International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:1, 2009

An Advanced Method for Speech Recognition

Meysam Mohamad pour, Fardad Farokhi

***Abstract*—**In this paper in consideration of each available techniques deficiencies for speech recognition, an advanced method is presented that's able to classify speech signals with the high accuracy (98%) at the minimum time. In the presented method, first, the recorded signal is preprocessed that this section includes denoising with Mels Frequency Cepstral Analysis and feature extraction using discrete wavelet transform (DWT) coefficients; Then these features are fed to Multilayer Perceptron (MLP) network for classification. Finally, after training of neural network effective features are selected with UTA algorithm.

***Keywords*—**Multilayer perceptron (MLP) neural network, Discrete Wavelet Transform (DWT) , Mels Scale Frequency Filter ,UTA algorithm.

I. Introduction

SPEECH is one of the most important tools for communication between human and his environment. Therefore manufacturing of Automatic System Recognition (ASR) is desire for him all the time [1]. In a speech recognition system, many parameters affect the accuracy of the Recognition System. These parameters are: dependence or independence from speaker, discrete or continues word recognition, size of vocabulary book, language constrains, colloquial speeches and recognition environment conditions. Problems such as noisy environment, incompatibility between train and test conditions, dissimilar expressing of one word by two different speakers and different pronouncing of one word by one person in several times, is led to made system without complete recognition; So resolving each of these problems is a good step toward this aim. A speech recognition algorithm is consisted of several stages that the most significant of them are feature extraction and pattern recognition. In feature extraction category, best presented algorithms are zero crossing rate, permanent frequency, cepstrum coefficient and liner prediction coefficient [2].

Generally, there are three usual methods in speech recognition: Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Artificial Neural Networks (ANNs) [3].

F. A. Author is with the Department of Electrical Engineering ,Azad University of Central Tehran Branch, Ponak, Hamila Ave, CO 80305 IRAN (corresponding author to provide phone: +98-21-66917877; mob: +98-911-193-5579; e-mail: meisam.mohamadpoor@gmail.com).

S. B. Author, Electrical Engineering Faculty, Azad University of Central Tehran Branch, Ponak, Hamila Ave, CO 80305 IRAN (corresponding author to provide; mob: +98-912-209-5782;e-mail: f_farokhi@iauctb.ac.ir).

Dynamic time warping (DTW) is a technique that finds the optimal alignment between two time series if one time series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series [4].

In speech recognition Dynamic time warping is often used to determine if two waveforms represent the same spoken phrase. This method is used for time adjustment of two words and estimation their difference. In a speech waveform, the duration of each spoken sound and the interval between sounds are permitted to vary, but the overall speech waveforms must be similar [5]. Main problem of this systems is little amount of learning words high calculating rate and large memory requirement.

Hidden Markov Models are finite automates, having a given number of states; passing from one state to another is made instantaneously at equally spaced time moments. At every pass from one state to another, the system generates observations, two processes are taking place: the transparent one, represented by the observations string (feature sequence), and the hidden one, which cannot be observed, represented by the state string [6,7]. Main point of this method is timing sequence and comparing methods.

Nowadays, ANNs are utilized in wide ranges for their parallel distributed processing, distributed memories, error stability, and pattern learning distinguishing ability [8]. The Complexity of all these the systems increased when their generality rises. The biggest restriction of two first methods is their low speed for searching and comparing in models. But ANNs are faster, because output is resulted from multiplication of adjusted weights in present input. At present TDNN (Time-Delay Neural Network) is widely used in speech recognition [9].

Since we are going to present an algorithm for speech recognition and the identity of the speaker is not important, our method should be word sensitive compare to the speaker sensitive ones.

II. Architecture of System

The overall architecture of our speech recognition system has been shown in the figure below. Our speech recognition process contains four main stages:

1- Acoustic processing that main task of this unit is filtering of the white noise from speech signals and consists of three parts, Fast Fourier Transform, Mels Scale Bank pass Filtering and Cepstral Analysis.

2- Feature extraction from wavelet transform coefficients.
3- Classification and recognition using backpropagation learning algorithm.
4- Feature selection using UTA algorithm [10].

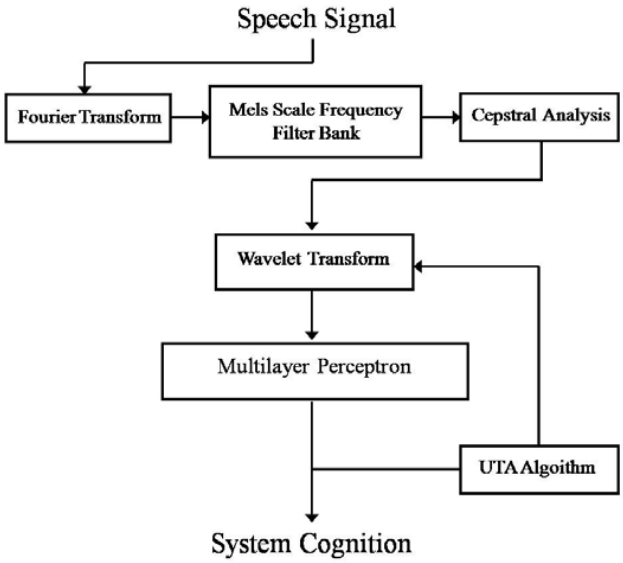


Fig. 1 System Architecture

III. Analysis of Data

Ann's performance depends on the pattern classification method. So Before feeding this speech data to the ANN for classification, we analyze whether any correlation exists between the spoken digits. The correlation of a spoken digit with itself should give a relatively high value than the correlation to the other digits. A high value in correlation signifies the similarity and a smaller value signifies that there is a marked difference in data allowing the ANN to easily classify the data. This correlation matrix is presented in Appendix 1 for our database.

IV. Preprocessing

The digitized sound signal contains relevant, the data, and irrelevant information, such as white noise; therefore it requires a lot of storage space [11]. Most frequency component of speech signal is below 5KHz and upper ranges almost include white noise that directly impact on system performance and training speed, because of its chromatic nature. So speech data must be preprocessed.

A. Mel Frequency Cepstral Coefficients System

To simplify the subsequent processing of the signal, useful features must be extracted and the data should be compressed. The power spectrum of the speech signal is the most often used method of encoding. The human ear performs something very similar to a Fourier Transform on incoming sound signals before passing the information on to the brain for analysis [12]. Mel Frequency Cepstral Analysis is used to encode the speech signal. Mel scale frequencies are distributed linearly in the low range but logarithmically in the high range, which corresponds to the physiological characteristics of the human ear [13]. Cepstral Analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the speech signal. The sequence of processing includes for each chunk of data:

- Shift raw data into FFT order,
- Find the magnitude of the FFT,
- Convert the FFT data into filter bank outputs,

So first, speech signal was transferred to frequency domain by Fast Fourier Transform (FFT). Resultant vector was converted to 25 sub vectors, respectively 5 sub vector with linear intervals under 2 KHz frequency and 20 sub vector with logarithmic intervals on upper frequencies. Then the set of Mel scale filter banks is shown below was implemented on it and energy values of upper frequencies are decreased. Sub arrays are combined with each other and Inverse Fast Fourier Transform (IFFT) is performed.

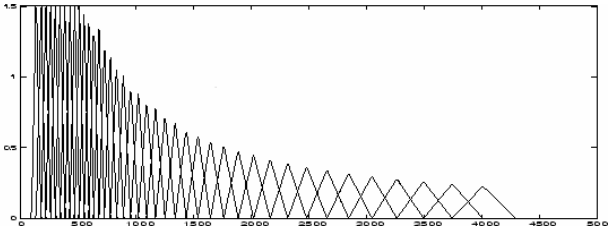


Fig. 2 Mel's Scale Frequency Filter Bank

B. Discrete Wavelet Transform

Wavelet transform can be viewed as the projection of a signal into a set of basis functions named wavelets. Such basis functions offer localization in the frequency domain. Compare to STFT which has equally spaced time-frequency localization, wavelet transform provides high frequency resolution at low frequencies and high time resolution at high frequencies [14]. Figure 3 provides a tiling depiction of the time-frequency resolution of wavelet transform.

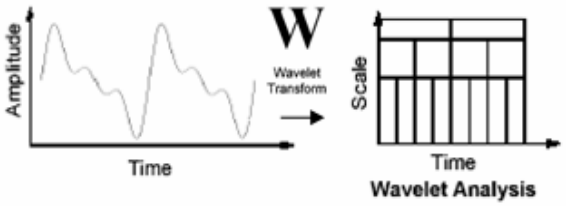


Fig. 3 Time-frequency tiling for DWT

The discrete wavelet transform (DWT) of a signal x[n] is defined based on so-called approximation coefficients, $w_{\varphi}[j_0,k]$, and detail coefficients, $w_{\psi}[j,k]$, as follows [15]:

$$w_{\varphi}[j_0,k] = \frac{1}{\sqrt{M}} \sum_n x[n]\, \varphi_{j_0,k}[n] \quad (1)$$

$$w_{\psi}[j,k] = \frac{1}{\sqrt{M}} \sum_n x[n]\, \psi_{j_0,k}[n] \qquad for\ j \geq j_0$$

where n = 0,1,2,..., M – 1, j = 0,1,2,..., J – 1, k = 0,2,..., 2j – 1, and M denotes the number of samples to be transformed. The basic functions $\varphi_{j,k}[n]$, and $\psi_{j,k}[n]$ are defined as:

$$\varphi_{j,k}[n] = 2^{\frac{j}{2}} \varphi\ [2^j n - k]$$
$$\psi_{j,k}[n] = 2^{\frac{j}{2}} \psi\ [2^j n - k] \quad (2)$$

$\varphi\ [n]$ is called scaling function and $\psi\ [n]$ wavelet function. For the implementation of DWT, the filter bank structure is often used. The approximation coefficients at a higher level are passed through a highpass and a lowpass filter followed by a down sampling by two, to compute both the detail and approximation coefficients at a lower level. This tree structure is repeated for a multi-level decomposition.

In the last part of preprocessing stage, we used Discrete Wavelet Transform (DWT) to extract speech features from Filtered signal and the outcome was coefficients array.

Therefore total energy value of each interval was calculated to provide requirement features for ANN feeding. In this part of the algorithm, we found that calcuted energy values of some intervals have negligible variation in different classes and sometimes energy value of an interval has a lot of variation in one class. So we thought it would be a good idea to select useful features before feeding them to ANN.

V. Feature Selection

Due to the good performance of the UTA algorithm, we selected it for feature selection. In this algorithm, average of one feature in all instances is calculated. Then the selected feature in all input vectors has been replaced by the calculated mean value. Then trained network are tested with the new features and data matrix. If the system cognition is decreased, that features was effective but if result didn't change or improve (noisy feature) that feature are considered as ineffective and should be removed from the input vector.

VI. MLP Neural Network

A neural network (NN) is a massive processing system that consists of many processing entities connected through links that represent the relationship between them [15]. A Multilayer Perceptron (MLP) network consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of multiple neurons. An artificial neuron is the smallest unit that constitutes the artificial neural network. The actual computation and processing of the neural network happens inside the neuron [16]. In this work, we use an architecture of the MLP networks which is the feedforward network with backpropagation training algorithm (FFBP). In this type of network, the input is presented to the network and moves through the weights and nonlinear activation functions toward the output layer, and the error is corrected in a backward direction using the well-known error backpropagation correction algorithm [17]. The FFBP is best suited for structural pattern recognition. In structural pattern recognition tasks, there are N training examples, where each training example consists of a pattern and a target class (x,y) [18]. These examples are assumed to be generated independently according to the joint distribution P(x,y). A structural classifier is then defined as a function h that performs the static mapping from patterns to target classes y=h(x). The function h is usually produced by searching through a space of candidate classifiers and returning the function h that performs well on the training examples during a learning process [19]. A neural network returns the function h in the form of a matrix of weights.

The number of neurons in each hidden layer has a direct impact on the performance of the network during training as well as during operation [20]. Having more neurons than needed for a problem runs the network into an over fitting problem. Over fitting problem is a situation whereby the network memorizes the training examples. Networks that run into over fitting problem perform well on training examples and poorly on unseen examples. Also having less number of neurons than needed for a problem causes the network to run into under fitting problem. The under fitting problem happens when the network architecture does not cope with the complexity of the problem in hand. The under fitting problem results in an inadequate modeling and therefore poor performance of the network.

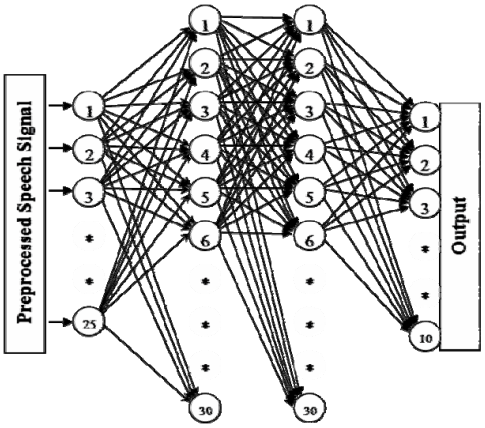


Fig. 4 MLP Neural Network architecture

Unfortunately, coming up with the right number of hidden layers and neurons for each hidden layer can only be achieved by trial and error. Many experiments have been conducted to get the optimum number of hidden layers and neurons [21].

VII. Result Analysis

Our database contains 600 recorded Persian spoken digits by a mono-speaker and the sampling frequency was 22050. Each word has been repeated six times by ten different male speakers. Half of the data randomly selected for training purpose and half rest were used for testing. The database was developed under normal conditions using an ordinary quality microphone directly hooked to a pc.

The FFT with 22050 point was used to find the power spectrum of each frame. the filter bank processing intervals, with calculated mels scale coefficients has been shown in appendix 2. After perform IFFT with 22050 point, DWT which based on wavelet Daubechies4 with the same point was implemented. Output was an Array with the length of 22050. This Array was divided to 30 sub array with nonlinear intervals which five of sub arrays were connived for the reason that will be described.

In this research, a new approach has been applied to build MLP neural network. LabVIEW, powerful graphical programming software developed by National Instruments, which has, so far been successfully used for data acquisition and control, has been used here for building neural network.

Neural networks are parallel processors. They have data flowing in parallel lines simultaneously. LabVIEW has the unique ability to develop data flow diagrams that are highly parallel in structure. So LabVIEW seems to be a very effective approach for building neural network.

MLP neural network was successfully developed using LabVIEW and MATLAB and we compared two process time. Process time of our network in LabVIEW was significantly less than the process time in MATLAB, so LabVIEW has been selected for our network implementation.

The input feature vector had 30 DWT energy values per frame. The first layer followed by two hidden layers of 30 elements. The network had 10 outputs as classes, which each output uniquely representing one category.

We used adoptive learning rate method to enhance the training speed and so the learning rate has been decreased during learning process.

After learning process with implementation of UTA algorithm on outcome features, acquired features for intervals 750-900, 1500-1800, 3000-3300, 5100-5700, 11800-13300 were removed because of their poor consistency detected by used feature selection algorithm so the dimension of the input vector has been reduced to 25.

System performance was evaluated in online speech recognition. In this mode system should decide on a processed speech signal vector. Result shows excellent system's performance in real time mode as it can be seen in appendix 3.

VIII. Conclusion

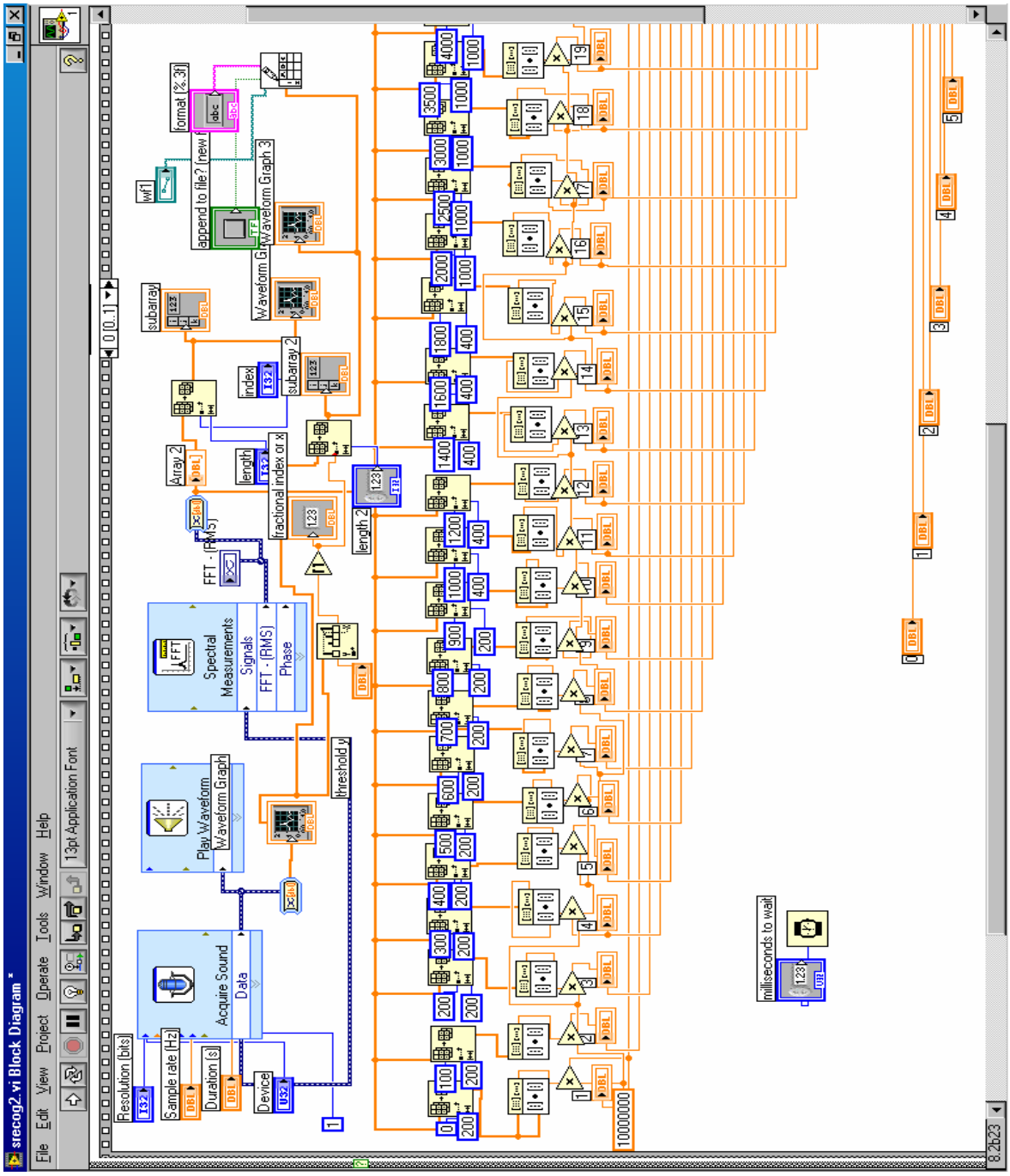
Currently development of speech recognition is widely used in industrial software market. In this paper, we presented a new method that developed an automatic Persian speech recognition system performance. Using UTA algorithm redounded to increase system learning time from 18000 to 6500 epoch and system accuracy average value to 98%. Considerable specifications of this system are excellent performance with minimum training samples, fast learning and wide range of recognition and online classification of the receiving signals.

Our goal is to ultimately design a network which would be able to do continues speech recognition on a larger vocabulary.

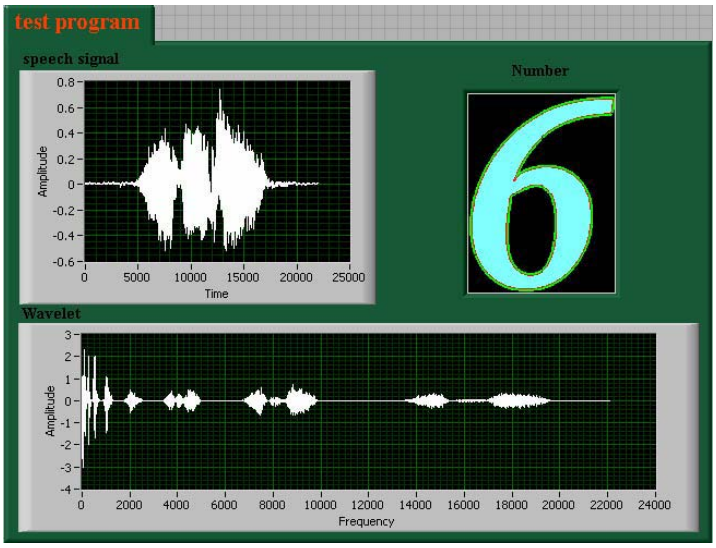
Appendix

CLASS	0	1	2	3	4	5	6	7	8	9
0	0.97	0.48	0.44	0.37	0.55	0.72	0.27	0.73	0.58	0.79
1		0.99	0.74	0.63	0.73	-0.01	0.67	0.65	0.67	0.64
2			0.98	0.56	0.71	-0.02	0.73	0.65	0.66	0.69
3				0.95	0.63	-0.10	0.65	0.37	0.40	0.42
4					0.91	0.09	0.76	0.7	0.64	0.69
5						0.93	-0.14	0.47	0.35	0.62
6							0.98	0.52	0.71	0.56
7								0.97	0.76	0.78
8									0.99	0.72
9										0.95

Appendix 1 Correlation of the data that was used to train MLP



Appendix 2 A view of preprocessing program



Appendix 3 Recognition of system for number 6

Digit	0	1	2	3	4	5	6	7	8	9
Sensevity	96.6%	90%	96.6%	100%	90%	93.3%	93.3%	90%	100%	93.3%
specifity	99.6%	100%	99.2%	99.6%	99.2%	100%	100%	99.6%	97%	99.2%
accuracy	99.3%	99%	99%	99.6%	98.3%	99.3%	99.3%	98.6%	97.3%	98.6%

Appendix 4 System performance

ACKNOWLEDGMENT

The author is grateful to Dr. K .Kangarlo for his constant and patient support and useful suggestions. The invaluable inspiration and contributions from our friends are gratefully acknowledged.

REFERENCES

[1] Abdul Ahad, Ahsan Fayyaz, Tariq Mehmood. "Speech Recognition using Multilayer Perceptron" . IEEE trans. pp.103,2002.
[2] Karina Vieira, Bogdan Wilamowski, and Robert Kubichek " Speaker Verification for Security Systems Using Artificial Neural Networks". IEEE trans. pp.1102-1105,2003.
[3] Song Yang, Meng Joo Er, and Yang Gao. "A High Performance Neural-Networks-Based Speech Recognition System". IEEE trans. pp.1527,2001.
[4] Keogh, E. & M. Pazzani. "Derivative Dynamic Time Warping". In Proc. of the First Intl. SIAM Intl. Conf. on Data Mining, Chicago, Illinois, 2001.
[5] Abdulla, W., D. Chow, and G. Sin, "Cross-words reference template for DTW-based speech recognition systems", in Proc. IEEE TENCON, Bangalore, India, 2003.
[6] Corneliu Octavian DUMITRU, Inge GAVAT. "Vowel, Digit and Continuous Speech Recognition Based on Statistical, Neural and Hybrid Modelling by Using ASRS_RL ". EUROCON 2007, The International Conference on "Computer as Tool", pp.858-859.
[7] i.Gavat, O.Dumitru, C. Iancu, Gostache, "Learning strategies in speech Recognition", Proc. Elmar 2005, pp.237-240, june 2005,Zadar, Croatia.
[8] Bahlmann. Haasdonk. Burkhardt. "speech and audio recognition" . IEEE trans. Vol 11. May 2003.
[9] Edward Gatt, Joseph Micallef, Paul Micsllef, Edward Chilton. "Phoneme Classification in Hardware Implemented Neural Networks ". IEEE trans, pp.481, 2001.
[10] Redondo, M.F. Espinosa, C.H. "A comparison among feature selection methods based on trainednetworks." IEEE trans.Aug1999
[11] Kirschning. 1. "Continuous Speech Recognition Using the Time-Sliced Paradigm", MEng.Dissertation, University Of Tokushinia, 1998.
[12] Tebelskis. J. "Speech Recognition Using Neural Networks", PhD. Dissertation, School Of ComputerScience, Carnegie Mellon University, 1995.
[13] J. Tchorz, B. Kollmeier; "A Psychoacoustical Model of the Auditory Periphery as Front-end forASR"; ASAEAAiDEGA Joint Meeting on Acoustics; Berlin, March 1999.
[14] Cory L. Clark "LabVIEW Digital Signal Processing and Digital Communications". McGraw-Hill Companies.2005
[15] " Digital Signal Processing System-Level Design Using LabVIEW " by Nasser Kehtarnavaz and Namjin Kim University of Texas at Dallas. 2005.
[16] M. Kantardzic. Data Mining Concepts, Models, Methods, and Algorithms. IEEE, Piscataway, NJ, USA, 2003.
[17] R.P. Lippmann, "An Introduction to computing with neural nets." IEEE ASSP Mag. , vol 4, Apr.1997
[18] H. B. D. Martin T. Hagan and M. Beale. Neural Network Design. PWS Publishing Company, Boston, MA, USA, 1996.
[19] T. G. Dietterich. Machine learning for sequential data: A review. In Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, pp.15–30, 2002. Springer-Verlag, London, UK.
[20] MathWorks. Neural Network Toolbox User's Guide, 2004.
[21] S.M Peeling, R.K Moore and R.J.Tomlinson, "TheMulti Layer Perceptron as a tool for speech pattern processing research." in Proc. IoA Autumn Conf.Speech Hearing. 1986.