.Kim, J. Lee, and J. Nam, "***Comparison and Analysis of SampleCNN Architectures for Audio Classification,***" Journal of Latex Class Files, vol. 14, no. 8,pp.1-13, August, 2015.

[23] D.Salvati, C.Drioli and G.L.Foresti, "***End-to-End Speaker Identification in Noisy and Reverberant Environments Using Raw Waveform Convolutional Neural Networks,***" InterSpeech,pp.4335-4339, September 15–19, 2019.

[24] S.Bunrit, T. Inkian, N.Kerdprasop, and K. Kerdprasop, "T***ext-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network,***" International Journal of Machine Learning and Computing, vol. 9,no. 2,pp.143-148, April, 2019.

[25] S. S Kodaganur and S.V. Chakrasali, "***ImplementAtion Of Speaker Recognition System On FPGA Using Logicore IP,***"International Journal Of Advance Research In Science and Engineering,vol.no.5,speacial Issue no.1,pp.176-182,May,2016.

[26] K. R. Ghule and R. R. Deshmukh, "***Feature Extraction Techniques for Speech Recognition: A Review,***" International Journal of Scientific & Engineering

Research, vol. 6, Issue 5,pp.143-147, May-2015.

[27] A. K. Panda and A. K. Sahoo, "***STUDY OF SPEAKER RECOGNITION SYSTEMS,***" Bachelor's thesis,Department for Electronics & Communications,National Institute Of Technology, Rourkela,2007-2011.

[28] Q.Jin, "***Robust Speaker Recognition,***" PhD thesis,Department foPhilosophy in Language and Information Technologies, Carnegie Mellon University,2007.

[29] S.Furui, " ***An Overview of Speaker Recognition Technology,***" ESCA Workshop on Automatic Speaker Recognition,Identification and Verification,Martigny,pp.1-9,1994.

[30] D. S. Rodríguez, "***Text-Independent Speaker Identification,***" Master Science Thesis, Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, AGH University Of Science And Technology Krakow,2008.

[31] F. Zheng , G. Zhang and Z.Song, "***Comparison Of Different Implementations Of MFCC,***" J. Computer Science & Technology, 16(6): 582-589, Sept. 2001.

[32] K. N. Van, T. P. Minh, T. N. Son, M. H. Ly,T. T. Dang, and A. Dinh," ***Text-dependent Speaker Recognition System Based on Speaking Frequency Characteristics,***" Springer Nature Switzerland AG ,LNCS 11251, pp. 214–227,2018.

[33] S. K. Singh," ***Features And Techniques For Speaker Recognition,***" M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03.

[34] T. Chauhan, H.Soni and S. Zafar," ***A Review of Automatic Speaker Recognition System,***" International Journal of Soft Computing and Engineering ,ISSN: 2231-2307, Vol.3, Issue-4, September, 2013.

[35] K. R. Farrell, R. J. M. Senior and K. T. Assaleh," ***Speaker Networks Recognition Using Neural and Conventional Classifiers,***" IEEE Transactions On Speech And Audio Processing, vOL. 2, nO. 1, part 11, January, 1994.

[36] S. Balakrishnama," ***Speech Recognition using Mel cepstrum, delta cepstrum and delta-delta features,***" ECE 8993: Fundamentals of Speech Recognition, for Department Of ElectricalL And Computer Engineering, Mississippi State University ,December, 1998.

[37] A. Mantri, M. Tiwari and J. Singh," ***Performance Evaluation of Human Voice Recognition System based on MFCC feature and HMM classifier,***" Journal of Engineering Research and Applications ,ISSN : 2248-9622, Vol. 4, Issue 2(Version 1), pp.715-719, February 2014.

[38] H.Y. Khdier,W.M. Jasim and S. A. Aliesawi," ***Neural Networks Based Voiceprint Recognition System and Verification,***" REVISTA AUS 26-2,pp.348-357,2019.

[39] M. S. Alghamdi," ***SPEAKER RECOGNITION: EVALUATION FOR GMM-UBM AND 3D CONVOLUTIONAL NEURAL NETWORKS SYSTEMS,***" Master's thesis,Department for Engineering Computer Science,University of Colorado Colorado Springs,2019.

[40] S.Gupta, J. Jaafar, W. F. w. Ahmad and A. Bansal, "***Feature Extraction Using MFCC,***" Signal & Image Processing : An International Journal , vol.4, no.4, pp.101-108,August, 2013.

[41] R. Mukherjee," ***Speaker Recognition Using Shifted MFCC,***"Master's thesis, Department for Electrical Engineering,,College of Engineering,University of South Florida,2012.

[42] E.L.Campbell, G. Hernández, and J. R.Calvo," ***Feature extraction of Automatic Speaker Recognition, analysis and evaluation in real environment,***" 6th International Workshop, IWAIPR 2018, Havana, Cuba, September 24–26, 2018.

[43] R. Ranjan and A.Thakur," ***Analysis of Feature Extraction Techniques for Speech Recognition System,***" International Journal of Innovative Technology and Exploring Engineering , ISSN: 2278-3075,Volume-8, Issue-7C2,pp.197-200, May 2019.

[44] Ch.J. Devi , B.S. P. Reddy, K.V.Kumar,B.M.Reddy and N.R. Nayak," ***ANN Approach for Weather Prediction using Back Propagation,***" International Journal of Engineering Trends and Technology, Vol.3,Issue.1, 2012.

[45] N.Chauhan," ***Speaker recognition using pattern recognition neural network and feedforward neural network,***" International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March-2017.

[46] A. Shrestha and A. Mahmood***," Review of Deep Learning Algorithms and Architectures,***" IEEE Access, Volume XX, February 2017.

[47] S. POUYANFAR, S. SADIQ , Y. YAN, H. TIAN, Y. TAO, M. P. REYES, M. SHYU, S. CHEN and S. S. IYENGAR,"***A Survey on Deep Learning: Algorithms, Techniques, and Applications,***" ACM Computing Surveys, Vol. 51, No. 5, Article 92. Publication date: September 2018.

[48] M. COŞKUN1, Ö.YILDIRIM, A. UÇAR and Y. DEMIR,"***An Overview Of Popular Deep Learning Methods,***" European Journal of Technic, Vol.7, No. 2, 2017.

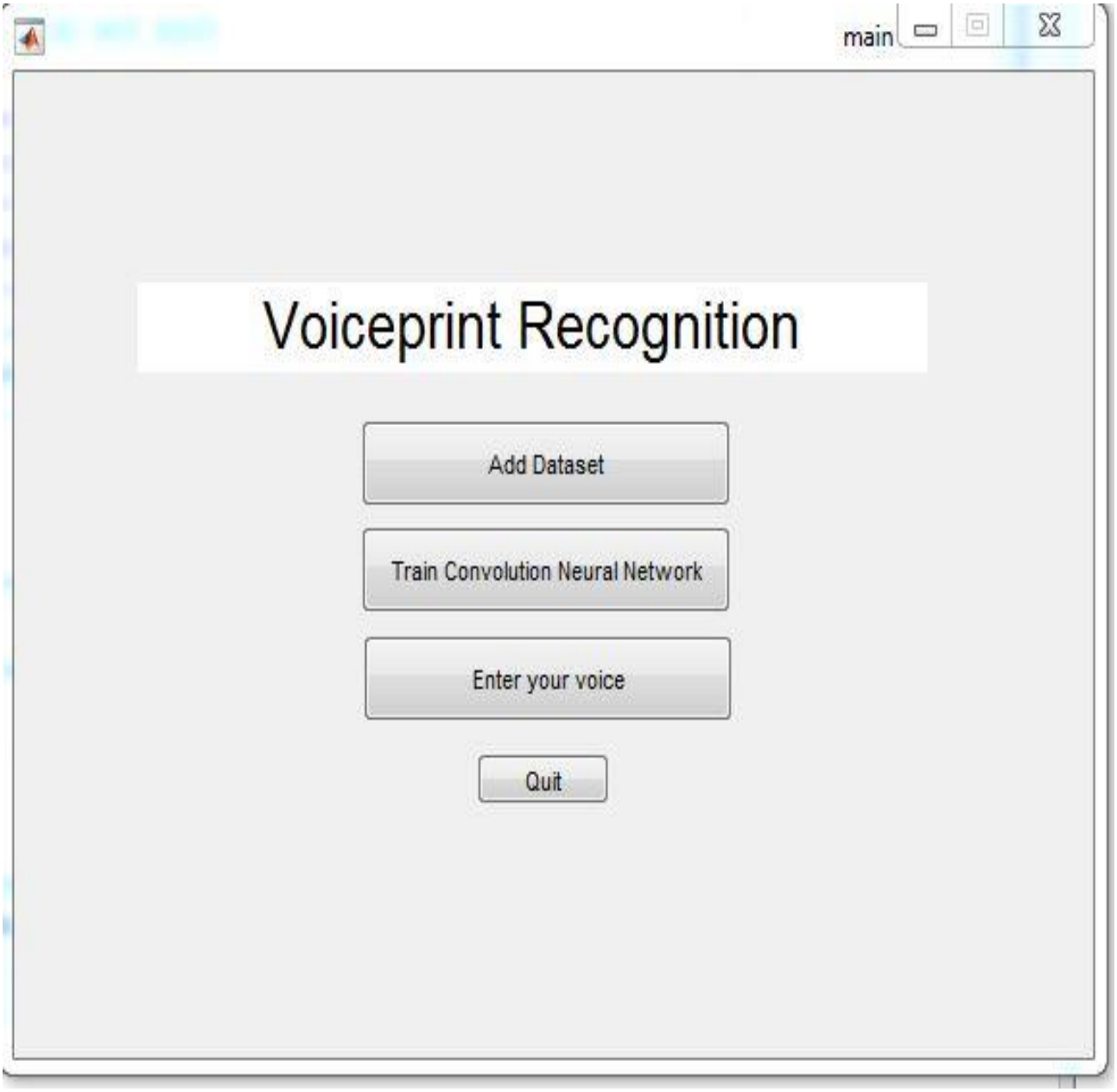
[49] I. Namatēvs,"***Deep Convolutional Neural Networks: Structure, Feature Extraction and Training,***" Information Technology and Management Science, ISSN 2255-9086 ,vol. 20, pp. 40–47 ,December 2017.

[50] L.Deng and D.Yu," ***Deep Learning: Methods and Applications,***" Foundations and TrendsR in Signal Processing, vol. 7, nos. 3–4, pp. 197–387, 2013.

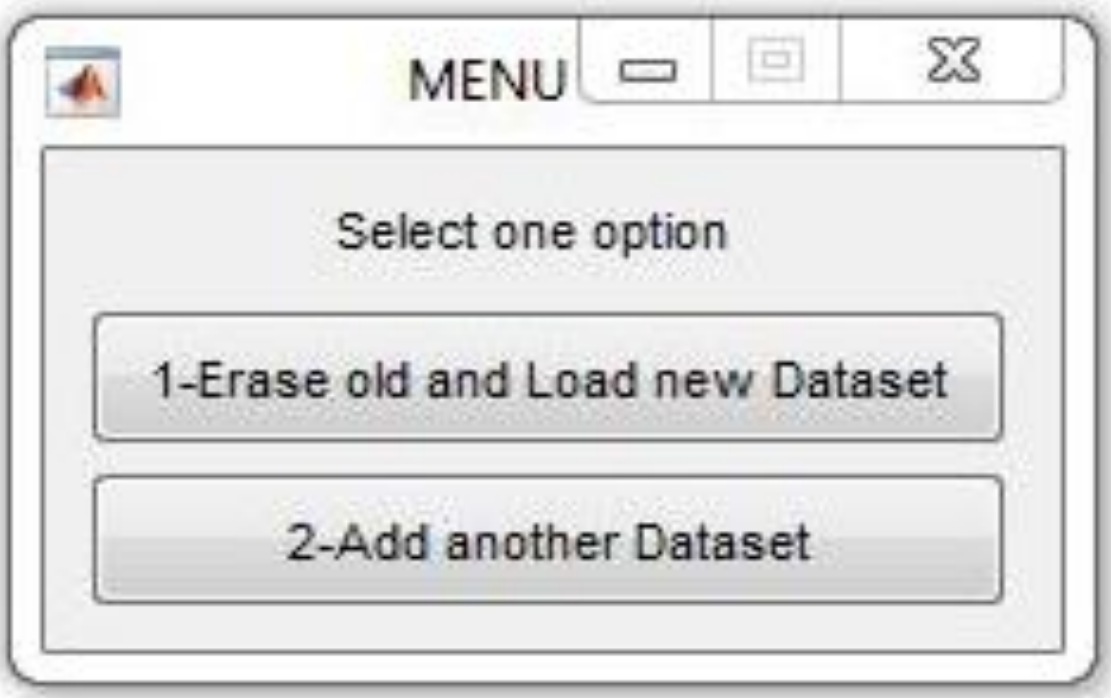
[51] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit/.

[52] S.K. Shetty and A. Siddiqa," ***Deep Learning Algorithms and Applications in Computer Vision,***" International Journal of Computer Sciences and Engineering,Vol.7, Issue-7, July, 2019.

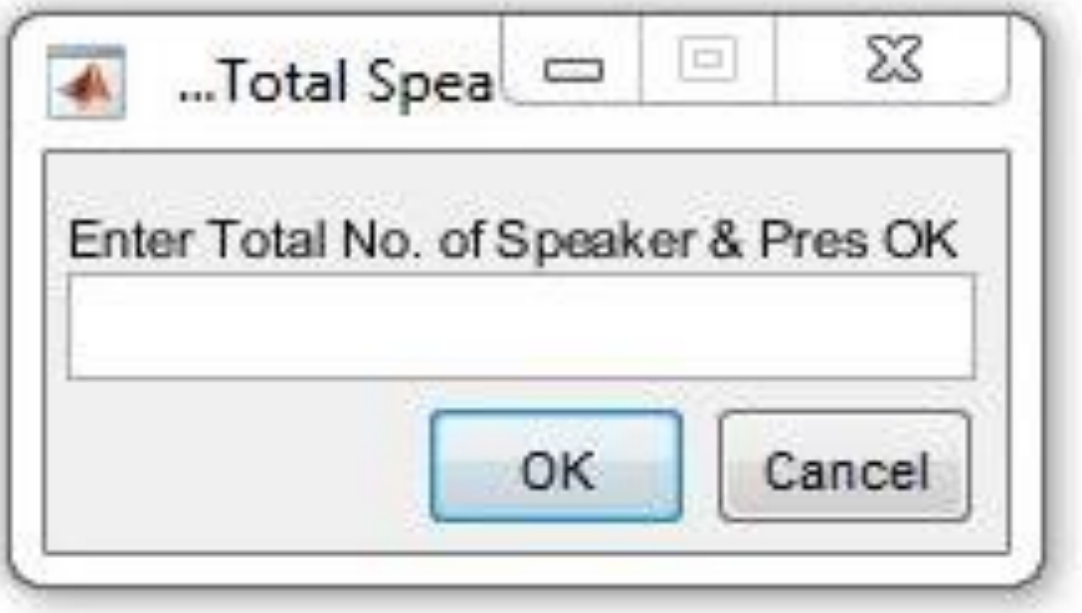
Appendix



Main form of the system



Choosing One of The Options For Add a Dataset



Determined the total of speakers

Choosing one of the options for recognition process

Entering the name of recorded voice

الخلاصة

التعرف على البصمة الصوتية (VPR) هو الآلية التي يتم من خلالها تحديد هوية المستخدم المزعومة باستخدام خصائص مأخوذة من صوتهم، حيث تعد هذه التقنية واحدة من أكثر تقنيات التعرف على القياسات الحيوية شيوعًا وفائدة في العالم خاصة المجالات ذات الصلة بالأمن. يمكن استخدامها للمصادقة والمراقبة وتحديد الطب الشرعي للمتحدثين ومجموعة متنوعة من الأنشطة ذات الصلة.

الهدف من هذه الرسالة هو تصميم استراتيجية التعلم العميق ، والتي ستوفر طريقة للتعلم ضمنيًا التعرف على البصمة الصوتية في البيئات الصاخبة. تم استخدام طريقتين لـ VPRS واستخدمت نفس بنية الشبكة العصبية الالتفافية للطريقتين ومحاولة زيادة دقة النظام من خلال التعامل مع مجموعة بيانات ضخمة تضيف إلى خلفيتها ضوضاء عشوائية لإثبات كفاءة النظام في الظروف الصاخبة.

في هذه الرسالة ، يتم تطبيق محاولة لإنشاء نظام يتعرف على هوية المتحدث البشري باستخدام الشبكة العصبية التلافيفية (.(CNNتستخدم CNN لكل من استخراج الميزات وخوارزمية التعلم العميق ، وبالتالي ستعزز قدرة النظام على أن يكون أكثر دقة وأكثركفاءة. تم تصميم بنية CNN للعمل مع كل من MFCC-CNN و RW-CNN.في كلتا الحالتين ، تكون مدخلات CNN المقترحة عبارة عن صور ، أي الشبكة التي تعاملت مع الصور ، حيث يتم استخدام نفس بنية CNN لكلتا الطريقتين. تظهر النتائج التي تم الحصول عليها أن كلا الطريقتين متشابهتان في دقتها ٠,٩٦ ومتوسط خطأ مربع ٣,٢٠٠٠e-08 النتائج لكن تختلف في الأداء حيث تظهر نتائج الوقت أن RW-CNN أفضل من MFCC-CNN سواء مع الضوضاء أو بدونها. وبعبارة أخرى ، فإن RW-CNN أكثر كفاءة في البيئات النظيفة والصاخبة من .MFCC-CNN

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة الانبار

كلية علوم الحاسوب وتكنولوجيا المعلومات

قسم علوم الحاسبات

خوارزميات التعلم العميق لأنظمة تمييز البصمة الصوتية في البيئات الصاخبة

رسالة مقدمة الى

قسم علوم الحاسبات – كلية علوم الحاسوب وتكنولوجيا المعلومات - جامعة الانبار
وهي جزء من متطلبات نيل درجة ماجستير علوم في علوم الحاسبات

قدمت من قبل

هاجرياس خضير

بإشراف

أ.د. صلاح عواد سلمان

أ.د. وسام محمد جاسم

١٤٤١هـ **2020 م**