Iraqi Journal of Statistical Science (20) 2011
The Fourth Scientific Conference of the College of Computer Science & Mathematics
pp [291-308]

Sound Encryption Using Feature Extraction and Neural Network

Alyaa Moufaq[*] Ielaf O. Abdul-Majjed[**] Melad Jader[***]

ABSTRACT

In recent years, secure communication techniques have increased widely and unexpectedly . In order to establish reliable communication technology and to ensure that the data (sound) reaches its intended end and to be accessible to all through the shared network , there is a need to encrypt the transmitted information.

The research was divided into three stages, the first stage included the process of extracting sound features for the file to be sent .In the second stage neural networks were used in the encryption process of the properties resulting from the first phase. In the final phase encryption algorithms were used to encrypt the result from the previous phase. The speech signal of male and female were coded and encrypted. The measures (SNR, PSNR, NRMSE) were used to improve the results. Besides that the Matlab were used as a programming language in this paper.

التشفير باستخدام استخلاص الخواص والشبكات العصبية لملفات الصوت

المستخلص

في السنوات الأخيرة تزايد الاهتمام بصورة هائلة وغير متوقعة في تامين تقنيات الاتصال ولغرض تحقيق عملية التناقل وضمان وصول البيانات (صوت) الى الجهات المطلوبة كان لابد ان تكون في متناول الجميع عبر الشبكة المشتركة وهنا تبرز الحاجة الى تشفير المعلومات المرسلة.

تم انجاز البحث بثلاث مراحل تضمنت المرحلة الاولى عملية استخلاص خواص الصوت (features) للملف المراد ارساله وفي المرحلة الثانية تم استخدام الشبكات العصبية في عملية التشفير للخواص الناتجة من المرحلة الاولى اما المرحلة الاخيرة فقد تم استخدام احدى خوارزميات التشفير لتشفير نتائج المرحلة السابقة. تم ترميز وتشفير اشارات لصوت رجل و امرأة وبعدها تم استخدام المقاييس (SNR، PSNR،

[*] Lecturer\ College of Computers Sciences and Math.\ University of Mosul.
[**] Lecturer\ College of Computers Sciences and Math.\ University of Mosul.
[***] Lecturer\ College of Computers Sciences and Math.\ University of Mosul.

Received:1/10 /2011 ____________________Accepted: 21 /12 / 2011

(NRMSE لغـرض اثبـات صـحة النتـائج وكفاءتهـا فـضلا عـن ذلـك، تـم اعتمـاد Matlab كلغة برمجية في هذا البحث.

1.Introduction

In this research, we proposed a new approach for encrypting and compressing sound signals using:

- Linear predicative coding (LPC).
- Elman neural network .
- XOR Encoding.

LPC is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. It provides extremely accurate estimates of speech parameters, and is relatively efficient for computation.

It digitally encodes analog signals using a single-level or multilevel sampling system in which the value of the signal at each sample time is predicted to be a linear function of the past values of the quantized signal. LPC is related to adaptive predictive coding (APC) in that both use adaptive predictors. However, LPC uses more prediction coefficients to permit the use of a lower information bit rate than APC, and thus requires a more complex processor.[16]

A particular type of neural network is the recurrent neural network. This network is a dynamical system, in which the output depends on the inputs and the internal state, which evolves with the network inputs.

2. **Related work:**

Neural network used in several researches, [3] uses modular neural network(MNN) to identify the speaker by the characteristics extracted. but [10] uses the N.N for recognition and considering the case of speaker recognition by analyzing the sound signal with the help of intelligent techniques. However [6] uses mean and variance of the discrete wavelet transform in addition to other features that have been used previously for audio classification and used multilayer perceptron (MLP) neural networks as a classifier .The aim of [1] is to develop a system for encoding good quality speech at a low bit rate using linear predictive coding also, [9] used the linear predictive for better interpretation of spoken words. While, our research has Neural Network as well as LPC to encrypt the sound.

3.Why Use LPC

Under normal circumstances, speech is sampled at 8000 samples/second with 8-bits used to represent each sample. This provides a rate of 64000 bits/second. Linear predictive coding reduces this to 2400 bits/second by breaking the speech into segments and then sending the voiced/unvoiced information, the pitch period and the coefficients for the filter that represent the vocal tract for each segment. At this reduced rate the speech has a distinctive synthetic sound and there is a noticeable loss of quality. However, the speech is still audible and it can be easily understood. Since there is information loss in linear predictive coding, it is a lossy form of compression. This low bit is also appealing to government because of its resistance to jamming and channel noise.[17]

4.LPC Modeling

A. Physical Mo

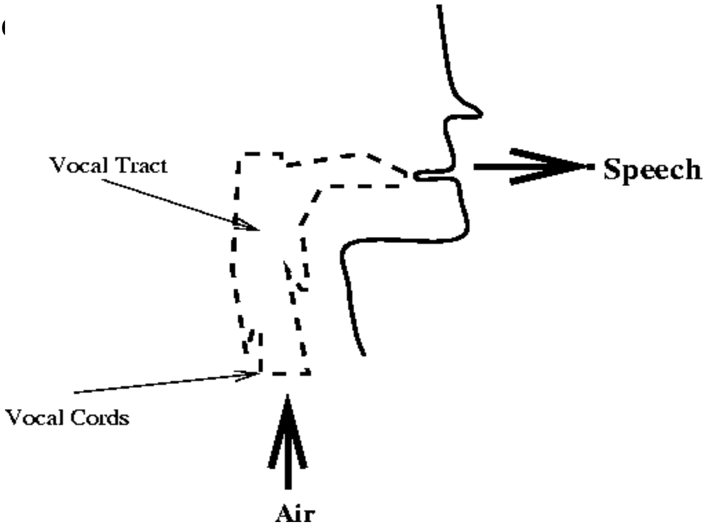


Figure (1): Physical Model

When you speak:

- Air is pushed from your lung through your vocal tract and out of your mouth becomes speech.
- For certain voiced sound, your vocal cords vibrate (open and close). The rate at which the vocal cords vibrate determines the pitch of your voice. Women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration).
- For certain fricatives and plosive (or unvoiced) sounds, your vocal cords do not vibrate but remain constantly opened.

- The shape of your vocal tract determines the sound that you make.
- As you speak, your vocal tract changes its shape producing different sounds.
- The shape of the vocal tract changes relatively slowly (on the scale of 10 msec to 100 msec).
- The amount of air coming from your lung determines the loudness of your voice.[1]

B. **Mathematical Model:**

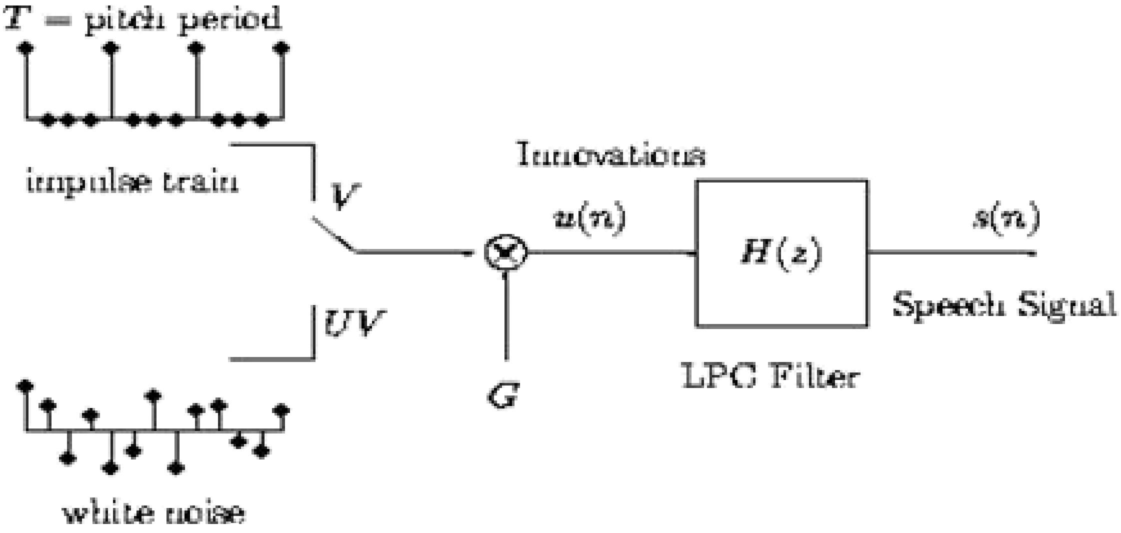


Figure (2): Mathematical Model

- The above model is often called the LPC Model.
- The model says that the digital speech signal is the output of a digital filter (called the LPC filter) whose input is either a train of impulses or a white noise sequence.
- The relationship between the physical and the mathematical models:[1][16]

Vocal Tract	⇔	H(z) (LPC Filter)
Air	⇔	u(n) (Innovations)
Vocal Cord Vibration	⇔	V (voiced)
Vocal Cord Vibration Period	⇔	T (pitch period)
Fricatives and Plosives	⇔	UV (unvoiced)
Air Volume	⇔	G (gain)

The LPC filter is given by:

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{10} z^{-10}} \quad \ldots\ldots\ldots\ldots(1)$$

which is equivalent to saying that the input-output relationship of the filter is given by the linear difference equation:[1]

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n) \quad \ldots\ldots\ldots\ldots(2)$$

The LPC model can be represented in vector form as:

$$\mathbf{A} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V/UV, T) \quad \ldots\ldots\ldots\ldots(3)$$

- $\mathbf{A}$ changes every 20 msec or so. At a sampling rate of 8000 samples/sec, 20 msec is equivalent to 160 samples.
- The digital speech signal is divided into frames of size 20 msec. There are 50 frames/second.

The model says that:

$$\mathbf{A} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V/UV, T)$$

is equivalent to

$$\mathbf{S} = (s(0), s(1), \ldots, s(159))$$

Thus the 160 values of $\mathbf{S}$ are compactly represented by the 13 values of $\mathbf{A}$. [5][13]

There's almost no perceptual difference in $\mathbf{S}$ if:

- For Voiced Sounds (V): the impulse train is shifted (insensitive to phase change).
- For Unvoiced Sounds (UV):} a different white noise sequence is used.

LPC Synthesis: Given $\mathbf{A}$, generate $\mathbf{S}$ (this is done using standard filtering techniques).

LPC Analysis: Given $\mathbf{S}$, find the best $\mathbf{A}$ (this is described in the next section).[15]

5.LPC Analysis

Consider one frame of speech signal:

$$\mathbf{S} = (s(0), s(1), \ldots, s(159)) \quad \ldots\ldots\ldots\ldots(4)$$

The signal $s(n)$ is related to the innovation $u(n)$ through the linear difference equation:

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n) \quad \ldots\ldots\ldots\ldots(5)$$

The ten LPC parameters $(a_1, a_2, \ldots, a_{10})$ are chosen to minimize the energy of the innovation:

$$f = \sum_{n=0}^{159} u^2(n) \qquad \text{……….…(6)}$$

Using standard calculus, we take the derivative of f with respect to a_i and set it to zero:

$$\begin{aligned} df/da_1 &= 0 \\ df/da_2 &= 0 \\ &\ldots \\ df/da_{10} &= 0 \end{aligned}$$

We now have 10 linear equations with 10 unknowns:[16]

$$\begin{bmatrix} R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) & R(9) \\ R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) \\ R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) \\ R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) \\ R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) \\ R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) \\ R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) \\ R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) \\ R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) \\ R(9) & R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix} = \begin{bmatrix} -R(1) \\ -R(2) \\ -R(3) \\ -R(4) \\ -R(5) \\ -R(6) \\ -R(7) \\ -R(8) \\ -R(9) \\ -R(10) \end{bmatrix}$$

where

$$\begin{aligned} R(k) &= \textstyle\sum_{n=0}^{159-k} s(n)s(n+k) \\ &= \text{autocorrelation of } s(n) \end{aligned} \qquad \text{……….…(7)}$$

5.1 Levinson-Durbin Recursion:

$$\begin{aligned} E^{(0)} &= R(0) \\ k_i &= [R(i) - \textstyle\sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)]/E^{(i-1)} \qquad i = 1, 2, \ldots, 10 \\ \alpha_i^{(i)} &= k_i \\ \alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad j = 1, 2, \ldots, i-1 \\ E^{(i)} &= (1 - k_i^2) E^{(i-1)} \end{aligned}$$

Solve the above for $i = 1, 2, \ldots, 10$, and then set

$$a_i = -\alpha_i^{(10)}$$

To get the other three parameters $(V/UV, G, T)$, we solve for the innovation:

$$u(n) = s(n) + \sum_{i=1}^{10} a_i s(n-i) \qquad \text{……….…(8)}$$

Then calculate the autocorrelation of $u(n)$:

$$R_u(k) = \sum_{n=0}^{159-k} u(n)u(n+k)$$

Then make a decision based on the autocorrelation.[17]

5.2 2.4kbps LPC Vocoder

The following is a block diagram of a 2.4 kbps LPC Vocoder:

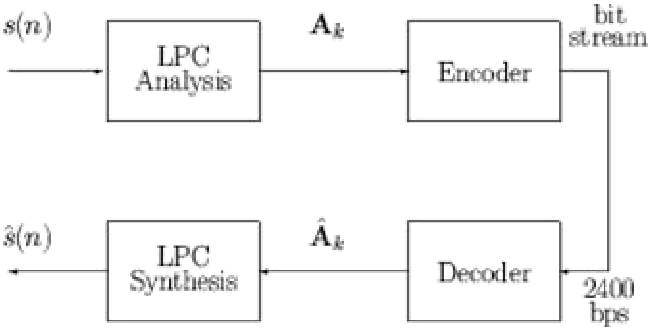


Figure (3): 2.4 kbps LPC Vocoder

- The LPC coefficients are represented as line spectrum pair (LSP) parameters.
- LSP are mathematically equivalent (one-to-one) to LPC.
- LSP are more amenable to quantization.
- LSP are calculated as follows:[11]

$$\begin{aligned} P(z) &= 1 + (a_1 - a_{10})z^{-1} + (a_2 - a_9)z^{-2} + \ldots + (a_{10} - a_1)z^{-10} - z^{-11} \\ Q(z) &= 1 + (a_1 + a_{10})z^{-1} + (a_2 + a_9)z^{-2} + \ldots + (a_{10} + a_1)z^{-10} + z^{-11} \end{aligned}$$

- Factoring the above equations, we get:

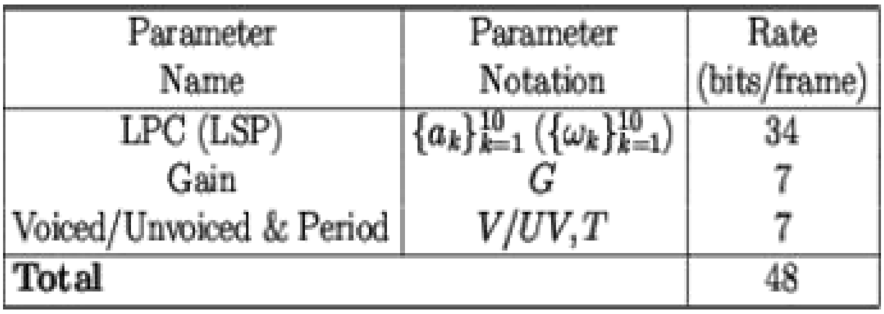
$$\begin{aligned} P(z) &= (1 - z^{-1}) \prod_{k=2,4,\ldots,10}(1 - 2\cos\omega_k z^{-1} + z^{-2}) \\ Q(z) &= (1 + z^{-1}) \prod_{k=1,3,\ldots,9}(1 - 2\cos\omega_k z^{-1} + z^{-2}) \end{aligned}$$

$\{\omega_k\}_{k=1}^{10}$ are called the LSP parameters.

- LSP are ordered and bounded:

$$0 < \omega_1 < \omega_2 < \ldots < \omega_{10} < \pi$$

- LSP are more correlated from one frame to the next than LPC.
- The frame size is 20 msec. There are 50 frames/sec. 2400 bps is equivalent to 48 bits/frame. These bits are allocated as follows:

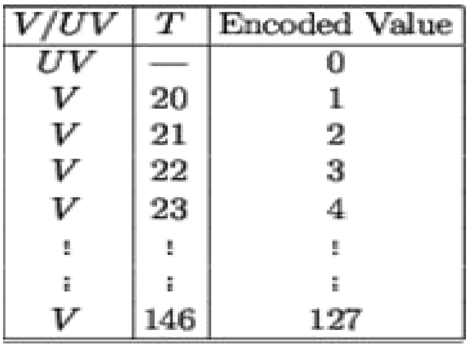


Parameter Name	Parameter Notation	Rate (bits/frame)
LPC (LSP)	$\{a_k\}_{k=1}^{10}$ ($\{\omega_k\}_{k=1}^{10}$)	34
Gain	G	7
Voiced/Unvoiced & Period	$V/UV,T$	7
Total		48

Table (1): 2.4 kbps LPC Vocoder

The gain, G , is encoded using a 7-bit non-uniform scalar quantizer

- For voiced speech, values of T ranges from 20 to 146. $(V/UV, G, T)$ are jointly encoded as follows:[16]
-



V/UV	T	Encoded Value
UV	—	0
V	20	1
V	21	2
V	22	3
V	23	4
⋮	⋮	⋮
V	146	127

Table (2): voiced and unvoiced encoded

6. Elman Neural Network Structure

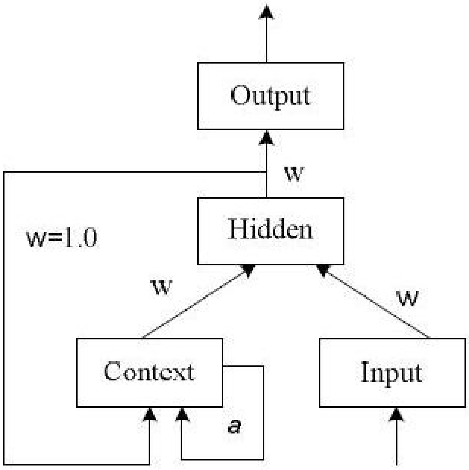


Figure (4): structure o the ENN

In Figure (4) after the hidden units are calculated, their values are used to compute the output of the network and are also all are stored as "extra inputs" (called context unit) to be used when the next time the network is operated.

Thus, the recurrent contexts provide a weighted sum of the previous values of the hidden units as input to the hidden units. As shown in Figure (4), the activations are copied from hidden layer to context layer on a one for one basis, with fixed weight of 1.0 (w=1.0). The forward connection weight is trained between hidden units and context units as well as other weights. If self-connections are introduced to the context unit when the values of the self-connections weights (a) are fixed between 0.0 and 1.0 (usually 0.5) before the training process, it is an improved ENN as proposed [2]. When weights (a) are 0, the network is the original ENN.

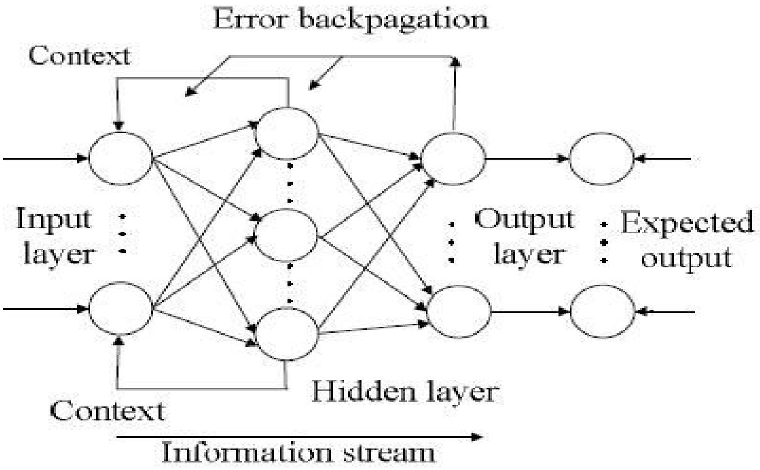


Figure (5): Internal Process Analysis of ENN

From Figure (5) we can see that training such a network is not straightforward since the output of the network depends on the inputs and also all previous inputs to the network. So, it should trace the previous values according to the recurrent connections(Figure 4). So, the calculation of the functional derivatives is not straightforward and it leads to low efficiency to deal with various signal problems.

Figure (6) shows that a long ENN where by a back propagation is used to calculate the derivatives of the error (at each output unit) by unrolling the network to the beginning. At the next time step t+1 input is represented.

The context units contain values which are exactly the hidden unit values at time t (and the time t-1, t-2 …) and these context units provide the network with memory [8]. Therefore, the ENN network is converted into a dynamical network that is efficient in the use of temporal information of the input sequence, both for classification as well as for prediction [7][14]. However, the efficiency of the ENN is limited to low order system due to the insufficient calculation of the derivatives in some degree.

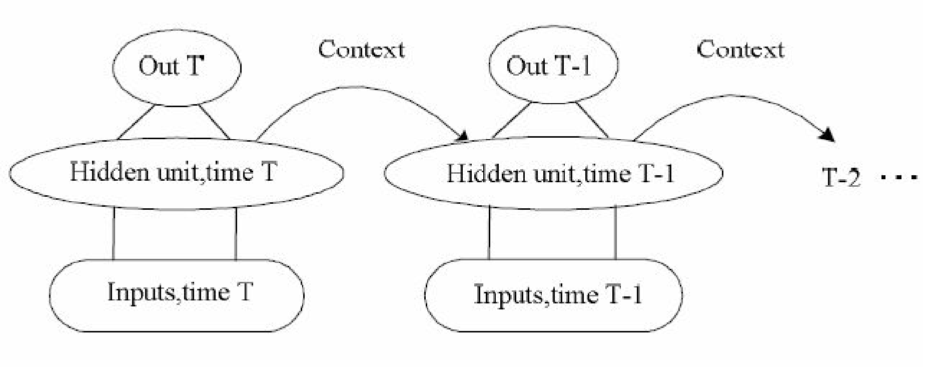


Figure (6): Unroll the ENN through Time

The learning commonly used in the ENN is back propagation algorithm since it adjusts the network parameters (weights and thresholds) to minimize the error measure function equation below. Therefore it needs to compute the differentials of the error measure, activation function and their analog multiplications[12].

$E = \sum_{t=0}^{T} E|_p$ where p indexes over all the patterns for the training set in the time interval [0, T]. In our paper, the time element is updated by the next input of the pattern from training set. So we can get the following equation.

$$E = \sum_{t=0}^{T} E_P = \sum_{p=1}^{P} E_p$$

E_p is defined by

$$E_p = \frac{1}{2}\sum_j (t_{pj}^k - o_{pj}^k)^2$$

where t_{pj}^k j is the target value (desired output) of the j-th component of the output for pattern p and o_{pj}^k is the j-th unit of the actual output pattern produced by the presentation of input pattern p at the time k, and j indexes all the output units[12]]].

7. XOR-Encryption

Exclusive-OR encryption works by using the Boolean algebra function exclusive-OR (XOR). XOR is a binary operator (meaning that it takes two arguments - similar to the addition sign, for example). By its name, exclusive-OR, it is easy to infer (correctly, no less) that it will return true if one, and only one, of the two operators is true.

The logical operation exclusive disjunction, also called exclusive or (symbolized XOR or EOR), is a type of logical disjunction on two operands that results in a value of true if exactly one of the operands has a value of true. A simple way to state this is "one or the other but not both."[18]

8.Suggestion methods

In this research a wave-type audio signal was used , and the function of Linear predictive coding encoders is to break up the sound signal into different segments and then send information on each segment to the decoder. The encoder sends information on whether the segment is voiced or unvoiced and the pitch period for the voiced segment which is used to create an excitement signal in the decoder. Four matrices (coefficients, voiced, pitch and gain) were fed into the Elman neural network to train it to obtain the values of hidden nod and weight matrix. To ensure addition confidentiality the hidden nod and weight matrix was encrypted using the XOR encoder .Finally, the output was sent. Figure (7) shows the flowchart of encryption algorithm.

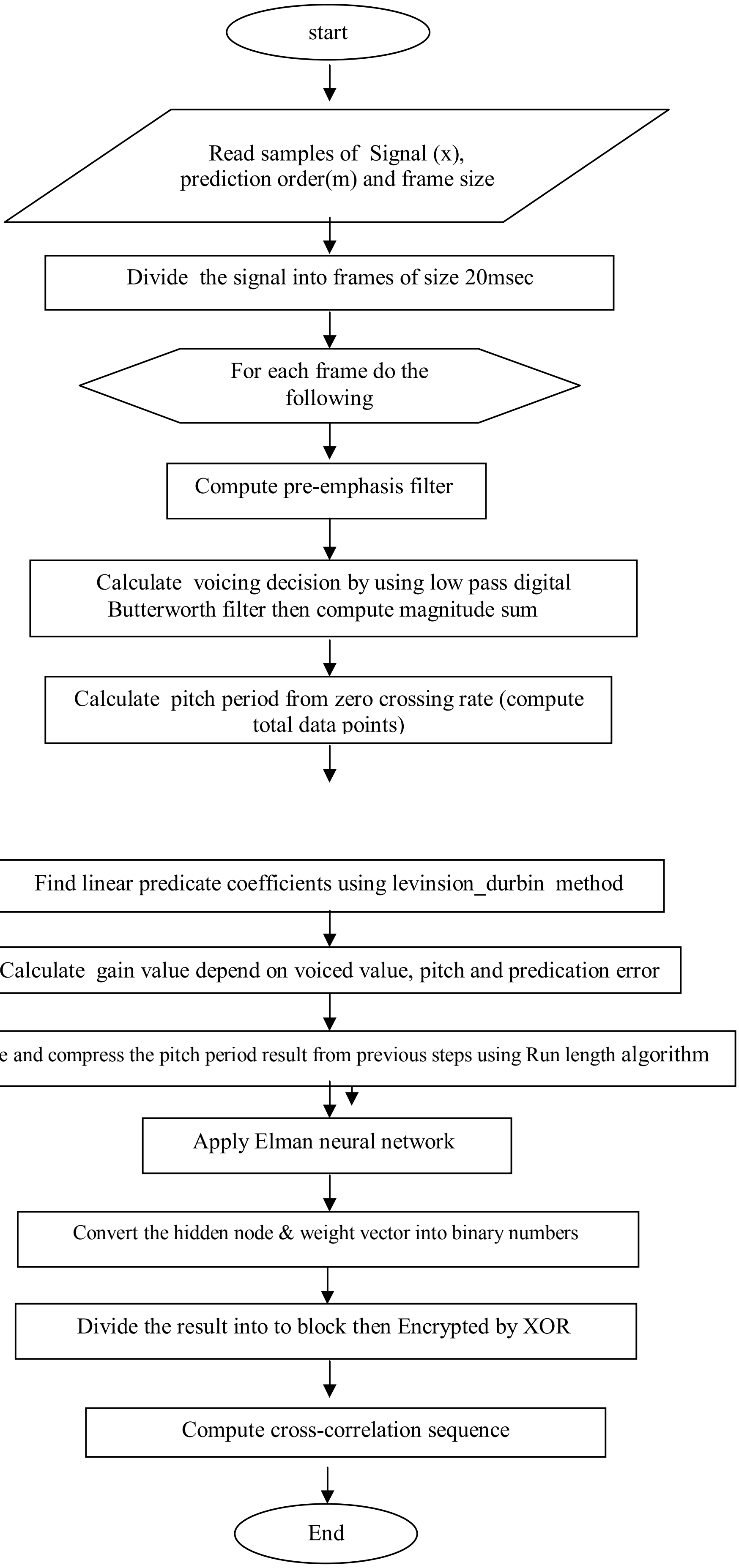

start
Read samples of Signal (x), prediction order(m) and frame size
Divide the signal into frames of size 20msec
For each frame do the following
Compute pre-emphasis filter
Calculate voicing decision by using low pass digital Butterworth filter then compute magnitude sum
Calculate pitch period from zero crossing rate (compute total data points)
Find linear predicate coefficients using levinsion_durbin method
Calculate gain value depend on voiced value, pitch and predication error
encode and compress the pitch period result from previous steps using Run length algorithm
Apply Elman neural network
Convert the hidden node & weight vector into binary numbers
Divide the result into to block then Encrypted by XOR
Compute cross-correlation sequence
End

Figure (7):Flowchart of suggestion method to Compression and Encryption

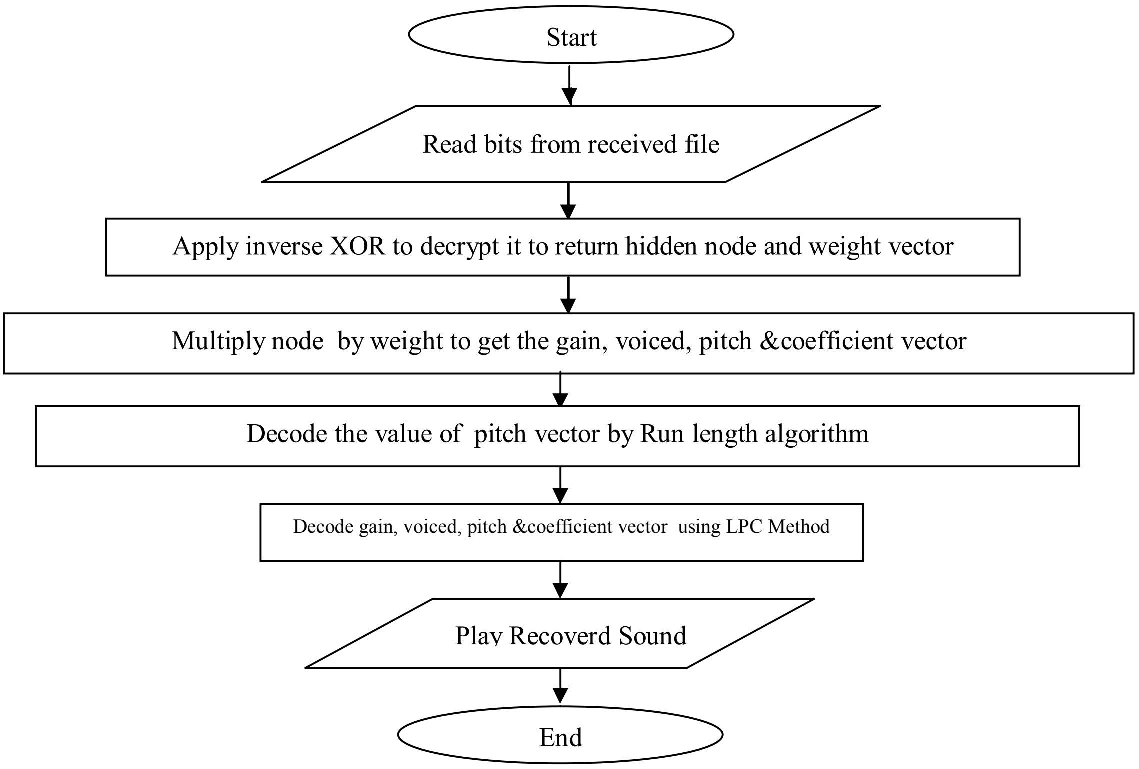

Start
Read bits from received file
Apply inverse XOR to decrypt it to return hidden node and weight vector
Multiply node by weight to get the gain, voiced, pitch &coefficient vector
Decode the value of pitch vector by Run length algorithm
Decode gain, voiced, pitch &coefficient vector using LPC Method
Play Recoverd Sound
End

Figure (8): Flowchart of suggestion method to Decryption and Decompression algorithm

9.Experimental Result

The program of the suggested method was tested on male and female speech files recorded using a microphone on a PC. The first (male) speech signal was sampled at 8000 samples/second and quantized at 8 bits/sample. Approximately 8 seconds of speech. Figure (9) shows a sample of original and recovered speech message, the second speech signal(female) sampled at 8000 samples/second and quantized at 8 bits/sample. Approximately 3 seconds of speech. Figure (10) shows a sample of original and recovered speech message and the third speech signal (female) sampled at 16000 samples/second and quantized at 16 bits/sample. Approximately 3.01 seconds of speech. Figure (11) shows a sample of original and recovered speech message. The quality of the recovered speech signal (male, female) is measured by using SNR and PSNR as shown in table (3).

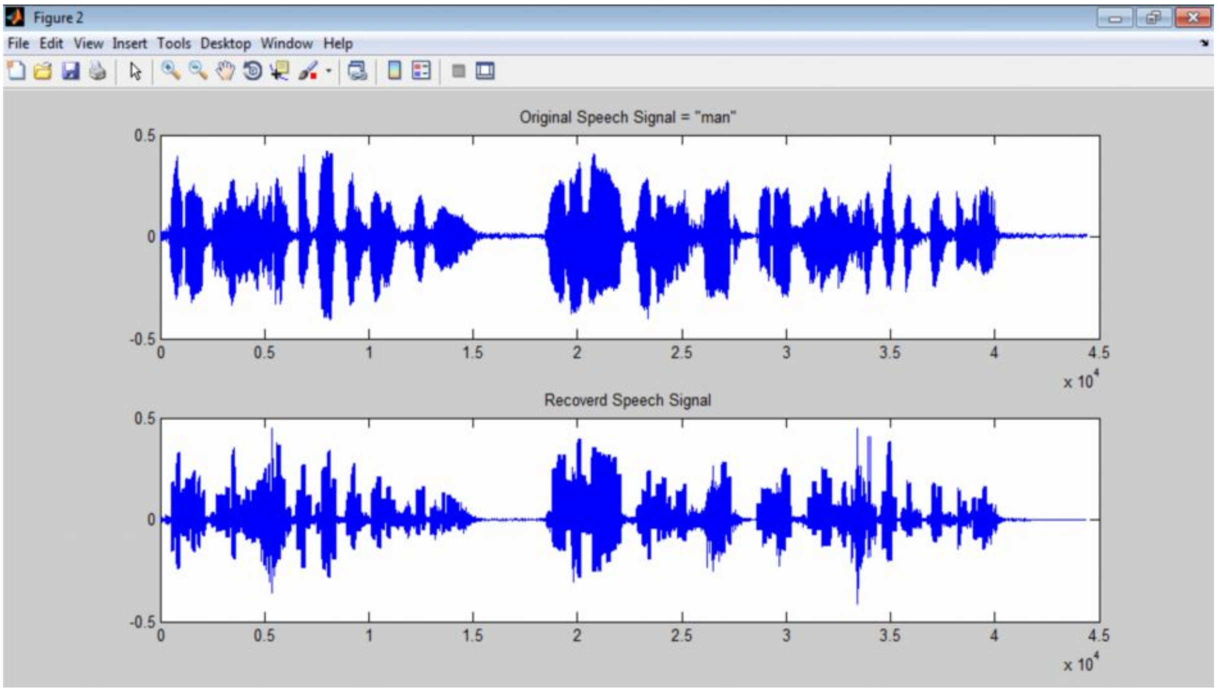


Figure (9): Samples of original and recovered signal(male)

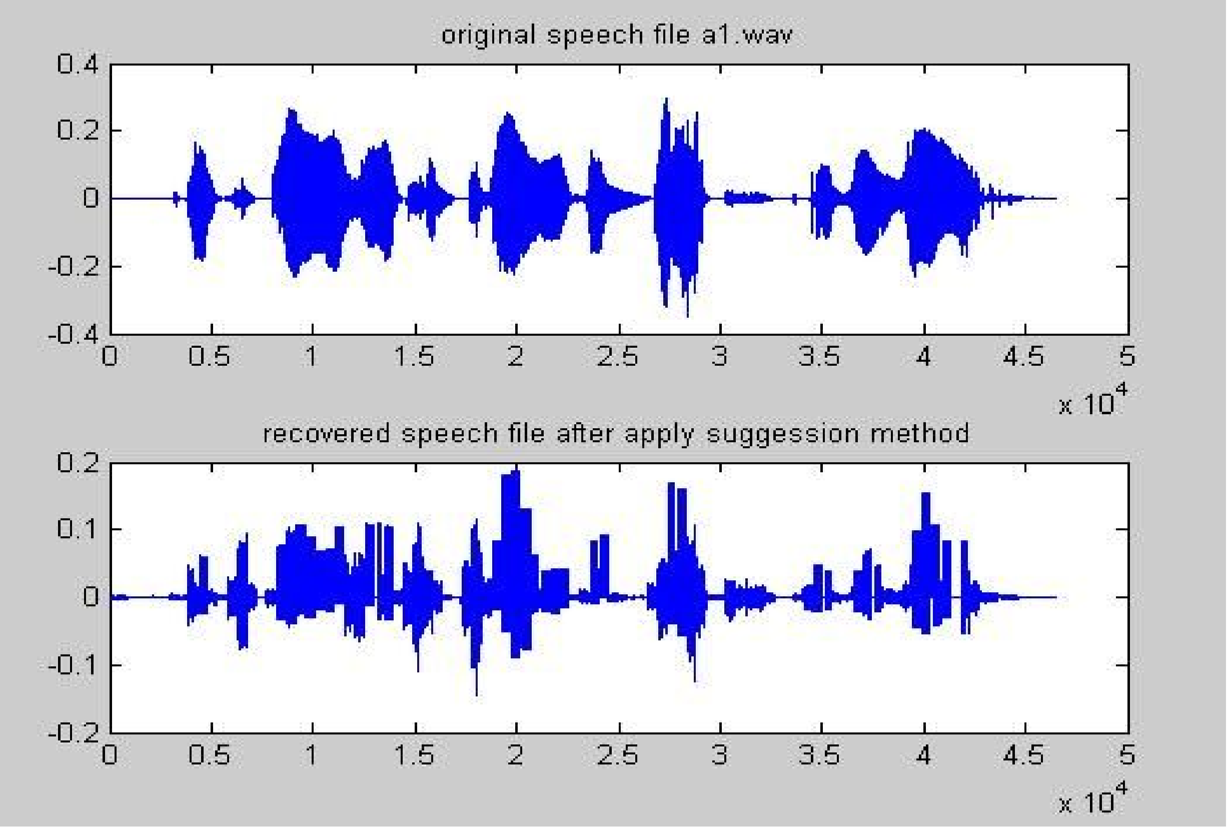


Figure (10): Samples of original and recovered signal (female)

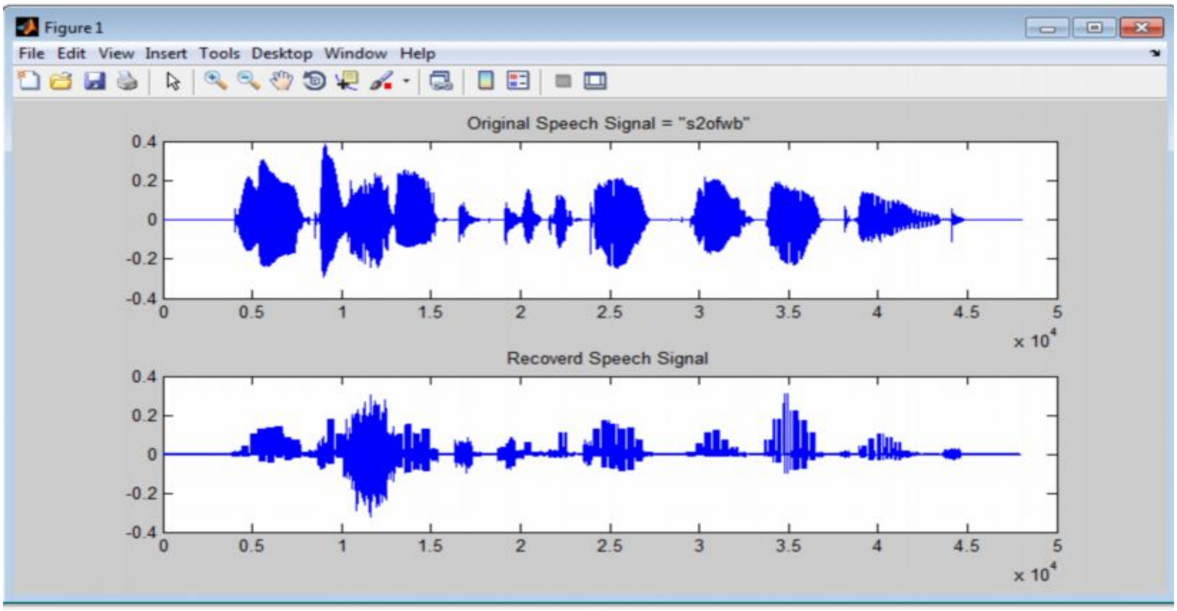


Figure (11): Samples of original and recovered signal (female sampling rate=16000 and 16 bit/sample)

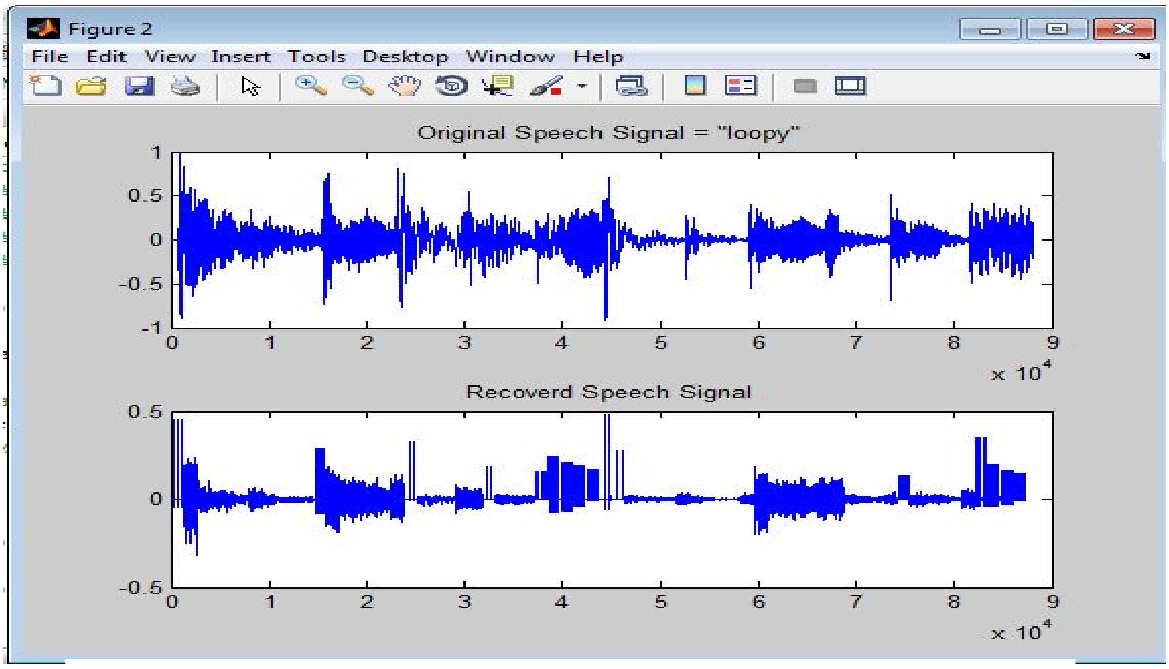


Figure (12): Samples of original and recovered signal (senfo sampling rate=44100 and 16 bit/sample)

Type of Signal	Sampling Rate	Resolution of sample	Performance Measures	
			SNR	PSNR
Signal1(male)	8000	8	58.789	67.9069
Signal2(female)	8000	8	48.736	41.8067
Signal3(female)	16000	16	33.057	39.955
Signal4(senfo)	44100	16	25.467	28.5932

Table (3) Results of applying the performance measures

10. Conclusions

1. Linear Predictive Coding achieves high compression rate by coding each 64000 bits/ second to bit rate of 2400 bits/second.
2. By using Linear Predictive Coding we achieve a coding level and security because we send the bits of human production of sound instead of transmitting an estimate of the sound wave
3. Using facilities of Elman neural network we increase the security by using 10 nods in the hidden layer. Finally we used XOR Encoding after dividing the result into two blocks.
4. After executing the above methods, it is concluded that the method is better and has a good performance for encrypted male speech signal , as it's clear from table (3) because women tend to have high pitch (fast vibration) while males tend to have low pitch (slow vibration).

References

1. Amol R.Madane,Zalak Shah,Raina shahand Sanket Thakur,"Speech Compression Using Linear Predictive Coding",Proceedings of the International Work Shop on Machine Intelligence Research,2009.
2. D.T. Pham and X. Liu, "Identification of linear and nonlinear dynamic systems using recurrent neural networks", *Artificial Intelligence in Engineering*, Vol.8,pp.90-97,1993.
3. Dr.R L K Venkateswarlu, Dr.Vasantha Kumari, A K V Nagayya,"Efficient Speech Recognition by Using Modular Neural Network", Int. J. Comp. Tech. Appl., Vol 2 (3), 463-470.
4. J.T. Conner, D. Martin, L.E. Atlas, " recurrent neural networks and robust time series prediction", *IEEE Trans. Neural Networks* 5(2) 240-253, 1994.
5. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals. Prentice Hall, EngleWood Cliffs, NJ 1978.
6. M. Kashif Saeed Khan, Wasfi G. Al-Khatib, Muhammad Moinuddin, "utomatic Classification of Speech and Music Using Neural Networks", *MMDB'04*, November 13, 2004, Washington, DC, USA. Copyright 2004 ACM 1-58113-975-6/04/0011.
7. M. M. El Choubassi, H. E. El Khoury, C. E. Jabra Alagha, J. A. Skaf , "Arabic Speech Recognition Using Recurrent Neural Networks ". Electrical and Computer Engineering Department Faculty of Engineering and Architecture – American University of Beirut 1107 2020, P.O.BOX: 11-0236,LEBANON.
8. Neural Network Library in Modelica Fabio Codec`a Francesco Casella Politecnico di Milano, Italy Piazza Leonardo da Vinci 32, 20133.
9. Omesh Wadhwani*, Amit Kolhe, Sanjay Dekate," Recognition of Vernacular Language Speech for Discrete Words using Linear Predictive Coding Technique", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-5, November 2011.
10. Patricia Melin, Jerica Urias, Daniel Solano, Miguel Soto, Miguel Lopez, and Oscar Castillo," Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic

Algorithms", Engineering Letters, 13:2, 13_2_9,August 2006.

11. Peter Morris, Audio Compression Manager, 2002.
12. Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Reviw, 1958, 65, 386-408.
13. S. Furui, Digital Speech Processing, Synthesis and Recognition.
14. S. Lawrence, C.L. Giles, S. Fong, "Natural language grammatical inference with recurrent neural network" ,*IEEE Trans. Knowledge Data Eng*. 12(1), 126-140, 2000.
15. Simon Haykin, Pearson Education, "Adaptive Filter Theory", 2002.
16. http://www.data-compression.com/index.shtml .
17. http://www.otolith.com/pub/u/howitt/lpc.tutorial.html.
18. http://www.cprogramming.com/tutorial/xor.html.