

DESIGN AND IMPLEMENTATION OF SPEECH TO SIGN LANGUAGE TRANSLATOR
SYSTEM

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

AHMED HASHIM HAMZA TAMEEMI

A MASTER OF SCIENCE THESIS
IN
THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

JUNE 2021

DESIGN AND IMPLEMENTATION OF SPEECH TO SIGN LANGUAGE TRANSLATOR
SYSTEM

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

ATILIM UNIVERSITY

BY

AHMED HASHIM HAMZA TAMEEMI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF SCIENCE

IN

TH THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

JUNE 2021

Approval of the Graduate School of Natural and Applied Sciences, Atilim University.

Prof. Dr. Ender Keskinlik
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of **Master of Science in Electrical and electronic engineering department, Atilim University.**

Assoc. Prof. Dr. Kemal Efe Eseller
Head of Department

This is to certify that we have read the thesis DESIGN AND IMPLEMENTATION OF SPEECH TO SIGN LANGUAGE TRANSLATOR SYSTEM submitted by AHMED HASHIM HAMZA TAMEEMI and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Hakan TORA
Supervisor

Examining Committee Members:

Asst. Prof. Dr. Emre SUMER
Computer Engineering Department
Baskent University

Asst. Prof. Dr. Hakan TORA
Electrical and Electronics Engineering Department
Atilim University

Asst. Prof. Dr. Baran USLU
Electrical and Electronics Engineering Department
Atilim University

Date:28/6/2021

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



AHMED HASHIM HAMZA TAMEEMI

Signature: _____

ABSTRACT

Design and Implementation Of Speech To Sign Language Translator System

Tameemi, Ahmed Hashim Hamza

M.S., Electrical and Electronics Engineering Department

Supervisor: Asst. Prof. Dr. Hakan TORA

June 2021, 60 pages

Numerous efforts have been to design interpretation strategies between the healthy and deaf people. One of the most important of these systems is what called as Human-Computer-Interaction (HCI) systems. Many simulation works of sign language are evolved in the last years in this field. At most, these efforts have been worked to convert Sign Language into English speech and text, and verse versa. Usually there is minimal interaction between healthy persons and the Arabic deaf persons. So it is very important to design a conversion system that can translate ArSL into Arabic speech or text and vice versa to provide a better communication between normal persons and Arab deaf community. In this thesis, a translation system for Speech to Sign language will be accomplished; the target is to design a translator scheme that can translate the speech signals of Arabic letters into display the standard Arabic sign language. This circuit can be used to realize a human-friendly program scheme that can be used for numerous applications such as communication between the deaf and the normal people. Several stages have been carried out to gratify the system design requirements, including: recording the signals of speech for data collection, feature extraction after pre-processing, and the recognition final step. Then, the system provides the recognition factors to the translator structure, to display image of movements of fingers that antipodean the spoken letter. In this work, Pattern Recognition Neural Network (PRNN) was used. It is a feedforward network that has an ability of classifying input samples into desired classes. The network was trained with back-propagation algorithm by 560 training samples, 20 samples for each letter of the entire 28 main Arabic letters.

Then, the trained model was validated and tested with 140 and 140 dataset respectively, to see how fit the PRNN model predicts the matching data set of output labels. The network is successfully trained for all 28 classes (letters). The recognition was achieved with an excellent diagnosis rate of up to 98%.



Keywords: Arab Sign Language ArSL, Speech Recognition and Pattern Recognition Neural Network (PRNN)

ÖZ

İşaret Dili Çevirmen Sistemi Tasarımı Ve Uygulaması

Tameemi, Ahmed Hashim Hamza

Yüksek Lisans, Elektrik- Elektronik Mühendisliği Bölümü

Danışman: Dr. Öğr. Üy. Hakan TORA

Haziran 2021, 60 sayfa

Sağlıklı ve sağır insanlar arasında yorumlama stratejileri tasarlamak için çok sayıda çaba sarf edilmiştir. Bu sistemlerin en önemlilerinden biri İnsan-Bilgisayar-Etkileşim (HCI) sistemleri olarak adlandırılan sistemlerdir. Bu alanda son yıllarda birçok işaret dili simülasyon çalışması geliştirilmektedir. Bu çabalar en fazla İşaret Dili'ni İngilizce konuşma ve metne, ayeti ise İngilizceye dönüştürmek için çalışıldı. Genellikle sağlıklı kişilerle Arap sağırları arasında minimum etkileşim vardır. Bu nedenle, normal kişiler ve Arap sağır topluluğu arasında daha iyi bir iletişim sağlamak için ArSL'yi Arapça konuşmaya veya metne çevirebilen bir dönüştürme sistemi tasarlamak çok önemlidir. Bu tezde, Konuşmadan İşaret diline çeviri sistemi yapılacaktır; hedef, Arap harflerinin konuşma sinyallerini standart Arap işaret diline çevirebilen bir çevirmen şeması tasarlamaktır. Bu devre, sağırlar ve normal insanlar arasındaki iletişim gibi çok sayıda uygulama için kullanılabilecek insan dostu bir program şemasını gerçekleştirmek için kullanılabilir. Sistem tasarım gereksinimlerini karşılamak için, veri toplama için konuşma sinyallerinin kaydedilmesi, ön işlemeden sonra özellik çıkarımı ve tanıma son adımı dahil olmak üzere çeşitli aşamalar gerçekleştirilmiştir. Daha sonra sistem, konuşulan harfin tersi olan parmak hareketlerinin görüntüsünü göstermek için çevirmen yapısına tanıma faktörlerini sağlar. Bu çalışmada Örüntü Tanıma Sinir Ağı (PRNN) kullanılmıştır. Girdi örneklerini istenen sınıflara ayırma yeteneğine sahip ileri beslemeli bir ağıdır. Ağ, toplam 28 ana Arap harfinin her harfi için 20 örnek olmak üzere 560 eğitim örneği ile geri yayılım algoritması ile eğitilmiştir. Ardından, PRNN modelinin eşleşen çıktı etiketlerini ne kadar uygun tahmin ettiğini görmek için eğitilen model sırasıyla 140 ve 140 veri seti ile doğrulandı ve test edildi. Ağ, 28 sınıfın (harflerin) tamamı için başarıyla eğitilmiştir. Tanıma, %98'e varan mükemmel bir teşhis oranıyla sağlandı.

Anahtar Kelimeler: Arap İşaret Dili ArSL, Konuşma Tanıma ve Örüntü Tanıma Sinir Ağı (PRNN)



DEDICATION

TO THE CANDLE THAT HIGHLIGHTED MY WAY, MY GREAT ... FATHER

TO THE SPRING OF KINDNESS, MY DEAREST ... MOTHER

TO MY BROTHERS AND SISTERS WITH LOVE AND GRATITUDE

AND FOR EVERY PERSON WHOSE HEART THROBBED FOR ME



ACKNOWLEDGMENT

In the Name of Allah, the Most Gracious, the Most Merciful. Peace and blessings of Allah be upon the noblest of the Prophets and Messengers, our Prophet Mohammed and upon his family, companions and who follows.

I would like to express my gratitude to all, who during the past two years have supported me in some way or another. To my supervisor Asst. Prof. Dr. Hakan TORA, a special thanks for giving me the opportunity to work with him and for his patience, tolerance, understanding, and encouragement. It has been a great pleasure working with him. I really appreciate the time and efforts he devoted for the completion of my dissertation. I'd like to express my gratitude to the head of institute represented by the director Prof. Dr. Ali Kara for helping by providing facilities, advice and support. I also thank the staff members of Atilim University for providing facilities. A special thanks to my brothers (Ayser and Aymen) and their wives on their unlimited supportive, big gratitude to my colleagues the engineers Yousif, Mustafa, Ahmed, Anwer, Zaid, for providing help and support. I am grateful to my parents for being my support and guide during most difficult times of my life. They are source of inspiration to me all the time. Without them it would not have been possible for me to become what I am today.

TABLE OF CONTENTS

ABSTRACT.....	iii
ÖZ.....	v
DEDICATION.....	vii
ACKNOWLEDGMENT.....	viii
TABLE OF CONTENTS.....	ix
LIST OF TABLES.....	xii
LIST OF FIGURES.....	xiii
CHAPTER 1.....	1
1.1 Introduction	1
1.2 Literature Survey:.....	2
1.3 Methodology	5
1.4 Aim of the work	7
1.5 Applications of the work	7
1.6 THESIS ORGANIZATION	7
CHAPTER 2	1
2.1 Introduction	1
2.2 THE SPEECH SIGNAL.....	1
2.3 Speech Signal Representation	2
2.3.1 Short-time Fourier Analysis	1
2.4 Arabic language.....	2
2.4.1 Sign Language	3
2.4.2 Arabic Sign Language ArSL:	3
2.5 Preprocessing:	4
2.5.1 Sound Level Measurement	5

2.5.2 Speech Signal Pre-Processing	6
2.6 Features Extractions:	9
2.6.1 Temporal Features	9
2.7 Speech Recognition:	11
2.7.1 Requirements Selection of Audio Classifier:	11
2.7.2 Relevant issues of ASR design:	11
CHAPTER 3	1
3.1. Introduction	1
3.2. Proposed Work for Speech to Sign Translator	1
3.2.1 Data Collections	1
3.2.2 Preprocessing	6
3.2.3 Feature Extraction	9
CHAPTER 4	11
4.1 Introduction	11
4.2. Recognitions	11
4.3 Design of PRNN	11
4.3.1 Training Process of the Proposed System	12
4.4 Results	13
4.4.1 Training Results	13
4.4.2 Testing Results	16
4.5 Discussion	21
CHAPTER 5	1
5.1 Conclusion	1
5.2 Future Work	2

LIST OF TABLES

Table 1.1 Relationship between sound intensity and perceived loudness.	68
Table 2.1 Main issues of ASR design [32]	12
Table 3.1 Structure of dataset elements	4
Table 4.1 Structure of class output	5
Table 5.1 Confusion matrix estimated values.....	20



LIST OF FIGURES

Figure 1.1 General construction of environmental audio recognition scheme	6
Figure 2.1 Speech recognition system by source-channel model.....	2
Figure 2.2 (a) Short-time spectrum of speech signal and its achieved with: (b) rectangular window of 10ms; (c) rectangular window of 25ms; (d) Hamming window of 10ms; and (e) Hamming window of 25ms.....	1
Figure 2.3 (a) Spectrogram of the speech signal (b) The original speech waveform	14
Figure 2.4 Arabic Sign Language ArSL.	4
Figure 2.5 Pre-processing to speech signal.....	7
Figure 2.6 Pre-emphasize Filter Responses [30]	9
Figure 2.7 The principle of Artificial Neural Network.....	124
Figure 2.8 The construction of Multi-layer Perceptron Neural Network	157
Figure 3.1 Block diagram of the Suggested Work for Talking to Sign Translator.....	29
Figure 3.2 Dettaille of Speech to Sign Translator Proposed Work.....	Error! Bookmark not defined.
Figure 3.3 Microphone used for Data Recording	Error! Bookmark not defined.
Figure 3.4 Preprocessing system.....	66
Figure 3.5 Figure (3.5) Signal after removing silent	77
Figure 3.6 Signal after normalization	88
Figure 3.7 Signal after pre-emphasis	39
Figure 3.8 Features extracted from signals	100
Figure 4.1 Seven inputs-Twenty eight outputs PRNN for two training iterations.....	43
Figure 4.2 Training Performance for two training iterations.....	144
Figure 4.3 Training actual output.....	155
Figure 4.4 Error between training outputs	155
Figure 4.5 Rounded training outputs	16
Figure 4.6 Training error for rounded output.....	16
Figure 4.7 Testing actual output	17
Figure 4.8 Error between testing outputs	17
Figure 4.9 Rounded testing outputs	188

Figure 4.10	Testing error for rounded outputs.....	188
Figure 4.11	Confusion matrix result	199
Figure 4.12	Curve of Receiver Operating Characteristic (ROC).....	50
Figure 4.13	Test Result for two different input letters.....	211



CHAPTER 1

GENERAL OVERVIEW

1.1 INTRODUCTION

Signing is continuously portion of human communications. Using of sign or gestures is not related to gender, age, or ethnicity. When the muscles of speech are not fully developed, children use motions as the principal meaningful of communication. For thousands of years, deaf persons have formed signs and used them between themselves. With the diversity of deaf cultures in countries, signature has advanced to shape sophisticated and complete languages. Subsequent groups of deaf children have been learned and elaborated these sign languages. Using the mutual sign language, deaf people communicate simply between themselves. But there is a problem when the communication starts between a deaf and other healthy person. Generally both will be disappointed after a brief period. Wherefore, numerous efforts have been to design interpretation strategies between the healthy and deaf people. One of the most important of these systems is what called as Human-Computer-Interaction (HCI) systems [1]. Many simulation works of sign language are evolved in the last years in this field. At most, these efforts have been worked to convert Sign Language into English speech and text, and verse versa.

Important inventions and innovations have been presented for persons with learning shortages to overcome their difficulties. These creations based on Personal Computer (PC) and use Information and Communication Technologies (ICT). Also, they can be explored to help deaf persons in their earning and communication difficulties.

There are over 7 Billion persons in world at 2012. Deaf are around 70 million persons consistent with the World Federation of the Deaf [2], more than 1 million just in Saudi Arabia [3]. So working on a study to grow a scheme to convert an Arabic speech to an avatar, will have an important effect to the deaf community in more than 25 Arabic

countries, where the official language is Arabic; and in more than 20 countries where a minority of the people speak Arabic language. Luckily, the Arabic sign language is not diglossic (standard and colloquial) as in the Arabic language. Where each Arabic country has its own dialect, in spite of its people speak Arabic such as in Iraqis, Libyans, Egyptians, and so on. Luckily, the same alphabets are used in all Arabic countries and consequently the same Arabic Sign Language (ArSL). This is a very helpful feature for the research work in this field.

Usually there is minimal interaction between healthy persons and the Arabic deaf persons, where Arab deaf community is nearly closed ones, and is mainly focused about families through deaf members [4]. So it is very important to design a conversion system that can translate ArSL into Arabic speech or text and vice versa to provide a better communication between normal persons and Arab deaf community. Several associated works that have been donated from other studies will be labeled in the next section.

Humans can simply recognize and separate one audio source from a sound combination, such as sure speech from a noisy environment containing music and speech from different people. The Analysis of sound for recognizing and separating of combination of sound causes existing in an acoustic sight is known as Computational Auditory Scene Analysis (CASA) [5, 6]. This analysis can make Personal Computer (PC) better than humans for hearing and understanding acoustic content. CASA can be divided into three key study areas [6]: 1) Recognition of audio context such as action or location happening in a specified environment, 2) Recognition of individual categorized sound events existing in the acoustic sight, 3) Recognition/classification of the contents of audio signals such as for indexing, audio content retrieval, and searching based on audio.

This work deals with speech recognition and detection event also mentioned as Speech to Sign language conversion.

1.2 LITERATURE SURVEY

Signing is continuously portion of human communications. Using of sign or gestures is not related to gender, age, or ethnicity [1]. In the latest years, numerous researches in sign language methods have been obtained [1].

As the development of speech recognition systems or Automatic Speech Recognition (ASR) whether dependent or independent speaker, becomes widely used in all daily life applications. Researchers follow their works to progress of this task in both software and hardware implementation. Some of these related studies are listed as follow:

Musa Hakan Asyali et al., 2011 [7] displayed a voice controlled dynamic hand prosthesis to play out a few essential movements. They planned a multi-fingered prosthetic hand with the capacity of grabbing and discharging objects. The second part of their work included the utilization of speech recognition to control the prosthetic hand.

Nashwa El-Bendary et al., 2011 [8] proposed an ArSLAT system to recognize the Arabic alphabets which consists of five stages: pre-processing, category and best-frame detections, extraction of features, and classification stage. Results show an accuracy of 83.7% and 91.3% for using Multi-Layer Perceptron (MLP) classifier and Minimum Distance Classifier (MDC) respectively.

Ahmed Abdul Haseeb and Asim Illyas, 2012 [9] presented an automatic and general system for conversion of Pakistan Sign Language (PSL) to English speech and vice versa. They concluded that the natural language processing, machine translation, and speech recognition methods can be basic components for such an automatic and general scheme. Validation of the system model is achieved frequently based on field experts.

S.M. Halawani et al., 2012 [10] presented an Arabic conversion scheme from speech to signing can be used by the Deaf folks. This system was composed of information of captured 3D Arabic signing gestures. The signing gesture was recorded by exploitation information gloves.

The digitized signing in a graphical translation was re-animated customary. The stress was to provide the first gesture as close as possible.

Alaa Hassan Mahmoud and Salma Alzaki Ali, 2014 [11] used Vector Quantization (VQ) and Mel-frequency Cepstrum Coefficient (MFCC) techniques to convert speech of Arabic word into text. This has been realized by first recording teachers' voices for each word in a noisy environment. Secondly these words have been used to extract their features using MFCC technique. Then VQ technique creates code for each word. Finally, each code word will be indexed with its equivalent text at the conversion stage. The system Word Error Rate was 20%.

Anetha K et al., 2014 [12] developed an associate degree Automatic Yankee Signing (AYS) recognition system exploited ANN to convert the sign language alphabets into sound and text. A glove circuit was implemented with 3-axis measuring device, flex sensors, and electromyogram sensors to pick up motions.

Jonathan Gatti, et al., 2014 [13] used the OpenSCAD to design a robotic hand prototype and a low-cost 3D printer to manufacture it. The prototype used voice based control to yields the sign language. They used a Raspberry Pi mini-computer to produce instructions for a set of servo motors that activated by an Arduino UNO controller. They discussed limitations of the prototype, its features and probable upcoming progresses.

Zaini Othman¹, et al., 2014 [14] used characters that consequent from Arabic and distinct in Malay as in a Jawi Unicode script for improved accuracy. This table is based on the mixture of Urdu Unicode, Persian, and Traditional Arabic. They used LPC-ANN combination to reach maximum accuracy upon controlled environment. Results of the proposed engine show an average recognition accuracy of 89%.

Attar Shahebaj, et al., 2015 [15] presented an Automatic Speech Recognition (ASR) system for speech or hearing impaired people communication. Their system recognizes

speech sign. Recognized spoken words are represented using American standard sign language via a robotic arm and also on the computer using visual basic. A software package is provided to convert the speech signal into the sign language.

Megha Agrawaland and Tina Raikwar 2016 [16] used a signal processing techniques to design a speech recognition system. They compared its performance with the adopted feature extraction method based ASR scheme for the specific language.

Beena M.V., et., al. 2017 [17] introduced a recognition system for static gestures of American Sign Languages (ASL) collected from Kinect sensor into the field of automatic sign language. CNN classifier was used from kinect Depth images. 24 alphabets and 0-9 numbers were classified with 33000 images for an efficiency of 94.6774%.

Ayushi Trivedi, et., al., 2018 [18] studied many Text To Speech (TTS) and Speech to Text (STT) conversion techniques, and their usage and applications. They concluded that Hidden Markov Model (HMM) works better than Cuckoo Search Optimization technique based on ANN as a STT converter although its disadvantage of computational feasibility. Also they concluded that the Formant Synthesis technique is the best TTS converter compared with Articulator Synthesis and Concatenative Synthesis techniques.

D. Nagajyothi, et., al., 2018 [19] presented an Airport enquiry system based on ASR for telugu language. The most commonly asked questions in an airport are collected to be as a training database for the Convolutional Neural Network (CNN). Their examination verified that speech recognition system based on the CNN was capable to realize better performance than the system based on other conventional NN methods.

Mariusz Kubanek, et., al., 2019 [20] presented a specific coding of frequency and time characteristics for speech recognition. Their approach used three separate CNN layers: traditional time convolution layer, spectrum frequency convolution layer

and Mel-Frequency Cepstral Coefficients (MFCC) frequency convolution layer. The proposed method used for dividing continuous speech into syllables. They concluded that the division into syllables is quite easy when speaking slowly.

M. M. Kamruzzaman, 2020 [21] proposed a vision-based system for the recognition of Arabic hand sign-based letters using CNN with a deep learning model then translation of them into a text which is finally output as an audio of the Arabic language by a speech engine. Accuracy of 90% was given in the proposed system.

Shailesh bachani, et., al., 2020 [22] developed a system for hand gesture detection based on features like status of fingers, center of mass (centroid), orientation, and thumb location in hand image. The proposed system used ANN as a classifier for alphabets and numbers of the American Sign Language (ASL). The system can capture hand gestures and make the words as in text format. Mute persons will understand the sentences after seeing these words.

1.3 METHODOLOGY

Audio classification, as in several pattern classification techniques, is consists of three basic constituents: (1) Measurement of the signal or sound event in the sensing constituent; (2) characteristic features extraction in the audio processing constituent; and (3) context recognition of the sound event in the classification constituent.

The microphone is usually used for the sensing in Context Awareness (CA) applications that based on audio. Features extraction is mostly used as processing for the recorded audio signal. For speech processing, the time-frequency analysis methods established for audio tasks can be used initially in several cases. Each frame in the audio signal is quantized in the feature extraction stage and it's transformed into several individual k dimensional vectors. These feature vectors the recognized by the classifier used to

decide what they represent as an audio event context. Figure 1.1 shows the overall construction of the sound classification technique, where the raw audio data represents the input and the context (activity) information yields as an output.

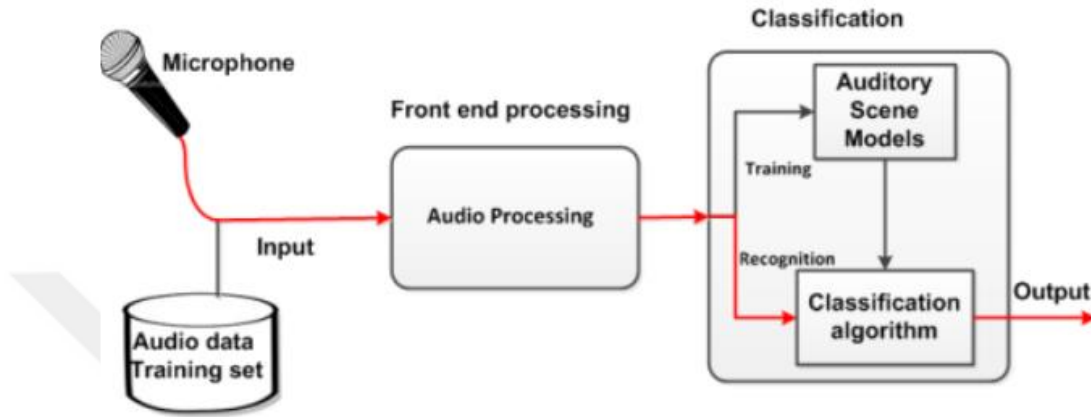


Figure 1.1 General construction of environmental audio recognition scheme

The Environmental Sound/audio Recognition (ESR) procedure has two stages: training and recognition stages. The system receives the collected training data sets through the training stage, and then creates characteristic models for each of the sound input. Where, through the recognition stage, the system uses the models created through the first stage, receives its sound signals directly from the input device (microphone), and then determines and matches the category of the received audio data. This data are processed directly (online) to provide real-time recognition continuous result for the customer.

The principal steps that have been processed through the design of the proposed ArSL translation system are as follow:

1. First, during literature reading of audio processing techniques, the classification/recognition techniques and feature extraction procedures are showed. The key objective of this step is to initially determine the finest combination of the proper classification algorithm and the effective feature sets that can deliver result of less computational complexity and maximum recognition accuracy.
2. As an imperative process, performance test is executed for the pre-selected methods in the literature study for comparison between them and for choosing the proper scheme in

an offline simulation process using Matlab codes.

3. ArSL translation system is designed based on the experimental comparison results using the most suitable techniques for the sound of Arabic letters, using another Matlab (offline) simulation process. The designed ArSL translation system can process online the audio input data and delivers real-time recognition output.

1.4 AIM OF THE WORK

In this thesis, a translation system for Speech to Sign language will be accomplished; the target is to design a translator scheme that can translate the speech signals of Arabic letters into display the standard Arabic sign language. This circuit can be used to realize a human-friendly program scheme that can be used for numerous applications such as communication between the deaf and the normal people. Several stages have been carried out to gratify the system design requirements, including: recording the signals of speech for data collection, feature extraction after pre-processing, and the recognition final step. Then, the system provides the recognition factors to the translator structure, to display image of movements of fingers that antipodean the spoken letter.

1.5 APPLICATIONS OF THE WORK

1- The current invention is a system of translation and conversion from the Arabic language into the standard Arabic sign language. The main application is to establish a link between the deaf and the outside world represented by the healthy people.

2 - In television programs that can be added to such a system to engage this segment (deaf people).

3 - In schools and universities to communicate with the hearing impaired.

1.6 THESIS ORGANIZATION

Chapter one introduced a general overview for the deaf community and signing, the related work of sign language systems, methodology and aim of the work. Chapter two will present a necessary overview about the analysis of speech signal implicated in recognition functions, Arabic language review, the ArSL, general preprocessing and speech features, in addition to Artificial Neural Networks principles as signal classifier. Chapter three will describe the proposed conversion system that can translate the speech in Arabic language into an ArSL. The proposed work includes Data Collections, Preprocessing (Removing Silence, Normalizing, and Pre-Accentuation), and Feature Extraction. Chapter four will present results for training and testing processes of the proposed model based on ANN classifier. Discussion of these results also will be introduced. Finally chapter five will contain the conclusion and future works.

CHAPTER 2

BACKGROUND & PRINCIPLES OF SPEECH PROCESSING

2.1 INTRODUCTION

To comprehend the required procedures and methods used during this work, the theoretical part that gives the necessary background about the analysis of speech needed to applications based recognition techniques is described in this chapter. The theoretical part is distributed into many subdivisions. The speech signal and its characteristics, Arabic language review, and the ArSL are labeled in the first division. The second subsection is an overview to preprocessing stage and all operations done on speech signal. The third subsection talking about important features of speech, features extraction and the analysis of automatic speech recognition. And finally general principles of artificial neural networks are presented.

2.2 THE SPEECH SIGNAL

As a starting point, how the speech signal is perceived after created in the human system will be introduced. Systems of speech recognition attempt to simulate the speech communication system in human, where the speech signal is transmit from speaker human into listener. Figure (2.1) show the source-channel model in the system of speech recognition [23, 24] with certain similarity to the system of human communication.

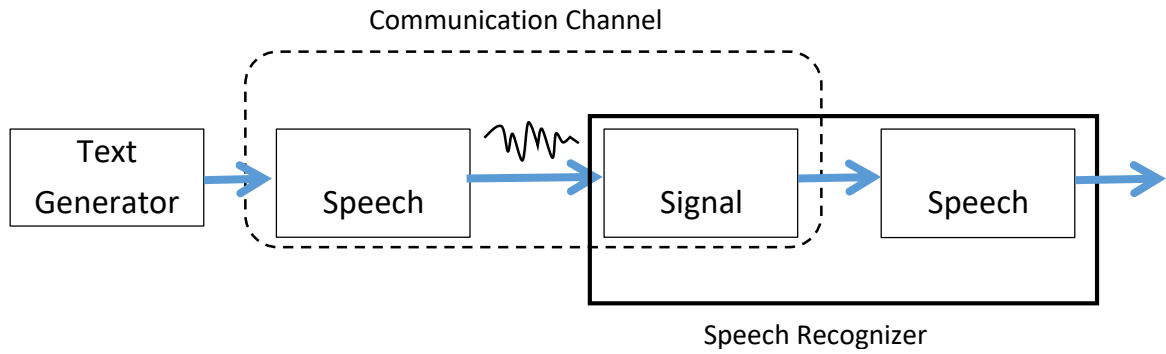


Figure 2.1 Speech recognition system by source-channel model

Transfer ideas are the main objective of human speech communication. The brain of speaker is made the ideas, perform the sequence of source word W , and then text generator will delivered it. The speech generator section converts the source word W into the signal waveform of speech X that is transported to the listener through the communication channel (air) with a possible external noise denoted as $\hat{X}$. The human hearing system perceives the acoustical signal, and then the brain of listener processes this signal to comprehend its meaning.

The speech recognizer components: the signal processing and the speech decoder are shape the perception process. They decode and process the received audio signal $\hat{X}$ into a word sequence $\hat{W}$. Hopefully $\hat{W}$ is as close as to the original W [23]. Thus, in the speech recognition system, the speech perception can be seen as inverse process of the speech production.

2.3 SPEECH SIGNAL REPRESENTATION

Plots of waveforms can be used for extract some feature about phonetic information. But for most important properties of speech, the resonances can be great significance for transporting the phonetic information and for displaying these features in what called as the short-time spectrum [25].

2.3.1 Short-time Fourier Analysis

When the signal is multiplied by a time window function with a suitable period, the amount of a Fourier Transform this waveform is called as the short-time [25].

The speech signal will be decay into frames (short speech sequences) to be analyzed each one individually in the short-time analysis. In this analysis the greatest significant factors are the shape and the width of the time window. For comparison rectangular and Hamming windows are used both of 10ms and 25ms in obtaining the short-time spectrum of spoken signal as in Figure (2.2) [25]. Rectangular window is more properly for better stationary resolution, while for better frequency resolution the Hamming window can be used. Based on short-time Fourier analysis, the Spectrogram is the competent two-dimensional representation of the speech signal. It shows frequency in the vertical axis and time in the horizontal axis. Then, a grey scale is usually used, for specifying the energy in each frequency/time point, where black denotes high energy, and white for low energy [24]. A color scale can be used to represent spectrograms as in Figure (2.3), in which lightest red portions signify high energy, and darkest blue portions for low energy.

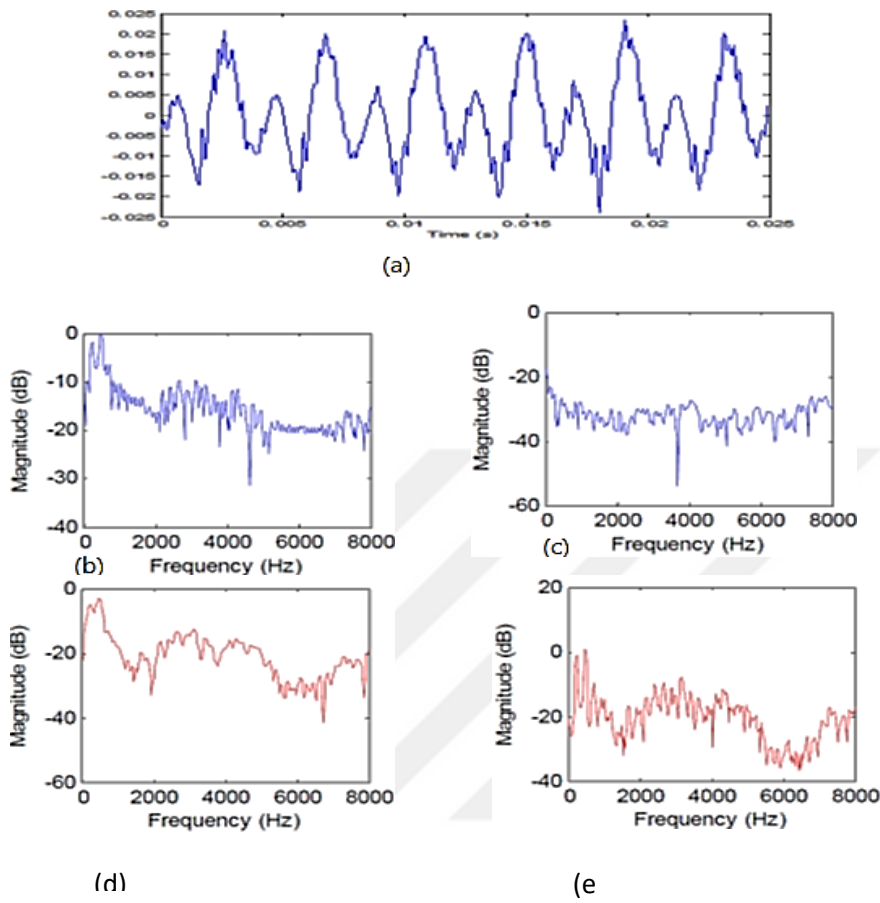


Figure 2.2 (a) Short-time spectrum of speech signal and its achieved with: (b) rectangular window of 10ms; (c) rectangular window of 25ms; (d) Hamming window of 10ms; and (e) Hamming window of 25ms [25].

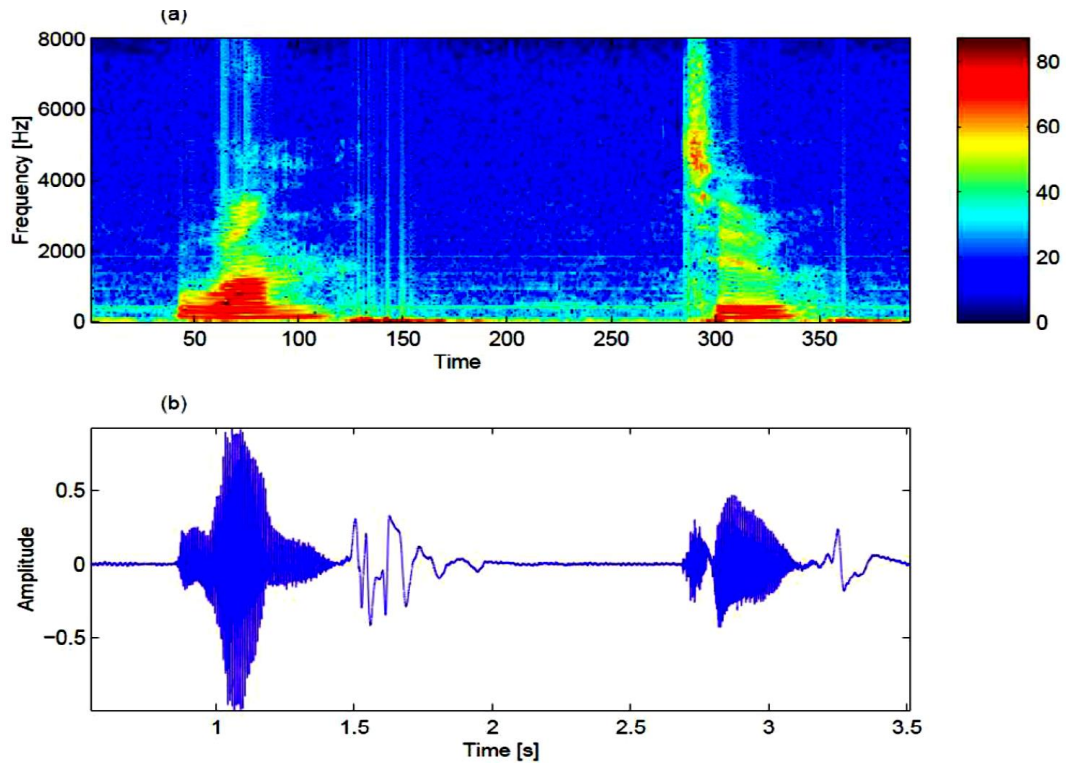


Figure 2.3 (a) Spectrogram of the speech signal (b) The original speech waveform [26]

2.4 ARABIC LANGUAGE

Arabic is one of the oldest Semitic languages in the world. Presently, in terms of the number of speakers it is the fifth language. In 25 countries, it is the native language with more than 250 million speakers counting Saudi Arabia, Egypt, Iraqi Jordon, Yemen, Oman, Lebanon, and Syria...etc. Also there are some languages such as Urdu and Persian use Arabic Alphabets. However there have been few studies on ASR of Arabic.

Most preceding searches on Arabic speech recognition have focused on recognizing of the formal linguistic standard called as Modern Standard Arabic (MSA). It is used in the broadcast media, courtrooms lectures, etc. in the Arabic-speaking world. Recent MSA recognition can be achieved with less than 15% of word error rates. On the other hand, at 1997 start few studies to develop Arabic speech recognizer for different dialectal forms on Call Home data [27].

2.4.1 Sign Language

Sign language is a mixture of shapes and movements of some portions of the human body such as hands and face used by mute persons and deaf. Signs are used to communicate sentences and words to spectators. Signs can be indicated by static shapes of the hand called posture. Gestures are specific movements of the hands with definite shapes, with some facial expressions, in sign language. Signs are provided for entire words or letters. So, sentences can be performing using the signs of word or letters [28]. Arabic Sign Language (ArSL) is chosen in this work.

So, gestures play a main function in the human daily events, especially in communication to make understanding be easier. Gesture recognition denotes to distinguishing human expressions of motion containing the, head, face, arms, hands, and/or body. Hand gestures show a significant role in helping us for more expression in less time. They are employed in many researches for Human-Machine interface and for controlling in many other applications [28], such as remote controllers, robots, media players, and virtual environments. Sign languages by hand gestures can be construed exclusively by examining the patterns of four elementary parts: hand configuration (shape), orientation, hand movement, and classification. [28]

2.4.2 Arabic Sign Language ArSL:

ArSLs are essentially manual languages that contain three features: hands shape (hand configuration), position/space (location of hand in relation to body), and movement (contacts and directions within space). In addition to the ArSLs manual features, they also use other non-manual shapes, such as mouth, tongue, and human face. All the avatars of the 35 Arabic alphabets should be storied for the ArSL sign database as shown in Figure (2.4) [29].

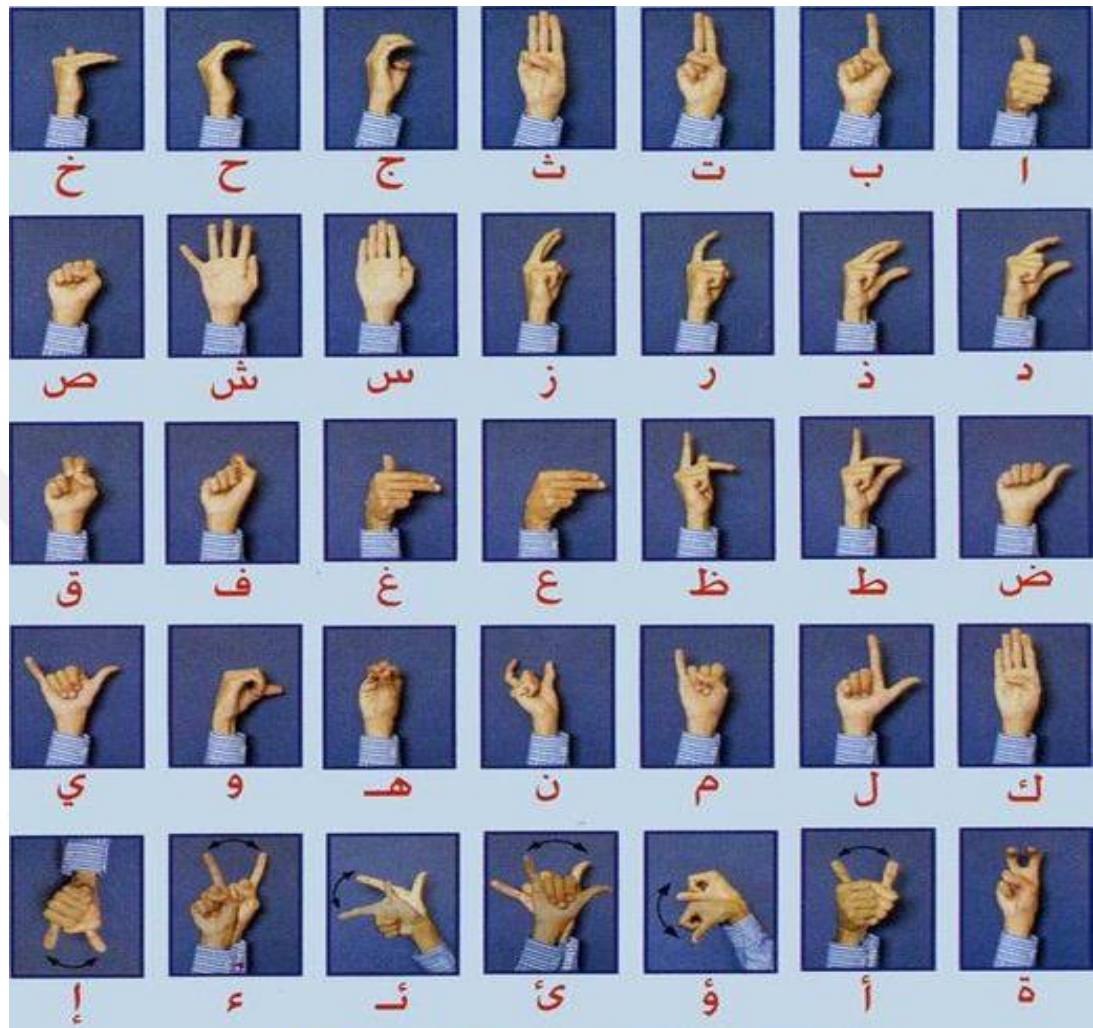


Figure 2. 4 Arabic Sign Language ArSL.

2.5 PREPROCESSING:

Speech signal contained large amount of information and has a time varying non-stationary property. In other words, it's a signal that has infinity of spectral information. The frequency bandwidth of a speech signal is about 16 KHz. However, most of speech energy is under 7 KHz. A speech signal is called orthophonic if all the spectral components over 16KHz are discarded. A telephonic lower quality signal is obtained whenever a signal does not have energy outside the band 300-3400 Hz.

Therefore, digital speech processing is usually performed by a frequency sampling between 8000 samples/sec. and 32000 samples/sec., these values correspond to the bandwidth of 4 KHz and 16 KHz respectively [30].

2.5.1 Sound Level Measurement

The machine simulates the human ear to understand the uttered words or commands. The range of human hearing is generally considered to be 20 Hz to 20 kHz, but it is far more sensitive to sounds between 1 kHz and 4 kHz. For example, listeners can detect sounds as low as 0dB Sound Power Level (SPL) at 3 kHz, but require 40 dB Sound Power Level (SPL) at 100 Hz (an amplitude increase of 100). Listeners can tell that two tones are different if their frequencies differ by more than about 0.3% at 3 kHz. This increases to 3% at 100 Hz [30].

The difference between the loudest and faintest sounds that humans can hear is about 120 dB. Listeners can detect a change in loudness when the signal is altered at about 1 dB (a 12% change in amplitude). In other words, there are only about 120 levels of loudness that can be perceived from the faintest whisper to the loudest thunder.

The relationship between sound intensity and perceived loudness is shown in Table 2.1, expressed as sound intensity on a logarithmic scale, called decibel SPL. On this scale, 0 dB SPL is a sound wave power of 10^{-6} watts/cm², about the weakest sound detectable by the human ear. Normal speech is at about 60 dB SPL, while painful damage to the ear occurs at about 140 dB SPL [30].

Table 1.1 Relationship between sound intensity and perceived loudness.

	Watts/cm ²	Decibels SPL	Sound Example
Louder	10 ⁻²	140 dB	Pain
	10 ⁻³	130 dB	
	10 ⁻⁴	120 dB	Discomfort
	10 ⁻⁵	110 dB	Jack hammers and rock concerts
	10 ⁻⁶	100 dB	
	10 ⁻⁷	90 dB	OSHA limit for industrial noise
	10 ⁻⁸	80 dB	
	10 ⁻⁹	70 dB	
	10 ⁻¹⁰	60 dB	Normal conversation
	10 ⁻¹¹	50 dB	
Softer	10 ⁻¹²	40 dB	Weakest audible at 100 hertz
	10 ⁻¹³	30 dB	
	10 ⁻¹⁴	20 dB	Weakest audible at 10kHz
	10 ⁻¹⁵	10 dB	
	10 ⁻¹⁶	0 dB	Weakest audible at 3 kHz
	10 ⁻¹⁷	-10 dB	
	10 ⁻¹⁸	-20 dB	

2.5.2 Speech Signal Pre-Processing

Speech signal that is captured by microphone, is in the form of continuous data, that is must be transformed into discrete data to prepare it for further processes to obtain the feature vectors. The popular processes transforming speech signal and accommodate to the next stage of feature extraction are shown in figure (2.5). According to Figure (2.5) each block function is described as:

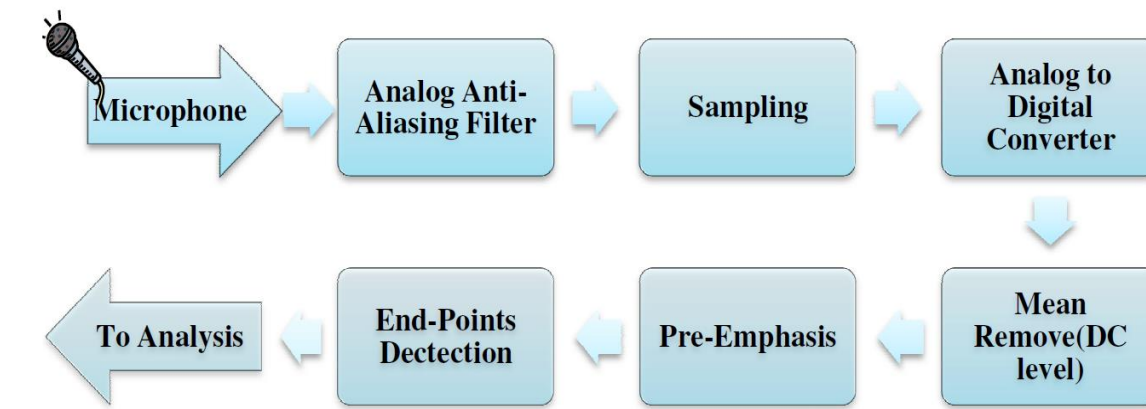


Figure 2. 5 Pre-processing to speech signal

2.5.2.1 Microphone:

Used as transducer to convert speech signal from sound pressure wave to analog data (continuous data).

2.5.2.2 Analog Anti-Aliasing Filter:

The analog signal (output of the microphone) should pass through an anti-aliasing filter which is really a low-pass filter with a cutoff frequency, which is less than the folding frequency (maximum frequency) defined as half of the sampling frequency ($f_s/2$), in order to restrict the bandwidth of a signal to satisfy approximately the sampling theorem. Once the anti-aliasing filter is done, the new analog signal has lost most of its high frequency components [30].

2.5.2.3 Sampling:

Major function of the sampling process, is to convert continuous time signals to discrete-time signals. According to sampling theorem, sampling frequency can't be smaller than twice the bandwidth of the signal.

Higher sampling frequency will lose less but has more data to process while lower sampling frequency (f_s) will lose more but has less sampled data, because the bandwidth of speech is not higher than about 4KHz-16KHz [30].

2.5.2.4 Analog to Digital Converter:

The main purposes of analog to digital (A/D) converter, is to quantize (digital representation of samples) each discrete sample $x(n)$, $n=0,1,...,N-1$ into a specific number. The resolution of the conversion represents a number of bits for each discrete sample, for example, when 16 bits are used to represent each sample. The values of the converted samples could vary from -32768 to 32767 , more bits used, more resolution obtained. [30]

2.5.2.5 Remove Mean (DC Level):

Microphone or A/D converter may introduce DC offset voltage to the input signal that should be removed. This function performs important things when calculating end and start points for the uttered voice. It is essential to remove any constant DC offset level from signal prior to calculating Short Term Zero Crossing Count technique, which is one of the methods to find end points of the speech signal. [30]

2.5.2.6 Pre-Emphasis:

The pre-emphasis is used to compensate the loss that the high speech signal frequencies suffer by effect of the propagation and radiation from the vocal cavity to the microphone. Pre-emphasis is performed by filtering the speech signal with a first order filter as shown in Figure (2.6).

This filter improves the efficiency of the stages used to calculate the speech spectrum, increasing, from the hearing point of view, the sensibility of the frequencies components is larger than 1KHz [30].

The High-Pass Filter is governed by the following equation:

$$S(n) = X(n) - aX(n - 1), \quad 1 \leq n \leq L \quad \dots \dots \dots (2.1)$$

$S(n)$ represents the signal that has been processed with pre-emphasis, while $X(n)$ represents the original signal, and L is the length of each audio frame (sampling number) [30].

The parameter (a) is chosen between 0.95 and 1.0, reflecting the degree of pre-emphasis [30]. A value of (a) is equal to 0.95, gives rise to more than 20dB amplification of the high frequency spectrum [30].

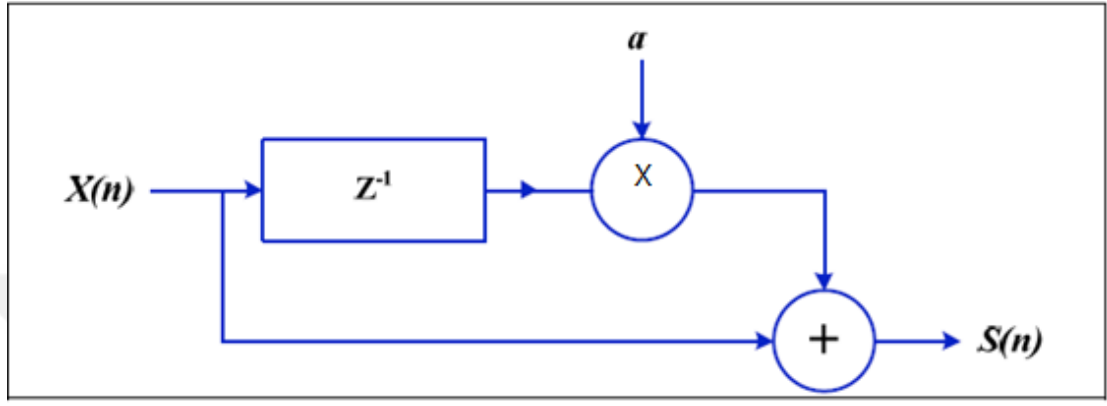


Figure 2. 6 Pre-emphasize Filter Responses [30]

2.6 FEATURES EXTRACTIONS:

Audio features can be mostly categorized into physical and perceptual features depending on their semantic explanation. Physical features define audio signals in terms of physical, statistical, and mathematical properties. Perceptual features deals with the perceived properties such as timbre, rhythm, loudness, and pitch. In most of literatures, physical features are categorized into spectral and temporal feature groups [31]

2.6.1 Temporal Features

The natural depiction domain for audio signals is called as the temporal domain. All features in temporal domain are extracted directly without any previous conversion for the raw audio signal. So, it has low computational complexity compared with that of features in spectral domain. The audio signal temporal features can include:

a- Mean Absolute Value

One of the greatest common features used for feature extraction is the mean absolute value ($\bar{M}$). It is basically the average of the EMG signal data values.

It is calculated as:

$$\bar{M} = \frac{1}{L} \sum_{l=1}^L M_l \quad \dots \dots (2.2)$$

Where L is the input vector length, M_L is the input data values.

b- Standard Deviation

To measure the difference from the mean for each component in EMG data, it should be use the standard deviation (STD) as follow:

$$STD = \sqrt{\frac{1}{L-1} \sum_{l=1}^L (M_l - \bar{M})^2} \quad \dots \dots \dots (2.3)$$

c- Variance

The variance is the mean value of the square of the deviation of EMG data signals. It can be defined by:

$$VAR = \frac{1}{L-1} \sum_{l=1}^L (M_l - \bar{M})^2 \quad \dots \dots \dots (2.4)$$

d- Root Mean Square

The RMS feature is the best degree for the characteristics of EMG signal, because of that the mean of this signal generally close to zero. The RMS relates to standard deviation. It can be defined as the square root of the mean square of EMG signal data. It is calculated by the following:

$$RMS = \sqrt{\frac{1}{L} \sum_{l=1}^L M_l^2} \quad \dots \dots \dots (2.5)$$

e- Minimum and Maximum Values

Minimum and Maximum values can be defined as significant features of an EMG data. They can be considered by measuring the smallest and the largest values of this signal as in follow:

$$P_{max} = \max(P_1, P_2, \dots, P_R) \quad (2.6)$$

$$P_{min} = \min(P_1, P_2, \dots, P_R) \quad (2.7)$$

f- Median Absolute Value

It is considered by obtaining the middle value for a set of absolute values after ascending sorted them. If there are two values are middle, the median is the average of these two values.

2.7 SPEECH RECOGNITION:

Speech Recognition (known as Computer Speech Recognition or Automatic Speech Recognition (ASR)) is the procedure of transforming a speech signal into a sequence of words, by using computer programs that implement the chosen recognition algorithm. [32]

2.7.1 Requirements Selection of Audio Classifier:

Recognition accuracy and computational complexity form the most principles for choosing of classification methods. Additionally, some applications, such as applications that are prone to noise, use robustness to noise as an additional criterion. [31]

Recognition accuracy of a classifier can be affected and improved by a selection of audio feature and a selection of a category of classifier.

Computational complexity is the total computational time that classifier needs to yield the classification outcome. It can also offer perception about the power consumption of classifier. Low computational complexity means faster classifier with lower power consumption.

Robustness to noise is the ability of classifier to disregard any feature differences produced by confusions (noise) such as amplitude scaling or bandwidth variations of an audio signal.

A trade-off between recognition accuracy and computational complexity is commonly required in the choice procedure of audio classifiers.

2.7.2 Relevant issues of ASR design:

Table 2.2 presents the important issues on which recognition accuracy depends.

Table 2.1 Main issues of ASR design [32]

Factor	Influence
Environment	working conditions and signal/noise ratio;
Channel	Distortion, band amplitude and echo
Transducer	Telephone and Microphone
Vocabulary	Specific or Generic: Characteristics of existing training data.
Speech styles	Speed(fast, normal, or slow) Production(speech read or spontaneous speech, isolated words or continuous) Voice tone (shouted, normal, or quiet)
Speakers	Speaker dependence/independence, age and sex

2.8 Artificial Neural Networks

In general, all biological functions and neural applications, such as memory and recognition, are stored in the very composite networks of process nodes (neurons) and interconnections between them (weights). Learning is a procedure of obtaining data from the environment and modifying the current weights to establishing relationships among the training as shown in Figure (2.7).

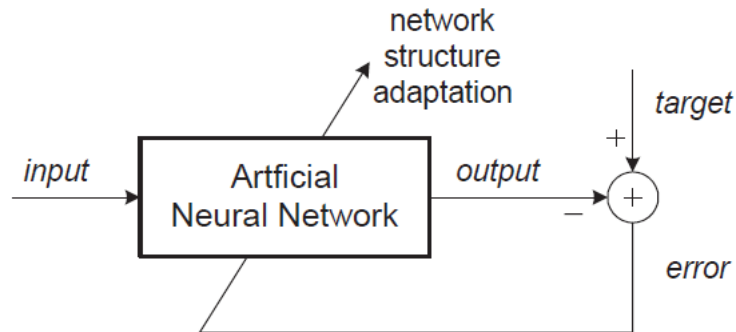


Figure 2.7 The principle of Artificial Neural Network

Neural networks are parallel implementations of nonlinear static or dynamic schemes, and their various processes can be executed concurrently. The purpose of neural network is determined by construction of elements or nodes,

the type of processing performed at neurons, and strengths of their connection.

Neural network uses separate samples, as a set of input-output pairs, and proper training method to periodically change weights of neuron interconnections and to select the suitable number of neurons to perform desired purpose. In regression tasks the output being predicted is a quantitative variable, while in classification problems, the output is a categorical variable. Neural network has the generalization capability to successfully achieve mapping or recognition of input data that have not been existed in the training data set [33].

McCulloch and Pitts in the 1940s begin the basis of neural theory; they showed that neural networks with appropriate number of artificial process nodes can compute some logical functions or arithmetic operations. Then Hebb presented a learning mechanism as mimic that for in biological neurons [33].

Rosenblatt designed the perceptron network and the related learning rule in 1950s as a first practical use of artificial neural networks to solve only a restricted type of problems [33]. Widrow and Hoff after few years presented a new learning algorithm to train adaptive neural structures as in the perceptron. In 1972 Anderson and Kohonen individually established neural networks as associative memories, by discovery the relationship between output and input training vectors. In 1982, associative memory also created by a certain type of recurrent networks as described by Hopfield. Rumelhart and McClelland discovered the backpropagation algorithm for training multilayer perceptron neural networks in 1986. Broomhead and Lowe in 1988 invented Radial basis function networks as an alternative to multilayer perceptron networks. In the 1990s, Vapnik designed support vector machines as one of most commonly used of supervised learning networks. In the last fifteen years many theoretical and practical algorithms in the field of neural networks has been rapidly developing [33].

Neural networks are used in a wide collection of applications, including function approximation, pattern classification, associative memory, data compression, prediction, control, nonlinear system modeling, and optimization. They are applied to a broad range of problems in many fields including chemistry,

banking, aerospace, electronics, finance, automotive, defense, engineering, speech, entertainment, manufacturing, telecommunications, games, medical, transportation robotics, and securities [33].

Commonly, Structure of artificial neural networks can be classified into three types: feedforward, feedback (recurrent) and competitive learning. In the feedforward configuration there is no feedback involved and the outputs are calculated straight from the inputs. Recurrent systems are dynamical configurations and have feedback connections from outputs into inputs. In the competitive learning systems the outputs are calculated based on specific amount of distance between the actual outputs and inputs. Feedforward networks are used for function approximation and pattern recognition; recurrent networks are used for optimization problems and as associative memories where competitive learning networks are used for pattern classification [33].

Learning with categorical outputs is called pattern classification or recognition. Categories are representative values assigned to the patterns and connect the pattern to the definite result which that pattern symbolizes. Only a predetermined number with possible values can be chosen for the categorical variables. The decision region for a given category is the union of all regions where it is predicted. Estimation of decision areas between the categories that are not perfectly divisible in the pattern space is the main difficult in pattern classification. There are two general groups of learning methods in neural network for pattern recognition: supervised and unsupervised learning. Through supervised technique set of inputs and wanted outputs (targets) is presented together, and each corresponding output is associated with an input. The network adjusts its weights to make matching between each input and the related desired outputs. This changing of weights continues over entire training set in order to minimize the error between the actual and desired outputs of the network. However, the desired outputs are not taken into account or not known during unsupervised learning methods, and network trains to detect inner structure or similarity of the input patterns, by modifying the network weights until analogous inputs start to yield similar outputs [33].

Generally the Multilayer Perceptron (MLP) is the greatest used structure of feedforward neural networks. It contains several layers of parallel nodes, typically input layer followed by one or more hidden layers and the last layer for output with linear/nonlinear transfer function neurons, as shown in Figure (2.8). For learning of composite nonlinear associations between input and output training data, the network needs to use nonlinear transfer functions in their hidden and output neurons. [33]

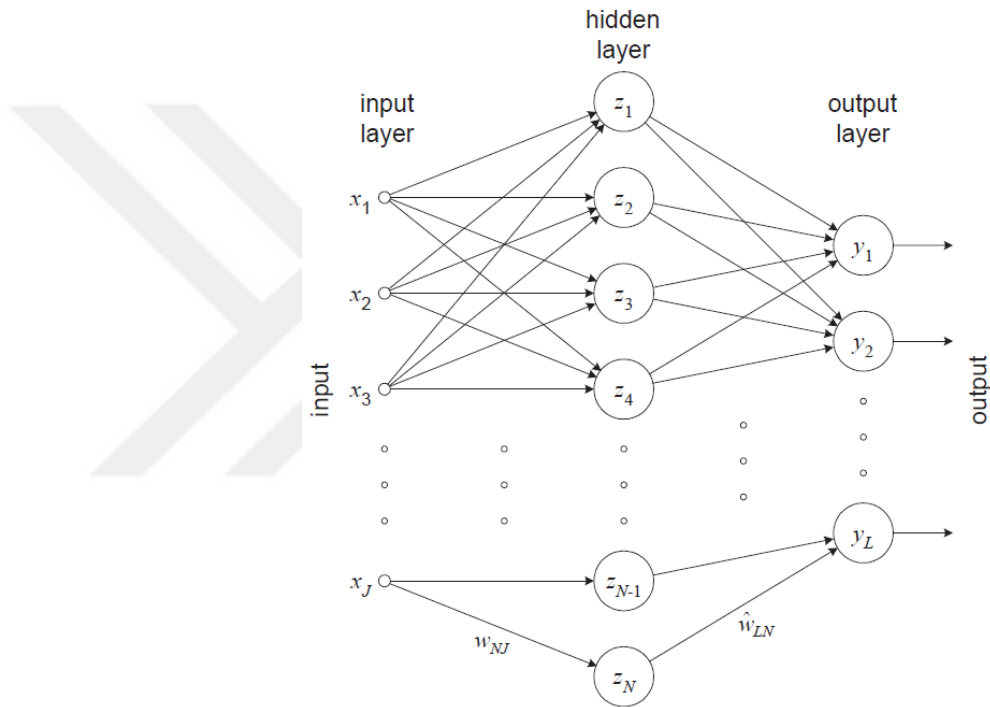


Figure 2.8 The construction of Multi-layer Perceptron Neural Network

CHAPTER 3

DESIGN OF PROPOSAL SYSTEM FOR SPEECH TO SIGN LANGUAGE TRANSLATOR

3.1. INTRODUCTION

This chapter presents the proposed conversion system that can translate speech in Arabic Language into an ArSL, for better communication of the deaf people with the normal community. The proposed work includes Data Collections, Preprocessing (Removing Silence, Normalizing, and Pre-Accentuation), and Feature Extraction.

3.2. PROPOSED WORK FOR SPEECH TO SIGN TRANSLATOR

To accomplish the goal of the work, two phases are required including offline phase for training and feature extraction, and online phase for the recognition process, as shown in figure (3.1). The proposed work is explained in figure (3.2), The sequence of columns are the spoken letters, which are captured by a microphone, then they are translated to signals by Matlab, after that the preprocessing, the features extraction and recognition process are done on the obtained signals, then using the display unit to give the necessary movements to represent the sign language.

3.2.1 Data Collections

It was a challenging step in this work for data generation and collection, because the Arabic letters sound very rare on the Internet, and also the works and research about spoken the Arabic letters are very insufficient.

The recording process is carried out by using a microphone. As the first step, the data set used in this work is collected; thirty samples are recorded from every person, and each subtending to the all 28 Alphabet sets (i.e. the main Arabic Letters).

For each of the recordings (data set), the speaker pronounces one letter each time. In this context, every sample is separated enough and thus does not interpose with each other. All samples are at one second sound symbols by using a MATLAB script. A total 840 sound samples are collected for the alphabets.

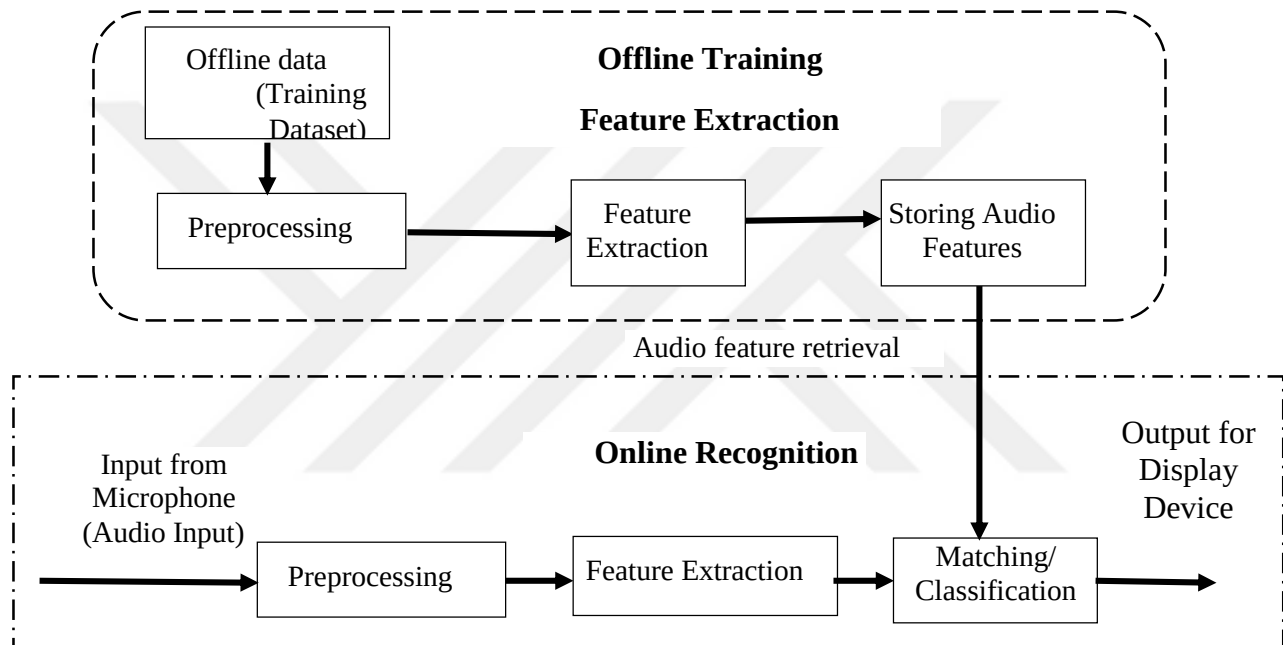


Figure 3.1 Block diagram of the Suggested Work for Talking to Sign Translator

Data Collections	Preprocessing	Features Extraction	Recognition	Decoder	Display Device
(٧) Audio Inputs 	1. Removing Silence 2. Normalizing 3. Pre-emphasis	1- RMS 2- STD 3-MAX & MIN 4- Energy Entropy 5- Median 6- SNR 7-THD 8- TOI	Artificial Neural Network	Generate the proper Signals to Recognized Letter	Image of Letter or Hand Gesture 

Figure 3.2 Detaile of Speech to Sign Translator Proposed Work

In effect the recording process, using the microphone of figure (3.3a), it gives good output signals, but with all surrounding effects (Noise) and the affect that the distance between the speaker and microphone not constant, and this will result different outputs that make the recognition process more difficult. Figure (3.3b) shows the recorded signals.

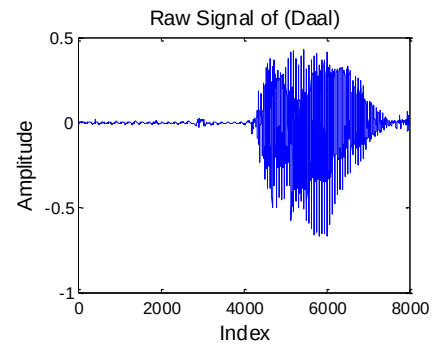
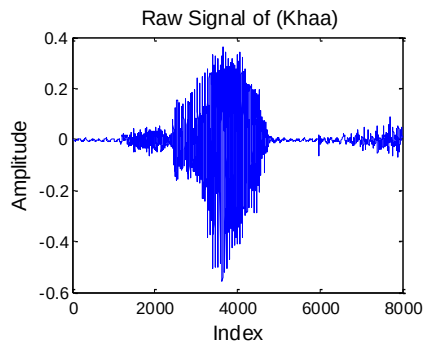
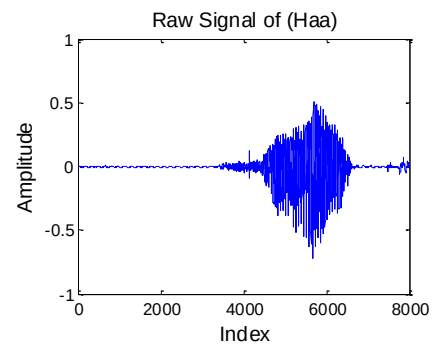
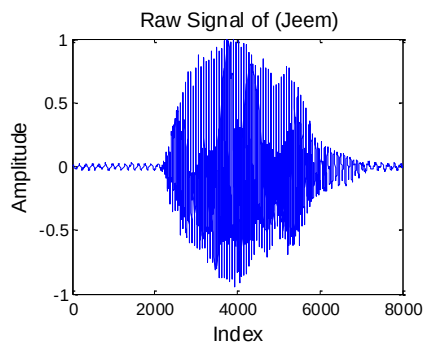
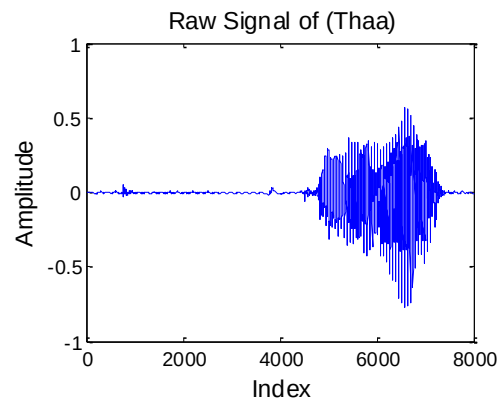
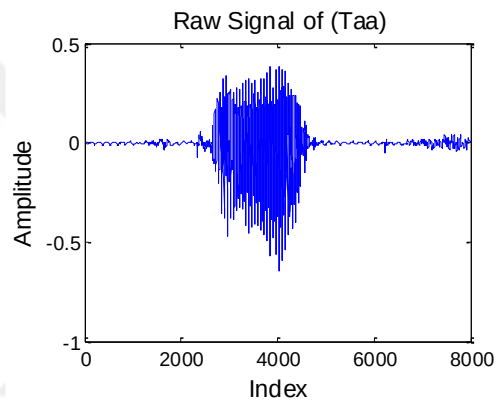
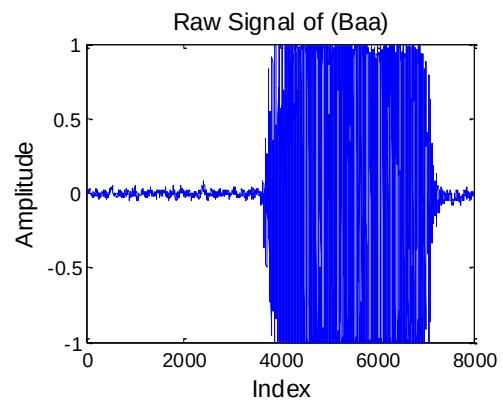
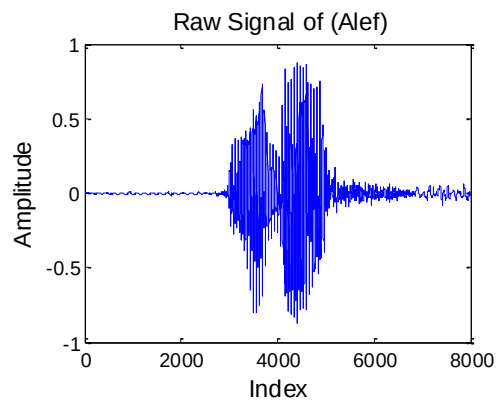


Figure 3.3 Microphone

used for Data Recording

In this work, the experimental data consists of a real voice dataset of Arabic letters which recorded by microphone. The dataset includes 840 voice records with the structure shown in Table 1. Twenty eight variables were selected for input data collection. Twenty nine variables were selected for input-output data classification. The first 28 variables represent the independent inputs whereas the last element is the dependent output.

The mainly goal of classification methods is to determine the all classes in which input data is to be categorized and to allocate each class label. According to the collected data from and the aim of this work, 28 classes were identified from “1” to “28” as shown in Table 2.



Original Signals recorded

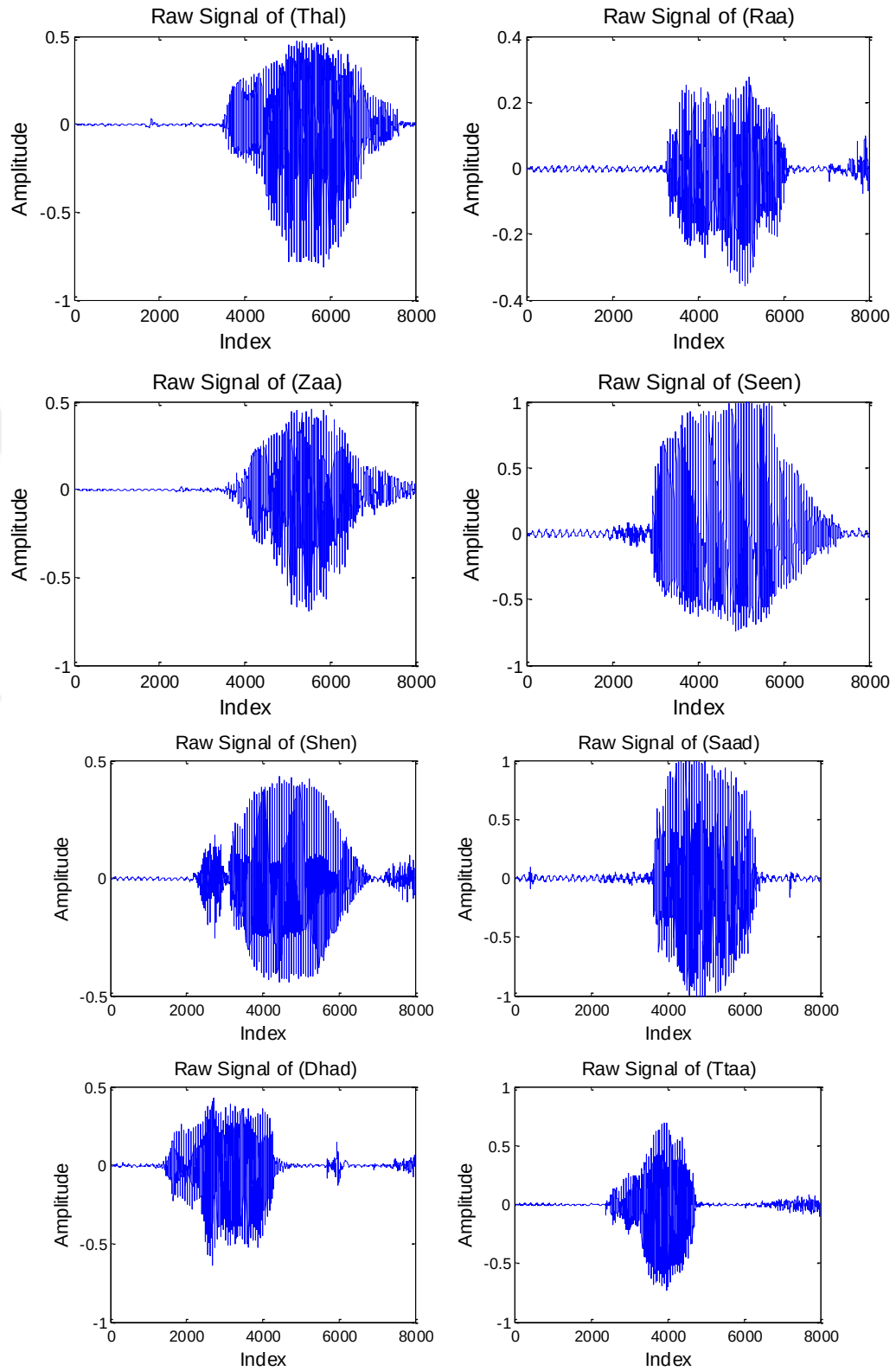


Figure (3.3b) Original Signals recorded (continued)

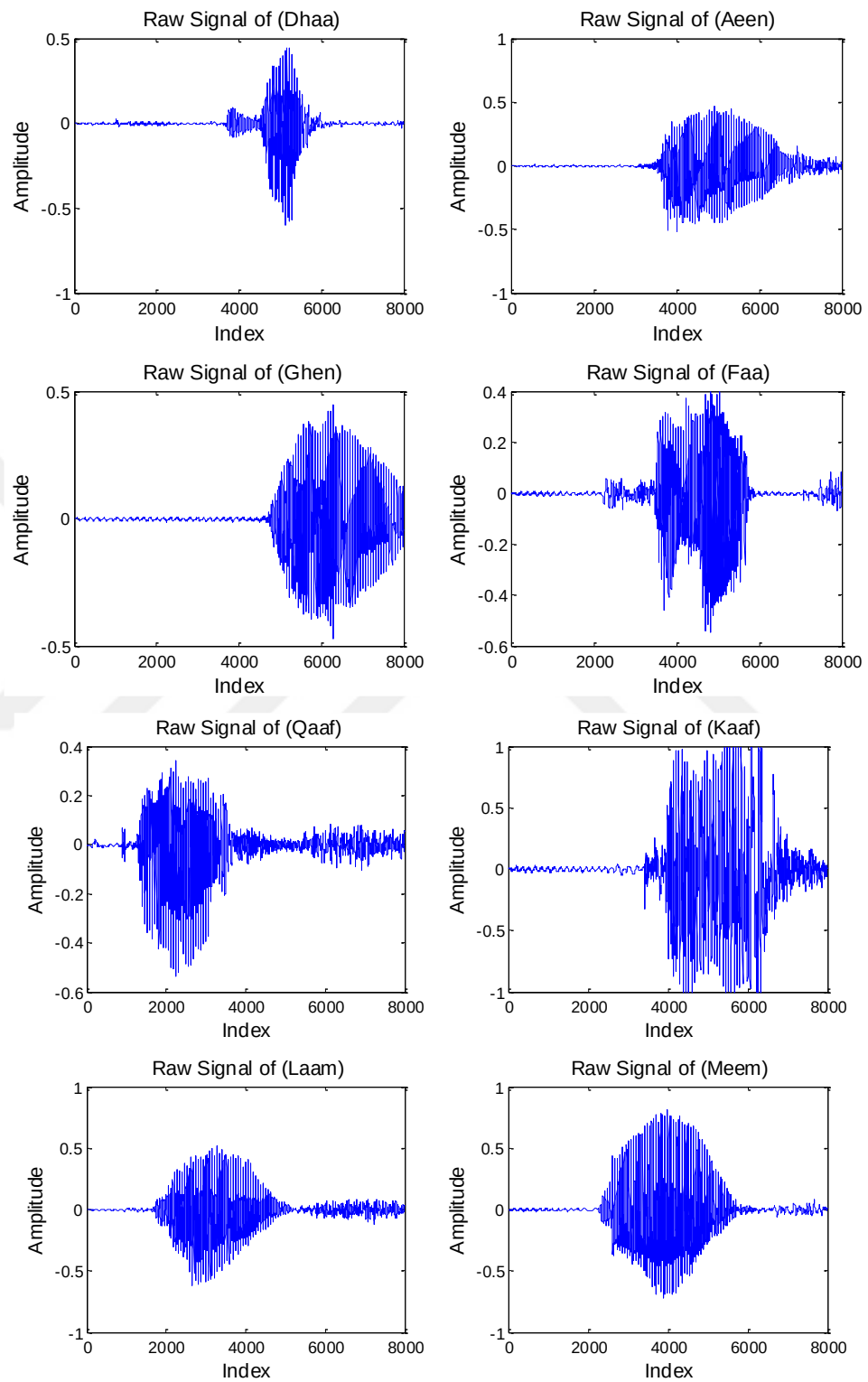


Figure (3.3b) Original Signals recorded (continued)

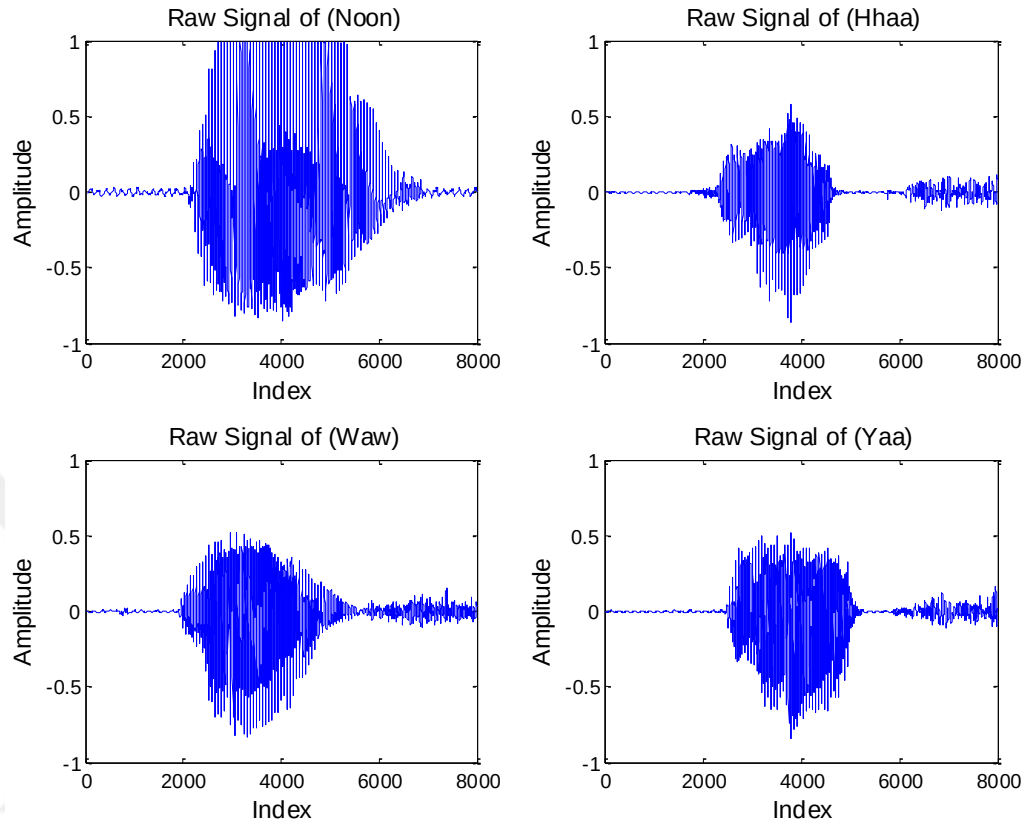


Figure (3.3b) Original Signals recorded (continued)

Table 3.1 Structure of dataset elements

Variable	Description	No of Data
x_Alef	x describes the input parameter and Alef denotes the Arabic letter 'أ'	30
x_Baa	x describes the input parameter and Baa denotes the Arabic letter 'ب'	30
⋮	⋮	
x_Yaa	x describes the input parameter and Yaa denotes the Arabic letter 'ي'	30

Table 3.2 Structure of class output

Output	1	2	3	4	5	6	7
Class Region	[0.5-1.49]	[1.5-2.49]	[2.5-3.49]	[3.5-4.49]	[4.5-5.49]	[5.5-6.49]	[6.5-7.49]
Class Description	أ	ب	ت	ث	ج	ح	خ
Output	8	9	10	11	12	13	14
Class Region	[7.5-8.49]	[8.5-9.49]	[9.5-10.49]	[10.5-11.49]	[11.5-12.49]	[12.5-13.49]	[13.5-14.49]
Class Description	د	ذ	ر	ز	س	ش	ص
Output	15	16	17	18	19	20	21
Class Region	[14.5-15.49]	[15.5-16.49]	[16.5-17.49]	[17.5-18.49]	[18.5-19.49]	[19.5-20.49]	[20.5-21.49]
Class Description	ض	ط	ظ	ع	غ	ف	ق
Output	22	23	24	25	26	27	28
Class Region	[21.5-22.49]	[22.5-23.49]	[23.5-24.49]	[24.5-25.49]	[25.5-26.49]	[26.5-27.49]	[27.5-28.49]
Class Description	ك	ل	م	ن	هـ	و	ي

3.2.2 Preprocessing

Preprocessing diminishes the measure of handling required in later stages, as appeared in the figure (3.4).

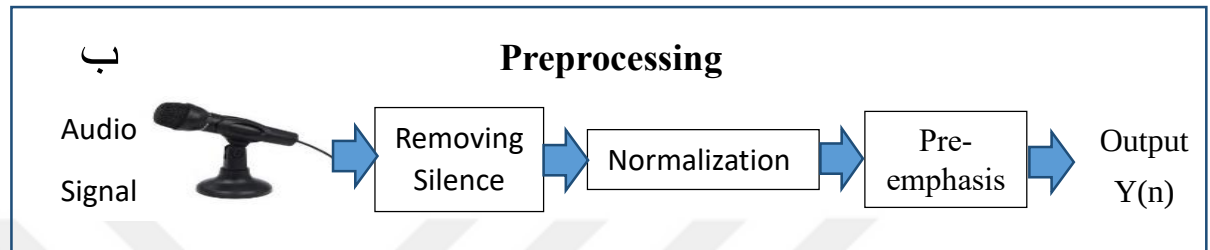


Figure 3. 4 Preprocessing system

3.2.2.1 Removing Silence

Amid the generation of non-speech, there is no speech yield because of there is no excitation delivered to the vocal band. Non-speech is distinguished by the non appearance of any signal qualities, most reduced vitality, "Energy" related to the speech waveform when contrasted with unvoiced and voiced speech. There are some number of zero intersections in the speech waveform when contrasted with unvoiced speech and no connection among progressive examples in the speech waveform.

The program expels the quietness interims from the information speech in view of an envelope edge. The information signal is up-sampled, divided to evacuate samples that fall beneath an edge, and afterward re-sampled back to the first inspecting rate, and sifted "filtered" to smooth through the discontinuities where stops in the dynamic speech happened. The utilized edge is a scaled capacity of the middle of the envelope. The default limit is one-fourth of the middle of the envelope, as shown in figures (3.5), where signals after remove silent for six different letters are displayed.

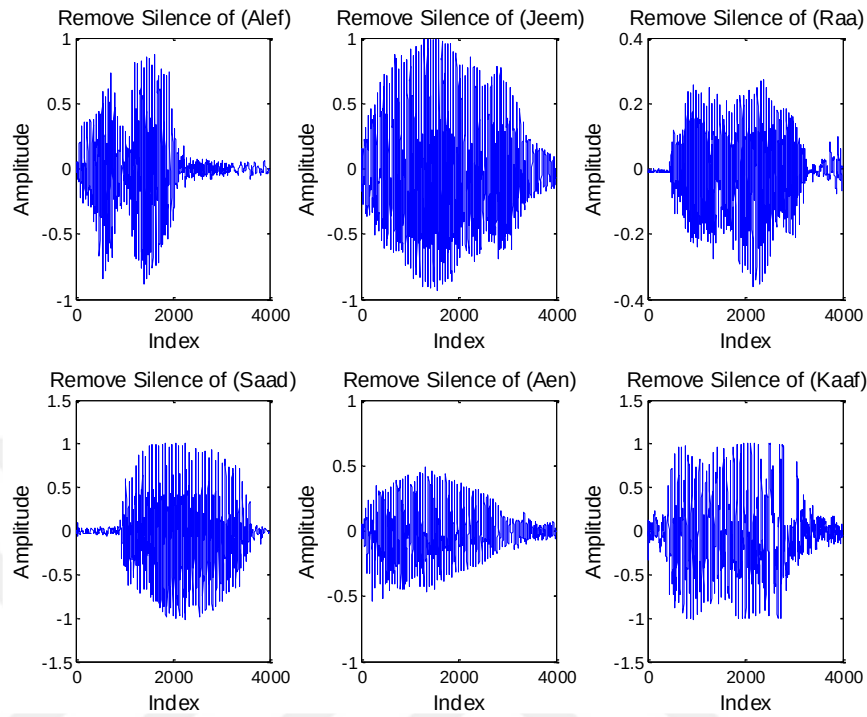


Figure 3. 5 Figure (3.5) Signal after removing silent

3.2.2.2 Normalizing

The extricated sampling signal is normalized to maintain a strategic distance from adequacy variety "amplitude variation" within catching. This standardization step makes the system to be strong to noise varieties. As compared with figure (3.3b), the normalized signal show less amplitude variation, with the normalization process as shown in figure (3.6).

3.2.2.3 Pre-Accentuation

The digitized talk signal $Y(k)$ is given to a "Finite Impulse Response (FIR)" Filter:

$$Y(k) = x(k) - \alpha x(k - 1) \quad \text{for } 0.9 \leq \alpha \leq 1.0 \dots \dots \dots (4.1)$$

Here, k is a number of tests, " $x(k)$ " represent a data talk banner 'signal', and $Y(k)$ is the yield prior-underlined banner, and " α " is an adaptable operator. The target of prior-emphasis is to compensate the high-repeat part that was covered in the midst of the sound age segment of individuals.

Moreover, pre- emphasis improves sound detection, and that are shown in figure (3.7). With the preprocessing stage results, where the salience is removed, and normalization, they give a further strengthened confidence in the statistical features as the classification tools.

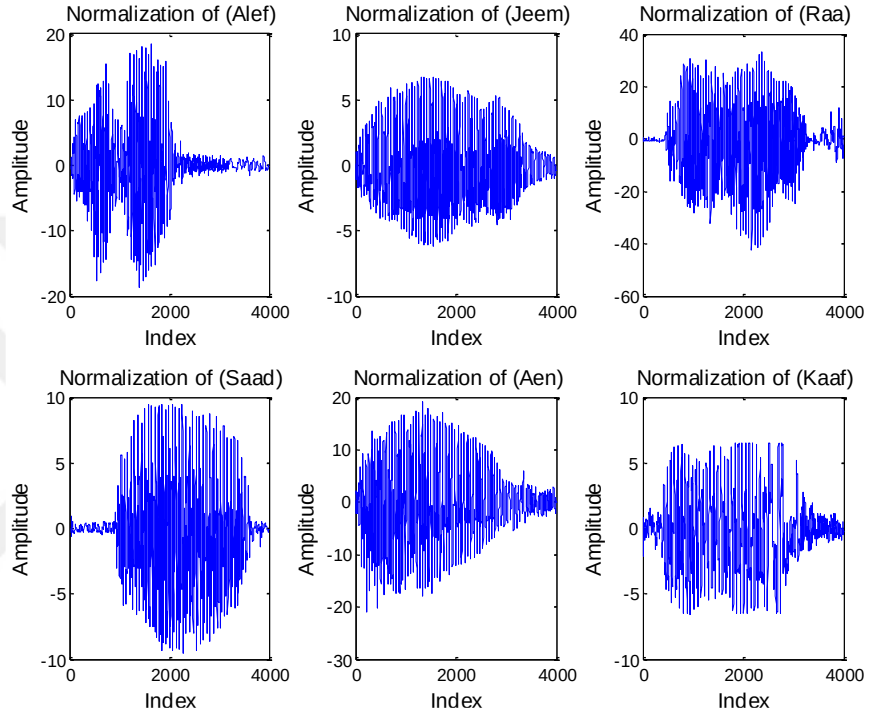


Figure 3. 6 Signal after normalization

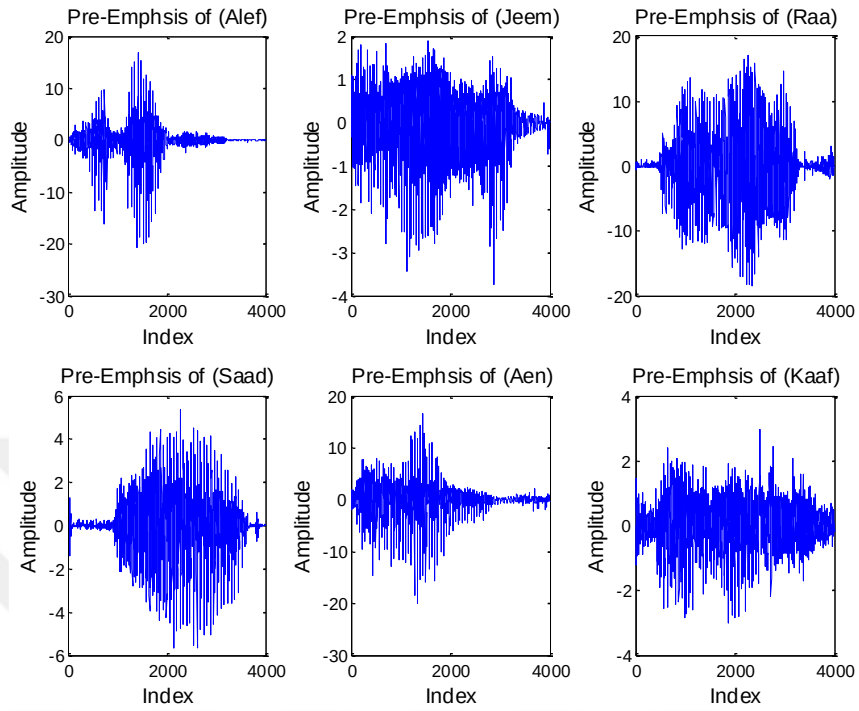


Figure 3.7 Signal after pre-emphasis

3.2.3 Feature Extraction

With second voice specimens, one still felt that the information was not in an acceptable formula. A statistical properties were applied, which represent the suitable features for sound signal. Many of the statistical features have been extracted in this work, such as, Root Mean Square (RMS), Mode, Median, Mean, Maximum (MAX), Minimum (MIN), Energy Entropy (EE), Variance (VAR), Standard Division (STD), Spectral Energy (SENG), Spurious Free Dynamic Range (SFDR), Total Harmonic Distortion (THD), Signal to Noise and Distortion Ratio (SINAD), Signal to Noise Ratio (SNR), and Third-Order Intercept (TOI) point, as shown in figure (3.8). After preprocessing steps, the statistical features have been extracted. The extracted features represents the core of the signal and will reduces the required memory size and then reduces the processing time. It is found that there are some similarity between some features results, and there are specific features effect more than others. Thus, as an alternative process, seven features are selected for training the network named: RMS, MAX, SENG, SFDR, SINAD, SNR, and THD.

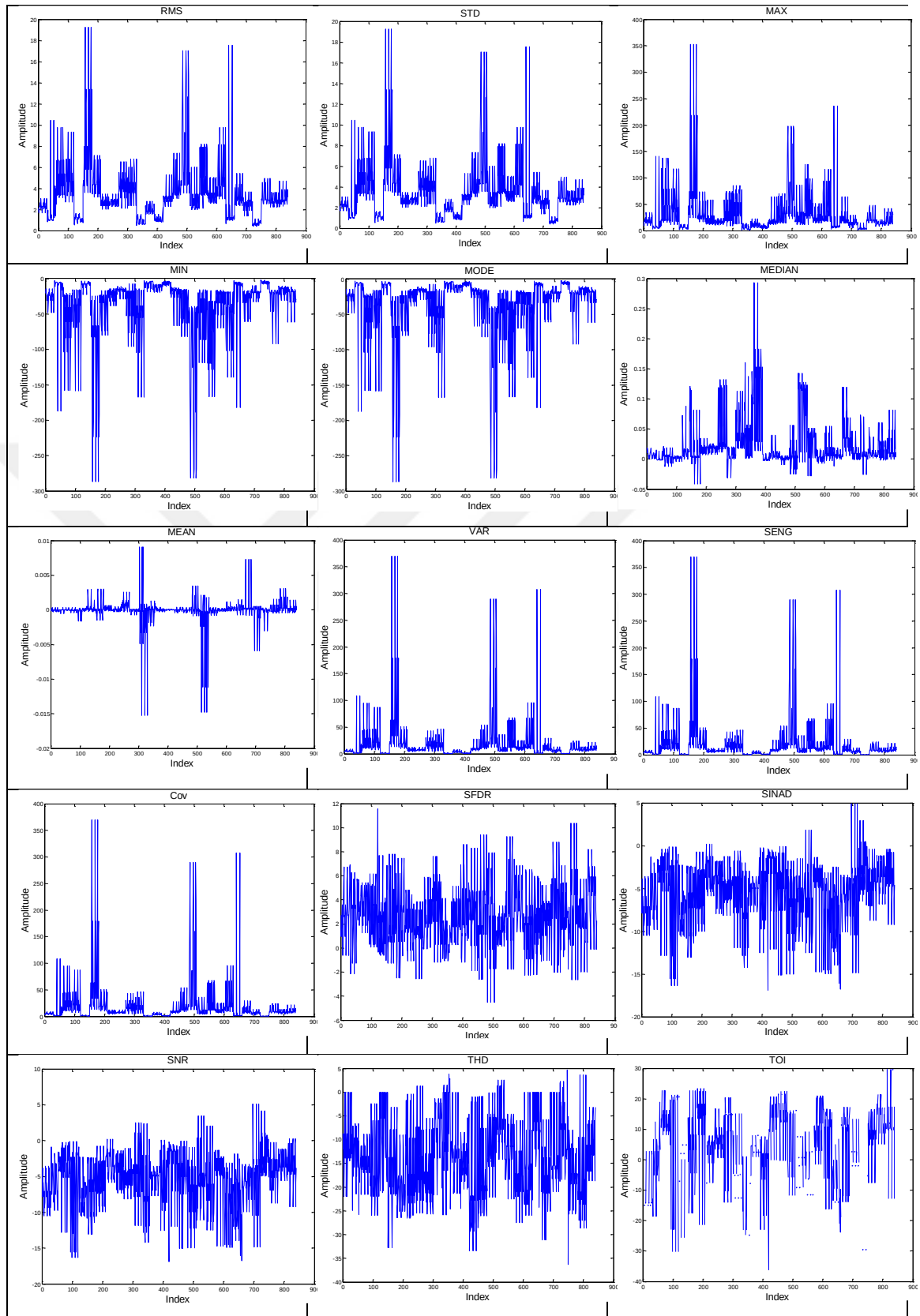


Figure 3.8 Features extracted from signals

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 INTRODUCTION

Both training and testing results will be presented in this chapter. Discussion of these results also will be introduced.

4.2. RECOGNITIONS

In this work, speech recognition setup being identified with appointed letters, mainly 28 categories, in our work; 28 letters have been taken, with 30 patterns for each letter. Artificial Neural Networks have been carried out for classification.

The modeling of the Neural Network (NN) is being decided on the quantity of "input, hidden, and output" neurons. The output neurons are just the quantity of classes that the system should work with, which for this situation will be 28, a class for each letter to be perceived. The information neuron amount is equivalent to the number that is pulled from every individual letter.

The recognition process has done on many cases,

- The speech letter, as it without feature extraction,
- The speech letter with feature extraction, also with three cases:
 - ALL features extracted
 - Some of effected features extracted

To do the recognition process, the Pattern Recognition Neural Network (PRNN) method was used.

4.3 DESIGN OF PRNN

While the classification step has been done by multilayer feed-forward neural network that have 7 distribution nodes in the first layer, 20 processing neurons in the middle layer and that needed because the complex nature of speech and finally, the output layer has 28 neurons as we have 28 classes (28 letter to be recognized), as shown in figure (4.1), with the 30 patterns for each letter (class).

4.3.1 Training Process of the Proposed System

In this work, Pattern Recognition Neural Network (PRNN) was used. It is a feedforward network that has an ability of classifying input samples into desired classes. In this pattern recognition network, the desired data should contain vectors of a 1 value in element k and all zero values in other elements, where k is the class that the network are try to characterize.

The proposed PRNN predictor is executed by Matlab2018b. This system based on neural learning requires the following steps:

1. Divide the collected data into 70% for training process, and 30% for testing and validation the model.
2. Generate N-H-M PRNN model, where N is the features extracted number from input letters, H represents hidden neurons, and M is the number of output classes (letters).
3. Train the designed PRNN by gradient decent method.
4. Validate and test the trained model to predict a new recognized letter.

The proposed PRNN is generated and trained with the training dataset as shown in figure (4.1). The training iterations continued until the RMSE is reduced to the predetermined value (set as 10^{-15}).



Figure 4.1 Seven inputs-Twenty eight outputs PRNN for two training iterations

4.4 RESULTS

The work results of the suggested system were divided into training and testing results.

4.4.1 Training Results

Numerous trials have been conducted to estimate the projected PRNN to produce the satisfactory results. The network was trained with back-propagation algorithm by 560 training samples, 20 samples for each letter of the entire 28 main Arabic letters. Then, the trained model was validated and tested with 140 and 140 dataset respectively, to see how fit the PRNN model predicts the matching data set of output labels.

The best modified PRNN predictor is selected based on the lowest RMSE achieved as shown in figure (4.2) for two training iterations. The Diagnosis Rate (DR) was calculated as:

$$DR = (M/T) * 100$$

Where M is the sum of matching (predicted) output result with the desired class, and T is the sum of test samples data.

The letter class is represented by the actual output from PRNN predictor.

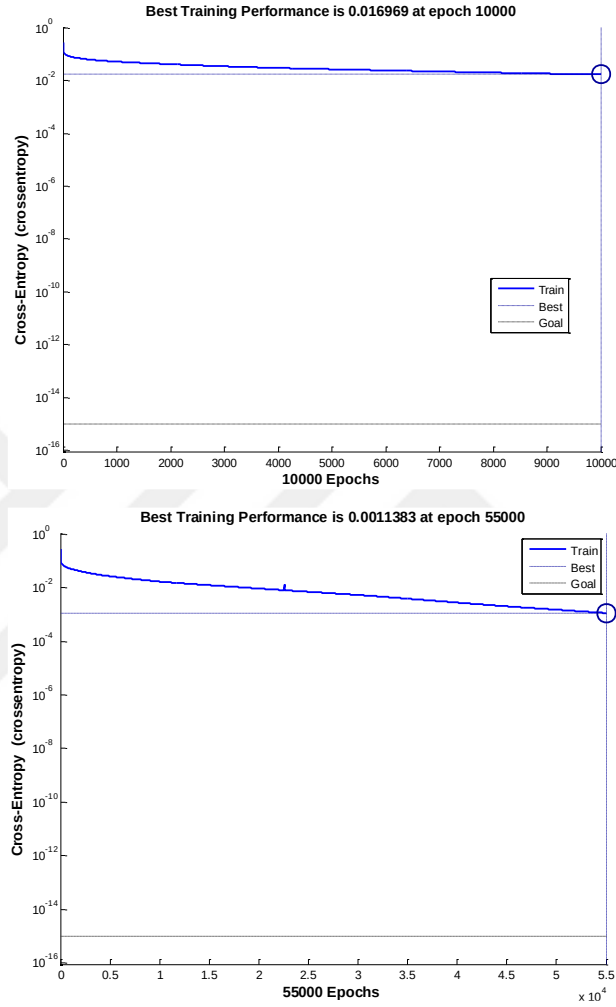


Figure 4.2 Training Performance for two training iterations

Figure (4.3) shows the actual predicted and the desired outputs at the training process, where the error between them is shown in figure (4.4). Then the actual output is rounded into the nearest integer value to show the closeness of it into the desired output class. The results are shown in Figures (4.5) and (4.6) for these outputs and the error between them respectively. RMSEs are 0 in training process for both actual and rounded output values. The network is successfully trained for all 28 classes (letters).

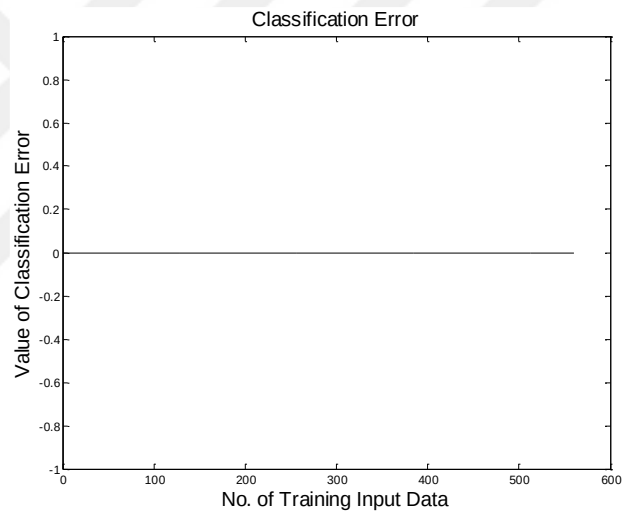
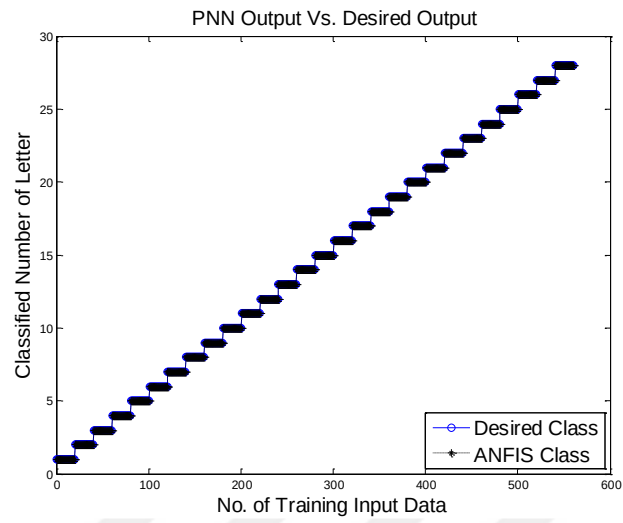


Figure 4.3 Training actual output

Figure 4.4 Error between training outputs

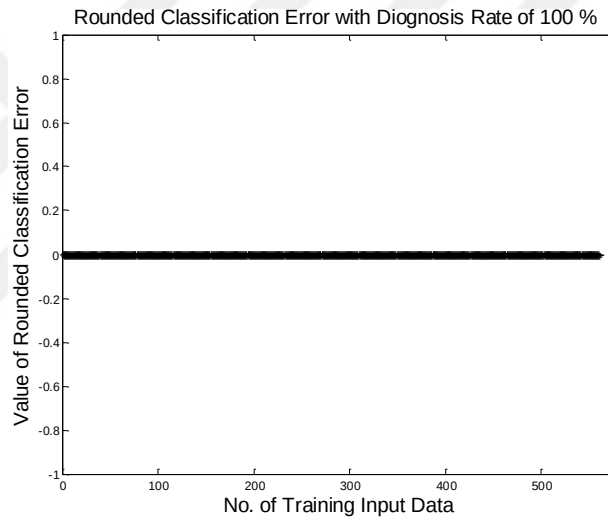
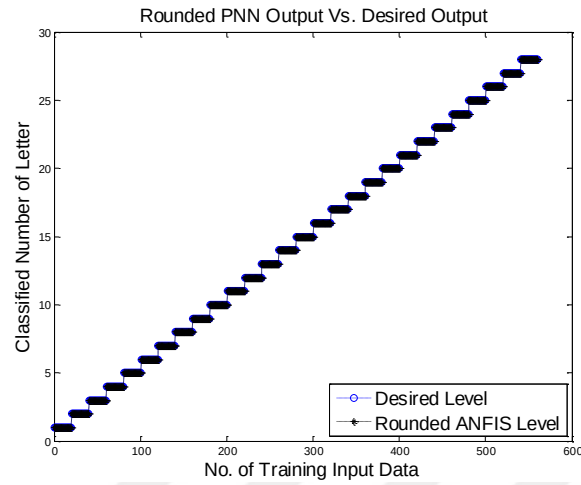


Figure 4.5 Rounded training outputs

Figure 4.6 Training error for rounded output

4.4.2 Testing Results

Figures (4.7-4.10) show the result for testing process with 280 unknown samples. The recognition was achieved with an excellent diagnosis rate of up to 98% with RMSE value of 0.0214. Figure (4.11) represents the confusion matrix for the test process. The trained PRNN correctly predicts all 10 test samples for all classes except classes (2, 5, 12, 14, 22, and 25) where it is capable of predicting 9 out of 10 samples; this provides an accuracy of 100% and 90% respectively. This yields 98% average accuracy for the proposed PRNN, which is an excellent result. It is also clear as shown in figure (4.12)

which represents the Receiver Operating Characteristic (ROC) curve analysis.

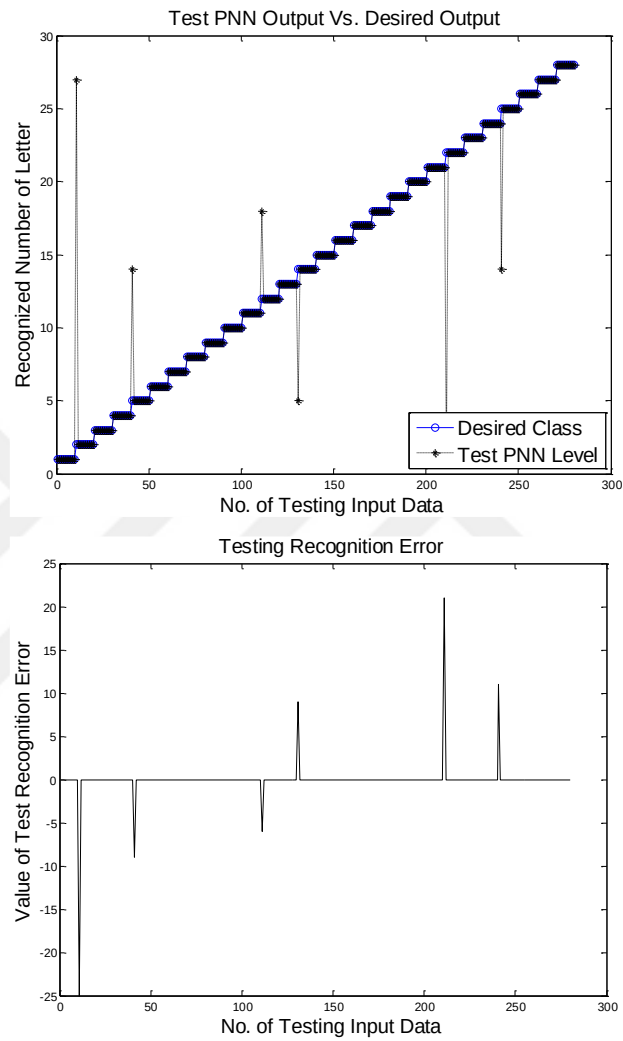


Figure 4.7 Testing actual output

Figure 4.8 Error between testing outputs

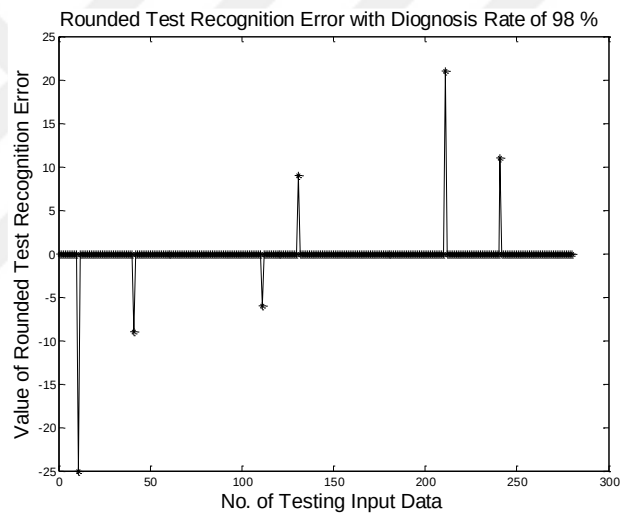
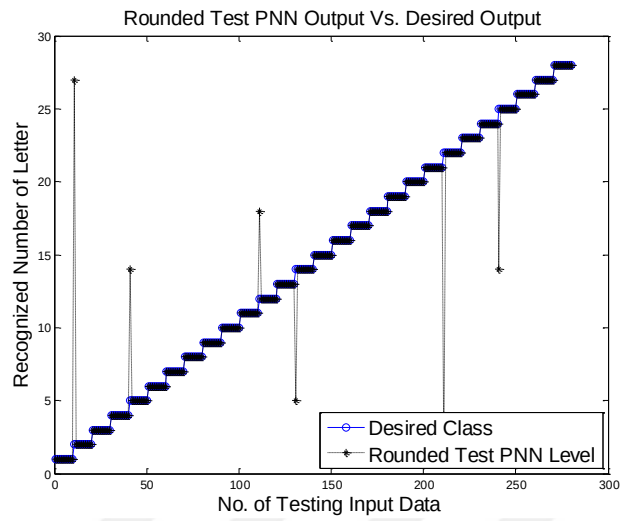


Figure 4.9 Rounded testing outputs

Figure 4. 10 Testing error for rounded outputs

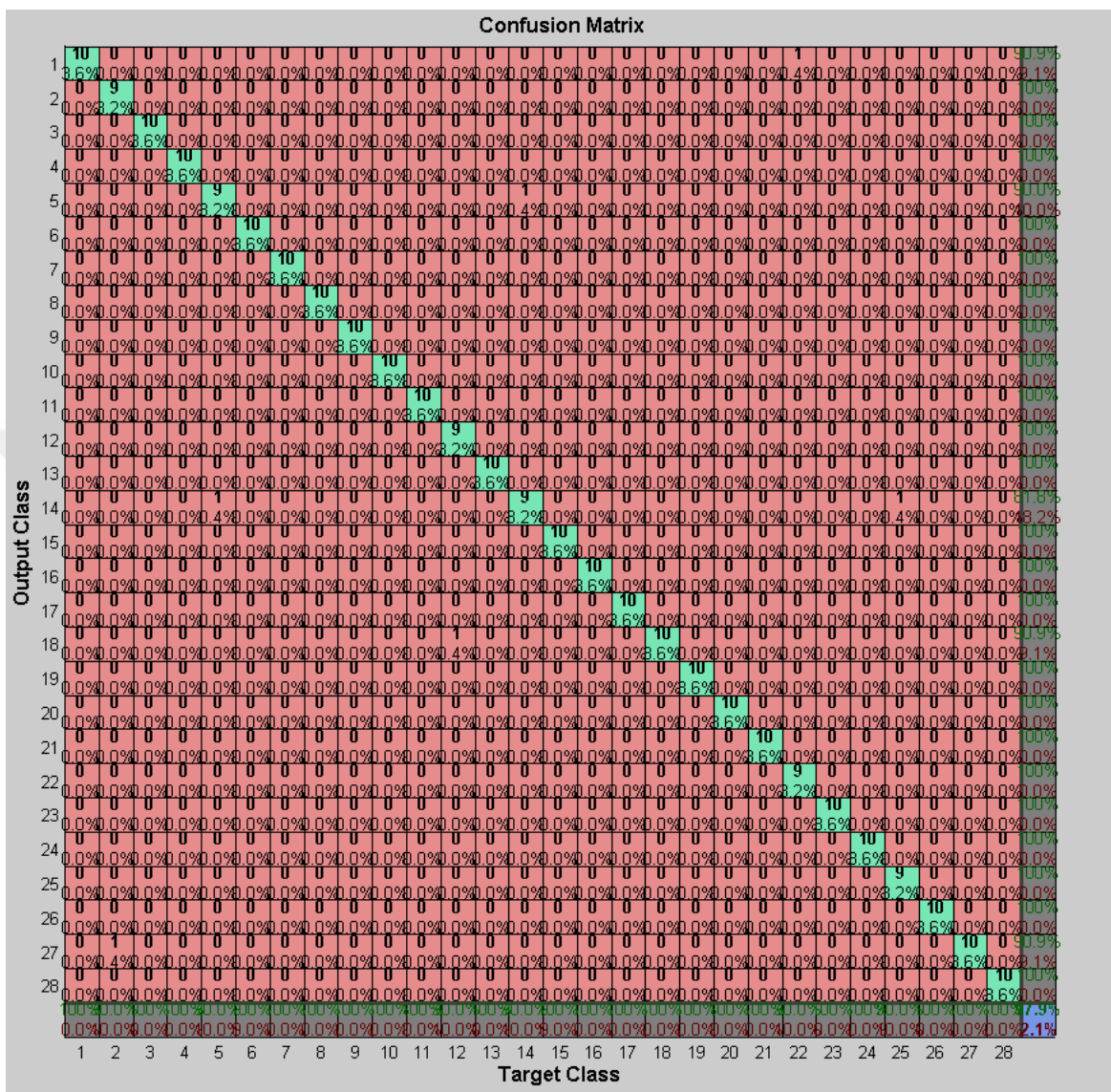


Figure 4.11 Confusion matrix result

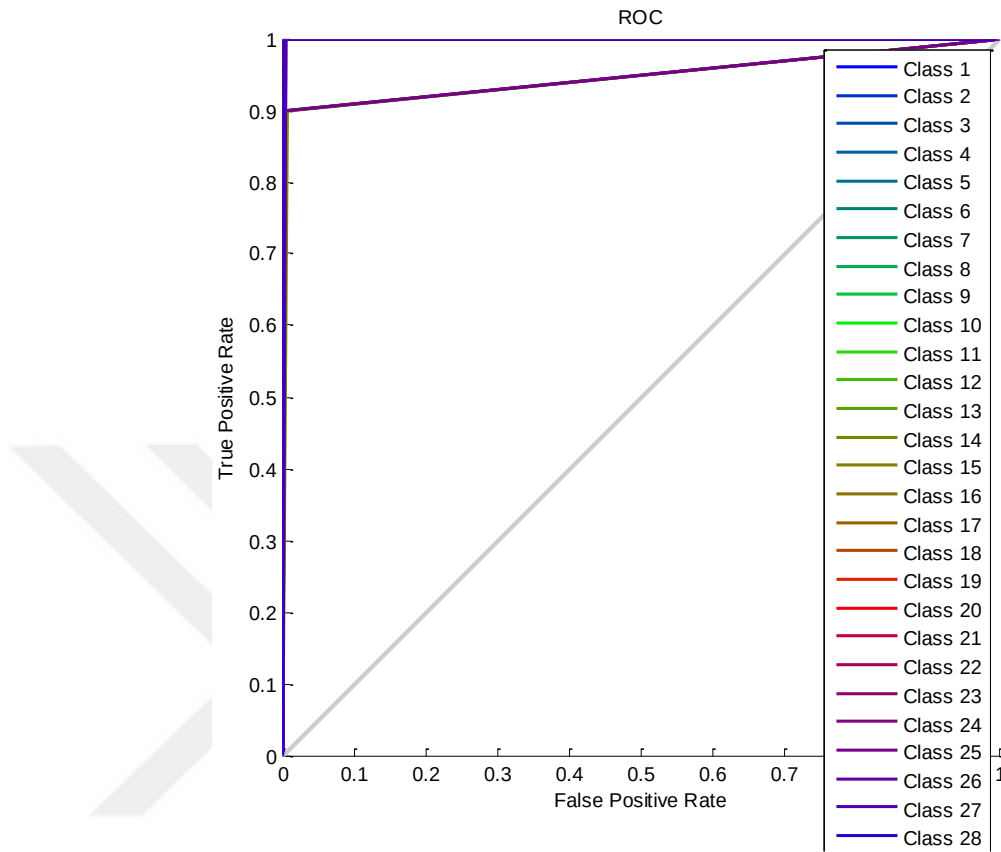


Figure 4.12 Curve of Receiver Operating Characteristic (ROC).

For analyzing the performance of the proposed system as a classification model, the Confusion matrix (also called Error matrix) is also displayed as the predictable number of true negatives, true positives, false negatives, and false positives as shown in Table (4.1), with associated specificity and sensitivity values,

Table 4 Confusion matrix estimated values

		Actual Class	
		Negative	Positive
Predicted Class	Negative	True Negative (TN)= 7554	False Positive (FP)= 6
	Positive	False Negative (FN)= 6	True Positive (TP)= 274

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{274}{274 + 6} = 0.9785$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{7554}{7554 + 6} = 0.9992$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{274 + 7554}{274 + 7554 + 6 + 6} = 0.9985$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{274}{274 + 6} = 0.9785$$

Finally the proposed system is experimentally tested with individually input voice for specific letter and display the recognition result as shown in figure (4.13).

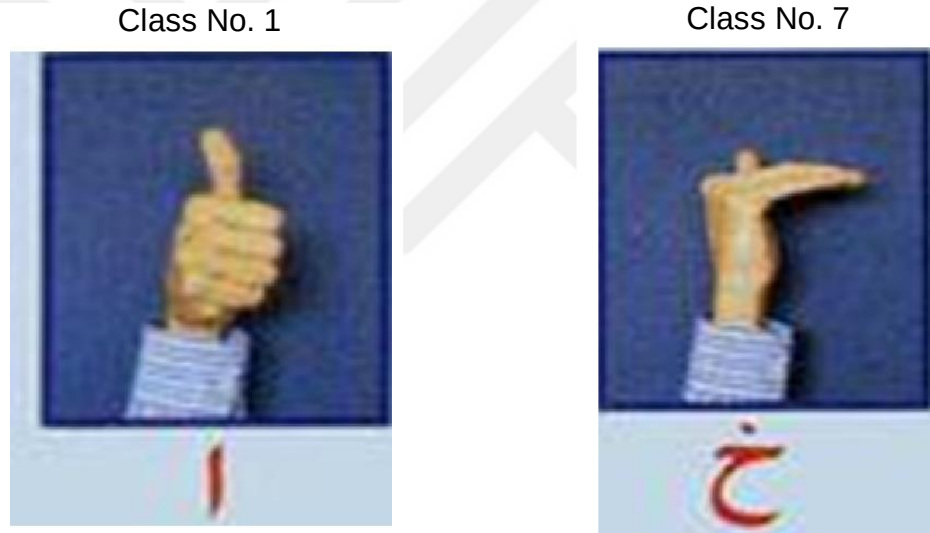


Figure 4.13 Test Result for two different input letters

4.5 DISCUSSION

This work presented the potential of using PRNN to recognize the voice of Arabic Letters mentioned as a class with the aim of help teach the deaf and dumb. The presented PRNN was realized with an accuracy of 98% and RMSE of 0.0214.

True Positive Rate, which is called as Sensitivity and Recall, is calculated up to 98%.

This means a very high ratio of actual positives that are properly predicted as positives. Specificity, called as True Negative Rate, is found to 99% which reflects an excellent amount of actual negatives that are properly predicted as negatives. The proposed ArSL system is work with an accuracy of 99% and precision of 98%.

Author believes that these results are excellent and sufficient to be used as a predictor of Arabic letters classification. This will enable the deaf and dumb organization to provide improved educational services and to support those involved in it.



CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

The main objective of this thesis is to suggest a PNN based ASR system for the deaf people to enhance their communication with normal people using Arabic sign language. The proposed ArSL transformer use PNN classifier to yield the Arabic hand sign based letters as an output after recognizing the audio of the Arabic letters. The proposed system will offer a tailored solution for people with a hearing impairment.

As known, changing number of hidden neuron can affects the classification results. A trial and error task is used to find the appropriate number of neuron. If suitable number of hidden nodes is not carefully chosen, the designed system can leads to incorrect results because of overfitting, which means the PNN based ArSL classifier may give well result at training process however it gives incorrect results at test process. It seems from the experimental results that the finest number of hidden neurons can be achieved by select it with less than or equal to the number of output classes at one hidden layer.

RMSE value is used to authenticate the designed classifier. If both training and testing RMSE values are accepted, the designed neural system can be used, else it cannot confirm.

The trained PNN classifier exhibited almost zero errors in recognizing most of the phonemic letters and a maximum recognition of the remaining letters. These unrecognized remaining characters weren't the same every time they trained. This means that the proposed neural classifier can be advance trained to recognize successfully all the training samples.

The effective Arabic letters speech recognition methodology established in this thesis can be employed in various other applications such as speech to text conversion. This system can also be used for Human-Computer Interaction (HCI) efficiently in Arabic speech recognition with accuracy higher than 98%.



5.2 FUTURE WORK

For additional increasing of the accuracy and quality of the suggested system, more advanced speech recognizing devices can be used such as Hidden Markov Models (HMMs) and also considering increasing the size of the training dataset in upcoming work. As a future work, the analysis and recognition procedures can be modified as a process to associate the distinguished letters to organize commonly used Arabic words and as well to extend data to recognize the Arabic numbers to be used for some voice based security and control applications.

The proposed PNN based ASR presents an efficient tool to classify individual phonemes (letters) as short-time units; however for more advanced classifying of isolated words or for continuous recognition tasks the PNN might suffer from their limited capability to model temporal dependencies. So it can be used as a feature transformation unit, pre-processing step, or as a dimensionality reduction net, prior to effective traditional technique based recognition such as Hidden Markov Models (HMMs).

Also that may investigate using another type of NN such as Long Short-Term Memory (LSTM) or Time Delay Neural Networks (TDNN's) and related Recurrent Neural Networks (RNNs) which have confirmed enhanced performance in ASR applications. Also using Deep Feedforward Neural Networks (DFNN) can be another solution which has several hidden process layers of neurons in the architecture of network. DFNN can constitute non-linear complex relationships and offer an enormous learning ability which give it the possibility of recognize composite speech patterns data.

REFERENCES

- [1] D. Bragg and J. Prasana, “Sign Language Recognition, Generation, and Translation :An Interdisciplinary Perspective”, *Int. Conf. Electr. Eng*, 2019, pp. 198-204.
- [2] M. Deriche, S. Aliyu, and M. Mohandes, “ An Intelligent Arabic Sign Language Recognition System using a Pair of LMCs with GMM Based Classification” *Int. J. Comp. Eng*, vol. 14, pp. 1-32, Feb. 2010
- [3] K. Bazargan, “FPGA '19: Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays”, Internet: <https://dl.acm.org/doi/proceedings/10.1145/3289602>, Jun. 19, 2019 [Jan. 22, 2021].
- [4] M. Zhao, “Real-Time Underwater Image Recognition with FPGA Embedded System for Convolutional Neural Network” Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6359710/>, Jul., 2017 [Mar.15, 2020].
- [5] M. A. Abdel-Fattah, “Arabic Sign Language using GMM Technique”, *Int. J. of Deaf. Edu*, vol. 10, pp. 2-8, 2010.
- [6] E. Enrique and G. Guerrero, “FPGA-based Chaotic Cryptosystem by Using Voice Recognition as Access Key”, Internet: <https://www.mdpi.com/2079-9292/7/12/414>, Dec. ,2018 [Jun 18, 2020].
- [7] M. Hakan, M. YILMAZ and M. TOKMAKC, “Design and implementation of a voice-controlled prosthetic hand”, *Turk. J. Elec. Eng*, Vol.19(1), pp. 7-32, 2012.

- [8] N. El-Bendary, M. Zawbaa and S. Daoud, "Arabic Sign Language Alphabets Translator", *Int. J. of Comp. Inf. Sys. and Indus. Man. App*, Vol. 3(2), pp. 498-506, 2013.
- [9] A. Haseeb and A. Illyas, "Speech Translation into Pakistan Sign Language Speech Translation into Pakistan Sign", M.A. thesis, University of Karlskrona, Sweden, 2011.
- [10] S.M. Halawani and A.B. Zaitun, "An Avatar Based Translation System from Arabic Speech to Arabic Sign Language for Deaf People", *Int. J. of Inf. Scie. and Edu*, Vol. 2(1), pp. 13-20, 2009.
- [11] A. Hassan and S. Alzaki Ali, "Speech To Text Conversion", *Int Conf. of Comp. and. Tech*, 2010, pp. 192-199.
- [12] K. Anetha, R. Parvin and H. Talk, "A Sign Language Recognition" *Int. J.of Innov. Res.in Comp. and. Commun. Eng*, Vol. 2(3), pp. 6-14, 2000.
- [13] J. Gatti, C. Fonda and L. Tenze, "Voice-Controlled Artificial Handspeak" *Int. J. of. Arti. Intel. And. App*, Vol. 5(1), pp. 22-35, 2004.
- [14] Z. Othman, N. Abdullah and Z. Razak "Speech to Text Engine for Jawi Language", *Int. Arab. J. of. Inf. Tech*, Vol. 11(5), pp. 44-49, 2014.
- [15] A. Shahebaj, D. Krupa and J. Madhavi, "American Standard Sign Language Representation Using Speech Recognition", *Int. J. of. Rec. Res. In. Elec. And. Elect. Eng*, Vol. 2, pp. 1-5, 1999.
- [16] M. Agrawal and T. Raikwar "Speech Recognition Using Signal Processing Techniques", *Int. J. of. Eng. And. Innov. Tech*, Vol. 5(8), pp. 3-9, 2008.

- [17] M. Beena and N. Agnisarman, "Automatic Sign Language Finger Spelling Using Convolution Neural Network", *Int. J. of. Pure. And. App. Math*, Vol. 117(20), pp. 9-15, 2015.
- [18] A. Trivedi, N. Pant and P. Shah, "Speech to text and text to speech recognition systems", *Int. J. of. Comp. Eng*, Vol. 20 PP. 36-43, 2017.
- [19] D. Nagajyothi and P. Siddaiah, "Speech Recognition Using Convolutional Neural Networks", *Int. J. of. Eng. and. Tech*, Vol. 7(4), pp. 133-137, 2011.
- [20] M. Kubanek , J. Bobulski, and J. Kulawik, "A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network", *Int. Conf. For. Comp. Eng*, 2013, pp. 52-61.
- [21] M. M. Kamruzzaman, "Arabic Sign Language Recognition and Generating Arabic Speech Using Convolutional Neural Network", *Int. J. OF. Wire. Commun. and. Mob. Comp*, Vol. 3(4), pp. 9-17,2001.
- [22] S. bachani, S. dixit and R. chadha, "Sing Language Recognition Using Neural Network", *Int. Res. J. of. Eng. and. Tech*, Vol. 7, pp. 11-20, 2013.
- [23] N. Alcaraz, "Speech Analysis for Automatic Speech Recognition", *Int. Conf. For. Elect. And. Tech*, 2009, pp. 55-66.
- [24] X. Huang, *Spoken Language Processing*, New Jersey: Wiley, 2001, pp. 123-66.
- [25] Holmes, *Speech Synthesis and Recognition*, London: Francis, 2001, pp. 42-65.
- [26] M. Nilsson, M. Ejnarsson and M. Kindy, "Speech Recognition Using Hidden Markov Model", *Int. Conf. For. Elect. and. Tech*, 2002, pp. 3-9.

[[27] M. Abdulfattah, A. Rasheed and M. ElAwady, “Phonetic Recognition of Arabic Alphabet letters using Neural Networks”, *Int. J. of. Elect. and. Comp* , Vol. 11(1), pp.32-40, 2009.

[28] A. Shahebaj, D. Krupa and J. Madhavi, “American Standard Sign Language Representation Using Speech Recognition”, *Int. J. of. Recent. Res. In. Elect. and. Elect. Eng*, Vol. 2, pp. 1-5, 2018.

[29] S.M. Halawani and A. Zaitun, “An Avatar Based Translation System from Arabic Speech to Arabic Sign Language for Deaf People”, *Int. J. of. Inf. Science. And. Edu*, Vol. 2(1), pp. 13-20, 2019.

[30] I. Trabelsi , D. Ayed, “On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels”, *Int. Conf. for. Elect. and. Tech*, 2012, pp. 1-9.

[31] T. Abreha, “Recognition System Using Smartphones”, M.A. thesis, University of Bangkok, Thailand, 2014.

[32] M. A. Anusuya and S. K. Katti, “Speech Recognition by Machine”, *Int. J. of. Comp. Science. And. Inf. Sec*, Vol. 6(3), pp. 22-33, 2016.

[33] S. Vasilic, “Fuzzy Neural Network Pattern Recognition Algorithm for Classification of the Events in Power System Networks”, M.A. thesis, university of Texas, USA, 2004.

