جامعة فيلادلفيا
نموذج تفويض

انا أيهم راسم فايز جعرون ، أفوض جامعة فيلادلفيا بتزويد نسخ من رسالتي للمكتبات او المؤسسات او الهيئات او الاشخاص عند طلبها.

التوقيع:
التاريخ:

Philadelphia University
Authorization Form

I, Ayham Rasem Fayz Jaroun, authorize Philadelphia University to supply copies of my Thesis to libraries or establishments or individuals upon request.

Signature:
Date:

SPEAKER RECOGNITION USING NEURAL NETWORK MODEL

By
Ayham Rasem Fayz Jaroun

Supervisor
Dr. Venus W. Samawi

This Thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

Deanship of Academic Research of Graduate Studies

Philadelphia University

February 2008

Successfully defended and approved on _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Examination Committee	**Signature**
Dr,_ _, Chairman. Academic Rank: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _ _ _ _ _ _ _ _ _ _
Dr,_ _, Member. Academic Rank: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _ _ _ _ _ _ _ _ _ _
Dr,_ _, Member. Academic Rank: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _ _ _ _ _ _ _ _ _ _
Dr,_ , External Member. Academic Rank: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ ( )	_ _ _ _ _ _ _ _ _ _ _ _

Dedication

To My Great Parents for their Endless Love and Support...

To My Friends for their Supporting Words and Smiles...

To My Doctors and Teachers with Respect...

To Those who were Always Behind me...

To Every One Put his Imprint on My Successfulness Path...

For Whose Going to Read my Words, and Believe That I Mean Him...

To My Dream... Which I Wish To Be Real...

Ayham R. Jaroun

Acknowledgement

First of all, my great thanks to ALLAH who help me and gave me the ability to achieve this work and all the good things in my life.

I would like to express my sincere gratitude and appreciation to my supervisor Dr. Venus W. Sammwi for her guidance, assistance and encouragement through the course of this project.

Special thanks to all of my friend and the staff of department of computer science for their support, interest and generosity.

Finally, I want to express my gratitude to my father and mother and all my beloved family for encouragement and understanding during the period of my study.

Ayham R. Jaroun

Table of Contents

Chapter Three (Methodology)

Chapter four (Assessment Results)

Chapter five (Conclusion and Feature Works)

List of Tables

List of Figures

List of Abbreviations

AANN	Feed forward Auto-Associative Neural Network
ACON	All-Classes-in-One-Network
ADALINE	Adaptive Linear Neuron networks
ANN	Artificial Neural Network
CWT	Continuous Wavelet Transform
DSP	Digital Signal Processing
IIR	Infinite Impulse Response
LMS	Least Mean Squares
LP	Liner Prediction
LPC	Linear Predictive Coding
LPCC	Liner Predictive Coding Coefficients
LVQ	Learning Vector Quantization
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
NN	Neural Network
OCON	One-Class-in-One-Network
SI	Speaker Identification
STFT	Short Time Fourier Transforms
SV	Speaker Verification
TIMIT	Instruments/Massachusetts Institute of Technology
VB	Voice Biometrics
VQ	Vector Quantization
WT	Wavelet Transformation

List of Symbols

Φ	Mother Wavelet Function
$W(x)$	Scaling Function
C_k	Wavelet Coefficients
δ	Delta is Function
$F(w)$	Fourier Coefficients
f(t)	Time of Signal
ψ	Continuous Wavelet Function
Δw	Weight Correction
$\theta_i(l)$	Bias Term
$f(.)$	Activation Function
δ_I	Error Signal
μ	Momentum Constant
W	Weight
L	Number of Layers
$\delta_j(l)$	Error Signal of a Lower Layer
$\delta_j(l+1)$.	Error Signal of the Upper Layer

Abstract

The use of biometric information has been known widely for both person identification and security application. It is common knowledge that each person can be identified by the unique characteristics of one or more of his/her biometrics. One of the biometrics that a person can be identified by is the unique characteristics of his/her voice. This work is concerned with the study of using voice as biometric information (i.e. speaker recognition) for controlling access to the facility that need to be protected from the intrusion of unauthorized persons.

The main significance of this work is to study a number of experimental investigations of using neural networks to recognize speakers and suggest the best neural network architecture leading towards the goal of higher accuracy for speaker recognition. Therewithal, attempt to draw a conclusion about the recommended feature-set (based on wavelet transformation) that could be used to discriminate speakers and improve the overall performance. To do this, features are extracted from different levels of continues wavelet transformation (three maximum values with their locations and level number in addition to mean and standard division of each level), these features are used to train three of speaker recognition neural networks (Adaptive Neural Network, Feed-forward Backpropagation Neural Network, and Learning Vector Quantization Neural Network). Experimental results that show the classification accuracy of each classifier were introduced. Then these results were analyzed and compared to arrive at the best neural network (as speaker recognizer), and decide which of the feature (or combination of them) among the above set leads to the minimum classification error rate (i.e. has the best discrimination ability).

A vocabulary of 240 speech samples is built for 4 speakers, where each authorized person is asked to utter every sample 10 times. Two different modes are considered in identifying individuals according to their speech samples, in the text-dependent speaker identification, the system gives identification rate in range of 80% to 91%, while for text-independent identification mode (presented with 100 trials), where the best obtained identification rate is 71.87%.

CHAPTER ONE
INTRODUCTION

Chapter One
Introduction

1.1 Introduction

Pattern recognition resulted from the need for automated recognition of objects, signals or images, or the need for automated discrimination system based on a given set of parameters. Over half a century of productive research, pattern recognition continues to be an active area of research since there are a lot of unsolved fundamental theoretical problems as well as increasing number of applications that can benefit from pattern recognition. One of the most important applications of pattern recognition is the authentication of person identity through biometrics.

A biometric is a measurement of a biological characteristic such as fingerprint, iris pattern, retina image, face or hand geometry; or a behavioral characteristic such as voice, gait or signature. Biometric technology uses these characteristics to identify individuals automatically. Ideally the characteristic should be universally present, unique to the individual, stable over time and easily measurable. It is common knowledge that each person can be identified by the unique characteristics of one or more of his/her biometrics. One of the biometrics that a person can be identified by is the unique characteristics of his/her voice [Parliamentary 2001].

In a world where authentication and privacy are taking a lot of the daily efforts, it is becoming more important for us to prove our identity to different systems every day so that we can access required and useful services. This research considers the problem of speaker recognition as it involves knowing the identity of a given speaker using a predefined set of samples [Wouhaybi 1999]. Different recognizers are used for speaker identification (such as neural networks, genetic algorithms, statistical approaches, etc) depending on different set of features extracted from the person voice (such as liner predictive coding, wavelet transformation (discrete and continues), DCT discrete cosine transform, etc). The main steps of this process starts with preprocessing the voice signal, perform sampling and quantization, then use wavelet transformation to extract feature from it. Finally, the extracted features are fed to a pattern recognition phase (classifier), in this work that

recognition phase is designed using Neural Network. This field is stile under intensive study at which the appropriate feature set that contains the best unique characteristic of each voice need to be investigated in addition to the appropriate classifier for each feature set.

1.2 Motivation

In computer-driven era, identity theft and the loss or disclosure of data and related intellectual property are growing problems. Maintaining and managing access while protecting both the user's identity and the computer's data and systems has become increasingly difficult.

Central to all security is the concept of authentication - verifying that the user is who he claims to be. We can authenticate an identity in three ways [Chief 2005]:

- By something the user knows (such as a password or personal identification number),
- Something the user has (such as a security token or smart card)
- Something the user is (such as a physical characteristic, fingerprint, voice, called a biometric) which can be identified using recognition systems that can recognize these human features.

One of the widely used systems is the voice recognition technique. Since every human have a unique feature in his voice it is useful to discriminate between two persons using their own voices. The idea of Speaker Recognition, which is different from Speech Recognition, is to verify the individual speaker against a stored voice pattern, not to understand what is being said. While speech recognition is concerned with understanding what is being said.

But the question appear what will happen when we are dealing with similar pronounced sound for example brother or even when we are dealing with two related subject that share the common account but speaking a different axons of the same language, For example there are many countries around the world that pronounce the same word in different way. This can be a problem especially in banks. When the person wants to enter his bank account for safety measure the bank ask him for authentication by talking to a system so they can verify the person, the difference in pronouncing the words will make some trouble when giving the authorization to access the account. For that a process called

two-factor authentication is suggested at which both voice recognition (to verify the individual speaker), and speech recognition (which is used assure the user personality throw a pronounced password or sentence) is used.

In recent years, there has been a booming interest in the use of biometric characteristics as a means of recognizing and identifying a person. Human voice is one of the most important biometric identifiers of a person, Speaker Recognition is concerned with extraction of the identity of the person speaking the utterance. The general area of speaker recognition can be categorized into Speaker Verification (SV) and Speaker Identification (SI). The task of speaker verification is to determine form voice samples whether a person is whom he or she climes. On the other hand, speaker identifications aims to determine which one of a set of known voices best matches the input voice samples. Both speaker verification and speaker identifications can be constrained Text-Dependent or unconstrained Text-Independent. A speaker verification system is said to be text-dependent if it knows that the speaker is supposed to say. On the other hand, a text-independent speaker verification system dose not has foreknowledge of what the speaker is supposed to say [Guojie 2004].

1.3 Literature Review

Several researches in the field of speaker recognition were developed. This section will describe some research efforts, with abstract for each one, concerning: speaker recognition, speaker verification, speaker identification.

- Saranli [Saranli 2000] studies the applicability of rank-based multiple classifier in speaker identification system as classifier algorithm, and uses the Fast Fourier transform spectral [FFTC], and Liner predictive coding coefficients [LPCC] for Features Extraction. From the FFTC, the first 12 coefficients were taken as features; from LPCC also the first 12 coefficients were taken as features. The identification accuracy for the classifier hybrid Fast Fourier transform spectral/liner predictive coding coefficients 95% instead of 88% for FFTC and 85% for LPCC.
- B. R. Wildermoth [Wildermoth 2001] uses the spectral coefficients to extract features for text-independent speaker identification. The achieved identification rate 94%. In addition to the features set, the pitch feature is add, where the obtained

identification rate are in the range of 6% to 14% and this is not helpful for speaker recognition because the speech is not always periodic.

- B. Yegnanarayana, K. S. reddy and S. P. Kishore [Yegnanarayana et, 2001] use linear prediction (LP) for features extraction, the features are extracted from the source and system component of speech production process on speaker recognition. The **source** and **system** components are derived using linear prediction analysis of short segments of speech. The ***source*** component is the LP residual derived from the signal, and the ***system*** component is a set of weighted linear prediction spectral coefficients, 19-dimensional weighted spectral coefficient feature vector is used to represent the vocal tract ***system*** characteristics. A block size of 40 samples is used to capture the ***source*** characteristics from the LP residual. The feed-forward Auto-Associative neural network (AANN) used as a classifier algorithm. A speaker recognition system for 20 speakers is built and tested using both models to evaluate the accuracy of source 77% and system 87% features. The study demonstrate the complementary nature of the two component to give accuracy about 90%
- Y. A. Taleb [Taleb 2003] two approaches used for features extraction, the first one linear predictive and the second phoneme based wavelet approach, the adaptive feature-bands subset selection technique used as classifier technique. The performance of the linear predictive approach was tested using random speech passage produced by 108 speakers. An identification rate up to 100% was obtained for 30 speakers, but for the 108 speakers 88% was obtained. The wavelet approach was tested using short speech segments produced by 20 speakers from the normal folder in the matlab speech toolbox database, where the obtained identification rate is about 80% and it is increased for 85%.
- T. Kinnunen [Kinnunen 2003] use several feature extraction methods. The first one Liner Predictive Coding Coefficients, Fast Fourier Transform spectrum, and line spectral frequency. Vector Quantization (VQ) used as recognizer of speakers. The obtained results are about 99% for Liner predictive coding coefficients features, 95% for Fast Fourier Transform spectrum, and 86% for line spectral frequency.
- Brain J. Love et, [Love 2004] exploit the multilayer Perceptron neural network as a classifier in the suggested speaker identification system. The input into the neural

network as a features extraction was 1, 12 Liner predictive coding coefficients with normalized mean pitch value (13 inputs), The result that obtained when recognize two speakers 89.583% in testing data and for three speakers 90.27% and for four speakers 85.41%.

- Muzhir Shaban Al-Ani, et, [Ani 2007] investigate the use of wavelet transform and neural networks together as speaker recognition system. Features are extracted by applying a discrete wavelet transform (DWT) from the 9th level wavelet, 22 coefficients obtained for each sampled speech, while a Multi-valued neurons (MVN) neural network (NN) is used as a classifier part. The neural network is trained using inputs. The research focus on text-independent speaker identification. The system was trained using 25-speakers, where the test shows that the system is able to recognize if the speaker is male or female with accuracy about 61%.

1.4 Major Contribution and Objective

The main contribution of this research is to give a push in the field Authentication System by proposing a technique for speaker recognition which can also in future extended to use speech recognition (to verify dialogs or passwords) technique based on language grammars for further security system, but since we have short in time and this subject is highly complicated subject so we will constraint our effort on speaker recognition contribution in authentication system using Neural Network as Artificial Intelligence technique, that are trained with features extracted from Wavelet Transformation (WT), then train the recognition methods with different features and study the best feature set suitable for building voice recognition system (with high recognition ability) to reach a simple but less expensive authentication system since most of the modern system in filed are complex and expensive system.

Also by this research we can make a decision on how to build a speaker recognition system by building it using different neural networks (three of them are chosen, ADALINE, Backpropagation and LVQ) and compare their recognition ability depending on recognition accuracy.

The research mainly will consist of three phases: Preprocessing Phase, Feature extraction phase, and Recognition Phase. Finally, a comparison study is introduced, and

final authentication system is suggested. The suggested system will be tested and evaluated k-fold cross validation method.

The main objective of this research project is the study of the behavior of using Wavelet transformation (Continues wavelet Transformation) based features (features are extracted from different levels), with neural network as classification part. Three different algorithms of neural networks are used to solve speaker verification problem and study the behavior of three different neural networks (ADALINE, Backpropagation, and LVQ (linear vector quantizer)) from classification accuracy point of view, and evaluate their performance with the used feature sets.

.

1.5 Organization of the Thesis

This thesis divides into five chapters:

- Chapter two introduced the (Theoretical Background For Speaker Recognition),
- Chapter three (Methodology) presents the design and implementation of the System,
- Chapter four (Assessment result) presents the experimental results for text-dependent and text-independent speaker recognition approach.
- Chapter Five (Conclusions and Future Work) demonstrates the remarks concluded, and the chapter include suggestions for future work.

CHAPTER TWO

THEORETICAL BACKGROUND

Chapter Two
Theoretical Background

2.1 Review of Speaker Recognition

Person voice contains information that could help in specifying his identity. A speech signal includes the presence and type of speech pathologies, the physical and emotional state of the speaker. Often, humans are able to extract the identity information when the speech comes from a speaker, which they are acquainted with. One of the main biometrics that could be used to identify persons is through their voice, which is the main core of this research. In this research, features are to be extracted from the voice converting it to wavelet form, and then a suitable classifier is constructed based on NN methodology.

This chapter will explore the main concepts of voice as a biometric in brief followed by that theoretical background for speaker identification concerning digital signal processing and feature extraction based on WT. The main concepts of NN are illustrated and how it could be used as recognition system. Due to the fact that voice signal could be associated with noise signal and non-speech information, a preprocessing operations will also be discussed on voice signal is needed.

2.2 Introduction of Voice Biometrics (VB)

Biometric voice verification is the process of comparing a voice sample with a stored, digital voice model, or voiceprint, for the purpose of verifying identity. A voiceprint is a digital representation of some of the unique characteristics of voice, including physiological characteristics of the nasal passages and vocal chords, as well as the frequency, cadence and duration of the vocal pattern [Diaphonics 2006].

If, for example, the person speaks “It’s me!” This pronouncement is usually made over the telephone or at an entryway out of sight of the intended hearer. It embodies the expectation that the sound of one’s voice is sufficient for the hearer to recognize the speaker. In short, “It’s me!” is the original real-world, speaker-recognition challenge.

Voice-biometrics (VB) systems can be categorized as belonging in two industries: **Speech Processing** and **Biometric Security** (as shown in figure (2.1)) [Markowitz 2000].

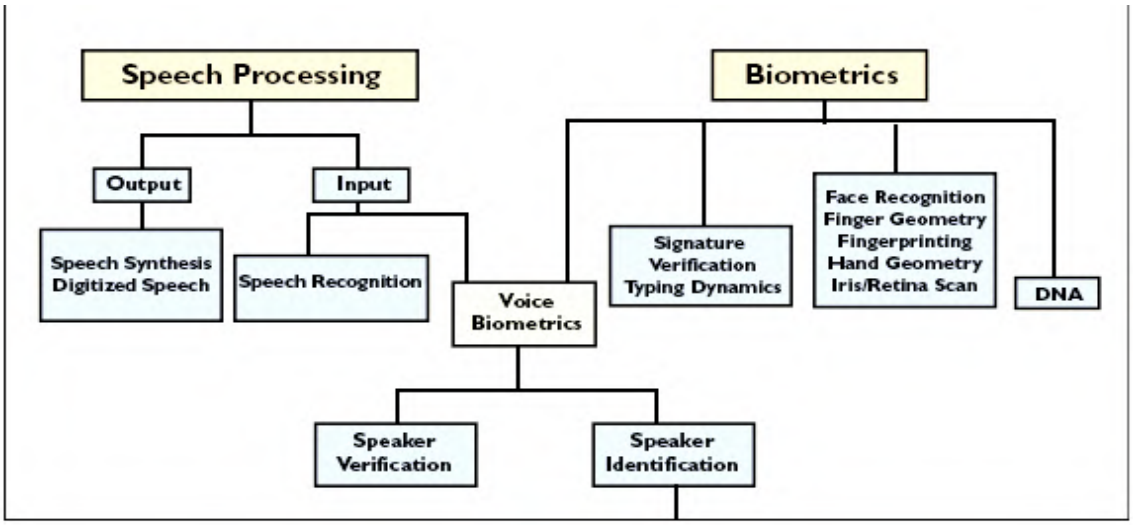


Figure 2.1: voice biometric system

2.2.1 Types of Voice Biometrics

Speaker or voice recognition is a biometric modality that uses an individual's voice for recognition purposes. It is a technology that differs from speech recognition, which recognizes words as they are articulated; it is not a biometric [NIST 2006].

There are two types of VB [Markowitz 2000]:

- **Speaker Verification:** Speaker-verification systems authenticate that a person is who she/he claims to be. If, for example, the person had shouted, “its Julie,” rather than, “It’s me,” the intended hearer would have to perform speaker verification based on that identity claim.
- **Speaker Identification:** Speaker identification assigns an identity to the voice of an unknown speaker. The assertion “It's me!” requires speaker identification, because the intended hearer is expected to assign the proper identity to the speaker based on the voice alone. In most cases, speaker identification is more difficult than speaker verification, because it involves multiple comparisons of utterances that are likely to be different from each other and may not have been recorded with comparable equipment.

There are several ways of interacting with speaker-verification and/or speaker-identification systems depending how the system could verify the person from the type of the pronounced sentence. Theses are [Markowitz 2000] [Pawar et, 2005]:

- **Text-dependent**: most commercial systems are text-dependent. They request a password, account number.
- **Text-dependent voice-only**: approach that uses the account number as both identity claim and password. Speech recognition decodes the input, and speaker verification uses the same input as the biometric sample it compares to the reference voiceprint.
- **Text-prompted systems**: This asks the speaker to repeat a series of randomly selected digit strings, word sequences, or phrases. Text prompting requires longer enrollment than text-dependent technology, because the reference voiceprint it generates must contain all the components that will be used to construct challenge-response variants.
- **Text-independent verification:** accepts any spoken input, making it possible to design unobtrusive, even invisible, verification applications that examine the ongoing speech of an individual. The ability of text-independent technology to operate unobtrusively and in the background makes it attractive for customer related applications, because customers need not pause for a security check before moving on to their primary business objective.

2.3 Identification Background

This section discusses theoretical background for speaker identification concerning digital signal processing theory. Then, the anatomy of human voice production organs and discuss the basic properties of the human speech production mechanism and techniques for its modeling is illustrated. This model will be used when discussing techniques for the extraction of the speaker characteristics from the speech signal.

2.3.1 Digital Signal Processing Fundamentals

Digital Signal Processing (DSP) is part of computer science, which operates with special kind of data – *signals*. In most cases, these signals are obtained from various sensors, such as microphone or camera. DSP is the mathematics, mixed with the algorithms

and special techniques used to manipulate with these signals, converted to the digital form [Karpov 2003].

Signal means here a relation of how one parameter is related to another parameter. One of these parameters is called *independent parameter* (usually it is time), and the other one is called *dependent*, and represents what are to be measured. Since both of these parameters belong to the continuous range of values, so it's called *continuous signal*. When continuous signal is passed through an *Analog-To-Digital converter (ADC)* it is said to be *discrete* or *digitized* signal. Conversion works in the following way: every time period, which occurs with frequency, called *sampling frequency*, signal value is taken and *quantized*, by selecting an appropriate value from the range of possible values. This range is called *quantization precision*, and usually represented as an amount of bits available to store signal value. Based on the *sampling theorem,* proved by Nyquist in 1940, digital signal can contain frequency components only up to one half of the sampling rate. Generally, continuous signals are natural signal while discrete signals exist mostly inside computers. Signals that use time as the independent parameter are said to be in the *time domain*, while signals that use frequency as the independent parameter are said to be in the *frequency domain* [Karpov 2003].

One of the important definitions used in DSP is the definition of *linear System. System* means here any process that produces *output* signal in a response on a given *input* signal. A system is called linear if it satisfies the following three properties:

- Homogeneity of a system means that change in the input signal amplitude corresponds to the change in the output signal.
- Additively means that the output of the sum of two signals results in the sum of the two corresponding outputs.
- And finally, shift invariance means that any shift in the input signal will result in the same shift in the output signal.

2.3.2 Human Speech Production Model

Ability to speak is the most important way for humans to communicate between each other. Speech conveys various kind of information, which is essentially the meaning of information speaking person wants to impart, individual information representing

speaker and also some emotional filling. Speech production begins with the initial formalization of the idea which speaker wants to impart to the listener. Then speaker converts this idea into the appropriate order of words and phrases according to the language. Finally, his brain produces motor nerve commands, which move the vocal organs in an appropriate way [Karpov 2003]. The sound is an acoustic pressure formed of compressions and rarefactions of air molecules that originate from movements of human anatomical structures (see appendix (A) for more information about human anatomical structure).

2.4 Speaker Recognition System

Speaker recognition system mainly consists of three phases: the preprocessing phase, feature extraction phase, and recognition phase (discrimination algorithm or classifier). These phases are illustrated below.

2.4.1 Preprocessing Phase

As with any real world digital application, any voice samples that are recorded will be corrupted by a finite amount of noise. In addition, since this subject will not have a perfect reaction time, there will be durations of signal at the beginning and end of the recording where no speech will be presented. In general, it is expected that both noise and the analysis of non-speech information would degrade the quality of our analysis and results [love et, 2004].

Many filters could be used to ***remove noise***, like low pass filter, high pass filter, band pass filter. ***Low-pass filter*** is a circuit offering easy passage to low-frequency signals and difficult passage to high-frequency signals, ***high-pass filter's*** task is just the opposite of a low-pass filter: to offer easy passage of a high-frequency signal and difficult passage to a low-frequency signal, ***finally, band-pass*** filter is resulted from combining the properties of low-pass and high-pass into a single filter [Kuphaldt 2007].

For ***non-speech information*** (silent area) removal, there are many commercial packages like: jet audio, sound forge, ashampoo Audio, etc.

2.4.2 Feature Extraction Phase

The process of feature extraction is to extract a set of essential characteristics that can identify or represent whole speech signal [Pawar et, 2005].

Speech contains many unique characteristics that are specific to each individual, and contain information that allow listener to determine both gender and speaker identity.

Different methods could be used to obtain feature from voice such as: Short Time Fourier Transforms (STFTs), Linear Predictive Coding (LPC) techniques, etc. under certain situation, these methods may not be suitable for representing speech; they assume signal stationary within a given time frame and may therefore lack the ability to analyze localized events accurately [long and datta 1996].

Wavelet Transform overcomes some of these limitations; it provides a useful decomposition of a signal, so that faint temporal structure can be revealed and handled by nonparametric models [murtagh 2003]. In this work, features are extracted using wavelet transformation.

2.4.3 Recognition Phase

There are several methods for speaker recognition which classify coordination of the voiceprint and speech model. They can be divided into [Zimmermann 2004]:

- Template methods: Into template methods we rank Dynamic Time Warping, Vector Quantization and the combination of both mentioned methods, Nearest Neighbors.
- Stochastic methods: like Hidden Markov Model.
- Methods which use artificial neural networks or genetic algorithms.

In this work, we are concerned with Neural Network.

2.5 Wavelet Transformation

Wavelet analysis allows the use of long time intervals where more precise low-frequency information and shorter regions where high-frequency information needed [Michel et, 2007]. Wavelet is a windowing technique with variable-sized regions; it is scale-based analysis when it started to become clear that an approach measuring average

fluctuations at different scales might prove less sensitive to noise [Graps 1995]. Figure (2.2) shows wavelet technique.

Figure (2-2): Wavelet Technique

Wavelet Transforms comprise an infinite set. There are several families of wavelet like (Haar, Dabechies, coiflets, symlets, meyer), the different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. Some of the wavelet bases have fractal structure. The Daubechies wavelet family is one example as shown in figure (2-3).

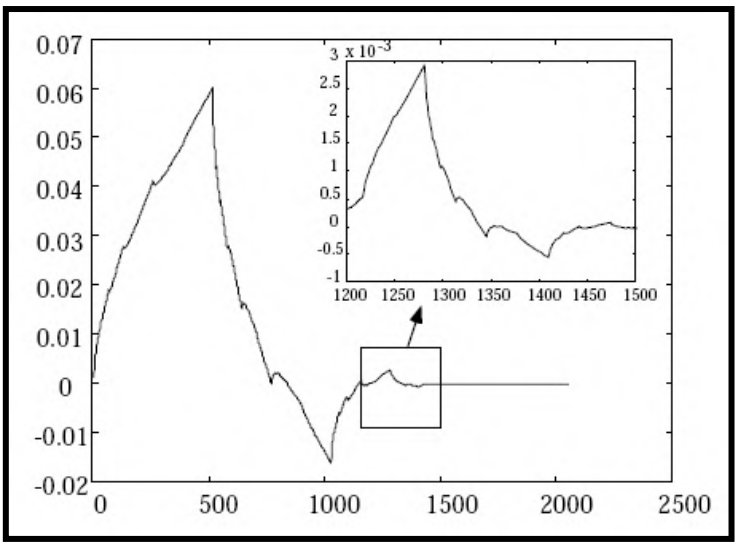


Figure (2-3): The fractal self-similarity of the Daubechies mother wavelet.

Within each family of wavelets (such as the Daubechies family), are wavelet subclasses distinguished by the number of coefficients and by the level of iteration. Wavelets are classified within a family most often by the number of vanishing moments. This is an extra set of mathematical relationships for the coefficients that must be satisfied,

which is directly related to the number of coefficients. Figure (2-4), illustrate several different wavelet families.

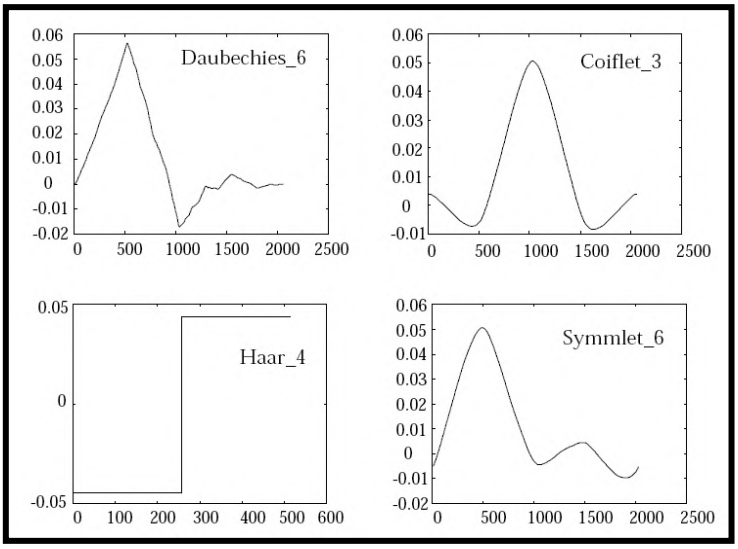


Figure (2-4): Different Wavelet Families

Wavelet transformation could be discrete or continuous transformation.

2.5.1 The Discrete Wavelet Transform

Dilations and translations of the Mother function, or analyzing wavelet $\Phi(x)$, define an orthogonal basis, the wavelet basis:

$$\Phi_{(s,l)}(x) = 2^{\frac{-s}{2}} \Phi(2^{-s}x - l) \qquad \text{..... (2-1).}$$

The variables *s* and *l* are integers that scale and dilate the mother function Φ to generate wavelets, such as a Daubechies wavelet family. The scale index *s* indicates the wavelet's width, and the location index *l* gives its position. Notice that the mother functions are rescaled, or \dilated by powers of two, and translated by integers. What makes wavelet bases especially interesting is the self-similarity caused by the scales and dilations. Once we know about the mother functions, we know everything about the basis.

To span our data domain at different resolutions, the analyzing wavelet is used in a scaling equation:

$$W(x) = \sum_{k=-1}^{N-2} (-1)^k C_{k+1} \Phi(2x + k) \qquad \ldots\ldots\ldots\ldots \text{(2-2).}$$

Where $W(x)$ the scaling functions for the mother function Φ; and C_k are the wavelet coefficients. The wavelet coefficients must satisfy linear and quadratic constraints of the form:

$$\sum_{k=0}^{N-1} C_k = 2, \quad \sum_{k=0}^{N-1} C_k C_{k+2l} = 2\delta_{l,o} \qquad \ldots\ldots\ldots\ldots \text{(2-3).}$$

Where δ the delta is function and l is the location index.

2.5.2 Continuous Wavelet Transform

Mathematically, the process of Fourier analysis is represented by the Fourier transform:

$$F(w) = \int_{-\infty}^{\infty} f(t) e^{-jwt} dt \qquad \ldots\ldots\ldots\ldots \text{(2.4).}$$

Which is the sum over all time of the signal f(t) multiplied by a complex exponential. The results of the transform are the Fourier coefficients $F(w)$, which when multiplied by a sinusoid of frequency w yield the constituent sinusoidal components of the original signal. Graphically, the process looks like.

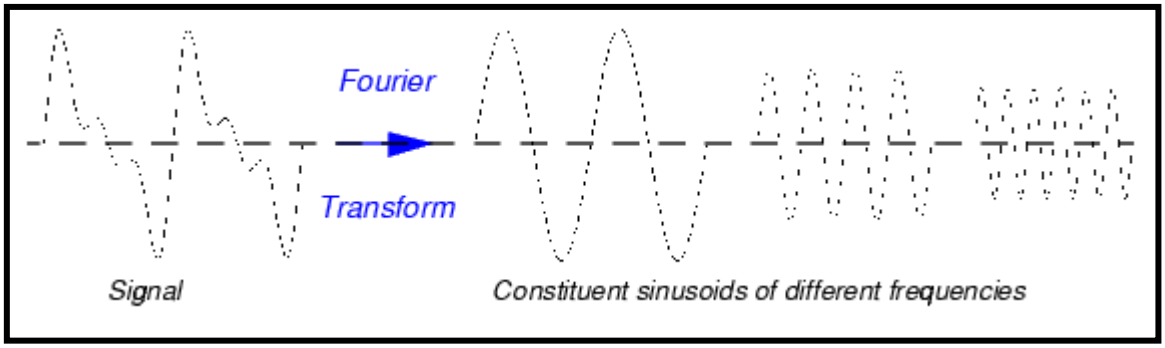


Figure (2-5): Signal and its Fourier Transform

Similarly, the Continuous Wavelet Transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function (ψ):

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt \quad \ldots\ldots\ldots\ldots \quad (2\text{-}5).$$

The results of the CWT are many wavelet coefficients C, which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal.

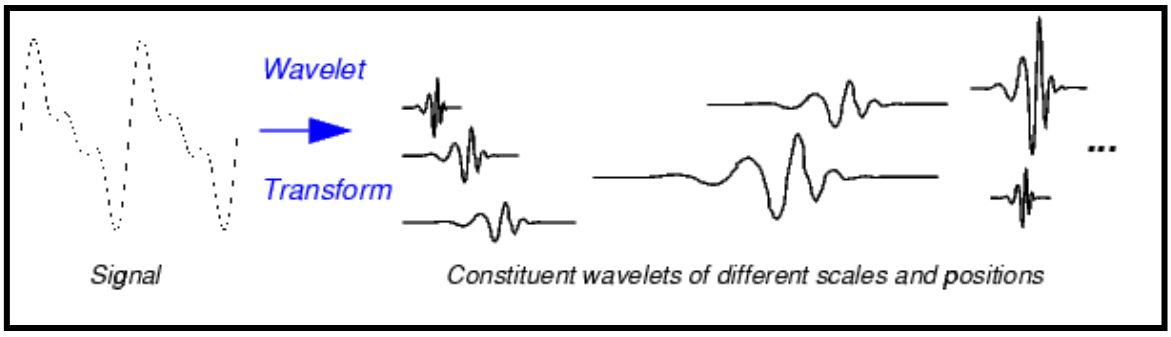


Figure (2-6): signal and it Wavelet Transform

2.6 Neural Network

An artificial neural network (ANN) is an interconnected group of artificial neurons that uses a computational model for information processing based on a connectionist approach. ANN is widely used in many fields of engineering, in particular for pattern recognition, classification and prediction [Gemello et, 2006].

Neural network, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions, other advantages include:

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time.
3. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

Any NN is characterized by its:

1. **Architecture**: Which represents the connection pattern between neurons? The behavior of the network highly depends on how neurons are arranged into layers (number of layers and number of neurons per layer), in addition to the type of connection (feedforward, feedback, lateral connection).
2. **Learning or Training algorithm:** It is a procedure used for modifying synaptic weights. The following table (2-1) provides a summary of the most popular five learning rules and their properties. These rules are tabulated and compared in terms

of the single weight adjustment formulas, supervised versus unsupervised learning mode, weight initialization, and required neuron activation functions.

Table (2-1): Summary of Learning Rules and their Properties [Zurada 1996]

Learning Rule	*Single weight adjustment* Δw_{ig}	*Initial Weights*	*Learning*	*Neuron Characteristics*
Hebbian	$\alpha y_i x_j$ $j=1,\ldots,n$	~0	U	Any
Perceptron	$\alpha(d_i - \text{sgn}(\mathbf{w}_i^T\mathbf{x}))x_j$ $J=1,\ldots,n$	Any small value between 0, & 1	S	Binary bipolar, or unipolar
Delta	$\alpha(d_i - f(\mathbf{w}_i^T\mathbf{x}))\ f'(\mathbf{w}_i^T\mathbf{x})$ x_j $j=1,\ldots,n$	Any small value between 0, & 1	S	Continuous
Widrow-Hoff	$\alpha(d_i - \mathbf{w}_i^T\mathbf{x})\ x_j$ $j=1,\ldots,n$	Any small value between 0, & 1	S	Any
Winner-take-all	$\alpha(x_j - w_{ij})$ i-winning neuron, $j=1,\ldots,n$	Random Normalized	U	Continuous

2.6.1 ACON versus OCON Approaches

Two possible network structures were suggested to be used for handling multi-category classification problem, these are [Kung 1993]:

- **All-Classes-in-One-Network (ACON):** The ACON structure is adopted by the conventional multilayer perceptron (MLP), where all the classes are lumped into one super-network. The super-net has the burden of having to simultaneously satisfy all the teachers, so the number of hidden units is expected to be very big. Empirical results confirm that the convergence rate of ACON degrades drastically with respect to the network size because the training is influenced by conflicting signals from different teachers. Therefore, it is sometimes advantageous to decompose a huge network into many subnets.

- **<u>One-Class-in-One-Network (OCON)</u>:** In the OCON structure, one subnet is devoted to one class only.

Although the number of subnets in the OCON is relatively large, each individual subnet has considerably smaller size than the ACON super-net which may offer computational savings in the training phase and performance improvements in the retrieving phase. Figure (2-7) shows the structure differences between ACON and OCON.

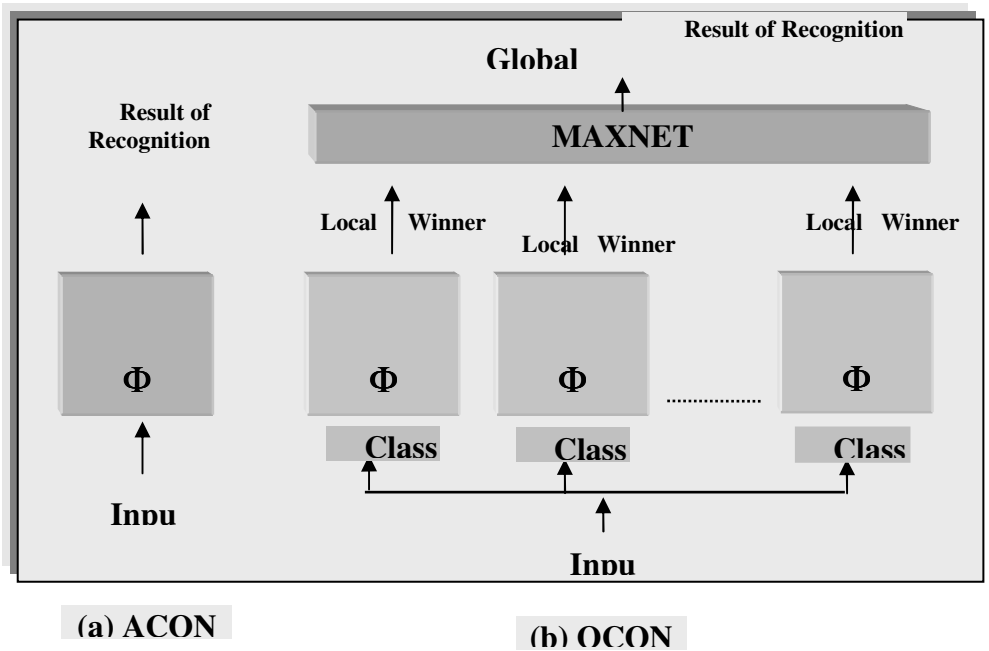


Figure (2-7): the (supernet) structure of ACON model. (b) An OCON structure viewed as a result of partitioning a single supernet into many small subnets.

2.6.2 Weight Adaptation Methods

One of the neural networks features is their learning ability which is accomplished by adaptively updating the synaptic weights that characterizes the strength of the connections. Usually, the optimal weights are obtained by optimizing (minimizing or maximizing) a certain "energy function". For example, in supervised learning, the popular criterion is to minimize the square error between the teacher (desired output) and the actual output value. To design a neural network training phase, the choice is between <u>stepwise training</u> (data adaptive), or <u>batch learning</u> (block adaptive). Before discussing the two training choices, it is important to know the differences between *iteration* and *sweep*. An

iteration involves presenting one training input vector (pattern) to the system, while, a *sweep* covers the presentation of an entire block of training data (or a set of patterns) to the system (as shown in figure (2-8)). In most training practices, multiple sweeps are used where the training patterns are repeatedly presented in a cyclic manner.

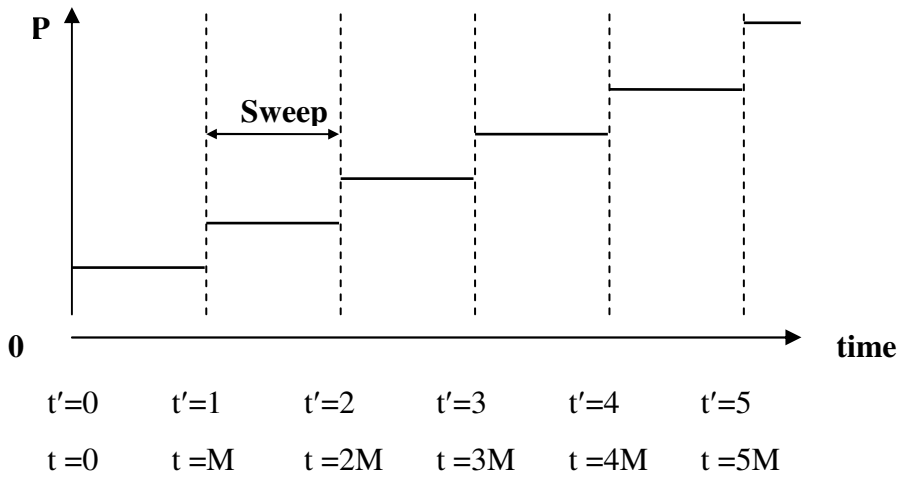


Figure (2-8): A block-adaptive updating rule is adopted only for purpose of analysis. The weights are assumed to be piecewise constant, as shown by the dotted lines.

2.6.3 Stepwise Training versus Batch Learning

The *stepwise training* (data adaptive) method updates the weights at each iteration *t*.

$$W^{t+1}=W^{t}+\Delta W^{t} \qquad \ldots\ldots\ldots\ldots \text{(2-6)}.$$

On the other hand, in *batch learning* (block adaptive), the weight correction term (Δw) is accumulated for several patterns (for an entire sweep) then make a single weight adjustment (equal to the average of the accumulated weight correction term) after each sweep.

2.6.4 Types of Neural Networks

In this section, three different neural network models described for pattern classification (as shown in figure 2-9), will be reviewed from architecture and designed algorithm points of view. These three nets are important building-blocks that can be used to construct more complex neural network systems.

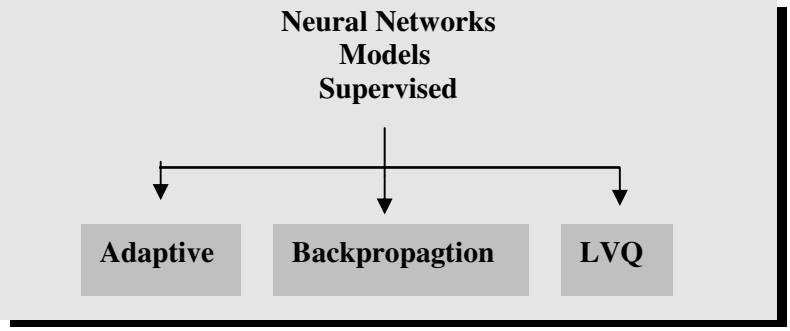


Figure (2-9): Three Considered Neural Networks Models

2.6.4.1 Feedforward Neural Network (Backpropagation)

Feed forward neural network consists of one or more layers of nonlinear processing elements or units. The processing elements belonging to the neighboring layers are connected by sets of synaptic weights. The earliest feedforward neural architectures focused on simple neural nets with single layer of linear or nonlinear output. Single layer nets based on linear model functions have very limited classification and approximation capabilities [Kung 1993]. For that, to enhance the classification and approximation capabilities, multilayer networks (together with the back-propagation training method) are usually adopted. In linear systems, there is no real benefit to cascading multiple layers of linear networks, since the equivalent weight matrix of the total system is simply the product of weight matrices of different layers [Patterson 1996].

In multilayer networks, nonlinear hidden neuron units are inserted between the input and the output layer. In this case it seems natural to assume that the more layers used, the greater power the network possesses. But this is not the case in practice, increasing the number of layers may cause slower convergence especially when using *backpropagation*. Two possible reasons are that *error signals may be numerically degraded when propagating across too many layers* and *that extra layers tend to create additional local minima*. Thus it is essential to identify the proper number of layers. Generally speaking,

two or three layer networks (one hidden layer) should be adequate for most applications [Kung 1993].

Figure (2-10) shows the multilayer-feedforward neural network architecture, where it consists of at least two hierarchical layers of neurons (middle layer or hidden layer, and the output layer) in addition to the input layer. The network is constructed in such a way that each layer is fully connected to the next layer.

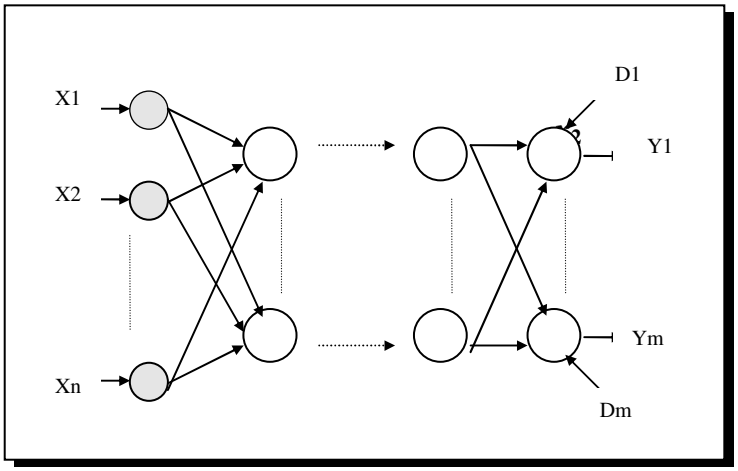


Figure 2-10: Multilayer-Feedforward Neural Network Architecture

The backpropagation algorithm offers an efficient computational speed up for training multilayer networks. The objective is to train the weights w_i so as to minimize the (energy function) *Least-Square-Error* between the desired and the actual output [Zurada 1996]:

$$E = \frac{1}{2}\sum_{i=1}^{m}(d_i - y_i)^2 \qquad \text{..... (2-7).}$$

The input layer receives the input and passes it along to each neuron in the next layer. Each neuron computes an output signal (activation) *u* as follows:

$$u_i(l) = \sum_{j=1}^{n_{l-1}} w_{ij}(l)\, x_j(l-1) + \theta_i(l) \qquad \text{..... (2-8).}$$

Where $\theta_i(l)$ is the bias term

$$x_i(l)=f(u_i(l)) \qquad 1\le i\le n_l,\ \ 1<l<L \qquad \text{..... (2-9).}$$

where L is number of layers

$f(.)$ is the activation function which is a form of sigmoid function, the output unit $y_i \equiv x_i(L)$. The basic gradient-type learning formula is:

$$w_{ij}^{t-1}(l)= w_{ij}^{t}(l) +\Delta w_{ij}^{t}(l) \qquad \text{..... (2-10).}$$

$$\Delta w_{ij}^{t}(l)= \alpha\delta_i^{t}(l) f'(u_i^{t}(l))x_j^{t}(l\text{-}1) \qquad \text{.....(2-11).}$$

The error signal δ_i can be obtained recursively by backpropagation, where:

$$\delta_i^{t}(L)= d_i^{t}\text{-}x_i^{t}(l), \qquad \text{..... (2-12).}$$

and

$$\delta_i^{t}(l) = \sum_{j=1}^{n_{l+1}} \delta_j^{t}(l+1) f'(u_{ij}^{t}(l+1)) w_{ji}^{t}(l+1) \quad \text{for } l=\text{L-1,...,1} \qquad \text{..... (2-13).}$$

Based on the above formulas, the synaptic weights between the l-th and (l-1)-th layers can be updated recursively (for $l=L,L$-1,…,1) as follows:

$$w_{ij}^{t+1}(l)= w_{ij}^{t}(l) +\alpha\delta_i^{t}(l) f'(u_i^{t}(l))x_j^{t}(l\text{-}1) \qquad \text{..... (2-14).}$$

The recursive formula is the key to the backpropagation learning. It allows the error signal of a lower layer $\delta_j(l)$ to be computed as a linear combination of the error signal of the upper layer $\delta_j(l+1)$. In this manner, the error signals $\delta_j(.)$ are backpropagated through all the layers from the top down.

Although Delta rule is widely known from its application in adaptive filtering, its simplicity and flexibility made it a practically attractive tool for the training of neural networks. However, the learning algorithms based on the Delta-rule are characterized by

slow convergence, and in some situations, can be trapped in local minima. Convergence can be improved by using the *momentum method*.

A momentum method is usually used to accelerate the convergence of backpropagation learning algorithm. This method involves supplementing the current weight adjustment with a function of the most recent weight adjustment. This is usually done according to the formula:

$$\Delta W^t = -\alpha\, \delta^t \mathbf{x}^t + \mu \Delta W^{t-1} \qquad \ldots\ldots\ldots\ldots \text{(2-15)}.$$

Where t and t-1 are used to indicate the current and the most recent training step, and μ is a user selected positive momentum constant. The second term (indicate a scaled most recent adjustment of weights) is called *momentum term*. Typically μ is chosen between 0.1 and 0.8. The momentum term technique can be recommended for a problem with convergence that occur too slowly or for cases when learning is difficult to achieve. Momentum has the effect of smoothing the error surface in weight space by filtering out high frequency variations.

The same algorithm of backpropagation with Delta rule can be used for backpropagation with momentum method, the only difference is the weight updating formula that should be changed to:

$$w_{jk}^{t+1} = \alpha \delta_k z_j + \mu \Delta w_{jk}^t \quad \text{where} \quad \Delta w_{jk}^t = w_{jk}^t - w_{jk}^{t-1} \qquad \ldots\ldots\ldots\ldots \text{(2-16)}.$$

$$v_{ij}^{t+1} = \alpha \delta_j x_i + \mu \Delta v_{ij}^t \quad \text{where} \quad \Delta v_{ij}^t = v_{ij}^t - v_{ij}^{t-1} \qquad \ldots\ldots\ldots\ldots \text{(2-17)}.$$

2.6.4.2 Adaptive Neural Network [Demuth 2007].

An adaptive neural network it's able to adapt its behavior according to changes in its environment or in parts of the system itself. The transfer function in adaptive neural network is linear; there is no hidden layer in adaptive neural network, the architecture for adaptive neural network show in figure (2-11).

The Adaptive Linear Neuron networks (ADALINE) networks are similar to the perceptron, but their transfer function is linear rather than hard-limiting. This allows their outputs to take on any value, whereas the perceptron output is limited to either 0 or 1.

However, IN ADALINE use of the Least Mean Squares (LMS) learning rule, which is much more powerful than the perceptron learning rule. The LMS or Widrow-Hoff learning rule minimizes the mean square error and, thus, moves the decision boundaries as far as it can from the training patterns.

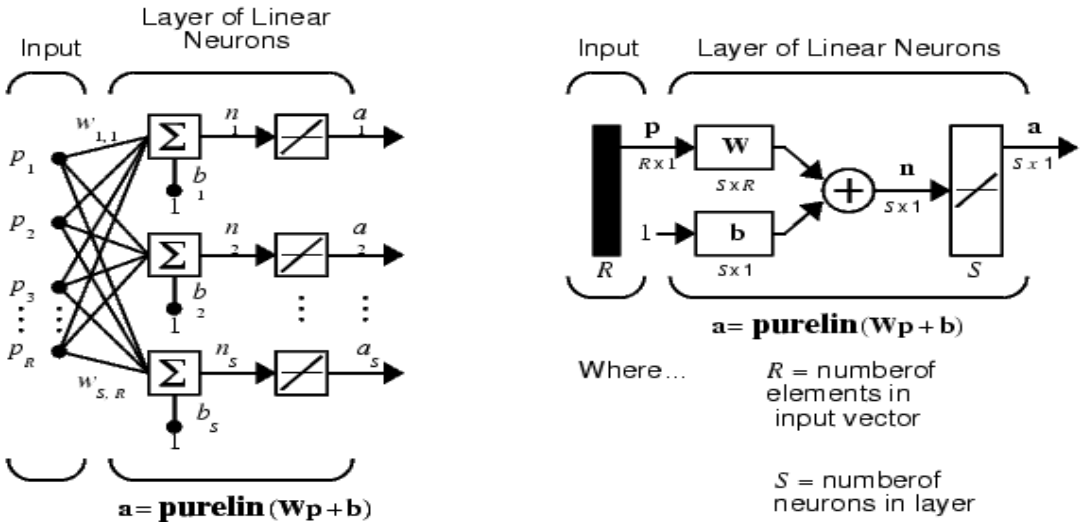


Figure (2-11): Architecture for Adaptive Neural Network

The ADALINE network shown below in figure (2-11) has one layer of S neurons connected to R inputs through a matrix of weights W.

This network is sometimes called a MADALINE for Many ADALINES. Note that the figure on the right defines an S-length output vector a.

The Widrow-Hoff rule can only train single-layer linear networks. This is not much of a disadvantage; however, as single-layer linear networks are just as capable as multilayer linear networks. For every multilayer linear network, there is an equivalent single-layer linear network. A linear neuron with R inputs is shown below in figure (2-12).

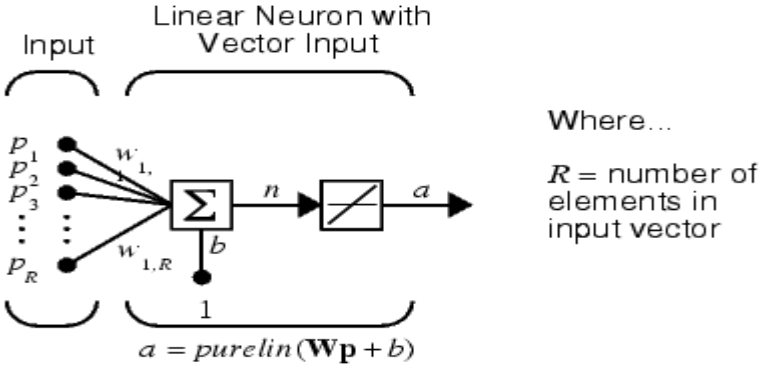


Figure (2-12): Liner Neuron Model

This network has the same basic structure as the perceptron. The only difference is that the linear neuron uses a linear transfer function, which we name purelin, the figure (2-13) show purelin transfer function.

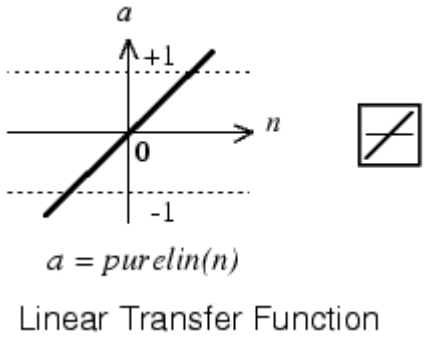


Figure (2-13): Purelin Transfer Function

The linear transfer function calculates the neuron's output by simply returning the value passed to it:

$$a = purelin(n) = purelin(\mathbf{W}\mathbf{p} + b) = \mathbf{W}\mathbf{p} + b \quad \text{..... (2-18).}$$

This neuron can be trained to learn an affine function of its inputs, or to find a linear approximation to a nonlinear function. A linear network cannot, of course, be made to perform a nonlinear computation.

2.6.4.3 Learning Vector Quantizer (LVQ) [Fausett 1994]

Learning vector quantization (LVQ) suggested by Kohonen in 1989. It is a supervised learning extension of the Kohonen network method. LVQ is a pattern classification method in which each output unit represents a particular class or category. During the training phase, the output units are positioned (by adjusting their weights through supervised training) to approximate the decision surface of the theoretical Bayes classifier. For that, it is important to know the classes (categories) of the training set in advance and make them part of the training set. The training procedure for LVQ rewards a winning neuron if it belongs to the correct class by moving it toward the input vector, and punishes it if the winning neuron does not belong to the correct class. After training, a LVQ net classifies an input vector by assigning it to the class of the output unit that has its weight vector closest (minimum Euclidean distance) to the input vector.

The architecture of a LVQ neural net, is essentially the same as that of the Kohonen neural net (see figure 2-14) except that sometimes several output neurons will be assigned to each class. The LVQ weight vectors are often referred to as reference (or codebook) vectors.

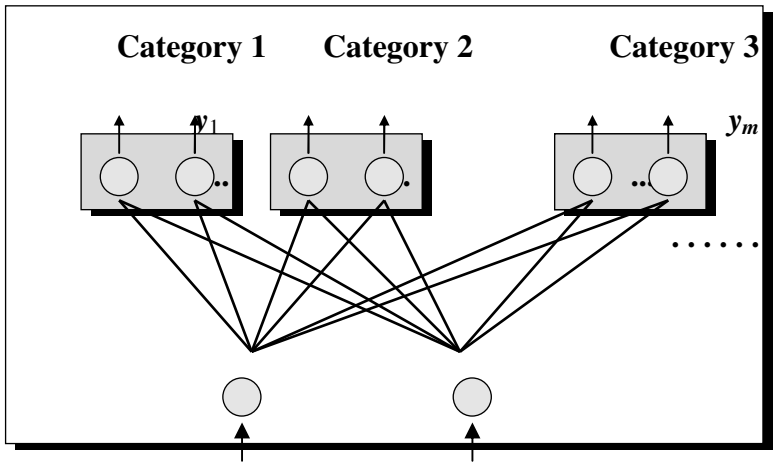


Figure (2-14): LVQ Neural Net Architecture.

Weight vectors could be initialized by using one of the following methods:

- Assign initial weight vectors and classification randomly.
- Take the first *m* training vectors (from different m classes) and use them as initial weight vectors; the remaining vectors are then used for training.
- Use K-means clustering, or the Self-Organizing map to place the initial weight vectors.

Each weight vector is then calibrated by determining the input patterns belonging to it.

CHAPTER THREE
METHODOLOGY

Chapter Three
Methodology

3.1 Introduction

The aim of this work is to build a speaker recognition system and evaluate the system capabilities and performance in sound recognition. Mainly, three different NN as classifiers are considered. (***Adaptive NN, Backpropagtion NN, Learning Vector Quantization NN***)

The data set consists of 240 sample recorded from 4 different person, each one say 6 words in different 10 utterances. Features are extracted after transforming the samples into wavelet form.

This chapter will focus on the methodology that is used in this work, at which the main phases used to construct suggest voice recognizer (preprocessing step for data preparation, feature extraction, training classifier and finally testing the system accuracy) are illustrated.

Matlab is used as a tool for feature extraction and classifiers (supervised NN), construction since it is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation.

3.2 Data Set

The data set used in this work consists of 240 different sounds recorded from four different persons (2 males, and 2 females) using Microsoft™ sound recorder, each one says 6 different words in different 10 utterances, and the words are (computer (C), software (S), printer (PR), operation (O), scanner (N), and training (T). With data from two males and two females, the networks should be more robust and able to distinguish between speakers of both sexes with equal accuracy. In addition, by recording each sample at separate times, we made sure that each voice sample was pronounced independently of the previous samples. Hypothesizing that by doing this, the network works more tolerant to the changes in a person's voice that occur over a short course of time.

3.3 The Suggested System Model

The main characteristic of any sound recognition system is the determination of the person who speaks. Sound recognition mainly consists of several phases in addition to the classification engine itself. In this work, a sound recognition system is developed using three different approaches of NN as classifiers. The behavior of these classifiers is to be investigated.

The developed system mainly consists of three phases as shown in figure (3.1), while figure (3.2) shows the flow control for the developed system.

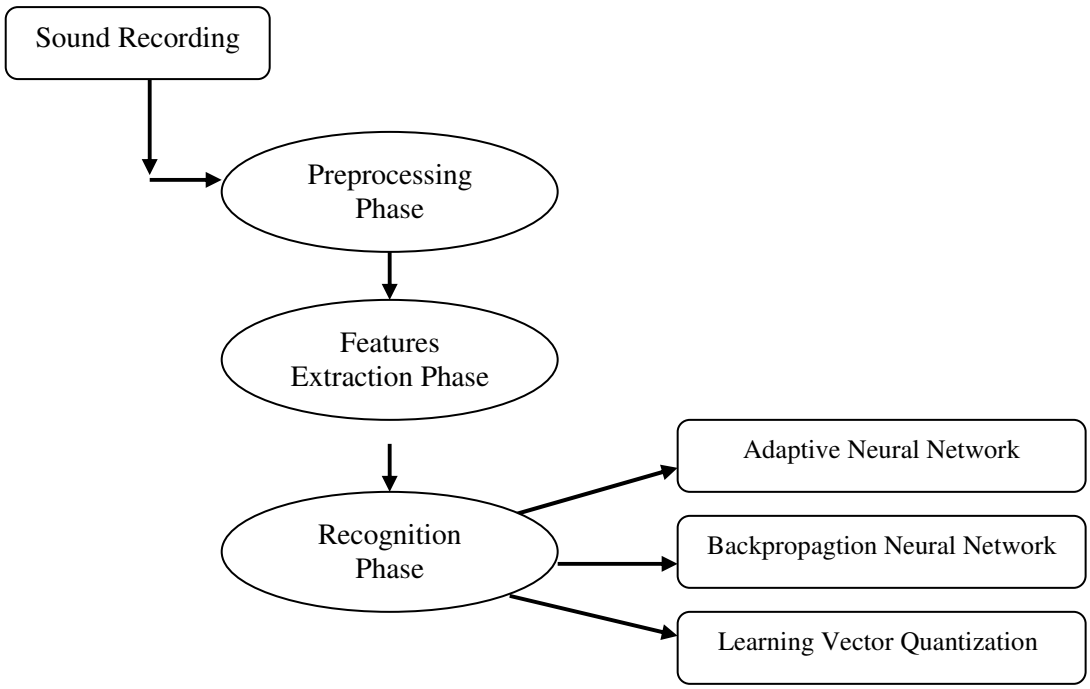


Figure (3.1): The Developed System Phases

In the figure (3-2) shown the developed system flow control, from the first when making sound recording after that applying the preprocessing phase then features extraction, finally the recognition phase which contain three classifier algorithms the first one Adaptive neural network, and the second the Backpropagtion neural network, and the last one the Learning Vector Quantization neural network.

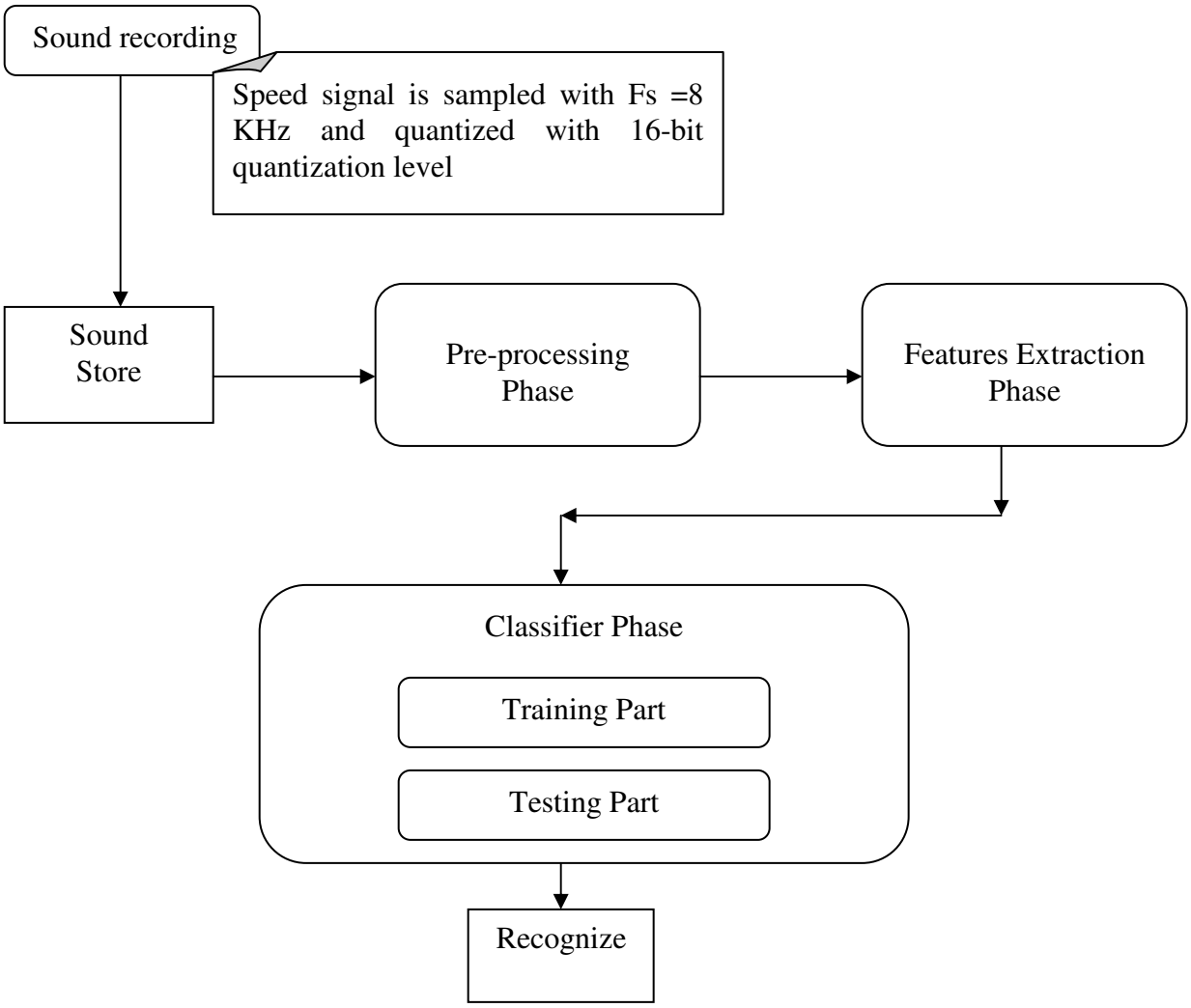


Figure (3.2): The Developed System Flow Control

3.3.1 Preprocessing Phase

Before starting of preprocessing phase, sound are recorded from different persons, these sounds is considered as the input for preprocessing phase, two processes are performed on the sounds (as shown in figure (3.3)):

1- Noise removal (removing unnecessary noise from the sound), the low pass filter for this purpose it used.

2- Non-speech information removal (removing the silent area from the sound), the Sound Forge for this purpose is used.

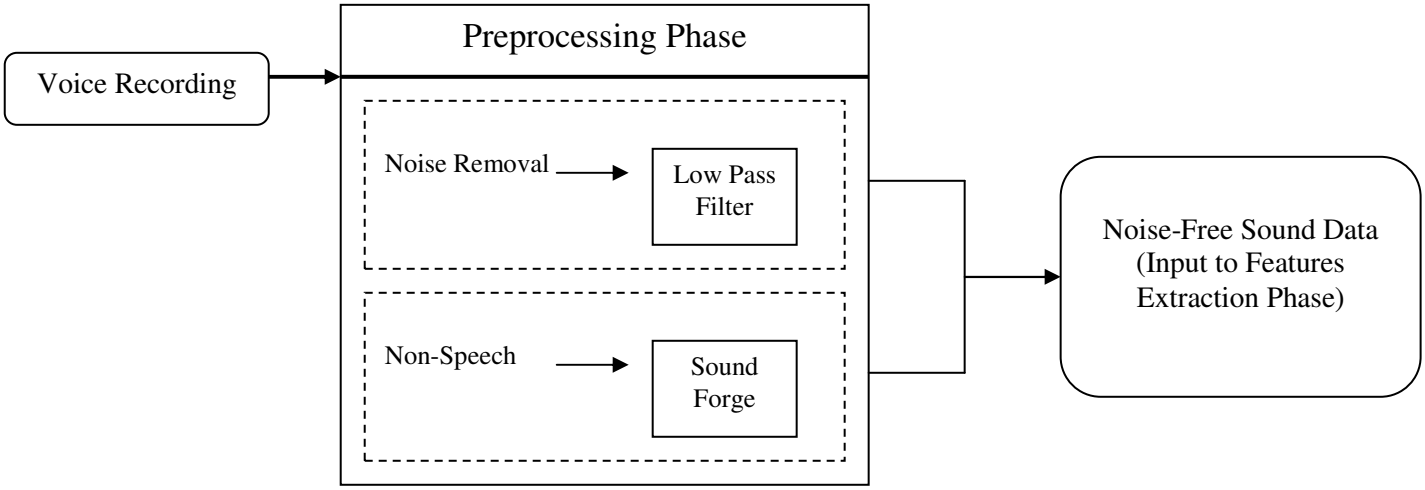


Figure (3.3): Preprocessing Phase

3.3.1.1 Voice Recording

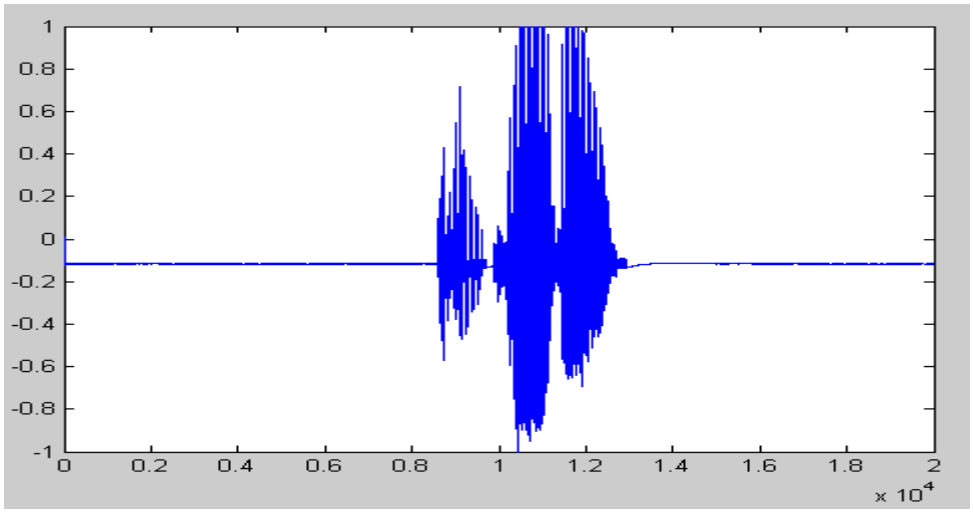
A microphone system is used to pick up the voiced signal and convert it into electric signal entered to the computer. To reduce the information lost of a speech signal, the parameter of the data acquisition should be selected according to the nature of the speech signal to be processed, for telephone-band speech: -3dB LPF at 3.4KHz with more than 24dB/octave attenuation, 8-10KHz sampling frequency, and 12-16 bits. Thus, the speech signal is sampled with Fs = 8 KHz and quantized with 16-bit quantization level.

3.3.1.2 Noise & Non-Speech Information Removal Procedure

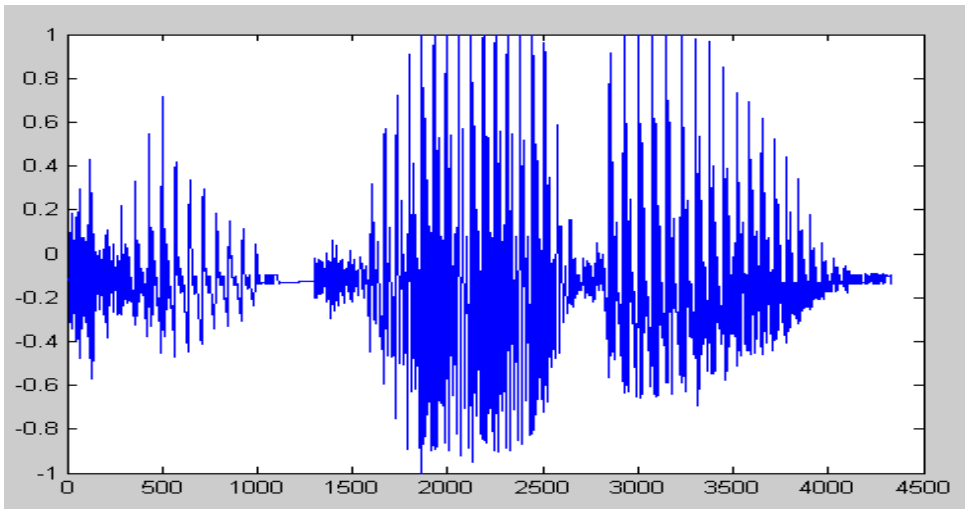
In order to remove ***noise***, Infinite Impulse Response (IIR) Butterworth low pass filter of a cutoff frequency about 500 HZ is applied to the signal (using MATLAB signal processing toolkit).

To remove the ***non-speech information***, the commercial package for this purpose its Sound Forge is used. This software provides a full audio processing and analyzing package.

The unwanted signals are removed and non-speech information also removed, the remaining is the pure human sound signal that will lead to good training results and so to good recognition. The example in figure (3.4) and figure (3.5) demonstrates the signal shape after applying non speech information removal and noise removal for the word "Computer".

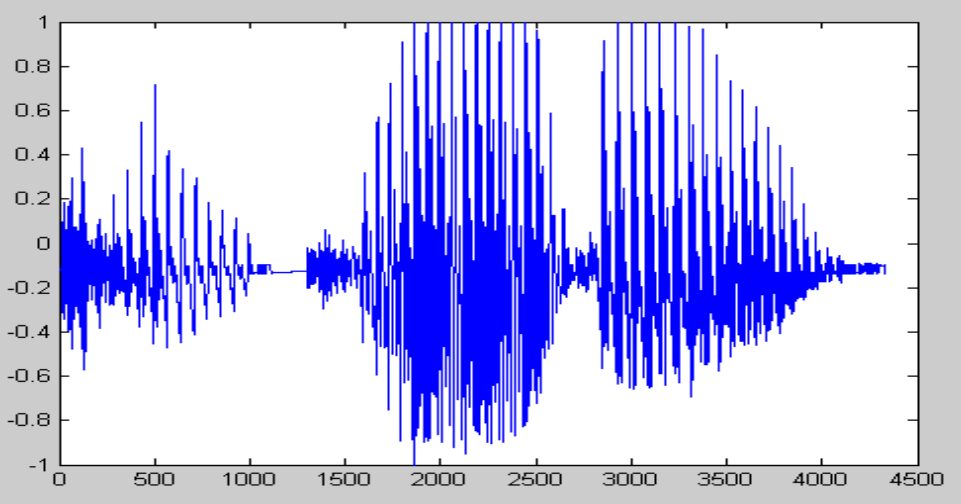


a) (Before removing non-speech information)

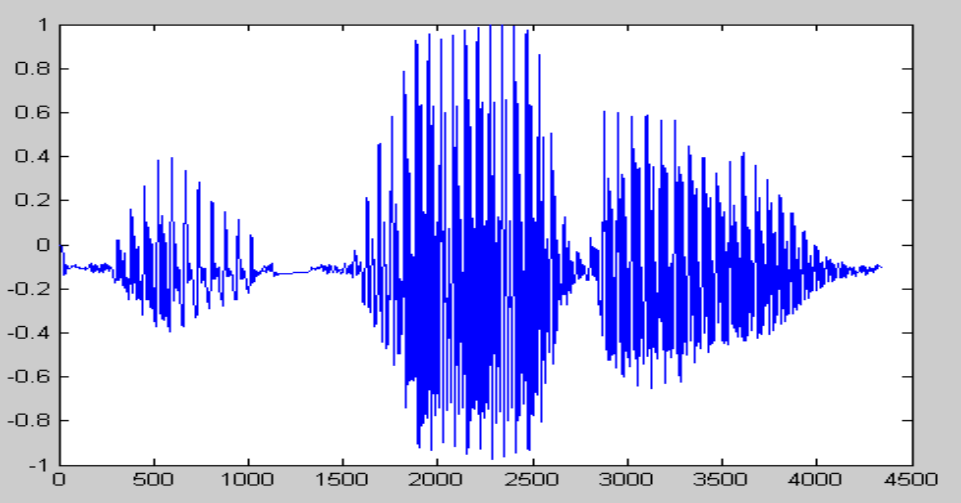


b) (After removing non-speech information)

Figure (3.4): Non-Speech Information Removals



a) (Before removing noise)



b) (After removing noise)

Figure (3.5): Applying Noise Removals

The output from this phase is a Noise-Free sound, to be entered as an input to the feature extraction phase.

3.3.2 Feature Extraction Phase

The second step is extracting the features from the sound, the features that categories and distinguish each sound from the other. Feature could be extracted using different methods, in this work features are extracted after applying WT. To do this, MATLB wavelet toolkit on noise-free sounds is used. Figure (3.6) describes features extraction process.

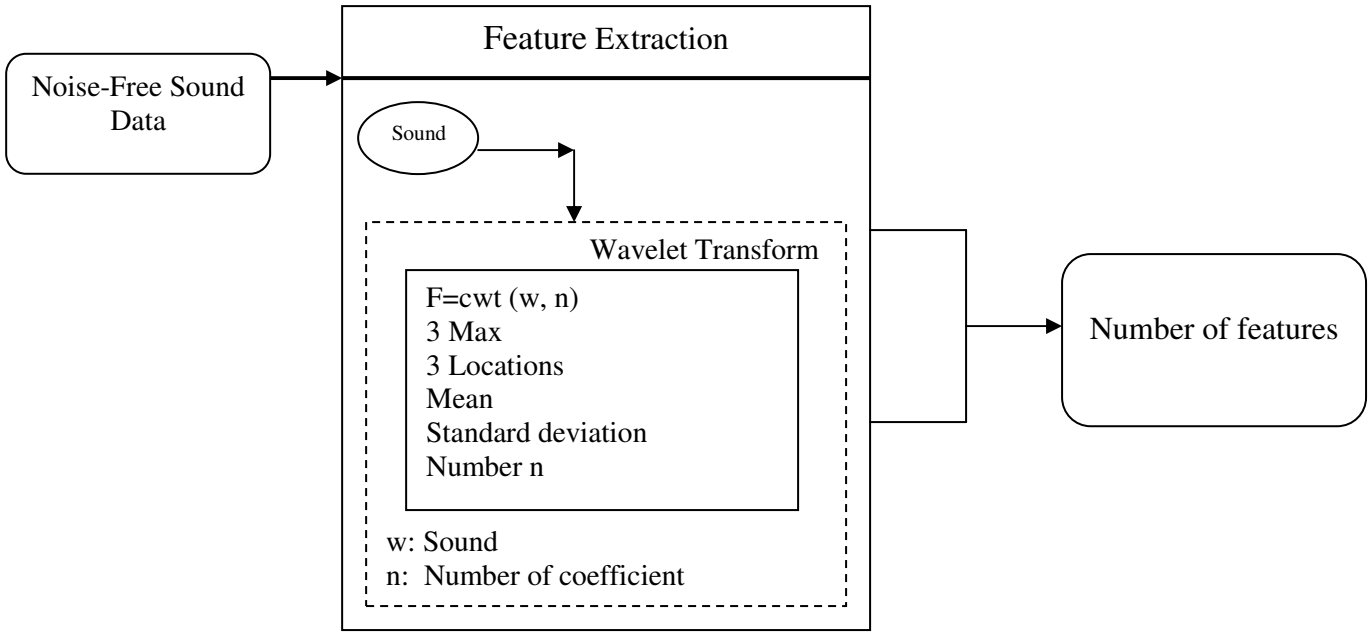


Figure (3.6) Feature Extraction Phase

After reading the sound file and removing noise and Non-speech information, CWT that have a fifth order Daubechies Wavelets is applied. Figure (3.7) shows the histogram for "computer" word pronounced by “men 1”, while figure (3.8) shows wavelet decomposition and WT tree for the word. The wavelet transformation is implemented using MATLAB. The result is an array of n×m dimension, **m** represents number of samples, **n** represents number of levels. In this work, n is assigned value (3, 5, or 7) for NN.

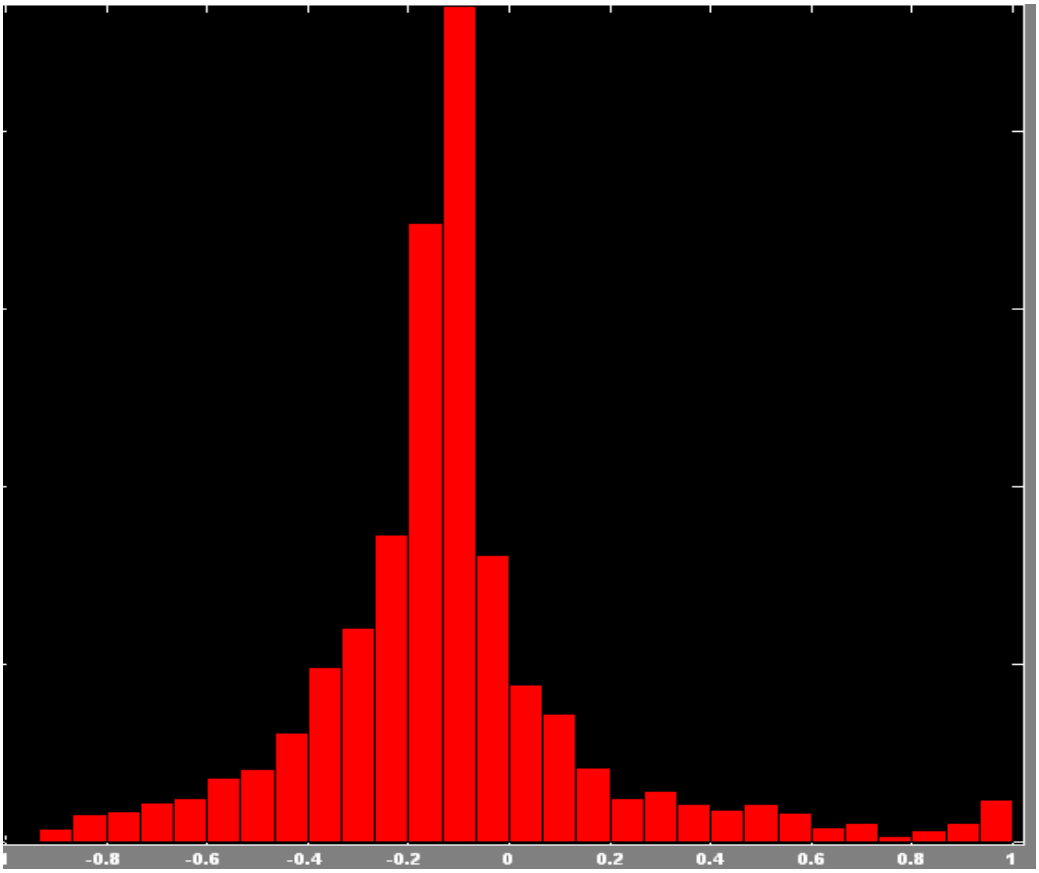
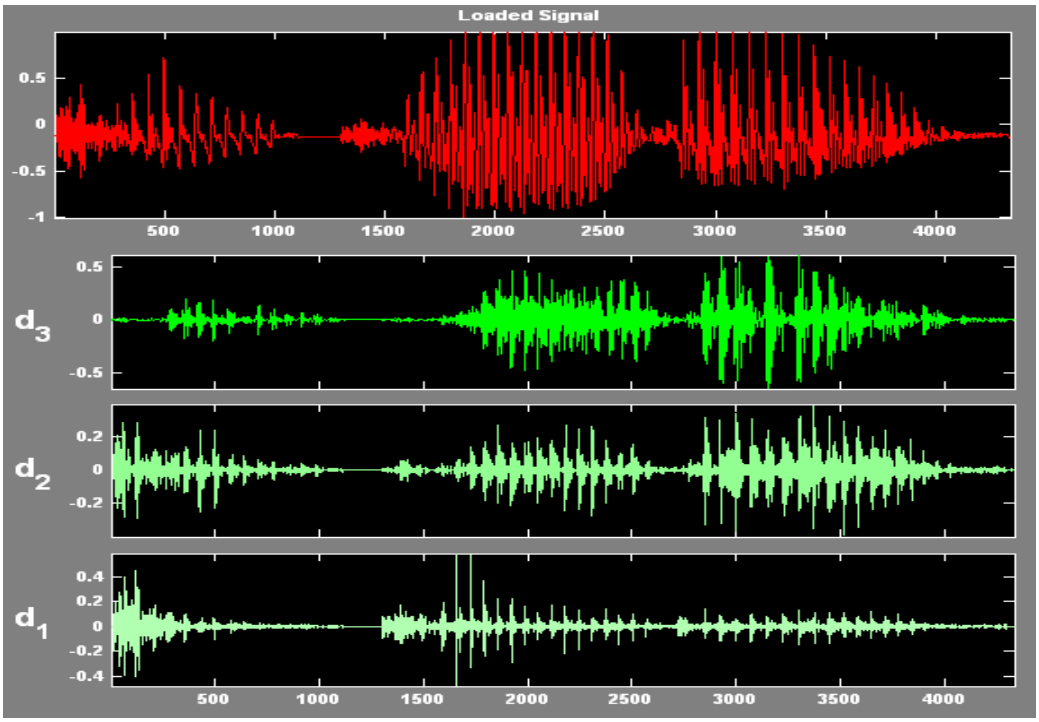
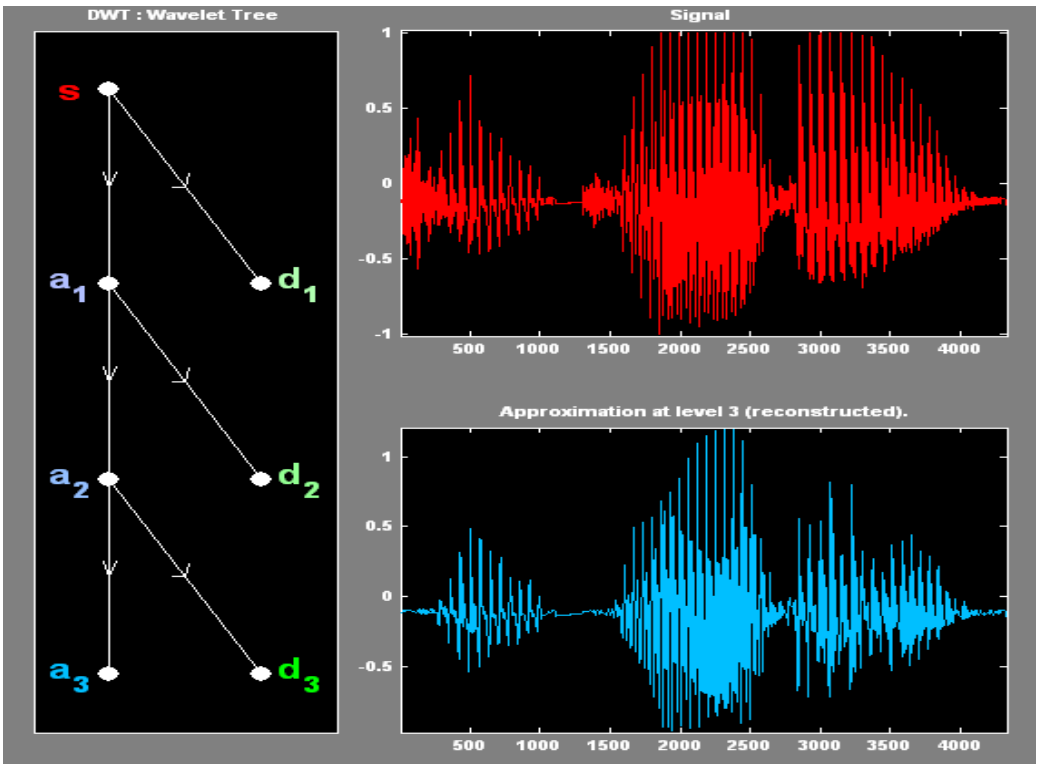


Figure (3-7): The Histogram for Computer Word from M1



(a) Wavelet Decomposition



(b) Wavelet Decomposition Tree

Figure (3.8): Wavelet Decomposition, Wavelet Decomposition Tree for the Word Computer

The resulted array represents the feature of the sound. To reduce feature array dimensions, one max and its location and number of level from each of the 5 levels, are extracted since most of the sound feature is localized at the max value of each level. Then, to improve the classification ability, three max and their corresponding locations and number of level are extracted from each level, but those trial did not lead to optimal result so, for further improvement mean and the standard deviation are calculated (for each level) and used as features in addition to the max three values with their corresponding locations, which means, 8 features are considered from each level.

For example we have 5000 sample with 5 level, number of values (if it is considered as features) that will enter the NN is about 5000*5=25000, this no is very large that will affect the speed and cost of NN processing time, but after taking the maximum three values from each level and their location and the mean and the standard devotion,

form each level the result will be 8 variables for each level, for further improvement, let the level number to be another input to the NN.

Finally, the result will be 9 features for each level, so the final input array will be as follow:-

[M11, M21, M31, L11, L21, L31, ME1, STD1, L1]

M11:-the first maximum number.

M21: the second maximum number.

M31: the third maximum number.

L11:-the location for M11.

L21: the location for M21.

L31: the location foe M31.

ME1: the mean for the first level.

STD1: the standard devotion fro the first level.

L1: Number of level.

3.3.3 Recognition Phase

To perform the recognition phase, three classifiers are considered, Adaptive Neural Network, Feed-forward Neural Network (Back propagation), and Learning Vector Quantization Neural Network. For all used classifiers, each classification operation mainly consists of two phases (training phase, and testing phase).

Training phase: each classifier will be trained by set of training patterns (set of feature vector) to partition the feature space in a way that maximizes the discrimination ability. For NN, training ability is to construct proper weight vectors that correctly classify the training set within some defined error rate.

Testing phase: the trained classifier assigns the unknown input pattern (feature vector) to one of the classes (persons) based on the extracted feature vector.

3.3.3.1 The Adaptive Neural Network

In this work, an Adaptive Neural Network that responds to changes in its environment as it is operating. Adaptive networks that are adjusted at each time step based on new input and target vectors can find weights and biases that minimize the network's

sum-squared error for recent input and target vectors. Figure (3.9) shows the architecture of the ADALINE Neural Network used in the developed system.

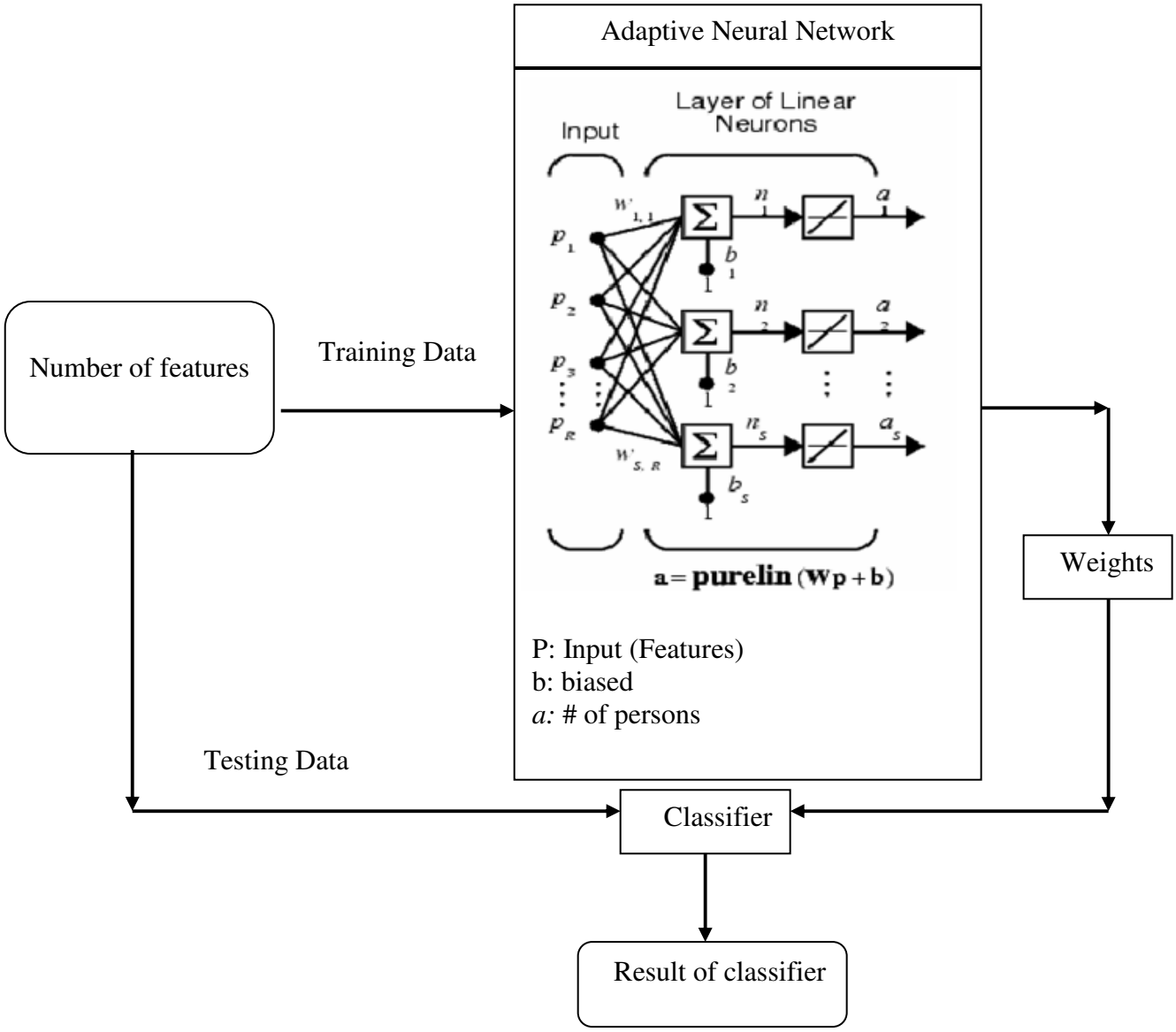


Figure (3.9): The Proposed Architecture for the Adeline Neutral Network

Adaptive Neural Network – Inner Process:

- **Input: Training Sound**

Initially, prepare the training data so that each input pattern (for already known person) is associated with the desired output pattern (i.e. each output node correspond to a person for that the desired pattern will set the node that represent that person to 1 and the other nodes to -1).

Finally an Adaptive neural network will be constructed out of those sounds features, a number of parameters will be provided in constructing and training the neural network, such as number of passes that was assigned to a value between 400 and 1000 pass (the value depends on recognition results), and the learning rate that was assigned to between 0.1 and 0.001.

The main output is the resulted weight matrix which represents the classifier and will be used in testing phase.

- **Input: Testing Sound**

Extract the feature set for the selected testing sounds (for known persons), then use the weight matrix resulted from the training phase. Count number of correctly classified inputs (feature vectors) and calculate the classification accuracy by using the formula (3.1):

Number of correctly classified patterns/ total number of patterns …… (3.1).

3.3.3.2 Feed-Forward Backpropagation Neural Network

Backpropagation Neural Network was created by generalizing the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function (associate input vectors with specific output vectors) or reach high classification accuracy. Networks with biases, a sigmoid layer, and a linear output layer are capable of approximating any function with a finite number of discontinuities, Figure (3.10) shows the architecture of the Backpropagtion Neural Network used in the developed system.

Backpropagtion Neural Network – Inner Process:

- **Input: Training sound**

The same steps used to train the ADALINE NN are used for training the Backpropagtion NN. The most distinguishing property in Backpropagtion neural network is the use of a hidden layer. ***The strategy used to choose the proper number of nodes in the hidden layer is by using pruning property*** (*starting from small number of nodes (start with*

8) to solve the problem, if the network does not converge or the classification rate is low, then increase number of nodes in the hidden layer by one. The process is repeated until reaching the best classification rate). Since there are hidden layer, then two ***activation function*** are used: between the first layer and the hidden layer, ”tansig” activation function is used, while between the hidden layer and the output layer, the "purelin" activation function or ”tansig” activation function is used.

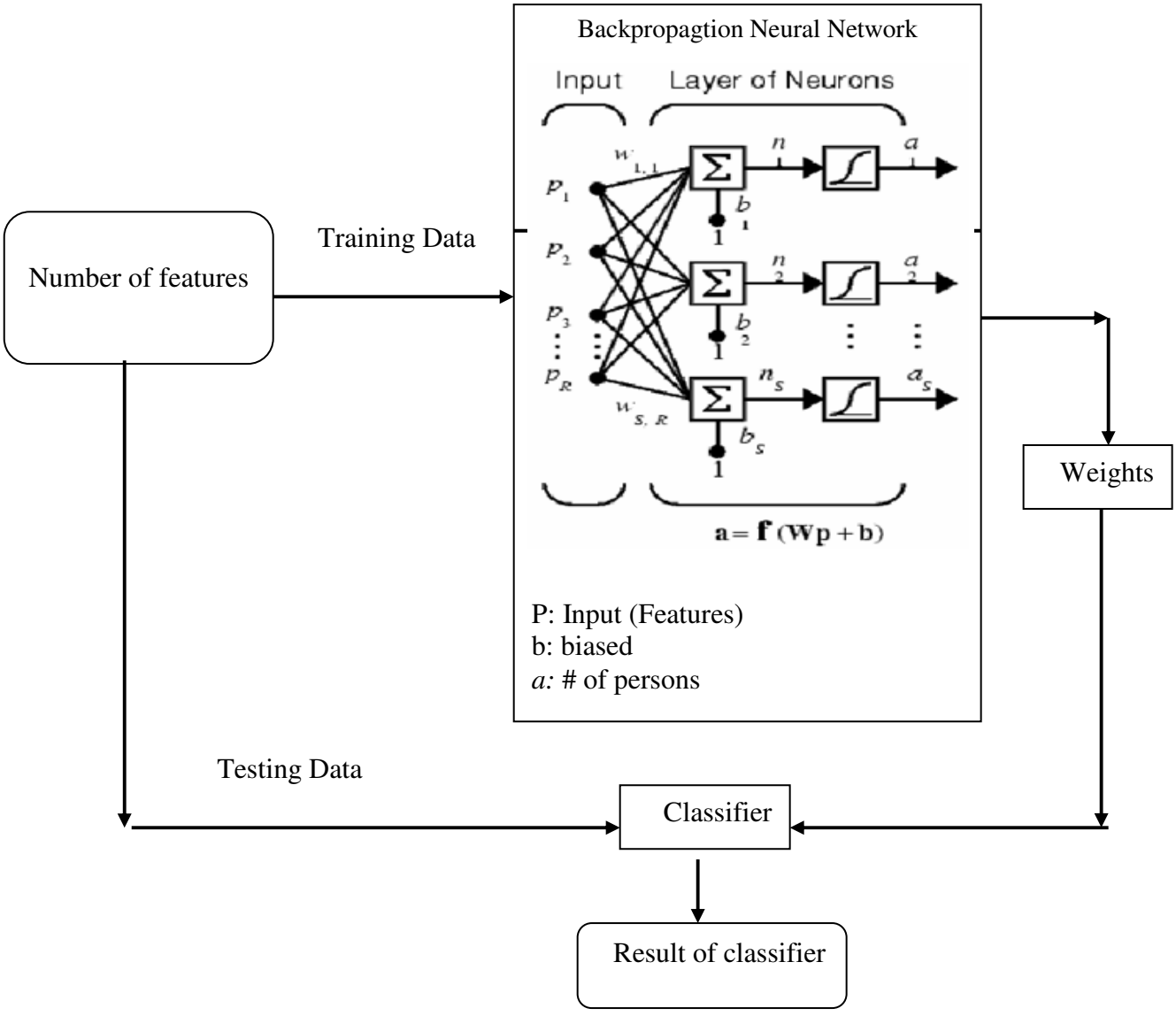


Figure (3.10): The Proposed Architecture for the Backpropagtion Neutral Network

As ***stopping condition***, The net will stop learning when it reach certain number of epochs (number of epochs that was assigned to a value between 400 and 1000 pass

depending on recognition results). The ***learning rate*** α was assigned to between 0.1 and 0.001.

The main output is the weights for these sounds that will be used in testing phase.

- **Input: Testing sound**

  Use the same strategy used with ADALINE NN.

3.3.3.3 Learning Vector Quantization Neural Network

LVQ can be understood as a special case of an NN, the network classify vectors into target classes by using a competitive layer to find subclasses of input vectors, and then combining them into the target classes (Liner layer), the network is given by prototypes W=(w(i),…,w(n)). It changes the weights of the network in order to classify the data correctly. For each data point, the prototype (neuron) that is closest to it is determined (called the winner neuron). The weights of the connections to this neuron are then adapted. An advantage of LVQ is that creates prototype that are easy to interpret for experts in the field. Figure (3.11) shows the architecture of the LVQ Neural Network used in the developed system.

Learning Vector Quantization -Inner Process:

The same steps used to train the ADALINE NN are used for training the LVQ NN. The most distinguishing property in LVQ neural network is the use of a hidden layer. ***The strategy used to choose the proper number of nodes in the hidden layer is by using pruning property*** (*starting from small number of nodes (start with 8) to solve the problem, if the network does not converge or the classification rate is low, then increase number of nodes in the hidden layer by one. The process is repeated until reaching the best classification rate*). Since there are just one hidden layer**.**

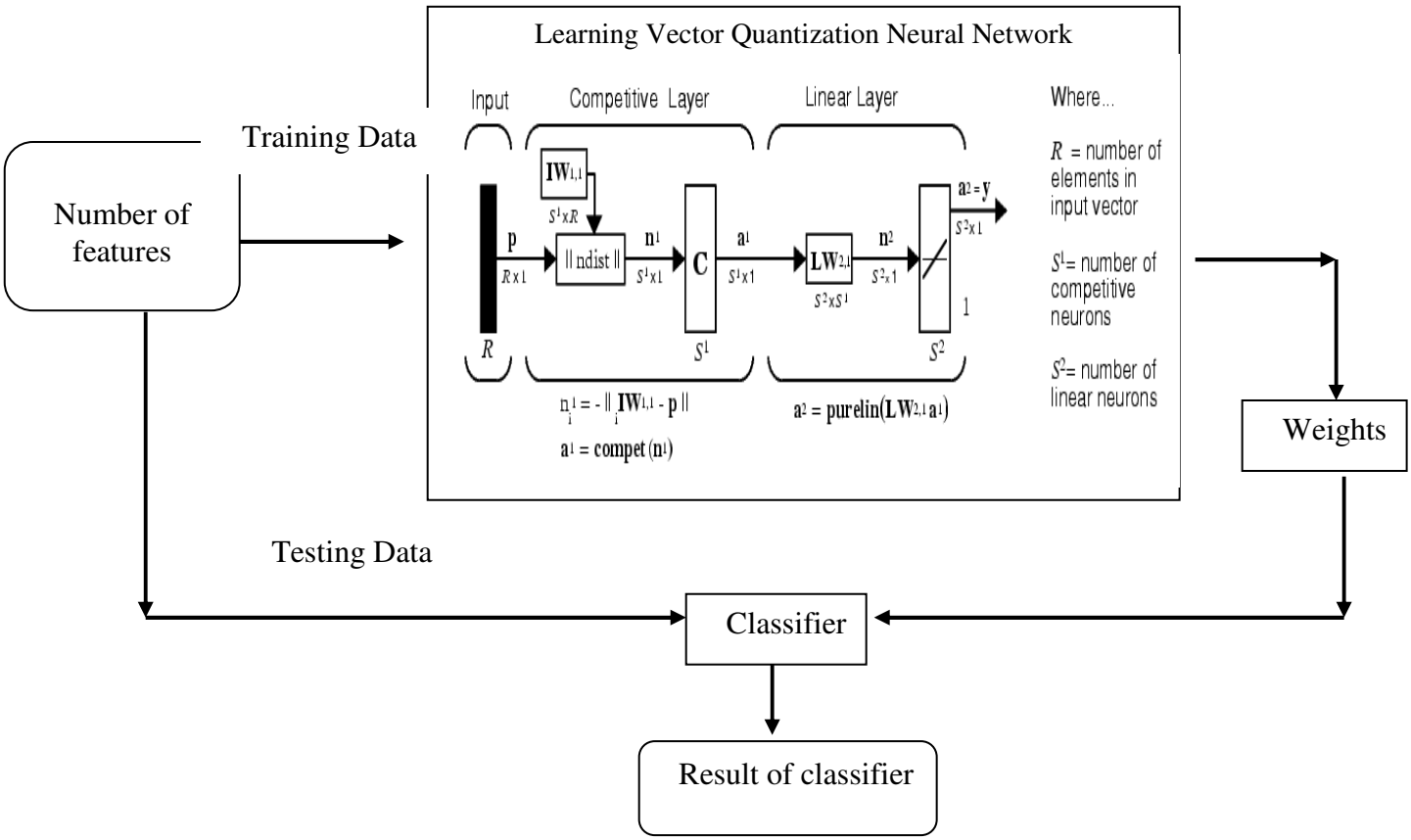


Figure (3.11): The Proposed Architecture for the LVQ Neutral Network

- **Input: Testing sound**

  Use the same strategy used with ADALINE NN.

CHAPTER FOUR
ASSESSMENT RESULTS

Chapter Four

Assessment Results

4.1 Introductions

The aim of this work is to implement a number of neural network architectures (proposed for speaker recognition purposes) and evaluate their applicability and performance in voice identification. These classification results need to be analyzed and evaluated to decide:

- Which feature-set has the best discrimination nature
- Which one of the neural networks (classifiers) gives the best classification accuracy for the voices used in this work

Three types of neural networks were considered be investigated. The chosen neural nets are:

- Adaptive Neural Network
- Learning Vector Quantization Neural Network
- Feed-forward Back-propagation Neural Network.

The classification ability of the above neural networks is studied and compared from classification accuracy point of view, at which each of these classifiers was trained with various types of features (one feature set at a time).The selected feature sets are extracted based on Wavelet transformation (Db5):

- Three Features (1 max with its corresponding location and level number) per level
- Seven Features (3 max with their locations and level) per level
- Nine Features (3 max with their locations correspondingly, mean, STD, level number) per level.
- For all experiments, features are extracted from three and five levels , but for Text-Independent (4 person) features are extracted from 7 level

Using different sets of features would help in determining the feature-set with best voice discrimination ability.

Convergence speed comparison will be ignored since the important factor in pattern classification problems is classification accuracy. To evaluate the performance of any neural network recognition system, the accuracy of the system result should be calculated as follows:

$$\frac{\textit{Number of correctly classified patterns}}{\textit{Total number of testing patterns}} \qquad (4.1)$$

This chapter will explore the experimental results and comparison between various NN classifiers. Analysis and assessment of the results will be illustrated.

4.2 Performance Evaluation

To design any pattern recognition system, two fundamental problems should be handled:

- **Model Selection** *(choose the model that provides the lowest error rate)*: for the three used neural networks, the main parameters are ***neural architecture*** (number of hidden layers, number of neurons/hidden layer), and ***training factors*** (learning rate, stopping condition). The modulation selection is done through trail and error, at which different trials were tested to tune the classier parameters. To find the best parameters, procedure below was followed:
  1. Divide the available data into training, and test set
  2. Select architecture and training parameters
  3. Train the model using the training set
  4. Evaluate the model using the testing set
  5. Repeat steps 2 through 4 using different architectures and training parameters
  6. Select the best model and train it using the train data
  7. Assess this final model using the testing sets
- **Performance Estimation** (*error rate is the true error rate*): Once the model selection and training is completed, its generalization performance needs to be evaluated on previously unseen data to estimate its true performance on field data. The most popular

methods for evaluating the generalization performance is to split the entire training data into two partitions, where the first partition is used for actual training and the second partition is used for testing the performance of the algorithm. The performance on this latter dataset is then used as an estimate of the algorithm's true (and unknown) field performance. Different approaches could be used:

- **One approach is to use the entire training data** to select the classifier and estimate the error rate. This approach suffers from the over-fit of training data, leading to error rate estimate which will be overly optimistic (lower than the true error rate).
- **Holdout Method** (split the training data into disjoint subsets): for a single train-and-test experiment, the holdout estimate of error rate will be misleading if "unfortunate" split happens. The limitations of the holdout can be overcome with a family of resampling methods at the expense of more computations cross-validation, random-subsampling, K-Fold cross-validation, leave-one-out cross-validation.

In this work, k-fold cross validation is used. The main advantages of this algorithm are that all data instances are used for training as well as for testing (in different times) allowing full utilization of the data. Since the procedure is repeated k times, the probability of an unusually lucky or unlucky partitioning is reduced through averaging.

To perform the k-fold cross validation, the entire available dataset is split $k > 2$ partitions, creating k blocks of data. Of these k ($k=5$ in this work) blocks, Repeat the procedure k times, using different block for testing in each case. The average of the k test performances is calculated, and is declared as the estimate of the true generalization performance of the algorithm.

4.3 Experimental Results: Compression

The data set is partitioned into two parts (training part and testing part). Three different Neural Nets are used as classifiers, each of which is trained and tested. For evaluation purposes, the classifiers obtained from training phase are tested using new sounds in addition to the training set. To evaluate efficiency of each suggested NN, two approaches are used:

- **Text-dependent** at which the speaker talks the same word in two or more different utterances. The classifier is tested with the same words used in training.
- **Text-independent** in which the speaker talks different words. The classifier is tested with new words different that the words used for training.

For ***text-dependent*** speaker recognition, 4 person are considered, each one says 6 words in different 10 utterances, the words are (computer(C), software(S), printer (PR), operation (O), scanner (N), and training (T)), the summation of all words is 240 word from two males, and two females, man is denoted as (M1, M2), and women in (W1, W2). These sets are divided into training data and test data as in table (4.1).

Table 4.1: Types of Data Files – Text Dependent

# of Samples	# of Persons	# of Training Voice %	# of New Words for Testing/Validation Voice %	Total Words
1	2 males (M1, M2)	60	60	120
2	2 females (W1, W2)	60	60	120
3	Male & female (M2,W2)	60	60	120
4	Two males, one females (M1, M2, W1)	90	90	180
5	One males, two females (W1, W2, M1)	90	90	180
6	Two males, two females (M1, M2, W1, W2)	120	120	240

For ***text-independent*** speaker recognition, the words are (hardware (H), testing (TE), programmer (P), information (I)), these word recorded 6 times in different utterances from 4 person two males (M1, M2,) and two females (W1, W2). The set of sounds is divided into training data and test data as in table (4.2)

Table 4.2: Types of Data Files – Text Independent

# of Samples	# of Person	# of Training Voice %	# of New Words for Testing/ Validation Voice %	Total Words
1	Two males &Two female	96	96	192

4.3.1 Text Dependent

This section present and compare the experimental results from the three networks: ADALIN, Backpropagation (one and two hidden layers), and LVQ. The compression will be from accuracy point of view. Taken in consideration different features sets calculated from continuous wavelet transform (CWT).

4.3.1.1 Adaptive Neural Network Text-dependent

The adaptive neural network describe in chapter three is used as classifier for speaker. To study its behavior, as a speaker recognizer, the neural net was trained with the features set illustrated in section (3.3.2), table (4.3) illustrates the classification accuracy of the net with each feature set.

From this table, it's clearly seen that the net has poor classification ability since it is always can recognize only one class (one class represent one person). The same results was reached with different feature sets (3, 7, 9), extracted from different levels (3, 4, 5, 6, 7), and with different learning rates [0.1, 0.001], and different number of epochs [500, 1000].

Table (4.3): Classification accuracy for ADALINE Text-dependent

Data sate	Training %	Testing %
2 males	50	50
2 females	50	50
1 male 1 female	50	50
2males 1 female	33.33	33.33
2females 1 male	25	25
2 males 2 females	25	25

4.3.1.2 Backpropagtion Neural Network Text-dependent

Chapter three illustrate the main architecture and inner process of Backpropagtion neural network used in this work. To study the behavior of the net, the neural net was trained with the features sets illustrated in section (3.3.2). Table (4.4) shows the classification accuracy with feature set one (three feature calculated from 5 levels i.e. 15 feature per sound except for (two males two females) seven level were token with total feature set 21 features) at which different number of hidden layers (1 and 2), and hidden neurons were tested ranging from 8 to 13 neurons. The best numbers of hidden layers and neurons per each layer are shown in the table.

Table (4.4): Classification accuracy for Backpropagtion for feature set one Text-dependent

# of Levels	Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%
5 levels	2 males	1	0.007	400	9	95.00	80
	2 females	1	0.003	700	9	93.33	81.66
	1 males 1 females	1	0.005	800	10	80	78.33
	2 males 1 females	2	0.009	700	[10 11]	93.33	70
	2 females 1 males	2	0.009	700	[11 10]	88.88	73.33
7 levels	2 males 2 females	2	0.01	600	[10 11]	88.33	69.166

Table (4.5) shows the classification accuracy with feature set two (seven features calculated from 5 levels i.e. 35 feature per sound except for (two males two females) seven level were token with total feature set 49 features) at which different number of hidden layers (1 and 2), and hidden neurons were tested ranging from 8 to 13 neurons. The best numbers of hidden layers and neurons per each layer are shown in the table. While Table (4.6) shows the classification accuracy with feature set three (nine features calculated from 5 levels i.e. 45 feature per sound except for (two males two females) seven level were token with total feature set 63 features) at which different number of hidden layers (1 and 2), and hidden neurons were tested ranging from 8 to 13 neurons. The best number of hidden layers and neurons per each layer are shown in the table. The table contains additional test at which the classification accuracy is calculated for all data set (training and testing) is used

Table (4.5): Classification accuracy for Backpropagtion for feature set two Text-dependent

Number of Levels	Data Sets	Hidden Layer	Learning Rate	Epoch	Hidden Neurons	Training%	Testing%
5 levels	2 males	1	0.007	400	9	100	85
	2 females	1	0.001	700	11	98.33	81.66
	1 male 1 females	1	0.005	900	11	86.6	80
	2male 1 females	2	0.009	700	[11 11]	96.66	80
	2 females 1 male	2	0.009	700	[11 11]	98.88	81.11
7 levels	2 males 2 females	2	0.1	600	[10 11]	92.50	75.83

Table (4.6): Classification accuracy for Backpropagtion for feature set three Text-dependent

# of Levels	Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%	All data Testing%
5 levels	2 males	1	0.007	400	9	100	91.66	97.50
	2 females	1	0.005	600	9	100	90	95.00
	1 male 1 female	1	0.06	800	13	100	88.33	94.16
	2 males 1 female	2	0.07	800	[10 5]	96.66	84.44	91.11
	2 females 1 male	2	0.1	850	[7 10]	97.77	84.44	90.55
7 levels	2 males 2 females	2	0.9	500	[10 5]	90	80.83	88.75

To specify the best feature set with Backpropagation neural net, the classification accuracy for the net with different feature sets is compared (see chart 4.1). From Chart (4.1) one can deduced the following:

- The best feature set that suites the backpropagation neural net (i.e. gave the best speaker recognition accuracy) is feature set three (9 features).
- Feature sets one and two give almost the same accuracy results for (two females), (one male one female). While feature set three gives batter classification accuracy for the reset data set.
- The worst result is with feature set one for (two males one females), (two males two females).

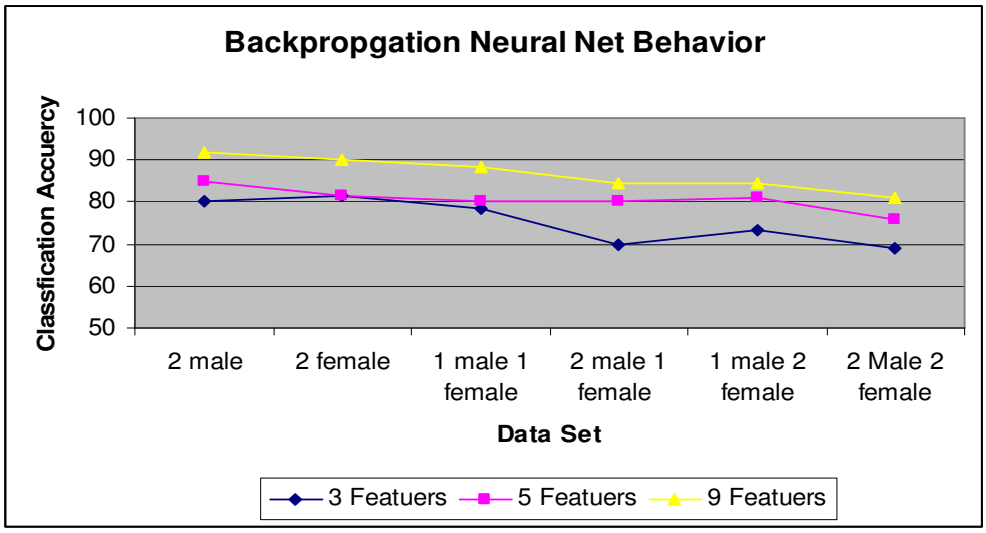


Chart (4.1): Backpropagation Neural Net Behavior

4.3.1.3 LVQ Neural Network Text-dependent

In the previous chapter we illustrated the main architecture and inner process of LVQ neural network used in this work. To study its performance, the LVQ net was trained with the features sets illustrated in section (3.3.2). Table (4.7) shows the classification accuracy with feature set one (three feature calculated from 5 levels i.e. 15 feature per sound except for (two males two females) seven level were token with total feature set 21 features) at which one hidden layers and different number of hidden neurons were tested starting from 6 as minimum number of hidden node to 120 as maximum number of hidden node. The best numbers of neurons per hidden layer are shown in the table.

Table (4.8) shows the classification accuracy with feature set two (seven features calculated from 5 levels i.e. 35 feature per sound except for (two males two females) seven level were token with total feature set 49 features) at which one hidden layers and different number of and hidden neurons were tested starting from 6 as minimum number of hidden node to 120 as maximum number of hidden node. The best numbers of neurons per hidden layer are shown in the table.

Table (4.7): Classification accuracy for LVQ with feature set one Text-dependent

# of Levels	Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%
5 levels	2 males	1	.001	3800	38	76.66	68.33
	2 females	1	0.001	4000	40	71.66	71.66
	1 male 1 female	1	0.001	3500	33	61.66	51.66
	2 males 1 female	1	0.001	3500	33	57.77	57.77
	2 females 1 male	1	0.001	4000	30	55.55	55.55
7 levels	2 males 2 females	1	0.001	1350	95	49.16	43.33

Table (4.8): Classification accuracy for LVQ with feature set two Text-dependent

# of Levels	Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%
5 levels	2 males	1	0.001	3000	30	80	75
	2 females	1	0.001	3500	33	70.00	71.66
	1 males 1 females	1	0.001	4000	33	61.66	56.66
	2 males 1 females	1	0.001	3500	38	54.44	55.55
	2 females 1 males	1	0.001	3500	37	55.55	53.33
7 levels	2 males 2 females	1	0.001	1100	80	51.66	45.8

Table (4.9) shows the classification accuracy with feature set three (nine features calculated from 5 levels i.e. 35 feature per sound except for (two males two females) seven level were token with total feature set 63 features) at which one hidden layers and different number of and hidden neurons were tested starting from 6 as minimum number of hidden

node to 120 as maximum number of hidden node. The best numbers of neurons per hidden layer are shown in the table.

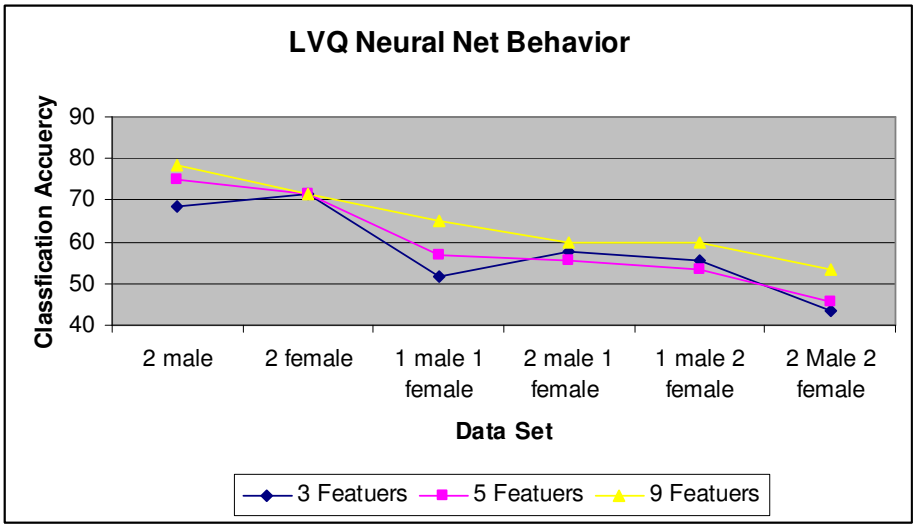
Table (4.9): Classification accuracy for LVQ with feature set three Text-dependent

# of Levels	Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%	All data Testing%
5 levels	2 males	1	0.001	3500	33	81.66	78.33	83.33
	2females	1	0.001	3000	30	70	71.66	78.33
	1 male 1female	1	0.001	3600	33	68.33	65.00	70
	2 males 1 female	1	0.001	3000	38	62.22	60	66.66
	2females 1 male	1	0.001	3500	35	62.22	60	68.33
7 levels	2 males 2females	1	0.001	1000	75	58.33	53.33	61.66

To study the behavior of LVQ with the different feature set, the classification accuracy with different feature sets is compared (see chart 4.2). From Chart (4.2) one can deduced the following:

- The best feature set that suites the LVQ neural net (i.e. gave the best speaker recognition accuracy) is feature set three (9 features). For (two females) all feature sets gives the same result.
- Feature set one and feature set two are comparable for all data set expect (two males) and (one male one female).
- Feature set one gives better result than feature set two with (two males one female) and (two females one male).

Chart (4.2): LVQ neural net behavior



4.3.1.4 Best Classifier Neural Net

From the result shown in the previous section its obvious that feature set three gives the best classification result for Backpropagtion and make no deferens for ADALINE. Therefore, the result for feature set three will be used for studying the behavior of three net (comparing the result). It's clearly seen from chart (4.3) that:

- The backpropagation neural net gives the best result for all data set.
- All three neural net give the best result for (two males) data set and the worst classification accuracy for (twp males two females) although the feature set were calculated from the 7^{th} level.

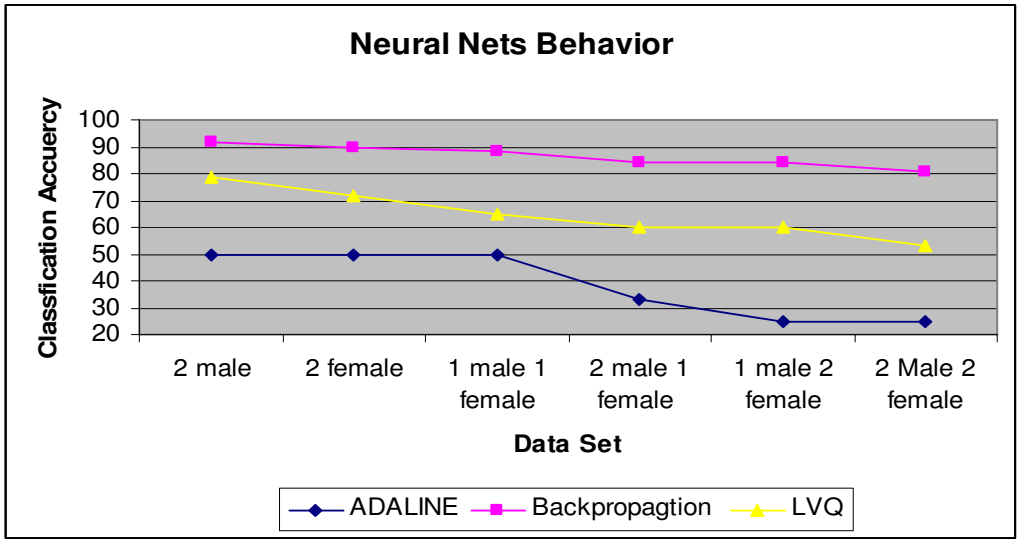


Chart (4.3): Neural Nets Behavior

4.3.2 Text-Independent

This section presents and compare the experimental results from the three networks: ADALIN, Backpropagation (one and two hidden layers), and LVQ for text-independent approach. The main different between the implementation of text-dependent and text-independent approach is through testing phase at which for text-dependent the testing was through using the same words with different utterances, while in text-independent the testing will be perform through using the words pronounced by the same person but differ than those used in training phase. (for the each person (computer(C), software(S), printer (PR), operation (O), scanner (N), and training (T)) are used in the training phase, while (hardware (H), testing (TE), programmer (P), information (I)) are used in the testing phase for the corresponding person), The compression will be from accuracy point of view, taken in consideration the classification accuracy with feature set three (since this feature give the best classification accuracy. In text-independent we make training and testing on (two males two females) data set.

4.3.2.1 Adaptive Neural Network Text-Independent

From table (4.10) illustrated below, it's clearly seen that the net has poor classification ability since it is always can recognize only one class. The same results was reached with 9 feature sets extracted from level number 7, and with different learning rates [0.1, 0.001], and different number of epochs [500, 1000].

Table (4.10): Classification accuracy for ADALINE Text-independent

Data Sets	Training%	Testing%
2 males 2 females	25	25

4.3.2.2 Backpropagtion Neural Network Text-Independent

Table (4.11) shows the classification accuracy with feature set three (nine feature calculated from 7 levels i.e. 63 feature per sound) at which different number of hidden layers (1 and 2), and hidden neurons were tested ranging from 8 to 13 neurons. The best numbers of hidden layers and neurons per each layer are shown in the table.

Table (4.11): Classification accuracy for Backpropagtion Text-independent

Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%
2 males 2 females	2	0.009	600	[11 5]	94.79	71.87

4.3.2.3 LVQ Neural Network Text-Independent

Table (4.12) shows the classification accuracy with feature set three (nine features calculated from 7 levels i.e. 35 feature per sound) at which one hidden layers and different number of hidden neurons were tested starting from 6 as minimum number of hidden node to 120 as maximum number of hidden node. The best numbers of neurons per hidden layer is shown in the table. While chart (4.4) shows the behavior of the three neural nets at which it is clearly seen that the Backpropagation gives the best classification accuracy**.**

Table (4.12): Classification accuracy for LVQ Text-independent

Data Sets	#of Hidden Layer2	Learning Rate	# of Epochs	# of Hidden Neurons	Training%	Leave 1 out Testing%
2 males 2 females	1	0.001	1500	80	45.83	40.83

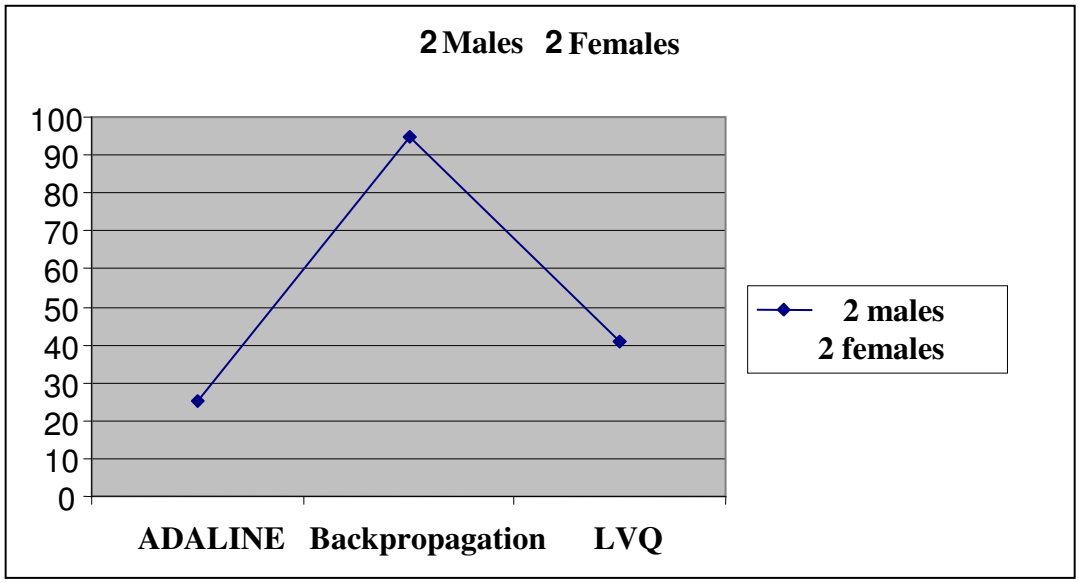


Chart (4.4): Neural Nets Behavior text-independent

CHAPTER FIVE

CONCLUSION & FUTURE WORK

Chapter Five
Conclusion & Future work

5.1 Conclusions

This work concerned mainly with studying the behavior of the neural network as classifier in the field of speaker recognition. In this work ADALINE, Backpropagation and LVQ neural networks were implemented and used to perform speaker recognition. These nets were trained with different feature sets based on wavelet transformation (Continuous Wavelet Transformation CWT). The first part of the suggested recognizer started from the preprocessing phase, which is based on the noise removal and non-speech information removal, then, For feature extraction, wavelet techniques discussed for reducing amount of feature extraction (instead of using all coefficients in each level, only the largest three (since it will contain most of the signal energy in the level) with its position, in addition to mean and standard deviation). Further, a study of the behavior of the suggested architectures of the three NN mentioned above along with the recognition accuracy is illustrated. As testing method, k-fold cross validation technique is used with k=5.

After all the extensive data acquisition, network training, and network testing, several conclusions have been drawn out of this work:

- After transforming the signal using CWT, the features were extracted from different wavelet levels, the most suitable levels were 5 levels and 7 levels.
- Three sets of features were extracted from each level, (1 max coefficient with its corresponding location and level number, three max coefficients with its corresponding locations and level number, 3 max with their locations correspondingly, mean, STD, level number), it was found that the third feature set much better than feature set two and feature set one.
- By comparing the three neural networks from recognition accuracy point of view, Backpropagtion NN is the best classifiers (since it has higher separable capability for nonlinearly separable data. On the other hand showed comparable behavior, while ADALINE NN showed bad recognition ability since it can only learn to recognize 1 class (person) only with all feature sets.

- The result accuracy increased when we adding mean value and Stander deviation value information on the features, the network is able to identify the speaker more accurately. Stander deviation and mean value information also makes the network more robust as more speakers are added into the system.
- The network obviously worked very well for text dependent speaker identification; however, it also showed very promising results for text independent speaker identification. In both text-dependent and text-independent the accuracy of backpropagation NN is better than the accuracy for LVQ NN and ADALONE NN.
- From the tables shown in chapter four, it was found that number of hidden neurons for LVQ should be relatively high, this is due to the fact that the data used is nonlinearly separable, this means that each class (person) need more than one cluster center (i.e. more than one cluster represent one person).
- Although our results were limited by the amount of training data we had, we still found the results to show potential. However, the problem of automatic speaker/voice recognition is very broad field with many problems yet to be solved. Further development on this work includes training the network to recognize unauthorized speakers; however, this is restricted again by the amount of training data that you obtain. We did not try to solve this problem because we simply could not collect enough data to characterize unauthorized speakers.

5.2 Future work

The following are some recommendation for future work to improve recognition ability and perform more studies:

- Add new feature sets calculated using the discrete wavelet transformation (DWT) and discrete cosine transformation (DCT) for feature extraction. In addition to the study of number of coefficient suitable for recognition and the affect of using energy of each level.
- Perform the study on a sentence not a word. Try to find out the person how said the sentence, mainly Identification of a male female, and specify the person.
- Use OCON structure Neural network with self organized or LVQ then a dissention based neural and study its recognition ability.

References

- [Ani 2007] Muzhir Shaban Al-Ani, Thabit Sultan Mohammed and Karim M. Aljebory, 2007, **Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform**, Journal of Computer Science 3 (5): 304-309, 2007 SSN 1549-3636,2007 Science Publications

- [Chief 2005], Chief Security Officer, 2005, **Biometric Authentication, http://www.csoonline.com.au/index.php/id;453581685;fp;8;fpid;8**

- [Demuth 2007] Howard Demuth, Mark Beale Martin Hagan, 2007, **Neural Network Toolbox 5**, User's Guide, 2007, http://www.mathworks.com/products/neuralnet/.

- [Diaphonics 2006], 2006, **ABOUT VOICE BIOMETRICS**, Diaphonics sound security, http://www.diaphonics.com/

- [Fausett 1994] L.Fausett, 1994, **Fundamental of Neural Network**, Printice-Hall International, Inc.

- [Gemello 2006] Roberto Gemello, Franco Mana, Dario Albesano, 2006, **Hybrid Hmm/Neural Network Based Speech Recognition In Loquendo**, ASR

- [Graevenitz 2003] Gerik Alexander von Graevenitz, 2003, **About Speaker Recognition Technology Bergdata Biometrics** GmbH, Bonn, Germany

- [Graps 1995] A.L. Graps; Summer 1995, **An Introduction to Wavelets**, IEEE Computational Sciences and Engineering, Volume 2, Number 2, , pp 50-61

- [Guojie 2004] li Guojie, 2004, **Radial Basis Function Neural Network For Speaker Recognition**, a Thesis Submitted To Nanyang Technology University In Fulfillment Of Requirements For Degree Of Master Of Engineering, School Of Electrical & Electronic Engineering

- [Karpov 2003] Evgeny Karpov, 2003, **Real-Time Speaker Identification, Master Thesis**, University Of Joensuu, Department Of Computer Science

- [Kinnunen 2003] T. Kinnunen, 2003, **Spectral Features For Automatic Text-Independent Speaker Recognition**, ph. D. Thesis, department of computer science, university of joensuu, finland, december 2003

- [Kuphaldt 2007] Tony R. Kuphaldt, July 25, 2007, **Lessons In Electric Circuits**, Volume II - AC Sixth Edition

- [Kung 1993] S.Y.Kung, 1993, **Digital Neural Networks**, PTR Printice Hall.

- [Long 1996] C.J Long. and S Datta, 1996, **Wavelet Based Feature Extraction for Phoneme Recognition**, ieeexplore.ieee.org, Publication Date: 3-6 Oct 1996 Volume: 1, On page(s): 264-267 vol.1.

- [Love 2004] Brian J Love, Jennifer Vining, Xuening Sun, 2004, **Automatic Speaker Recognition Using Neural Networks**, Projrct Submitted To Dr. Joydeep Ghosh, The University of Texas at Austin, http://www.lans.ece.utexas.edu/

- [Markowitz 2000] Judith A Markowitz, 2000, **Voice Biometric**, Communications of the ACM, /Vol. 43, No. 9, September 2000

- [Michel 2007] Misiti Michel, Misiti Yves, Oppenheim Georges, Poggi Jean-Michel, 2007, **Wavelet Toolbox**, http://www.mathworks.com/products/wavelet/

- [Murtagh 2003] F Murtagh, J.L Starck. O Renaud, 2003, **On Neuro-Wavelet Modeling, School Of Computer Sciences**, Queens University Belfast, Northen Ireland, UK,DAPNIA/SEI-SAP, CEA-Saclay, 91191 Gif sur Yveet, France,Faculte de Psychologie Et Sciences De L'education, Switzerland

- [NIST 2006] The national Institute of Standard and technology NIST, 2006, **Speaker Recognition**, united state of America, http://www.biometrics.gov.

- [Patterson 1996] D.W.Patterson, 1996, **Artificial Neural. Networks Theory and Applications**, Printice Hal.

- [Parliamentary 2001], 2001, **BIOMETRICS & SECURITY**, Parliamentary office of science and technology, nov 2001, http://www.parliament.uk/post/pn165.pdf

- [pawar 2005] r.v pawar, p.p kajave., and s.n mali, 2005, **Speaker Identification Using Neural Networks**, Proceedings Of World Academy Of Science, Engineering And Technology volume 7 august 2005 issn 1307-6884

- [Saranli 2000] A. Saranli, 2000, **An Unifying Theory For Rank-Based Multiple Classifier System, With Application In Speaker Identification And Speech Recognition**, Ph. D. Thesis, the Department of Electrical and Electronics Engineering, submitted to the graduate school of natural and Applied Sciences of the middle East technical University, January 2000

- [Taleb 2003] Y. A. Taleb, 2003, **statistical And Wavelet Approaches For Speaker Recognition**, MSc, Thesis, Department Of Computer Engineering, Al-Nahrain University, Iraq, June 2003

- [Wildermoth 2001] B. R. Wildermoth, 2001, **Text-Independent Speaker Recognition Using Source Based Features**, MSc, thesis, Griffith University, Australia, January 2001

- [Wouhaybi 1999] Rita H Wouhaybi, Mohamad Adnan Al-Alaoui1, 1999, **Comparison Of Neural Networks For Speaker Recognition**, IEEE Transactions on Computers.

- [Yegnanarayana 2001] B. Yegnanarayana, K. S. reddy and S. P. Kishore, 2001, **Source And System Features For Speaker Recognition Using Aann Models**,

Speech And Vision Laboratory, Department Of Computer Sciences And Engineering Indian Institute Of Technology Madras, Chennai-60036, India 2001

- [Zimmermann 2005] Július Zimmermann & Július Zimmermann jr.2005, **Stochastic Speaker Recognition Model**, The Slovak Association for the Study of English, as a National Member of ESSE, http://www.skase.sk/index.html

- [Zurada 1996] J.M.Zurada, **Introduction to Artificial Neural Systems**, JAICO publishing House, 1996.

Appendix

Appendix A

Anatomy

Anatomy [Karpov 2003]

The sound is an acoustic pressure formed of compressions and rarefactions of air molecules that originate from movements of human anatomical structures. Most important components of the human speech production system are the *lungs* (source of air during speech), *trachea* (windpipe), *larynx* or its most important part *vocal cords* (organ of voice production), *nasal cavity* (nose), *soft palate* or *velum* (allows passage of air through the nasal cavity), *hard palate* (enables consonant articulation), *tongue*, *teeth* and *lips*. All these components, called *articulators* by speech scientists, move to different positions to produce various sounds. Based on their production, speech sounds can also be divided into consonants and voiced and unvoiced vowels.

From the technical point of view, it is more useful to think about speech production system in terms of acoustic filtering operations that affect the air going from the lungs. There are three main cavities that comprise the main acoustic filter. According to they are *nasal*, *oral* and *pharyngeal* cavities. The articulators are responsible for changing the properties of the system and form its output. Combination of these cavities and articulators is called *vocal tract*.

Speech production can be divided into three stages: first stage is the sound source production, second stage is the articulation by vocal tract, and the third stage is sound radiation or propagation from the lips and/or nostrils. A v*oiced sound* is generated by vibratory motion of the vocal cords powered by the airflow generated by expiration. The frequency of oscillation of vocal cords is called the *fundamental frequency*. Another type of sounds - *unvoiced sound* is produced by turbulent airflow passing through a narrow constriction in the vocal tract.

In a speaker recognition task, we are interested in the physical properties of human vocal tract. In general it is assumed that vocal tract carries most of the speaker related information. However, all parts of human vocal tract described above can serve as speaker dependent characteristics.

الخلاصة

من المعروف أن الإنسان يمكن أن يتم تميزه او التعرف عليه عن طريق الخصائص البشرية مميزة له. تعتبر بصمة الصوت واحدة من البصمات التي يمكن استخدامها لتميز أو التأكد من هوية إنسان. يهتم هذا البحث بدراسة أستخدم الصوت كبصمة في مجال تميز المتكلم لكي يتم التأكد من هوية الشخص عند استخدامه لأنظمة تحتاج إلى الحماية من المستخدمين الغير مخولين حيث لا يسمح لغير المخول بالوصول إليها و استخدامها.

يعنى هذا البحث بدراسة امكانيه استخدام الشبكات العصبونيه لتميز المتكلم واقتراح معماريه شبكه عصبوينه التي لها ألقدره على تمييز المتكلم بأقل نسبه خطاء اضافه إلى دراسة امكانيه استخدام الخصائص(Features) المستخلصة من مستويات مختلفة من التحويل الموجي المستمر (Continuous Wavelet Transform) وقدرتها على تمييز الأصوات .تم اختيار أعلى ثلاث قيم في كل مستوى، مواقعها، رقم المستوى، الوسط الحسابي والوسيط الحسابي ، تم استخدام هذه الخصائص في تدريب ثلاثة أنواع من الشبكات العصبونية

(Adaptive Neural Network, Feed-forward Backpropagation Neural Network, and Learning Vector Quantization Neural Network)

ودراسة فاعليه كل منها في تمييز الأنماط الصوتية وذلك من خلال مقارنه نتائج التمييز لكل شبكه.

ولإغراض الدراسة، تم تسجيل 240 مفرده صوتية من قبل أربعة أشخاص (بحيث يلفظ كل شخص مخول المفردة 10 مرات، وقد تم أعتماد أسلوبين في تحديد هوية المتكلم وفقا للمفردات الصوتية المسجلة. للمجموعة المغلقة (Text-Dependent) تم الحصول بواسطة هذا النظام على نسب تتروح بين 80% و 91% في تحديد هوية المتكلم، اما في سياق المجموعة المفتوحة (Text-Independent) فأن الاختبارات أظهرت نسبة تحقق بحدود 71.87% من خلال مئة محاولة.

تمييز المتكلم
بأستخدام الشبكات العصبونية

من قبل

ايهم راسم فايز جعرون

باشراف

د. فينوس سماوي

قدمت هذه الرسالة أستكمالا لمتطلبات

الحصول على درجة الماجستير في علوم الحاسوب

عمادة البحث العلمي والدراسات العليا

جامعة فيلادلفيا

2008/02