A novel Arabic Speech Recognition method using neural networks and Gaussian Filtering

Nidal M. turab [#1], Khalaf Khatatneh[*2], Ashraf odeh[#3]

[#]*IT Faculty & CS Department & Isra University*
Amman, Jordan

[1]Nedalturab@ipu.edu.jo
[2] kfkhatatneh@ipu.edu.jo
[3] ashraf.odeh@ipu.edu.jo

Abstract— **Phoneme recognition is a very popular area of research, especially in the field of machine intelligence, as it adds a distinguish touch in technology world in many fields and provides many features,**
This paper discussed a phoneme recognition using neural network technique by using a strong algorithm in preprocessing stage to get good results. Moreover, it explained the stages of phoneme recognition, preprocessing procedures such as getting a signal, sampling, quantization, determine an energy then using neural network to get a good results and enhancements. Gaussian LowPass filtering was used to get enhanced results of voice signal and lowering the noisy; and then use neural network in a training stage to train a system to recognize the speech signal, this paper has benefited from some of the previous results for other researchers and entered good ideas to enhance Arabic voice recognition.

Keywords— **Neural networks, voice recognition, phoneme, wavelet, Gaussian filtering**

I- Related Work

A. Waibel T. Hanazawa G. Hinton in [1] have presented a Time Delay Neural Network for phoneme recognition. They used two hidden layers, an input layer and output layer. . They stated that their TDNNs: was able to invent without human interference meaningful linguistic abstractions and the network does not vary with translation. They also stated that their TDNN's performance achieved 96.5% with HMR that achieved 93.7% on a speaker-dependent phoneme recognition task. In [2] F. Freitag, & E. Monte presented a phoneme recognition system based on predictive neural networks. They used h feed-forward and recurrent neural networks for the speech frames observation vectors. They conducted experiments using the Gaussian and Rayleigh distribution to study the discriminative quality of the prediction error as distortion measure. The results they obtained are compared with results obtained by a continuous density HMM system.in [3] Mohammed l,& Foyzul Hassan combined articulatory dynamic parameters (Δ and ΔΔ). With recurrent neural network (RNN) based phoneme recognition method. They stated their proposed method provides a higher phoneme correct rate , and reduces mixture components in HMM for obtaining a higher performance. In [4] Chiraz, Najet Arous and Noureddine presented a variant method of the Growing Hierarchical Self-Organizing, they called it Map GHSOM GH-Ad-RSOM, it is tree evolutionary recurrent self-organizing model. The proposed GHSOM variant is composed of independent RSOMs. The method shows better high vowel classification rates, In [5] N.Uma, ,A.P.Kabilan and R.Venkatesh proposed a system that combines the advantages of ANN's and MM's for further speaker independent speech recognition. They stated that such a system could integrate multiple ANN modules, such a system can classify 98% of the phonemes correctly.

II. Introduction

Speech recognition systems based on Short Time Fourier Transforms (STFTs) or Linear Predictive Coding (LPC) techniques can carry out speech recognitions based on speech features. However, these methods may not be suitable for representing speech, they assume signal stationary within a given time period and may therefore lack the ability to analyze localized accecnts accurately [8].

The Linear Predictive Coding (LPC) technique assumes some linearity of the speech model, while other Cohens approaches based that falls under general class of time frequency transforms such as Cone-Kernel and Choi-Williams find some use in speech recognition applications; it should be mentioned that their disadvantage of introducing unwanted cross-terms into the representation.[9]

Speech recognition depend on phoneme like units is attractive since it is free from vocabulary limitations, this is very important in highly inflected languages such as Arabic, what makes speech recognition complicated is that there exist some variations of the same root word according to word size units [10].

Humans can recognize speech easily although they came from several variations in a language, even with different speaking rates, different regional accents and different vocal effort of the received speech. Theses robust speech variations is a major achievement of human speech recognition are not been well understood yet. Many studies exist on sources of variation in spoken languages such as gender and age of the talker, the effect of certain speaking styles such as speaking clearly to reach a higher intelligibility and the influence on accent on speech intelligibility[10], [11],[12]. Other factors that have significant influence of speaking are emotion, stress, fatigue, and health condition; these sources of variations are correlated. Deletions, insertions, and articulation are examples influence factors of speaking rate on pronunciation Nowadays the sound recognition is an important technique that adds a distinguish touch in technology world in many fields as it provide many features, enhancements, qualities, solutions to many problems around the world.

Gaussian LowPass Filter (GLPF) is a low-pass filter passes acoustic low-frequency signals and reduces the amplitude of high frequency signals (such as noise). It is used for smoothing sets acoustic barriers, blurring they provide a smoother form of an acoustic signal, removing the fluctuations.

II-I System Architecture for Phoneme Recognizer

The speech recognition method composed of three steps: acoustic signal processing, phoneme recognition and word recognition, the input speech waveform can be digitized using phoneme; phonemes are recognized using artificial neural network (high level and low level) and subsequently words

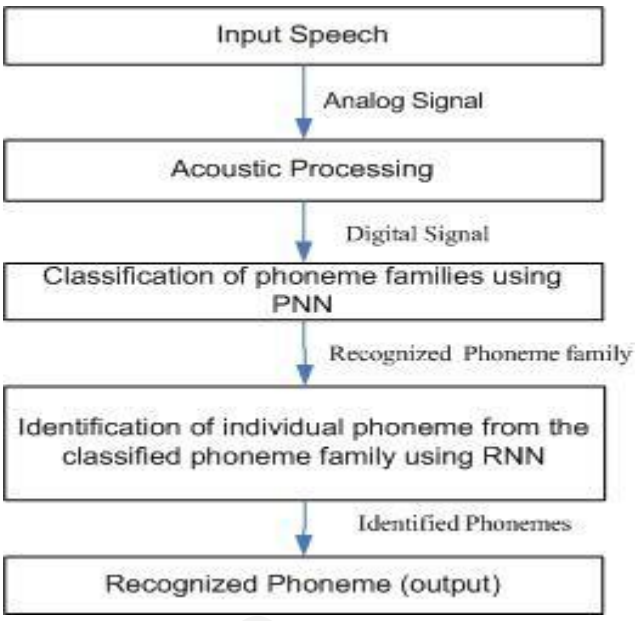

Figure1. System Architecture for Phoneme Recognizer

II-II SPEECH Recognition System

The general model for speech recognition system composed of five major phases [6]:

- Analog to Digital A/D Conversion phase: The input speech signal converted into an electrical signal, before performing A/D conversion.
- Segmentation: to reduce large memory usage and accelerating the computation complexity, speech segmentation is used
- Preprocessing: this stage includes filtering and scaling of the incoming signal in order to reduce any interfering noise and other external effect, filtering speech signal before recognition is very important process to remove noise related to speech signal that may be either low or high frequency noise. Figure 2 shows the effect of preprocessing on the speech signal of the Arabic word " Basrah".
- Feature Extraction: The goal of this stage is to represent any speech signal by a finite number of features, because the total amount of information in the speech signal is too much to process, and not all the information related to speech recognition process.

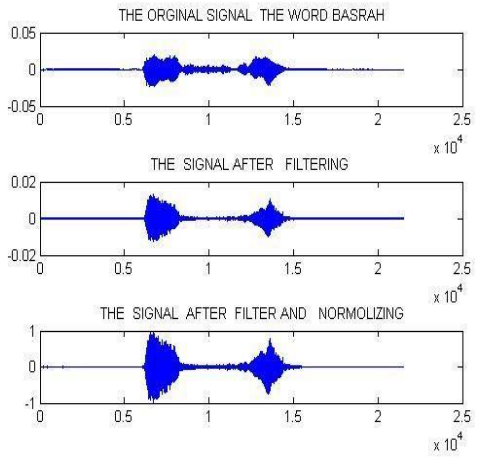

Figure2. The word "Basrah" before and after filtering and normalizing

II-III Role Of Neural Network Models For Developing Speech Systems

Neural networks are well known for capturing complex nonlinear relations present in speech signal. Neural networks can be classified into three categories: FeedForward Neural Network (FFNN), FeedBack Neural Network (FBNN) and the combination of both; the last one is known as Competitive Learning Neural Network (CLNN). Feedforward neural networks can be used for classification and function approximation; feedback neural networks can be used for pattern storage and pattern association [12].The number of nodes in hidden layers, hidden layers training for the network and processing units have significant impact on the model performance. When the input and output patterns are the same the network performs the autoassociation, otherwise it runs heteroassociation. An autoassociative neural network (AANN) captures the distribution of the data depending on the constraints of the structure. It should be mentioned that heteroassociative neural network (HANN) mainly used for some sort of mapping function between the input and output patterns sets of the network [12].

II-IV Feedforward Neural Network And Multi-Layer (FFNN)

As mentioned earlier, FFNN is expected to capture the relationship between the input and output feature vectors of the given training data; it is known that a neural network with two hidden layers can realize any continuous vector valued function. The activation function for the units is linear at the input layer is linear, and nonlinear at the hidden layers. The three influence factors of the network are the size of the training set, the architecture of the neural network, and the complexity of the problem [14].

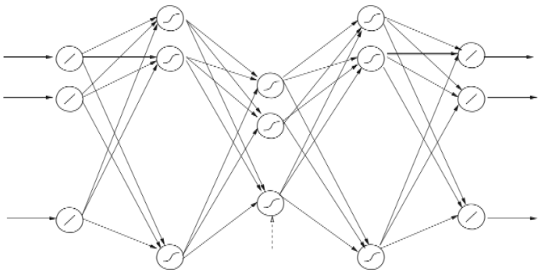

Figure 3. Five Layer Symmetric FFNN Model [15].

III. Wavelet Transform for Feature Extraction

Discrete Wavelet Transform (DWT) had been suggested for phoneme extraction in speech recognition, and high-energy wavelet coefficients were used as features. These features are not very reliable sense the DWT is a shift variant and requires additional processing for shift adjustment, and to exceed this problem, total energy of the wavelet coefficients in each frequency band was proposed as vector. However, it can only decomposes lower frequency bands.

Wavelet Packet (WP) decompose both lower and higher frequency bands, and is suitable for speech recognition applications. Alternatively; Admissible Wavelet Packet (AWP) can decompose either lower or higher frequency, and can be useful in benefiting from better time resolution rather than calculating the total energy in each band.

Wavelet transform is a time frequency analysis that is suitable for non-stationary signal analysis, while discrete wavelet transform and wavelet packet analysis that used for speech recognition applications. Wavelet decomposition is applied to extract the features of the signal [16].

$$\widehat{w}[f(t)] = F_p(t) = \int f(T)\widehat{w}p(t,T)dT$$

$$\widehat{w}\,p(t,\tau) = \frac{1}{\sqrt{|P|}}\widehat{w}\left(\frac{t-\tau}{p}\right) \quad \text{…….. Equ.1}$$

Where $\widehat{w}[f(t)]$ is the wave of the time-to-frequency relation, t: time in microseconds and τ is the shift time estimated that declared in equ.1.1

$$W[T_x(a,\tau)] = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-\tau}{a}\right)dt$$

$$W[T_x(a,\tau)] =< x(t), \varphi_{a\,\tau}(t) \quad ..$$

Equ.1.1

Where $W[T_x(a,\tau)]$ is a function of scale and position and contains many wavelet coefficients, dt is the diversion time in original time in neural network.

The wavelet transform is a convolution transform in with a factor P that provides expansion and translation properties into the convolution shown in equation 1; this gives us the ability to split signals into sub-bands of different bandwidth [16],The above equations was used to extract the Arabic phoneme "sh" from the word "Basrah" as shown in figure 4.

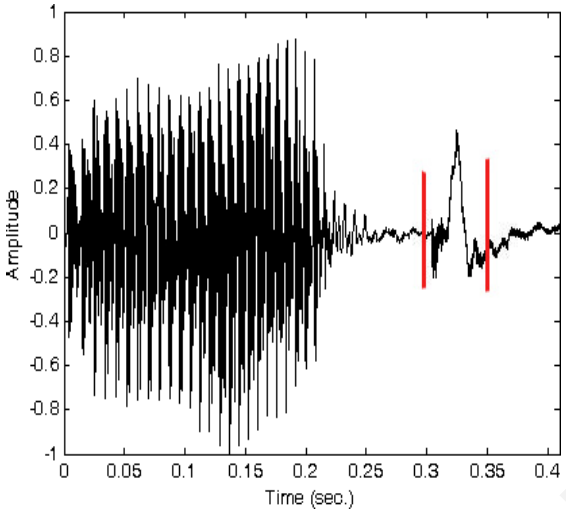

Figure 4. Short duration of the Arabic phoneme/sh/.

In wavelet sub-band based, it is important to consider influence effects of the neighbouring phonemes. A phoneme is divided into three equal segments, the first segment is influenced by the previous phoneme, the second segment has specific information about the current phoneme with minimum influence of previous or subsequent phoneme, and the third phoneme has transition information from current to next phoneme.

The output of each sub band is divided into three segments and the total energy (Eij) of the wavelet coefficients (sijk) in ith segment corresponding to jth sub band is calculated as [16]:

$$E_{ij} = \sum_{k=1}^{N_{ij}} S_{ijk}^2 \qquad i = 1,2,\dots\dots,L \qquad \text{.....}$$

Equ.2

Equation 2 is a general summation formula with Nij where i represent the segment number, j is the number of sub-bands in speech entered in our technique and L is the number of sub-bands; and Nij is the number of wavelet coefficients in the ith segment and the jth sub-band. The energy from *i* to *j* calculated in equation 3

$$FE_{ij} = log_{10}(w_i E_{ij}) \qquad \text{....}$$

Equ.3

Where w_i is the weighting factor for the ith segment.

Finally, a Discrete Cosine Transform (DCT) is applied to the log energy values of the ith segment as:

$$\text{F(u)}=c(u)\sum_{j=0}^{L-1}\cos\left(\frac{(2j+1)u\pi}{2L}\right) \qquad \text{... .Equ. 4}$$

$$c(0)=\sqrt{\frac{1}{L}}\,,c(u)=\sqrt{2/N}\;\; 1\le u\le(L-1)$$

we have developed a new technique by using Neural network methodology that includes all critical data and try remove redundancies and remove noise and distortion based on automatic speech recognition.

The speech waveform is sampled at frequency ranges from 6.6 kHz to 20 kHz to represent the voice speech. The vectors typically used between 10 and 20 parameters (i,j) in neural network which usually computed every 10 or 20 millisecond .Our neural network learned speech knowledge automatically to represent this new result in graphs in the next paragraphs for Arabic recognition words.

The result is L × 3-phoneme matrix of the DCT coefficients corresponding to the L sub bands and three segments, these phonemes give the variation of the energy in sub bands within each segment.

III-I Gaussian LowPass Filter (GLPF)

GLPS can be used to define the symbol time of the word to be measured in microseconds and the number of impulse response (filter span) as shown in figure 5, where D(u,v) is the distance from the centre of the frequency rectangles.

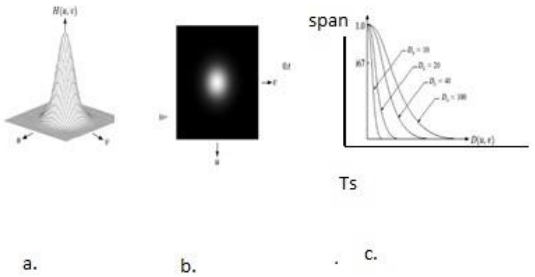

Figure 5 (a) Plot of GLPF (b) 2-D GLPF (c) plot of GLPF for different values of D(u,v) [2]

Now the parameters in equation 4 were used in

them Matlab Rectangle to measure the frequency of rectangles as shown in figures 6 that shows Gaussian low pass filter frequency response To study the effect of these parameters (filter span and symbol time.

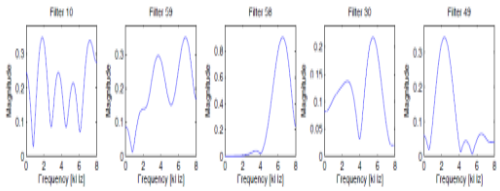

Figure 6. Frequency responses of filters learned in the first convolutional layer

Original wavelet coefficient transformed from sound wave of the phoneme sh", relation between Frequency and Time is shown in figure7

Figure 7

Filtered wavelet coefficient transformed from sound wave of the phoneme "sh" (using Gaussian low pass filter) is shown in figure 8

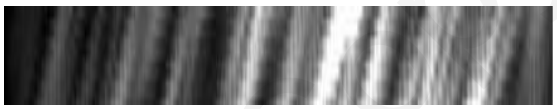

Figure 8

Sampling wavelet coefficient transformed from filtered sound wave of the phoneme "sh" is shown in figure 9

Figure 9

Figures 7,8 and 9 represent the frequency before the new technique was used in Matlab to get frequency voice results shown in figure 10

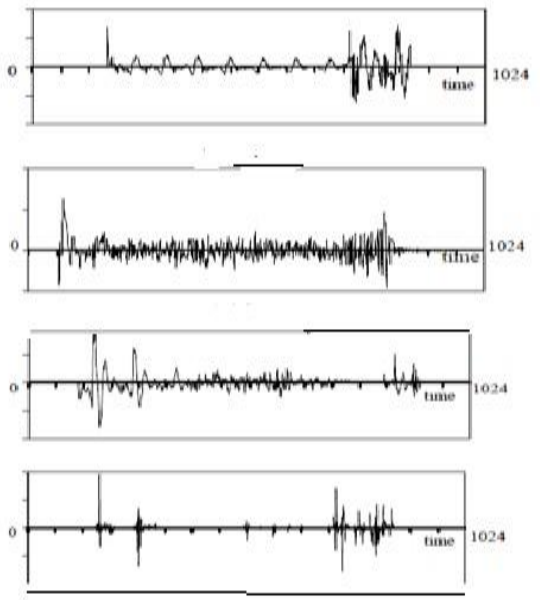

Figure 10: Wavelet Transforms of Arabic word “Basrah”

This new technique enhances the recognition of any Arabic voice word. We have used the stages to filter any noise that interfere with voice. These new results we obtained using neural networks helps us to use some coefficients to change the voice entered each time to this technique.

Conclusions

Sound recognition is an importance technique that it add a distinguish touch in technology world in many fields it provide many features, enhancements, qualities, solutions to many problems.

This paper discussed phoneme recognition using neural network technique, it explained the stages of phoneme recognition, getting a signal, sampling, quantization, determine an energy. It is discussed the use of Gaussian LowPass Filtering to get an enhanced results or signals, lowering , and finally it used the neural network -after training stage -the phoneme recognition to get very good speech recognition results.

References

1. 0 A. Waibel T. Hanazawa G. Hinton * IC Shikan and K.lang “ Phoneme Recognition: Neural Networks vs Hidden Markov Models “ http:// isl. anthropomatik.kit.edu/cmu-kit/downloads/ Phoneme_Recognition _Neural_ Networks (1).pdf
2. F. Freitag, & E. Monte “Phoneme recognition by means of predictive neural networks” Biological and Artificial Computation: From Neuroscience to Technology Lecture Notes in Computer Science Volume 1240, 1997, pp 1136-1143
3. - Mohammed Rokibul Alam Kotwal,& Foyzul Hassan and et la “Recurrent Neural Network Based Phoneme Recognition Incorporating Articulatory Dynamic Parameters Advances in Computing and

Communications Communications in Computer and Information Science Volume 192, 2011, pp 349-356
4. Chiraz Jlassi, , Najet Arous and Noureddine Ellouze “Phoneme Recognition by Means of a Growing Hierarchical Recurrent Self-Organizing Model Based on Locally Adapting Neighborhood Radii” Cognitive Computation September 2010, Volume 2, Issue 3, pp 142-150
5. N.Uma Maheswari, ,A.P.Kabilan and R.Venkatesh “SPEAKER INDEPENDENT PHONEME RECOGNITION USING NEURAL NETWORKS” Journal of Theoretical and Applied Information Technology Volume 2 no.6 pages 230-235
6. Ghassan S. Mosa and adbuallah Ali “ Arabic Phoneme recognition using Neural Fuzzy Petri Net and LPC Feature Extraction ” Signal Processing: An International Journal (SPIJ) Volume (3) : Issue (5) pages 161-171
7. Mahmoud.A.Osman, Nasser Al, “Speech compression using LPC and wavelet” 2010 2nd International Conference on Computer Engineering and Technology volume 7 pages 93-99
8. C.J.Long and S.Datta “Wavelet Based Feature Extraction for Phoneme Recognition”, http://www.asel.udel.edu/icslp/cdrom/vol1/239/a239.pdf
9. K. Elenius and G. Takács “Phoneme Recognition Using Multi-Layer Perceptrons” http://www.speech.kth.se/prod/publications/files/1168.pdf
10. Bernd T. Meyer, Tim Jürgens,” phoneme recognition depending on speech-intrinsic variability” J. Acoust. Soc. Am., Vol. 128, No. 5, November 2010, Pages: 3126–3141.
11. K SREENIVASA RAO “Role of neural network models for developing speech systems” Sadhan ¯ a¯ Vol. 36, Part 5, October 2011, pp. 783–836.c Indian Academy of Sciences.
12. BT Meyer” Human and automatic speech recognition in the presence of speech-intrinsic variations” http://d-nb.info/1007410892/34.
13. Randall Matignon” Neural Network Modeling Using Sas Enterprise Miner” 2005
14. Anil Kumar Vuppala “Neural Network based Feature Transformation for Emotion Independent Speaker Identification” Springer International Journal of Speech Technology Report No: IIIT/TR/2012/-1, pp335-349.
15. O. FAROOQ, S. DATTA “wavelet sub-band based temporal features for robust hindi phoneme recognition” Int. J. Wavelets Multiresolut Inf. Process. Volume 08, Issue 06, November 2010
16. N.UmaMaheswari,A.P.Kabilan , R.Venkatesh, "SPEAKER INDEPENDENT PHONEME RECOGNITION USING NEURAL NETWORKS",India, 2005 - 2009 JATIT
17. L.Mesbahi and A.Benyetto,”Continuous speech recognition by adaptive temporal radial basis function,” in IEEE international conference on systems,Manand Cybernetics,2004,pp.574-579.
18. Bernd T. Meyer, Tim Jürgens, Thorsten Wesker, Thomas Brand, and Birger Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability", Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, D- 26111Oldenburg, Germany , September 2010.
19. O. FAROOQ, S. DATTA, M.C. SHROTRIYA, " Wavlet Sub-Band Based Temporal Features For Robust Hindi Phoneme Recognition ", India.
20. Mahmood Yousefi Azar • Farbod Razzazi," A neural predictive coding feature extraction scheme in DCT domain for phoneme recognition", 15 September 2010.
21. C.J.Long and S.Datta, " Wavelet Based Feature Extraction for Phoneme Recognition", Loughborough LE11 3TU, UK.
22. Wesfried, E; Wickerhauser, M.V. “Adapted local trigonometric transforms and speech processing,” IEEE SP41,3597-3600 (1993).
23. Rafael C.Gonzalez, Richard E. Wood, "Digital Image Processing", third edition.
24. K SREENIVASA RAO, " Role of neural network models for developing speech systems", Indian Institute of Technology Kharagpur.
25. YIMING HUANG, Janusz A. Starzyk, Dennis Irwin, " Phoneme Recognition Using Neural Nework And Sequence Learning Model ", March 2009

Author profiles

Dr. Nidal Turab received his BSc degree in communication engineering from university of Garounis, Benghazi, Libya 1992 and M. Sc. in Telecommunication Engineering from University of Jordan, Amman, Jordan, 1996. Ph.D. in Computer Science. Polytechnic University of Bucharest, Faculty of Automatic Control and Computers, 2008. His research interests including Wireless Networks Security Issues, cloud computing, Encryption and Decryption Systems. Networking security. Currently, He is working at Isra University as an assistant Prof.

Dr. Khalaf Khatatneh is an expert in Information Technology (IT); he obttined his PhD degree from France university of Rouen in 2005, his field of research Artificial in Intelligence. Now he is in Sabbatical vacation at Al-Isra University for the year 2013/2014.

Dr. Ashraf A. Odeh an Assistant Professor in Computer Information System at Isra University-Jordan. He received a B.Sc. degree in Computer Science in 1995 and M.Sc. degree in Information Technology in 2003. With a Thesis titled "Visual Database Administration Techniques", He received PhD from department of Computer Information System in 2009 with a Thesis titled "Robust Watermarking of Relational Database Systems". His research interests include image processing, Watermarking, Relational Database, E-copyright protection, E-learning and E-content. He has submitted a number of conference papers and journals. Also he has participated in a number of conferences and IT days.