Speech Recognition Using Neural Networks

Bassam M. El-Zaghmouri

***Abstract*—** Speech recognition is an important part of human-machine interaction which represents a hot area of researches in the field of computer systems, electronic engineering, communications, and artificial intelligence. While speech signal is very complex and contains huge number of sampling points, the extraction of features from its time and frequency domain is very complex by analytical methods. The neural network capabilities to estimate the complex functions make it very reliable in such applications. This paper presents speech recognizer based on feed forward neural network with multi layer perceptron structure. The speech is preprocessed by two methods; discrete wavelet transformation (DWT) and principal component analysis (PCA). The results and structure are presented and comparison is make over them.

***Keywords*—** Neural Networks, MLP, Voice, Sound Recognition, Wavelet Transformation, Principal Component Analysis.

I. Introduction

THE problem of speech recognition deals with determining the identity of a given speech segment using a predefined set of samples. Different recognizers could be used to recognize the speech sentence (i.e. neural networks, genetic algorithms, statistical approaches, fuzzy logic, etc) depending on different set of features that could be extracted from the speech segment, such as LPC (Linear predictive coefficient), wavelet transformation (discrete and continues), DCT discrete cosine transform, etc. [1].

The main steps of speech recognition starts with preprocessing the sound signal to perform sampling and quantization; this depends on the sound acquisition tool that is being used; and then performs feature extraction. This research concerned with the features that extracted after wavelet transformation (DWT) and principal component analysis (PCA). Finally, the extracted features are fed to a pattern recognition phase (classifier). This field is still under intensive study of modern researches 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.

The voice biometrics has a main place in computer systems and access controls. Speech and speaker recognition has a role of protecting the user's identity, commanding in additions to the computerized data. Such systems have become increasingly difficult. The main concept of security is authentication identifying or verifying the voice command [3].

Bassam M. El-Zaghmouri is with University of Jerash. (Email ID – el_zaghmouri@ yahoo.com).

The biometrics is the concept of measuring unique features of human depending on bio-analysis. Such as a fingerprint, voice, face, etc [2].

The accumulated problems of the traditional methods of voice commands cause a major importance of intelligent methods. The shortcoming in these methods that the keypad or input switches that the user could use in touch to enter a specific command to the computer or embedded system. Sometimes, the user is not able to use his hands to reach the input keys or switches, or even the interactive screen, also, in the rise technology trends, the user demands grows more and more in human machine interaction [4].

One of the widely used systems for human machine interaction is the speech recognition technique. Since the human can easily use his / here voice for commanding or inputting, it is useful to discriminate between commands using speech sentences directly. The idea of speech recognition is to verify the individual speech sentences against a stored sentences patterns, and not to understand what is being said while speech understanding or speech-to-text conversion is concerned with understanding what is being said. In the field of speech recognition many techniques have been developed such as Hidden Markov Models, Neural network, Fuzzy logic and Genetic algorithms [6].

Human voice has two types of information high-level information and low level information. High-level information is values like dialect, an accent (the talking style and the subject manner of context).

Voice recognition deals with low-level information from the human speak voice, like pitch period, tone and spectural magnitude, rhythm, in addition to the bandwidth and frequency of an individual sound, is considered to be features. For voice recognition, another information attributes can be taken as features such like Mel-frequency Cepstrum Coefficients (MFCC) and Linear Predicative Cepstral Coefficient (LPCC). For robust voice recognition system, wavelet transform coefficients are used continuous or discrete. The concentration is on discrete wavelet [3].

The use of wavelet transform decomposition was chosen based on its high coefficients determination which contains recognizable features of the speech and voice, including the person identifiers. Many features can be extracted using wavelet analysis, where the purpose of this paper is to specify the best analysis method that gets the best voice features.

Also, the principal component analysis (PCA) is robust feature extraction method that could be used for feature extraction from speech signal. The PCA is statistical method that converts the speech samples into correlated variables.

II. PROBLEM STATEMENT

This paper goal is to develop the recognition process in the field of human machine interaction through voice or speech commanding by suggesting minimized number of features that would not affect the system accuracy and study the effect different preprocessing techniques; discrete wavelet transformation (DWT), principal component analysis (PCA), and the combination of both.

The recognition system that is used to select the minimized feature set using feed forward Neural Network (Multi-Layer Perceptron) and with back propagation learning algorithm. The suggested Neural Network will be trained with different sets of features extracted from deferent levels of DWT, PCA, or both then the trained recognition system will be tested to determine the accuracy of the proposed speech recognition system.

The proposed approach consists of three methodologies for feature extraction in preprocessing phase:

1. The input to the NNet is DWT of the signal.
2. The input to the NNet is PCA of the signal.
3. The input to the NNet is both DWT and PCA of the signal

Also, this paper presents a comparison between those three methods results.

III. PREPROCESSING AND FEATURE EXTRACTION

The preprocessing phase is consists of two functions; signal re-sampling and feature extraction. The speech signal that is being used either for training or testing, is recorded using custom function that is specialized for the algorithm presented in this paper. The sampling frequency is 8000sample per second, with 16bit buffer size, and stereo-channel recording. To make sure, all signals are fall withing the same time period, the re-sampling process re-meet the condition of the voice recorder, and make the period for all speech sentences to be the same as 1 second.

In fact, the 8000kpps is the lowest sampling rate that couldn't affect the nature of the signal. When trying to get down in that rate, the signal will start damage. Figure-1 shows the re-sampling effect on the speech signal, where figure-1a represents the signal before re-sampling and figure-1b represents the signal after re-sampling.

After re-sampling, the signal is subjected to be converted into different domain in terms of coefficients to be then entered to the neural network.

Two principles are used for feature extraction in this paper; wavelet transformation (DWT) and principal component analysis (PCA). This paper implements three separated methodologies for feature extraction as preprocessing of the speech samples before make them input for the neural network. Where the neural network input - depending the preprocessing methodology - could be:

1. Low frequency coefficients of DWT of the speech signal.
2. PCA of the speech signal.
3. PCA of the low frequency coefficients of DWT of the speech signal.

There are two main factors those should be selected in the design of wavelet based biometrics; number of levels, and the minimum set of coefficient extracted from level(s) that leads to better discrimination. This topic still under research since the most important thing is to keep the best recognition ability with minimum feature set to speed up the verification operation during searching huge voice dataset.

The bases functions that can be used in wavelet decomposition as the mother wavelet are including Haar wavelet, Daubechies wavelets, Coiflet1 wavelet, Symlet2 wavelet, Meyer wavelet, Morlet, Mexican Hat wavelet. When working with discrete signals, the selection is limited to Harr wavelet and Daubechies wavelet. Hence, the Haar wavelet causes significant leakage of frequency components and is not well suited to spectral analysis of speech, whereas, the Daubechies family of wavelets has the advantage of having low spectral leakage and generally produces good results [7].

The continuous wavelet transform (CWT) is defined as the sum over all the time of a signal that multiplied by scaled and shifted wavelet function. The result is a set of Wavelet coefficients, which are a function scale and position [1].

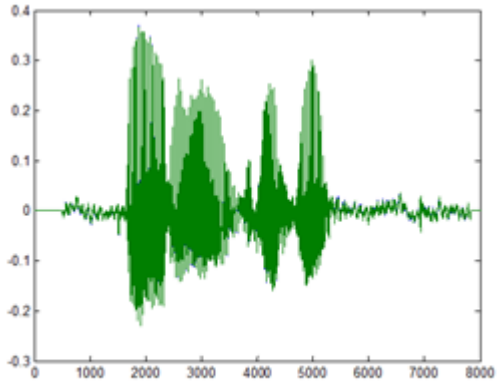
Dilation and translation of the Mother function, or analyzing wavelet Φ(x) defines an orthogonal basis, the wavelet basis is shown in equation -1 [5]:

$$\Phi_{(s,l)}(x) = 2^{\frac{-s}{2}} \Phi(2^{-s}x - l) \quad \ldots \quad (1)$$

Where,
"s" are integers that scale and dilate the mother function Φ(x) to generate wavelets, such as a Daubechies wavelet family.

"s" is wavelet width scale index.

"l" is the location index that gives its position.



(a)

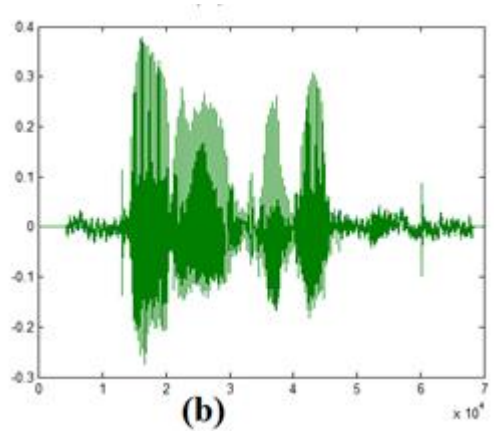


Fig. 1 Re-sampling of speech signal, (a) original signal, (b) re-sampled signal

It should be noticed that the mother functions are rescaled, or dilated by powers of two, and translated by integers. The wavelet bases is interesting because of the self-similarity caused by the scales and dilations. Once the mother functions are explained, everything about the basis will be clearer.

A special case of the wavelet transformation, the discrete wavelet transformation provides a compact representation of a signal in frequency and time. The discrete wavelet transform of specified signal can be computed by passing the signal through series of low-pass and high-pass filters to analyze the frequencies. The outputs that generated are then down sampled by two, so the output is half of original signal size [5].

The principal components analysis (PCA) is statistical analysis of time signal in the terms of orthogonal transformation that finds the correlation between set of data. The PCA converts the correlated variables set into uncorrelated variables set, where those uncorrelated set is commonly known as "principal components". As wavelet transformation minimizes the size of data set with every one level decomposition, the PCA minimizes the data samples size as the correlation level goes deeper. In some cases, the number of PCA components will be the same as the original signal samples, that could be causes just in a case where the data samples are originally uncorrelated.

The PCA in this paper enables to remove the redundant factors of the data set samples, thus, minimizing the data set at half (as selected in this paper), where it is enable to minimize the data samples size into less than one thenth.

IV. METHODOLOGY

This system is developed to recognize speech segments based on multi-layer perceptron (feed forward) neural network. The working samples are consists of five different sentences that is recorded fifteen times as different records. Those sentences are preprocessed and then divided into training samples and testing samples.

The main characteristic of any speech recognition is the determination of the speech segment ID. speech recognition system consists of several modules in addition to the classification engine. The proposed system consists of three main modules as shown in figure-2.

The neural network architecture is illustrated in figure-3. It structured form three layers; the input layer which contains 100 neuron and their activation function is tangential sigmoid, one hidden layer that contains 25 neuron and their activation function is logarithmic sigmoid, and the output layer consists of 5 neuron with tangential sigmoid activation function.

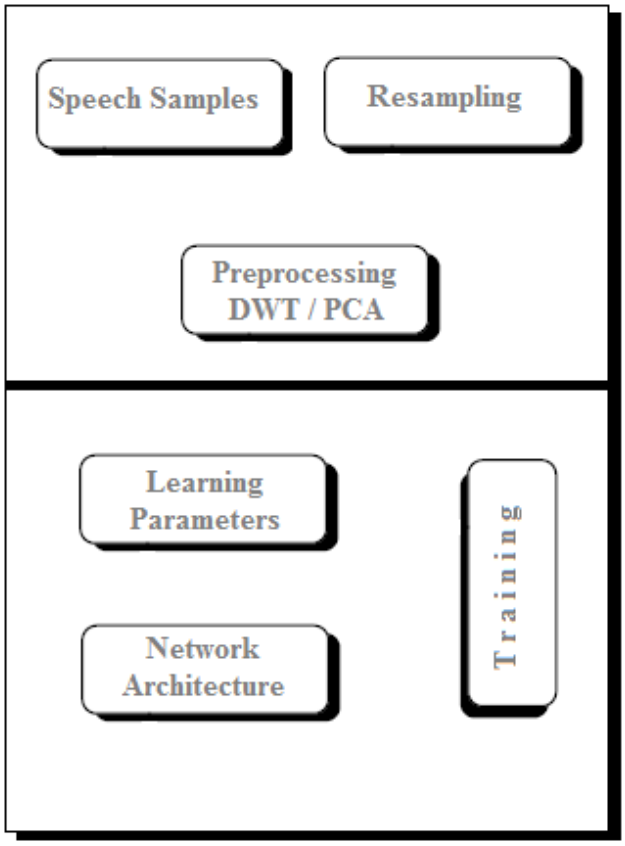


Fig 2. Block Diagram of the presented System.

The output layer ensures one separate neuron for each distinguishable output, thus, the sample that used for training and testing consists of five separate original speech sentence, so, each neuron of the 5 output units will be activated as 1 when it recognize its related speech sentence.

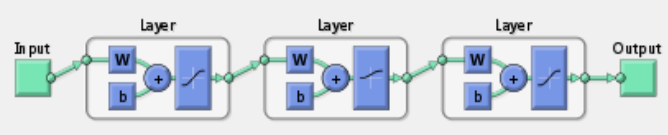


Fig 3. Structure of the presented neural network

Once, the speech sentence is recorded and preprocessed, discrete wavelet transformation (DWT) is being applied on debauched level-1 and the first network is trained based on them. Then, the PCA components are calculated and the second network is trained based on them. Then, the PCA were calculated for the DWT of the speech signal, and the third network is trained based on them. The result of this methodology will be illustrated in the next section.

V. RESULTS

In this paper, the data set consists of different speech sentences recorded directly from microphone. Those sentences are separated into two categories; training samples and testing samples. Once, each network of the three networks those described in the previous section is trained, the network then subjected to testing and evaluation.

The results of testing and evaluation is shown in table-1. The evaluation result on the same set that the networks are trained on it is 100%. This means that, the representational power of the network is capable to estimate the desired function in all different three cases. Also, it means that, the optimal set of network parameters (weights and biases) is capable to be gotten in reasonable training time. The test over non-trained data got different results than the training set. The accuracy is much lower over the testing set, and this comes from the fact that, the number of samples that was recorded for each speech sentence is not enough to make the network training represents the desired function for that sentence. This paper was based on 10 recorded samples for each speech sentence, and then changed the designed to be based on 15 recorded samples. The testing and evaluation was stated that, the testing error in case of 15 samples is much less than it in case of 10 samples, because of increase the number of training samples.

TABLE I
TESTING AND EVALUATION RESULTS.

Recognition / Network	Training set accuracy	Total set accuracy
Network 1 with DWT only	100%	88%
Network 2 with PCA only	100%	92%
Network 3 with PCA based on DWT	100%	89%

VI. CONCLUSION

This paper presented a speech recognition system that is based on neural network with signal preprocessing using two different techniques. The discrete wavelet transformation and the principal component analysis are used as signal preprocessing. Where the designed neural network is multi-layer perceptron.

The number of training samples is highly affecting the result of testing data evaluation. To get low validation and testing error, a reasonable number of training samples should be used.

The principal component analysis (PCA) gets results better than the discrete wavelet transformation (DWT) while simulation on non-trained set. Where both, outs the same result, with zero error over the trained data. This means that, the distinguishable data that could be easily estimated by the neural network becomes more clear when dealing with PCA as preprocessing than the DWT.

REFERENCES

[1] A.L. Graps, An Introduction to wavelets, IEEE Computational Science and Engineering, Vol 2. No. 2, pp. 50-61 1995 http://dx.doi.org/10.1109/99.388960

[2] Evgeny Karpov. Real-Time Speaker Identification, Master thesis. University of Joenssu, Department of Computer Science, 2003.

[3] Russell Kay, Biometric Authentication, Technical Report, CSO, the resource for security executives, 2005.

[4] R.V Pawar, P.P. Kajave, and S.N. Mali, Speaker Identification Using Neural network , WASET, Vol.12, No.7, PP. 31-35, 2005.

[5] chabane Djeraba, Hakim Saadane, Automatic Discrimination in Audio Documents , Nantes University, 2 rue dela Hossiniere, Bb 92208-44322 Nates Cedex3, France, 2000, pp.1-10.

[6] Belfast, Northern Ireland, France, 2003.

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

[8] Brain J. Love, Jennifer Vining, Xuening Sun, Automatic Speaker Recognition Using Neural Networks, Technical report, the University of Texas at Austin, 2004.

[9] F Murtagh, J.L Strack. O Renaud, On Neuro-Wavelet Modeling, School of Computer Science, Queens University

[10] George Tzanetrakis, Georg Essl, Perry Cook, Audio Analysis Using the Discrete wavelet Transform , Proceedings of WSEAS conference in Acoustics and music Theory Application , 2001.