International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:2, 2008

Speech Recognition Using Scaly Neural Networks

Akram M. Othman, and May H. Riadh

***Abstract*—**This research work is aimed at speech recognition using scaly neural networks. A small vocabulary of 11 words were established first, these words are "word, file, open, print, exit, edit, cut, copy, paste, doc1, doc2". These chosen words involved with executing some computer functions such as opening a file, print certain text document, cutting, copying, pasting, editing and exit.

It introduced to the computer then subjected to feature extraction process using LPC (linear prediction coefficients). These features are used as input to an artificial neural network in speaker dependent mode. Half of the words are used for training the artificial neural network and the other half are used for testing the system; those are used for information retrieval.

The system components are consist of three parts, speech processing and feature extraction, training and testing by using neural networks and information retrieval.

The retrieve process proved to be 79.5-88% successful, which is quite acceptable, considering the variation to surrounding, state of the person, and the microphone type.

***Keywords*—**Feature extraction, Liner prediction coefficients, neural network, Speech Recognition, Scaly ANN.

I. General Description

SPEECH conveys information, and what we are concerned with in computer speech processing is the transmission and reception of that information. This is not as simple as it might seem, because speech convey at least three different kinds of information simultaneously. The most important of these is what we might call linguistic information. This is the kind of information that is generally regarded as the meaning of an utterance. With the growth in the use of digital computers, the prospect of using speech as an input to a computer for entering data, retrieving information, or for transmitting commands led to renewed interest in the speech field [7].

II. Speech Recognition

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final result, as for applications such as command & control, data entry, and document preparation or retrieval [17]. The basic assumption of the whole word pattern matching approach is that different utterance of the same word by a particular talker result in similar patterns of sound. There will be variation in spectrum shape at corresponding parts of the patterns from the same word. There will also be variations in the time scale of the patterns, and this will make it difficult to compare corresponding parts [10].

A. M. Othman was with the SE Department, Faculty of Information Technology, ASU/Jordan, He is now with MIS Department , the Amman Arab University for Graduate Studies, AAU- JORDAN(corresponding phone: +962-796184806; fax: 962-6-55 16103; e-mail: akram.othman@aau.edu.jo).

M. H. Riadh., was with the Informatics institute for postgraduate studies, Baghdad-Iraq, She is now with IT department, Al-Hussein Bin-Talal University, Ma'an, Jordan, (corresponding phone: +962-799890287; fax: +962-3-2179050; e-mail: mayhr60@yahoo.com, may.riadh@ahu.edu.jo).

III. Speech Recognition System

Speech recognition is, in its most general form, a conversion from an acoustic waveform to a written equivalent of the message information. Fig. 1 shows a basic speech recognition system [13].

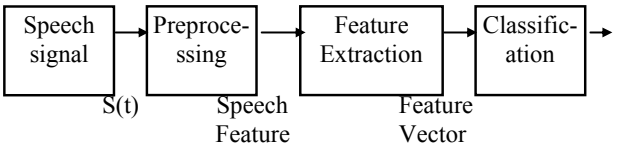

Fig. 1 The basic speech recognition system

IV. Speech Signal Processing and Feature Extraction

Speech signal processing and feature extraction is the initial stage of any speech recognition system, it is through this component that the system views the speech signal itself." Speech signal processing" refers to the operations we perform on the speech signal (e.g., filtering, digitization, spectral analysis, etc.) "Feature extraction "is a pattern recognition term that refers to the characterizing measurements that are performed on a pattern (or signal). These features form the input to the classifier that recognizes the pattern [10].

A. Sampling and Quantizing Continuous Speech

The acoustic speech signal exists as pressure variations in the air. A microphone converts these pressure variations into an electric current that is related to the pressure (similarly the ear converts the pressure variations into a serious of nerve impulses that are transmitted to the brain). To process the speech signal digitally, it is necessary to make the analog waveform discrete in both time (sample) and amplitude (quantize) [2]. The general nature of digital speech waveform representations is depicted in Fig. 2 [13].

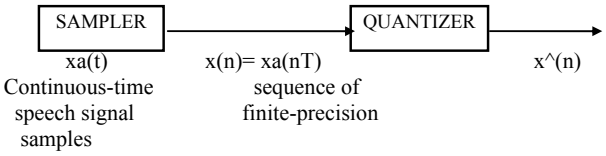

Fig. 2 General natures of digital speech waveform representations

B. Windowing

One way to avoid discontinuities at the ends of the speech segments is to taper the signal to zero or near zero and hence reduce the mismatch, is using windowing, there are two famous types of windows [21].

1) Rectangular Window

$$w(n)=\begin{cases} 1 & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2) Hamming Window

$$w(n)=\begin{cases} 0.54-0.46*\cos(2\pi\, n/(N-1)) & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

V. Linear Prediction Analysis

Linear prediction analysis (commonly called Linear prediction coding or LPC) is one of the most powerful speech analysis tools. It has become the predominant method for estimating the basic parameters of the speech signal (formants, fundamental frequency, vocal tract area function, and spectra). Other reasons are the speed of computation and accuracy of LPC methods [2],[10],[14].

Linear predictive analysis can be readily shown to be closely related to the basic model of speech production as in Fig. 2 in which the speech signal is modeled as the output of a linear, time-varying system excited by either qusi-periodic pulses (for voiced sounds) or random noise (for unvoiced sounds) [14]. Whose steady-state system function is of the form:

$$H(z)= G/1-\sum_{k=1}^{p} ak\, z-k \quad (3)$$

Thus, the parameters of this model are: Voiced/unvoiced classification, pitch period for voiced speech, gain parameter G, and the coefficients {ak }of the digital filter [13].

The LPC is a process of applying the linear prediction model to the speech signal by minimizing the sum of the squared difference between actual speech samples and the linearly predicted one, Unique set of predictor coefficients can be determined if the speech signal is assumed to be slowly varying within time.

The short-time average prediction error is defined in [13]:

$$En = \sum en2(m) \quad (4)$$

A. The Autocorrelation Method

The complete block diagram of speech recognition system using autocorrelation analysis is as in Fig. 3 [14].

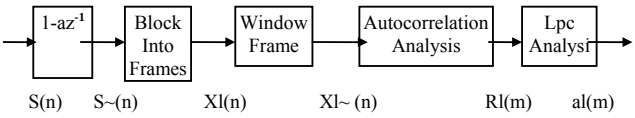

Fig. 3 Complete block diagram of speech recognition system

VI. Neural Networks

Neural networks model some aspects of the human brains, where thinking process is achieved in synaptic connections between neurons. The structure of the network is layered and capable of high parallelism. Neural networks are useful in classification, function approximation and generally in complex problems, which do not require accurate solution. Neural networks must be taught before they can be used, which corresponds to how humans learn. [19] .A neural network consists of units (processors, nodes) that are interconnected with several other such units; they function independently on the input they are given and their local data. Usually all of the units in a network are homogenous, but also heterogeneous networks exist. [4},[8],[18].

A. Neural Networks Architecture

Neural Networks use a set of processing elements (or nodes) loosely analogous to neurons in the brain (hence the name, neural networks). These nodes are interconnected in a network that can then identify patterns in data as it is exposed to the data. In a sense, the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order. The structure of a neural network looks something like in Fig. 4: [16]

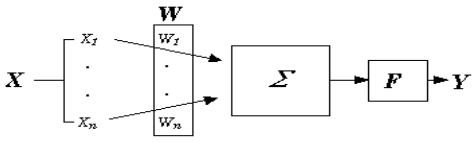

Fig. 4 General structure of a neural network

A set of inputs (X1 to Xn) is applied to each node representing the inputs from outside world or, alternatively, they may be outputs from other nodes. Each input is multiplied by a weight (W1 to Wn) associated with the node input to which it is connected and the weighted inputs are then summed together.

A threshold value (C) local for each node is added to the weighted summation and the resulting sum is then passed through a hard limiting activation function (F). The output of a node is therefore [16].

$$Y=F\left(\sum_{n=1}^{N} (Xn * Wn) + C\right) \quad (5)$$

Three common transfer functions are the sigmoid, linear and hyperbolic functions are widely used. The sigmoid

function produces values between 0 and 1 for any input from the resulting sum.[3]

B. Learning Methods

The learning exhibited by neural networks can be categorized to two sections:

1) Supervised Learning

Supervised learning requires the network to have an external 'teacher'. That tell the network how well it is performing (reinforcement learning) [20]. The algorithm adjusts weights using input-output data to match the input-output characteristics of a network to the desired characteristics [5].

2) Unsupervised Learning

If the network 'studies' on its own, the learning is unsupervised; the network attempts to reflect properties of some given data in its output just by examining the data, i.e. by guessing. This type of learning is also referred to as self-organization [20]. Hebbian learning is representative of unsupervised learning algorithm [5].

VII. Training

The networks are usually trained to perform tasks such as pattern recognition, decision-making, and motory control. The original idea was to teach them to process speech or vision, similarly to the tasks of the human brain. Nowadays tasks such as optimization and function approximation are common. [1] Training of the units is accomplished by adjusting the weight and threshold to achieve a classification. The adjustment is handled with a training rule (a learning rule), from which a training algorithm for a specific task can be derived.[18]

A. Learning Rules

A learning rule allows the network to adjust its connection weights in order to associate given input vectors with corresponding output vectors. During training periods, the input vectors are repeatedly presented, and the weights are adjusted according to the learning rule, until the network learn the desired associations. [9]

1) Hebb's Rule

This rule is due to Donald Hebb.It is a statement about how the firing of one neuron, this has a role in the determination of the activation of another neuron, affects the first neuron's influence on the activation of the second neuron, especially if it is done in a repetitive manner [22].

2) Delta Rule

This rule is also known as the least mean squared error rule (LMS).Takes the square of the errors between the target or desired values and computed values, and takes the average to obtain the mean squared error. This quantity is to be minimized [22].

VIII. Neural Networks for Speech Recognition

Various neural networks have been used for speech recognition. In the following some of these N.N:[11]

A. Kohenon Self-Organizing

The Kohenon Self-Organizing is a neural network trained by following a non-supervised algorithm [11].

B. Multi-layer Perceptrons (MLPs)

Back-Propagated Delta Rule Networks (BP) (sometimes known Multi-layer Perceptrons (MLPs) is well-known developments of the Delta rule for single layer, networks (itself a development of the Perceptron Learning Rule). MLP can learn arbitrary mappings or classifications. Further, the inputs (and outputs) can have real values. [6] Typical BP network architecture as in Fig. 5[16].

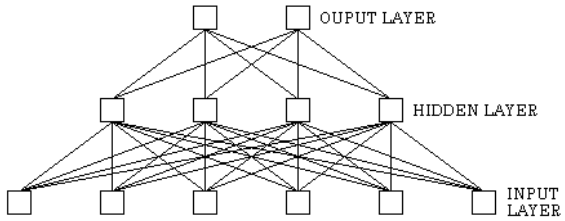

Fig. 5 Typical BP network architecture

The hidden layer learns to recode (or to provide a representation for) the inputs. More than one hidden layer can be used. The architecture is more powerful than single-layer networks: it can be shown that any mapping can be learned, given two hidden layers (of units) [6].

1) The Fully Connected Neural Networks

The most common form of neural network is the 3-layer, fully connected, feed forward MLP. The nodes are arranged in 3 layers; an input layer, a hidden layer and an output layer with inputs flowing in the forward direction from input layer to output layer through the hidden layer except during training. In this type of network, the inputs of every node in the hidden layer are connected to the outputs of every node in the input layer and the inputs of every node in the output layer are connected to the outputs of every node in the hidden layer [16].

2) The Scaly Neural Network Architecture

A problem with fully connected neural networks is their size. When it comes to the practical implementation of neural networks size becomes an important factor. In the case of computer simulation the problem comes with the large computational cost. The larger and more complex the network the longer it takes to train and once trained it takes longer for the network to perform its recognition task.

Fig. 6 shows an example of a neural network where scaly architecture has been applied between the input layer and the hidden layer. This is the approach adopted for the work here since the localized structure of the input zones is somewhat

analogous to the Cochlear processing, which occurs, in the human ear. A preliminary investigation into the ability of a scaly architecture neural network was carried out by A.D. Smith at the University of Newcastle upon Tyne. Smith' work suggested that further investigation of this network was required to better determine the effect of changing the parameters of the network [16].

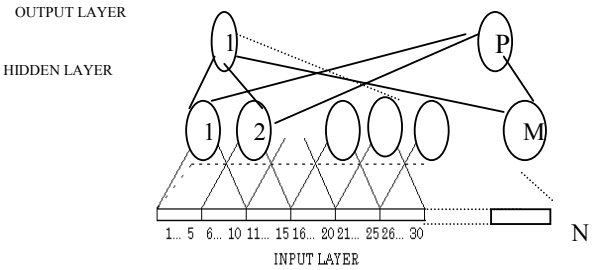

Fig. 6 Scaly Neural Networks

IX. System Description and Result

The system, which is divided into three parts:
1- Speech processing and feature extraction,
2- Neural networks for training and testing.
3- Information retrieval. The block diagram for the system is as illustrated in Fig. 7.

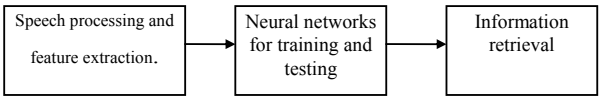

Fig. 7 The block diagram of the system

A. Speech Processing and Feature Extraction

The first step is to deal with speech utterance by using one of the methods of speech processing in order to provide feature extraction. The block diagram of this part is as illustrated in Fig. 8.

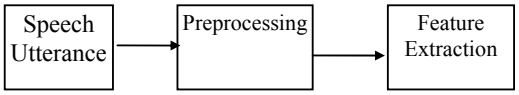

Fig. 8 Speech processing and feature extraction diagram

1) The Speech Utterance (Data Collection)

The source of data is a database consists of (11 words) spoken 8 times by 5 speakers; those are 2 males and 3 females of various ages. The data, which is speaker dependant, will be used for training and testing form.

In speaker dependent form, the first four utterances of each of the 11 words spoken by every speaker are used to train the network and the remaining utterances are used to test the network. Therefore, the speech database contains 220 utterances, which can be used for training the network, and 220 utterances, which are available for testing. These words are recorded by:-

1- Using Creative's application software "WAVE STUDIO", with sampling rate of 11KhZ, 8bit, and mono is used to record the utterance. 2- In a closed room, the same microphone is used to record the spoken words. 3-The files saved in a (*.wav) format.

2) Preprocessing

The speech signals are recorded in a low noise environment with good quality recording equipment. The signals are sampled at 11KHz using a 16-bit A/D converter. Accurate end pointing of speech data is a difficult task but reasonable results can be achieved in isolated word recognition when the input data is surrounded by silence.

a. Sampling Rate

Typical frame sizes are from (15 to 50 msec) i.e. 150 to 500 samples for a 11 KHz sampling rate [14]. Using sampling rate of (11 KHz), which is adequate to represent all speech sounds and then choose a frame size of N samples. For 25msec N= (256 samples) which seems to be the best size found to work with.

b. Overlap

Consecutive frames are spaced M samples apart. Clearly when M<N, there will be an overlap between adjacent frames. Such overlap inherently provides smoothing between vectors of features coefficient. (Typical values of M are M=N/3 or N/2) [14]. the overlap size chosen is to be N/2, which are 128 samples.

c. Window Type

A typical smoothing window used in LPC analysis system is the hamming window defined in equation 2 with 256-window length so as to taper the speech samples to zero at the end of the frame. Let the l th frame of speech be denoted as Sl(n) giving the windowed signal [14]

$$Sl^{\wedge}(n)=Sl(n) \cdot w(n), \quad n=0,1,\ldots,N-1$$

3) Feature Extraction

The digitized sound signal contains a lot of irrelevant information and requires a lot of storage space. To simplify the subsequent processing of the signal, useful features must be extracted and the data compressed. The linear prediction analysis is the most used method for speech recognition.

a. Linear Prediction Coding Coefficients

The autocorrelation method of the windowed frame of equation 6, will be used

$$Rl(m)=\sum_{n=0}^{N-1-m} Sl^{\wedge}(n)Sl(n+m) \quad m=0,1,..,P \qquad (7)$$

where, P is the order of the analysis system of the LPC. Typical values of P range from (8 to 16). Choosing P=12, will produce a feature set of 12 values for each frame. [14]

$$X(l)=\{Rl(0), Rl(1),\ldots,Rl(12)\} \qquad (8)$$

Then using equation (8) to solve it and find the values of LPC coefficient by using the Durbin's recursive equations and use the set αl(m) as the features of frame l.

$$\alpha(L)=\{\alpha l(1), \alpha l(2), \ldots \alpha l(12)\} \quad (9)$$

The LPC coefficients are calculated for 256 samples frames with successive frames being overlapped by 128 samples, so for each frame of 256 samples 12 LPC coefficients will be obtained and it will be used as the input to the recognizer (Neural Network). Fig. 9 shows the LPC coefficient for two speaker.

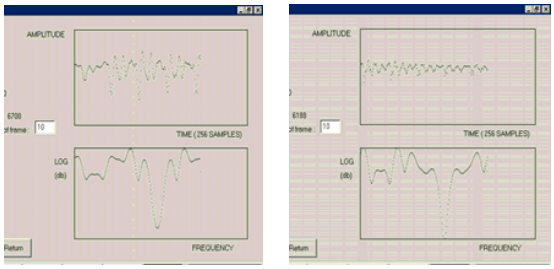

Fig. 9 The LPC coefficient for two speakers

B. Neural Networks for Training and Testing

The input data is processed (sec 4.2) and presented to the network so that successive feature vectors or frames are presented to the network inputs, each coefficient of the feature vectors being presented to one of the input nodes.

The data is used in the speaker dependant training and testing .involves using four utterances spoken by each speaker for training and the other four utterances for testing.

The mean length of an utterance over the entire data set is calculated and found to be 65 frames.

1) Scaly Connected Network

The architecture of the scaly connected networks requires an input layer of 780 input neurons to accommodate the 65 input feature vectors, each contains 12 coefficients. For the scaly architecture, the number of frames in a zone is taken as 10 frames with an overlap of 5 frames so that the number of nodes required in hidden layer is 132 nodes; the number of output nodes is 11 nodes.

2) Activation Function

Using the non-linear activation function of the type Sigmoid01 the function has an upper limit of 1 and a lower limit of 0 so the desired outputs of the network will be 0 and 1. The desired outputs are generated such that a +1 is at the output of the neuron of the correct category and a 0 on all the other output neurons.

3) The Learning Rate

The learning rate (η) is a parameter between 0 and 1 that changes the value of the weights starting from the output nodes back to the hidden nodes, and determine how fast the learning takes place. Trying to change the value of η from 0.9 to 0.1 to find the best learning and stable condition, the value of η is found to be 0.1.

4) Momentum Rate

Convergence is sometimes faster if a momentum rate (α) is added and weight changes are smoothed. The momentum is usually greater than zero, to ensure convergence, is also less than 1 [12]. The value of α is chosen to be 0.5 in order to obtain best performance for training.

5) Convergence

During training, the learning process stops when the error for all the cases reaches a specific tolerance. The error value is the least mean squared error between the actual outputs and the desired outputs of the network. After each iteration this error is compared to an accepted error (normally less than 0.1) until the difference between them reaches zero that means, the network is learning and the convergence has achieved.

6) Training Phase

Using the multilayer backpropagation algorithm to train the network for spoken words for each speaker (5 speakers), using the scaly network with input nodes=780, hidden nodes=132, and 11 output nodes each for one word, with the nonlinear activation function sigmoid01, learning rate= 0.1, momentum rate= 0.5, initialize the weights to random value between +0.1 and –0.1, the accepted error is chosen to be 0.009.

7) Testing Phase

Using the same multilayer backpropagation algorithm to test the network of the spoken words for the five speakers. Each speaker has to test the network by 11 words repeated four times.

Each speaker, tests the word four times and the node with the higher number in the output will be the winner node. The correct answer will be indicated by comparing this node with the input word to the network. So by testing the words said by each speaker the performance can be found by this equation; Table I contains the performance for the test phase for each speaker.

Performance = Total succeeded number of testing words/
Total number of words *100%. (10)

TABLE I
THE PERFORMANCE FOR THE TEST PHASE FOR EACH SPEAKER

Speaker	Performance
1 (male)	79.5%
2 (male)	84%
3 (female)	88%
4 (female)	86%
5 (female)	79.5%

X. Conclusion

The following conclusions can be pointed out:

The scaly type architecture neural network has been shown to be suitable for the recognition of isolated words for small vocabularies as it gave (79.5-88)% success.

The scaly type needs (426) iterations to reach acceptable error of (0.01) for three words only repeated four times for female speaker No. 3, while the fully connected type needs (2394) iterations.

Recognition of the words was carried out in speaker dependent mode. In this mode the tested data is presented to the network are different from the trained data.

The linear prediction coefficient (LPC) with 12 parameters from each frame improves a good feature extraction method for the spoken words since the first 12 in the cepstrum represent most of the formant information.

In all speech recognition systems, the data is recorded using a noise-canceling microphone, since this type of microphone is not available, the data was recorded using a normal microphone, but recorded in a closed room without any type of noise. Hence using the first type of microphone would give even better results.

Spoken language technology (SLT) is a multidisciplinary field, it requires two types of skills:

-Background skills like, mathematics, signal processing, acoustics, experimental skills, computing skills.

-Subject skills like, phonetics, auditory system, speech perception, ASR techniques, coding techniques.

References

[1] Accurate Automation Corporation, *"What are Artificial Neural Networks?"*, web site: http://www.accurate-automation.com/products/nnets.htm.

[2] C.H.Chen *"Signal Processing Handbook"* 1985.

[3] Christopher M.Fraser –California University, Hayward, 2000, "Neural Networks:Literature Review" web site, http://www.telecom.csuhayward.edu/~stat/Neural/CFProjNN.htm.

[4] Consortium for Virtual Operations Research, *"Artificial Neural Networks "*, 1997 web site http://cvor.pe.wvu.edu/neural.htm

[5] Da Ruan *"Intelligent Hybrid Systems, Fuzzy Logic, Neural Networks, Genetic Algorithms"* 1997 by Kluwer Academic Publishers.

[6] Dr. Leslie Smith, *"An Introduction To Neural Networks"*, Centre for Cognition and Computational Neuroscience, Department of Computing and Mathematics, University of Stirling, Website, http://www.cs.stir.ac.uk/~Iss/Nnitro/InvSlides.html

[7] Frank Fallside / William A. Woods *"Computer Speech Processing"*, Printic-hall, 1985.

[8] Gurney, K., *"Neural Nets"*, 12.6.1996 web site http://www.shef.ac.uk/psychology/gurney/notes/contents.html

[9] Ingrid F.Russell *"Neural Networks"*, Department Of Computer Science, University of Hartford, West Hartford CT 06117.web site,http://uhavax.hartfold.edu./disk$userdata/faculty/compsci…/neural-networks-tutorial.htm.

[10] J.C.Simon "Spoken Language Generation and Understanding", 1980.

[11] Jean Hennebert, Martin Hasler and Herve Dedieu "Neural Networks in Speech Recognition" Department of Electrical Engineering, Swiss Federal Institute of Technology, 1015 Lausanna, Switzerland.

[12] Kevin Gurney *"An Introduction to Neural Networks"* Ucl, Press Limited Taylor & Francis Group London, 1997.

[13] L.R.Rabinar /R.W.Schafer *"Digital Processing of Speech Signals"* 1978 Prentic-hall.

[14] Lawrence R. Rabiner & Stephene E.Levinson ***"Isolated and Connected Word Recognition Theory and Selected Applications"*** IEEE Transaction on communications , May 1981, vol. com-29, no.5 .

[15] Literature Review- "Speech Recognition for Noisy Environments" Web Site, http://www.dcs.shef.ac.uk/~jeremy/litrev.htm.

[16] Luna, "Neural Networks for Speech Recognition", Web Site, http://luna.moonstar.com/~morticia/thesis/chapter2.html.

[17] Ron Cole / Victor Zue "Spoken Language Input" 1998, web site http://cslu.cse.ogi.edu/HLTsurvey/ch1node2.html

[18] Sarle, Warren S., "Neural Net FAQ", web site ftp://ftp.sas.com/pub/neural/FAQ.html

[19] Satu Virtanen / Kosti Rytknen "Neural Network", Helsinki University of Technology, web site http://www.askcom/tlark.neural networks.html

[20] Siganos, Dimitrios & Stergiou, Christos, "Neural Networks", web site, http://www.dse.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[21] Tony Robins "Speech Vision Robotics Group", web site http://svr-www.eng.cam.ac.uk/~ajr .

[22] Valluru B.Rao and Hayagriva V.Rao "C++ Neural Networks and Fuzzy Logic" Managment Information Source, 1993.

Akram M. Othman, (Baghdad 1949) with over a career of 30 years in the professional and academic fields. Dr. Othman has completed over 19 professional case studies covering various aspects within the field of Information Technology (IT), and has published numerous academic papers. While Dr. Akram is a renowned expert in the field of IT, he has also held senior posts in research and development in economic studies, science and technology as well as human resources capacity-building. Additionally, Dr. Othman has held several honorary positions on the boards of various IT advancement institutions, from board member to vice president.

Currently he is a member of the Iraqi Science Academy (ISA), President of the Iraqi Computer Society (ICS) and Vice President of the Union of Arab ICT Associations (IJMA3).

Dr. Akram Othman has an MSc from Baghdad University and doctorate in Computer Sciences from University of Technology - Iraq and a BSc in Mathematical Sciences.

May. H. Riadh, (Iraq, Kurkuk ,1960), with over a career of 26 years in professional and academic field. She work as engineer for about 18 years and a lecturer in Informatics institute for postgraduate studies/ Iraq-Baghdad., Almansoor private university/ Iraq-Baghdad.

Currently she is a member in the Iraqi Computer Society (ICS) and a member in of the Union of Arab ICT Associations (IJMA3).

Dr. May has an Bsc. from university of technology in Baghdad in computer engineer, Msc from university of technology in Baghdad in computer science, and Phd in computer science from Informatics institute for postgraduate studies/ Iraq-Baghdad.