SPEAKER VERIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENT AND ARTIFICIAL NEURAL NETWORK

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Technology
In
Electronics and Instrumentation Engineering

By

Sujit Kumar Behera (108EI012)
Jatindra Kumar Singh (108EI018)

Under the guidance of

Prof. Samit Ari

Department of Electronics and Communication Engineering
National Institute of Technology
Rourkela- 769008
2012

NATIONAL INSTITUTE OF TECHNOLOGY
ROURKELA

CERTIFICATE

This is to certify that the Thesis Report entitled **"S*peaker verification using Mel Frequency Cepstral Coefficient and Artificial Neural Network"*** submitted by **Mr. Sujit Kumar Behera (108EI012)** and **Mr. Jatindra Kumar Singh (108EI018)** in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Electronics andInstrumentation Engineering during session 2008-2012 at National Institute Of Technology, Rourkela (DeemedUniversity) and is an authentic work by him under my supervision and guidance. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other university/institute for the award of any Degree or Diploma.

Dr. Samit Ari
Assistant Professor
Dept. of Electronics & Comm. Engg
National Institute of Technology
Rourkela-769008

Date: 14-05-2012

ACKNOWLEDGEMENT

First of all, we would like to express our deep sense of respect and gratitude towards our advisor and guide **Prof Samit Ari**, who has been the guiding force behind this work. We are greatly indebted to him for his constant encouragement, invaluable advice and for propelling us further in every aspect of our academic life. His presence and optimism have provided an invaluable influence on our career and outlook for the future. We consider it our good fortune to have got an opportunity to work with such a wonderful person.

Next, we want to express our respects to **Prof. L P Roy and Arunava Karmakar (M Tech)** for teaching and also helping how to learn. He has been great sources of inspiration to us and we thank him from the bottom of our heart.

We would like to thank all faculty members and staff of the Department of Electronics and Communication Engineering, N.I.T. Rourkela for their generous help in various ways for the completion of this thesis.

We would like to thank all our friends, classmates and especially **Abhijit Tripathy**, **Arghyapriya Choudhury, Debesh Kuanr, Sunil Barla** for their help and contribution throughout the time. We have enjoyed their companionship so much during our stay at NIT, Rourkela.

Sujit Kumar Behera **Jatindra Kumar Singh**

(108EI012) (108EI018)

ABSTRACT

Speaker recognition is defined as to make sure that if the person is the same person he claims to be or not. This technique is one of the biometric recognition techniques useful in all most all areas where security is a concern.

Speaker recognition can be divided into speaker identification and speaker verification. Speaker identification decides if a speaker is a specific person or is from a group. In speaker verification, a person makes an identity claim (e.g., by entering a pin number with the debit/credit card at ATM).

There are two main stages in this technique, feature extraction and feature matching. Feature extraction is the process in which we extract some useful data which can later to be used to represent the speaker. Feature matching involves identification of the unknown speaker by comparing the feature extracted from the voice with the enrolled voices of known speakers.

In this project we have extracted the MFCCs of the speech signal, which involve recording of the speech signal, windowing, framing, thresholding, STDFT (short time discrete fourier transform) calculation and then passing through mel frequency filter. Extracted features are then matched with the stored templates. Algorithms used in feature extraction are calculation of real cepstral coefficient calculation and mel frequency cepstral coefficient calculation. For feature matching we used multi-layer perceptron method in artificial neural network.

CONTENT

LIST OF FIGURES and TABLES

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

Speaker recognition maybe defined as the process of recognizing a person automatically using the information extracted from speech signal of the person. This technique uses the voice of the speaker to verify their identity to access to several services such as accessing the computer or server from remote place, voice dialing, accessing security services, mobile banking etc. where security is the primary concern.

In this project we have tried to make a simple automatic text dependent speaker recognition system. This speaker recognition system can help us to add an extra security level. For example we can install a speaker recognition system in domestic security like home, office, locker etc. so that we can unlock the door with either the voice signal or key. Even for more secure system we can take both key and voice verification compulsory.

1.2. MOTIVATION

For security application to crime investigations, speaker recognition is one of the best biometric recognition technologies. We can give our speech signal as password to the lock system of our home, locker, computer etc. Speaker recognition can also be helpful in verifying voice of criminal from the audio tape of telephonic conversations. The main advantage of biometric password is that there is nothing like forgetting, misplacing as knowledge-based password.

1.3. FLOW CHART

Here is a flow chart speaker verification showing all major steps involved in this project.

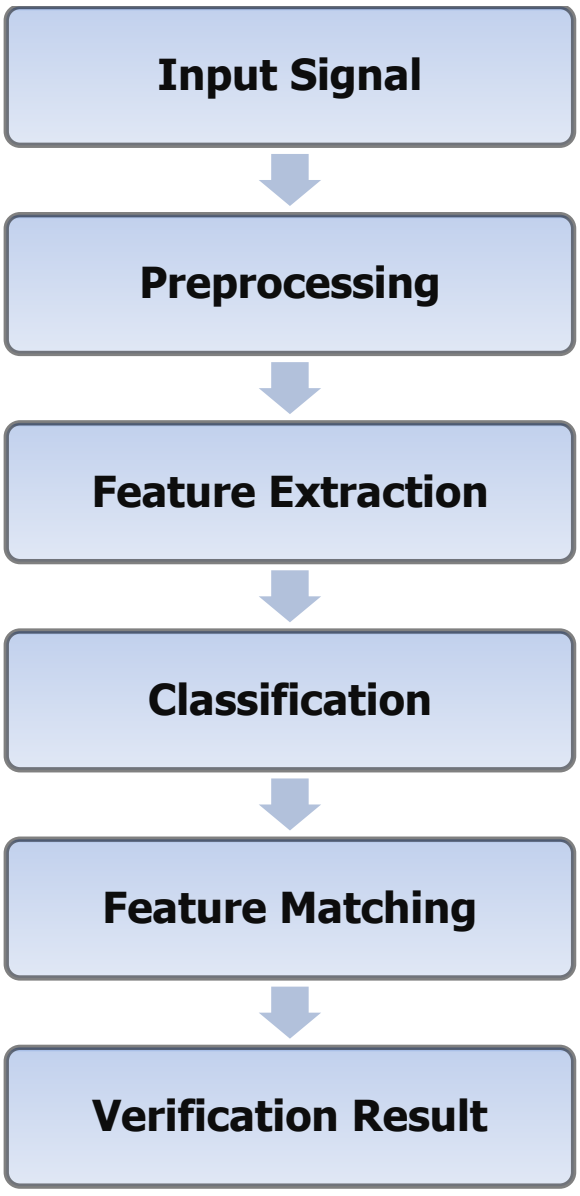

Fig 1.1 Flow chart of Speaker Verification System

1.3 Literature Review

Many universities, laboratories and industries have researched and designed several generation of speaker recognition system. In 1974, AT & T (American telephone and telegraph) have designed a text dependent system, in which cepstrum features are taken. In that system only 2% of verification and recognition error have been detected. STI in 1979 have designed a text independent system by taking LP (linear prediction) features. Long term pattern matching method is used in this system. After this in 1981 again AT and T have developed a text dependent system by taking normalized cepstrum features. Then in 1982 BBN (Bolt, Beranek and Newman) designed a text independent system. It used LAR (log area ratio) features and nonparametric pdf pattern matching technique. After then many other organization like Massachusetts Institute of Technology Lincoln Labs, National TsingHua University (Taiwan), Nagoya University (Japan), Nippon Telegraph and Telephone (Japan), Rensselaer Polytechnic Institute, Rutgers University, and Texas Instruments (TI) have developed much accurate speaker recognition system using different features.

Artificial neural network has successfully used for matching. Norton and Zahorian[11] have developed an ANN based speaker verification system. Zaki[10] have used a cascade neural network for speaker recognition. Radial basis function was used by Mark and Kung[12]. But there is a problem arises that reliability on ANN. To improve reliability Reddy and Das have developed Committee Neural Network (CNN) [9].

1.4 PRINCIPLES OF SPEAKER VERIFICATION

There are two major application of speaker recognition.

- **Verification**

  If the speaker claims to be the certain identity and the voice is used to verify this claim, the process is called Speaker Verification.

- Identification

  It is the task of determining an unknown persons' identity.

Speaker recognition system can be divided into two categories.

- Text dependent

  If the text must be the same for enrollment and verification, the system and process is said to be text dependent.

- Text independent

  In this procedure text-independent the technology does not compare what was said at enrollment and verification.

CHAPTER 2

FEATURE EXTRACTION

2.1. PREPROCESSING

Befor feature extraction we have to do a litle preprocessing with the speech signal. This include analog to digital conversion, resampling, windowing, thresholding according to our requirements.

2.1.1. ANALOG TO DIGITAL CONVERSION

We want digital signal to process (as we are working in matlab which deals with matrices) we have to convert theanalog signal to digital signal. As we have recorded the speech signal in matlab so we were not needed A/D conversion.

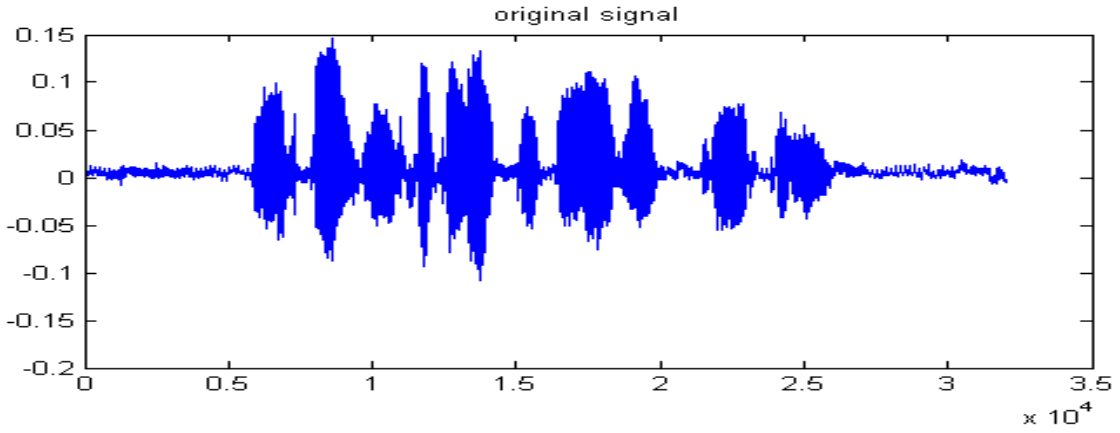

Fig 4.1 Original Speech Signal

In the above figure data aquired is shown of sampling frequency of 11025 samples per second.

2.1.2. RESAMPLING

Resampling is done according to the requirement. For example, to listen to the recorded speech we need to resample the recorded signal to 8000 samples per second.

2.1.3. WINDOWING

For windowing we used hamming window which acts like a filter which optimizes to minimize the nearest side lobe.

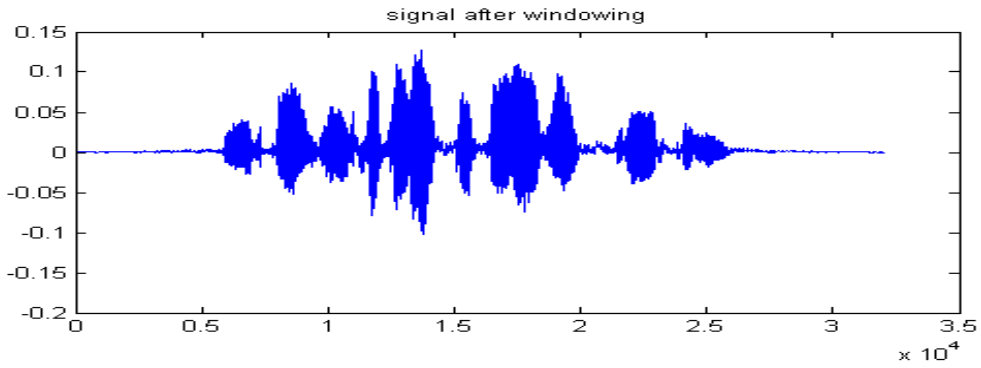

Fig 2.2 Signal after Windowing

2.1.4. THRESHOLDING

When we are interested in the utterances where we will get the data related to the voice characteristics of person and remaining data aquired are not needed or undesired then we can eliminate the undesired data by thresholding. In this process we need to set a value which will decide whether to keep or discard the data aquired.

Thresholding are two types:

- Hard thresholding

Hard thresholding is a normal process of setting to zero, the elements having having absolute value less than threshold value.

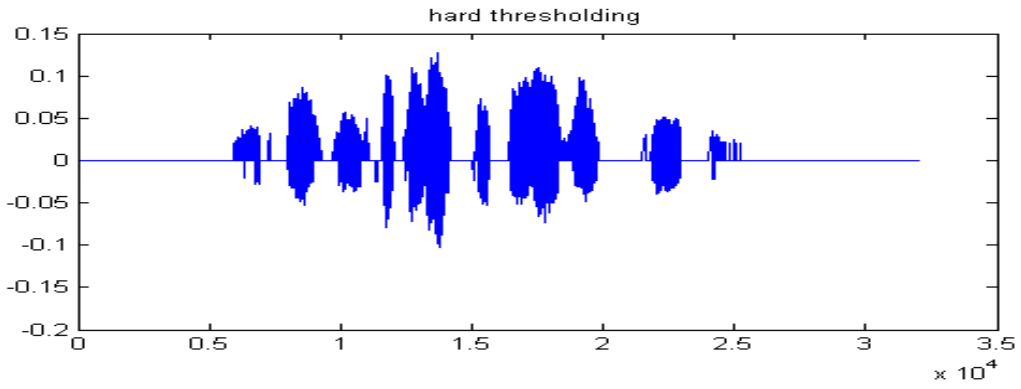

Fig 2.3 Signal after Hard Thresholding

- Soft thresholding

Soft thresholding is also the same process with a bit modification. It starts with setting to zero, the elements having having absolute value less than threshold value. And then setting to zero the elements whose absolute values are lower than the threshold, and then shrinking the nonzero coefficients toward 0.

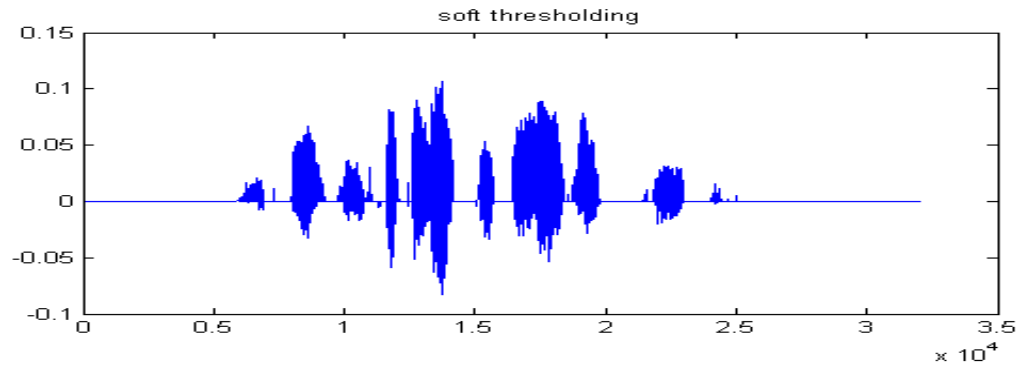

Fig 2.4 Signal after Soft Thesholding

In the figure 2.4 we have shown original signal, signal after windowing using hamming window and thresholding.

2.2. NORMALIZATION

Next we have done normalization. The main advantage of normalization is that it restricts the amplitude in the range from -1 to +1. This can be found by dividing the current value with the absolute value in the signal.

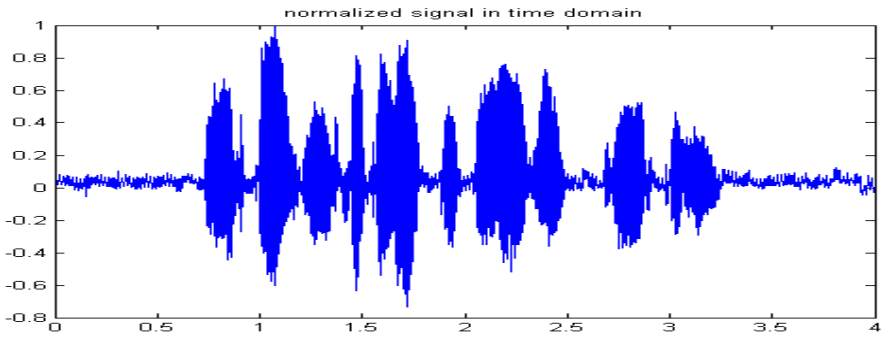

Fig 2.5 Signal after Normalization of original signal

In the figure above normalization of a speech signal can be seen.

2.3. SHORT TIME DICRETE FOURIER TRANSFORM (STDFT)

After normalization the next stage is STDFT, short time discrete cosine transform. It similar to DFT but the mail thing here to notice is the window within which we did the conversion. The advantage of STDFT is that we do not lose the property of time domain and the noise we got while recording the data gets localized within the window and does not get spread to the whole frequency spectrum.

2.4. CLACULATION OF CEPSTRAL COEFFICIENTS

After conversion to frequency domain the next stage is to filter the frequency spectrum through filters to get required cepstrum. We have used two types of filters, one is linear filter and another is mel filter.

2.4.1. REAL CEPSTRUM

The name "cepstrum" was derived by reversing the first four letters of "spectrum".Real cepstrum can be defined as inverse fourier transform of the logarithmic value of frequency spectrum of the speech signal.

$$Y_{ceptrum} = real(ifft(log(abs(fft(Y_{input})))))$$

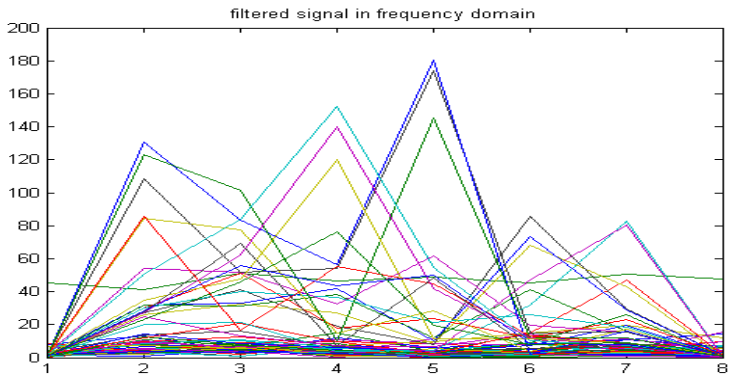

Fig 2.6 Absolute value of Real cepstrum

In the figure above absolute value of cepstral coefficients is plotted, and its logarithmic values are plotted in the figure below.

2.4.2. MEL CEPSTRUM

As we know cepstrum have a drawback that it does not matches with the frequency of human voice. To overcome this problem we used mel filter to calculate mel frequency cepstral coefficients.

MFCCs are commonly derived as follows:

1. Take the fourier transform of the signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take log of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.

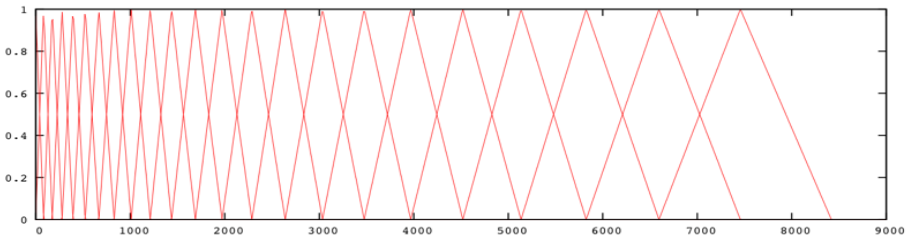

Fig 2.7 Mel filter bank

The relation between mel scale and linear scale can be from the equation

$$M = 2595\log10(1 + f/700)$$

5. The MFCCs are the amplitudes of the resulting spectrum.In the figure below values of the MFCCs are ploted.

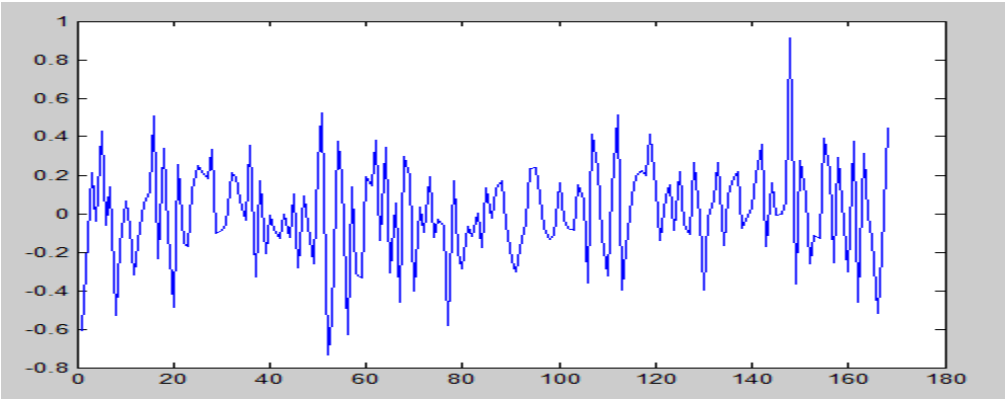

Fig 2.8 168 MFCCs from a single sample

CHAPTER 3

FEATURE MATCHING

3.1. FEATURE MATCHING

Feature matching involves assigning speech signals of each speaker a different class based on its feature. Features are taken from known samples and then unknown samples are compared with those known samples. Different techniques such as Neural Networks, Minimum distance classifier, Bayesian classifier, Quadratic classifier, Correlation are used for this purpose. In this project, we have opted for Artificial Neural Networks.

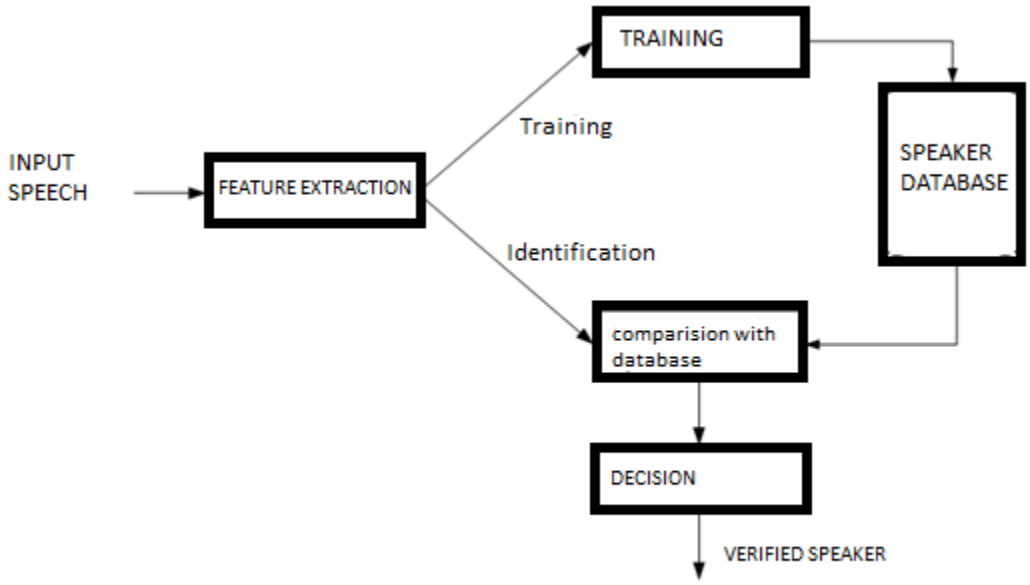

Fig 3.1 Block diagram feature matching

3.2. ARTIFICIAL NEURAL NETWOK

Neural network is used when we have large number of samples of each speaker with variations among them which are used to train the network and correspondingly weights are updated. Finally, the weights are applied to the testing samples to get the correct output. The main advantage of using Neural networks is that it is unaffected by the differing shape and style of testing samples as the network is already trained with large variations.

Back propagation algorithm is used to update the weights and bias matrix. Here, the learning parameter/step size ‘η’ has a major role as it controls the rate at which the error is reduced which further determines the time complexity.

An artificial neural network can be seen as a computer program that is designed to recognize patterns and learn "like" the human brains. The structure of a neural network is shown below.

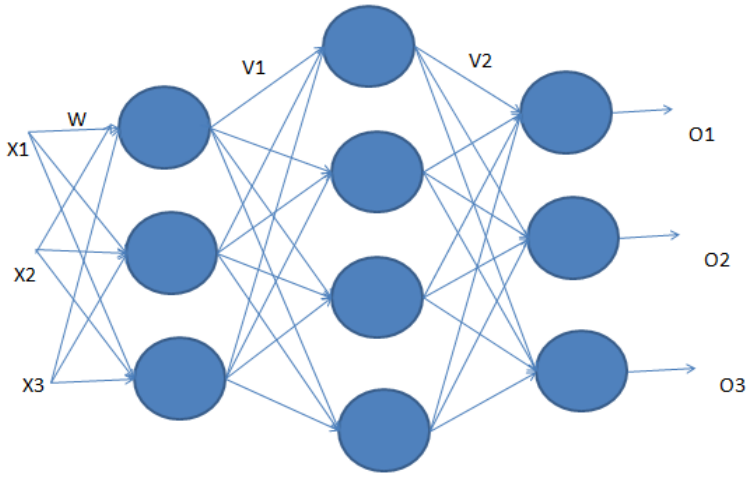

Fig 3.2 Structure of ANN

An ANN is composed of a large number of highly interconnected processing elements (artificial neurons) working in unison to solve a specific problem. An artificial neuron has (i) inputs X_1, $X_2 \ldots X_n$; (ii) a summing element (iii) a nonlinear element; (iv) connection weighing element, W_1, $W_2 \ldots W_n$ that are adjustable connection weights and (v)output Y. The factor$W_0.X_0$ = W_0 is the bias b, $X_0 = 1$ always.

$$net = \left(\sum_{i=1}^{n} W_i X_i\right) + b = \sum_{i=1}^{n} W_i X_i$$

$$Y = f(net)$$

Function of a neuron

Here, 'logsigmoid' function has been used as the activation function. The number of input layers is equals to the number of features in each the feature vector of each input character. The number of hidden layers has been taken as 10 and the output layers are equal to the no of class, here taken as 5 for 5 speakers.

3.3. BACK PROPAGATION ALGORITH

This algorithm is used to update the weights after one output is obtained. The output is compared with the target and error signal is generated. Then the weights are updated using the following formulas till the error becomes less than the goal error. In our case, the no of iteration is taken as 10000 and goal error is chosen as 10^{-5}.

Algorithm:

Consider the following diagram.

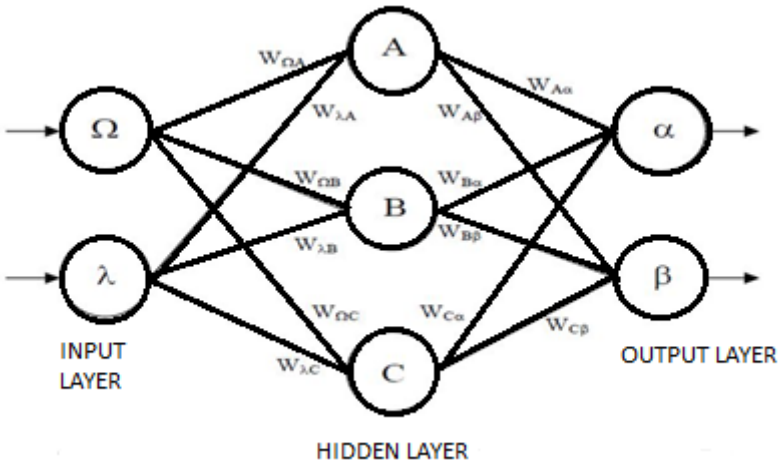

Fig 3.3 Back propagation Algorithm

1. Calculation of errors in output neurons

$$\delta_\alpha = out_\alpha (1 - out_\alpha) (Target_\alpha - out_\alpha)$$

$$\delta_\beta = out_\beta (1 - out_\beta) (Target_\beta - out_\beta)$$

2. Change in output layer weights

$$W^+_{A\alpha} = W_{A\alpha} + \eta\delta_\alpha out_A \qquad W^+_{A\beta} = W_{A\beta} + \eta\delta_\beta out_A$$

$$W^+_{B\alpha} = W_{B\alpha} + \eta\delta_\alpha out_B \qquad W^+_{B\beta} = W_{B\beta} + \eta\delta_\beta out_B$$

$$W^+_{C\alpha} = W_{C\alpha} + \eta\delta_\alpha out_C \qquad W^+_{C\beta} = W_{C\beta} + \eta\delta_\beta out_C$$

3. Calculation of (back-propagate) hidden layer errors

$$\delta_A = out_A (1 - out_A) (\delta_\alpha W_{A\alpha} + \delta_\beta W_{A\beta})$$

$$\delta_B = out_B\,(1 - out_B)\,(\delta_\alpha W_{B\alpha} + \delta_\beta W_{B\beta})$$

$$\delta_C = out_C\,(1 - out_C)\,(\delta_\alpha W_{C\alpha} + \delta_\beta W_{C\beta})$$

4. Change in hidden layer weights

$$W^+_{\lambda A} = W_{\lambda A} + \eta\delta_A in_\lambda \qquad W^+_{\Omega A} = W^+_{\Omega A} + \eta\delta_A in_\Omega$$

$$W^+_{\lambda B} = W_{\lambda B} + \eta\delta_B in_\lambda \qquad W^+_{\Omega B} = W^+_{\Omega B} + \eta\delta_B in_\Omega$$

$$W^+_{\lambda C} = W_{\lambda C} + \eta\delta_C in_\lambda \qquad W^+_{\Omega C} = W^+_{\Omega C} + \eta\delta_C in_\Omega$$

The constant η (called the learning rate, and nominally equal to one) is put in to speed up or slow down the learning if required.

CHAPTER 4

RESULTS & DISCUSSIONS

4.1. RESULTS

Fig. 4.1 shows a graphical representaion of speech signal after windowing.

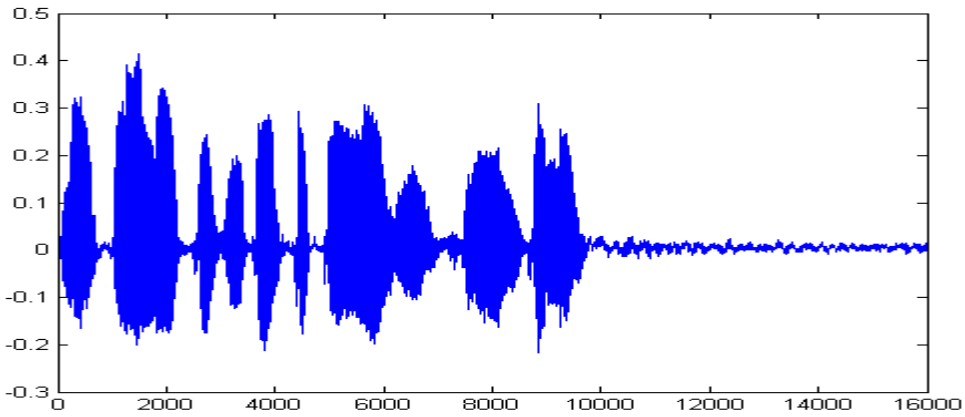

Fig 4.1 Speech acquired

In the Fig. 4.2 illustrates the speech signal after thresholding. We can see in the plot how thresholding sets all value to zero which are lower than thresholding value.

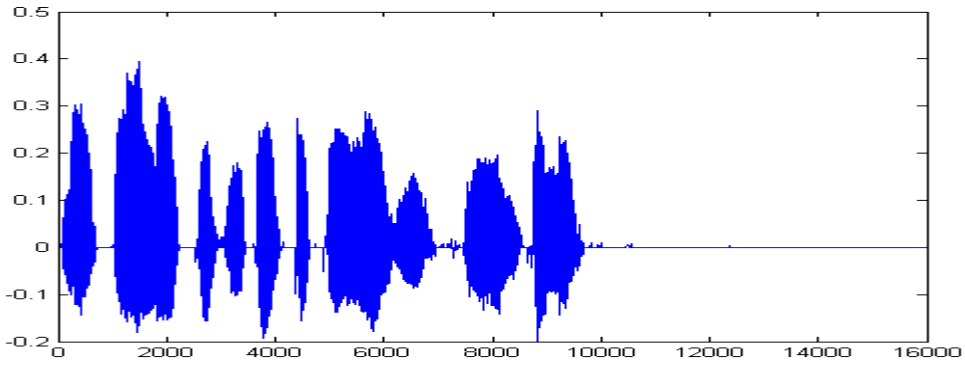

Fig 4.2 Thresholding

In the figure 4.3 graphical representation of truncated speech signal has done. The signal is truncated by taking the nonzero values from the signal after thresholding. The main advantage of truncating the signal is that, it minimizes the size of the signal to a great extent. In fact we do eliminate the portion of signal where there is no utterance but full of environmental noise.

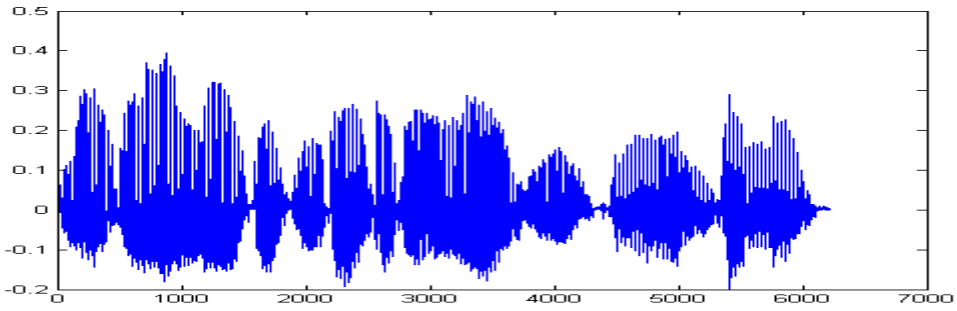

Fig 4.3 Truncating of data

In the figure 4.4 below plot of MFCCs of a speech sample is given. These coefficients are the element extracted from a speech signal which is used for enrollment of speakers.

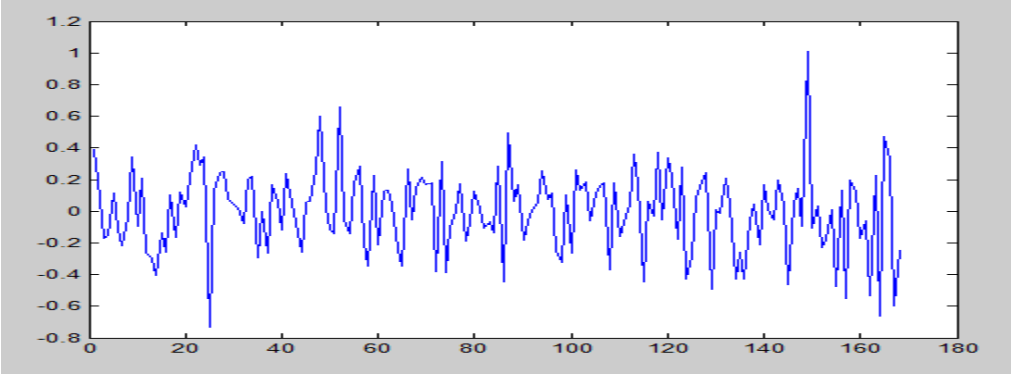

Fig 4.4 MFCCs

In the figure 4.5 first 24 elements of MFCCs of a signal are plotted.

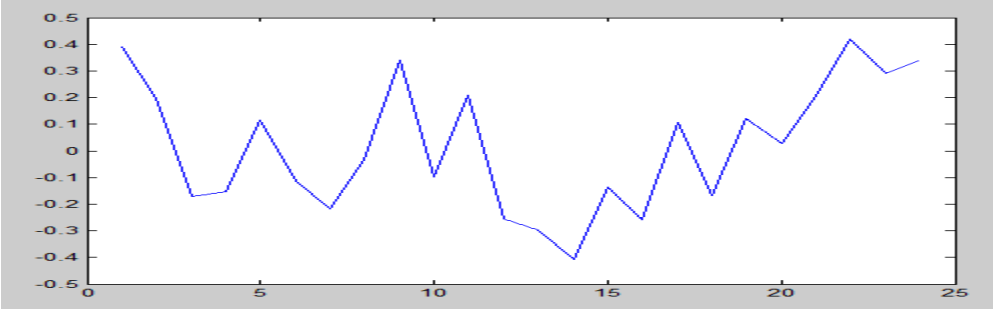

Fig 4.5 First 24 elements of Mel frequency cepstrum

In the feature matching we need to minimize the mean square error (**mse**) for better enrollment of the speech signals. In the figure below relation between **mse**and iteration has shown.

Table below illustrates the relation between ‘eta’ and number of iterations. We always have to keep the value of ‘eta’ such that the verification time will be less. Verification time is dependent on the number of iteration taking place which again depend on ‘eta’ value at the time of enrollment of the speech samples.

Value of Eta	Number of iterations	Value of eta	Number of iterations
0.005	1000	0.050	1175
0.006	1000	0.055	971
0.007	8311	0.060	939
0.008	8005	0.065	950
0.009	6533	0.070	865
0.010	5639	0.075	736
0.015	3833	0.080	727
0.020	2849	0.085	678
0.025	2263	0.090	648
0.030	1872	0.095	690
0.035	1810	0.095	690
0.040	1442	0.100	582
0.045	1304		

Table 4.1 Change in number of iterations with ‘eta’

Result of enrolment and testing is shown in the figure 4.7 below. It is a MATLAB command window showing number of iterations to minimize **mse**and the matrix showing the result. We have taken a [5 1] matrix to represent the serial number of the speaker with which the testing signal got matched. In this case the testing signal matched with 4th speaker's speech.

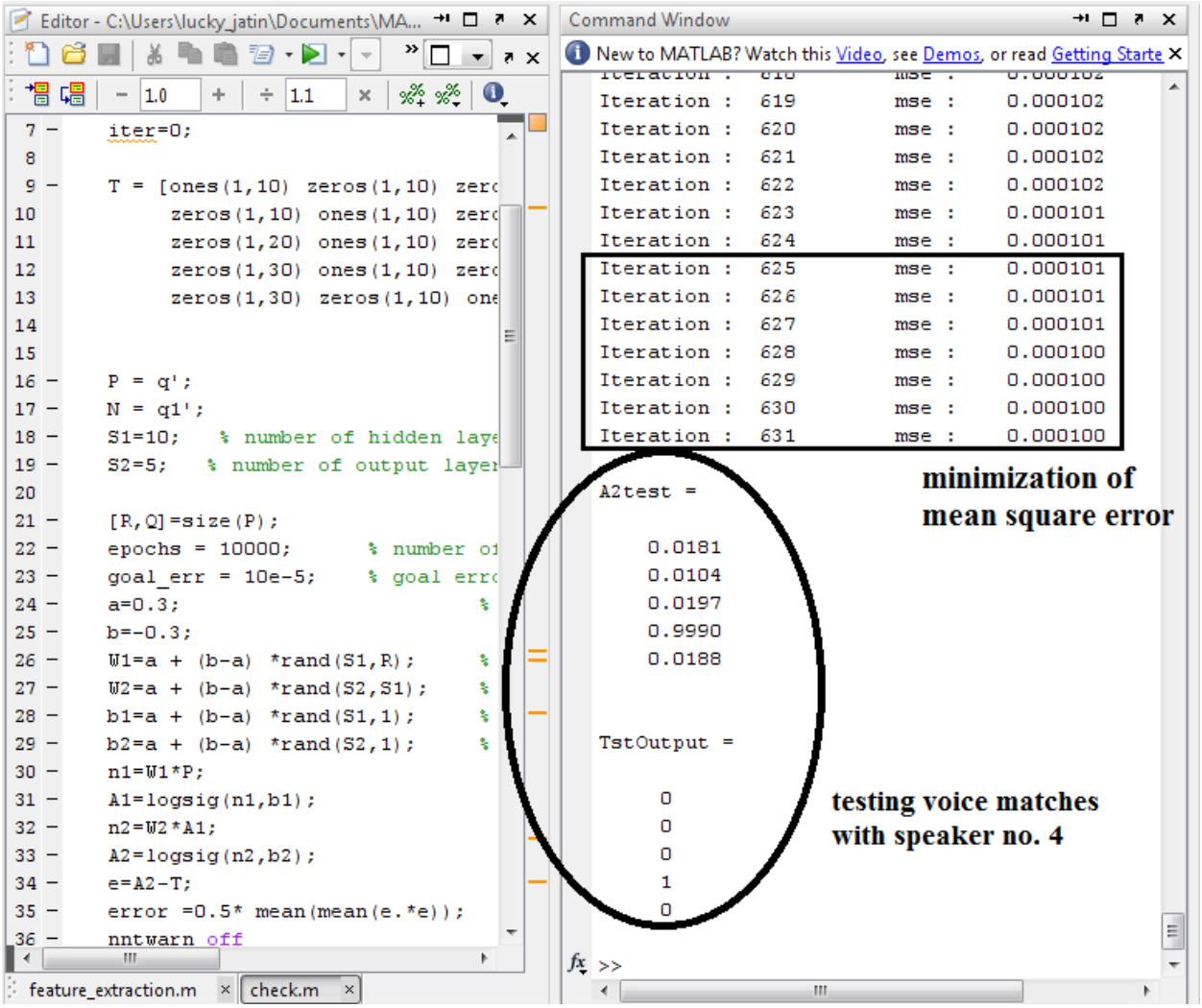

Fig 4.7 Matlab window showing result of speaker verification

We found out Receiver Output Characteristic (ROC) curve which is shown in figure 4.8 below. From the figure it is clear that the ROC curve we got is the ideal one. For a matched signal we should be getting the ROC curve in the upper triangular area where as for an unmatched signal the curve should be in the lower triangular area.

We got true positive ratio, false positive ratio and false acceptance ratio. Apart from this we got equal error rate (eer) which can be mathematically written as follows:

eer = (false acceptance ratio / false positive ratio) X 100

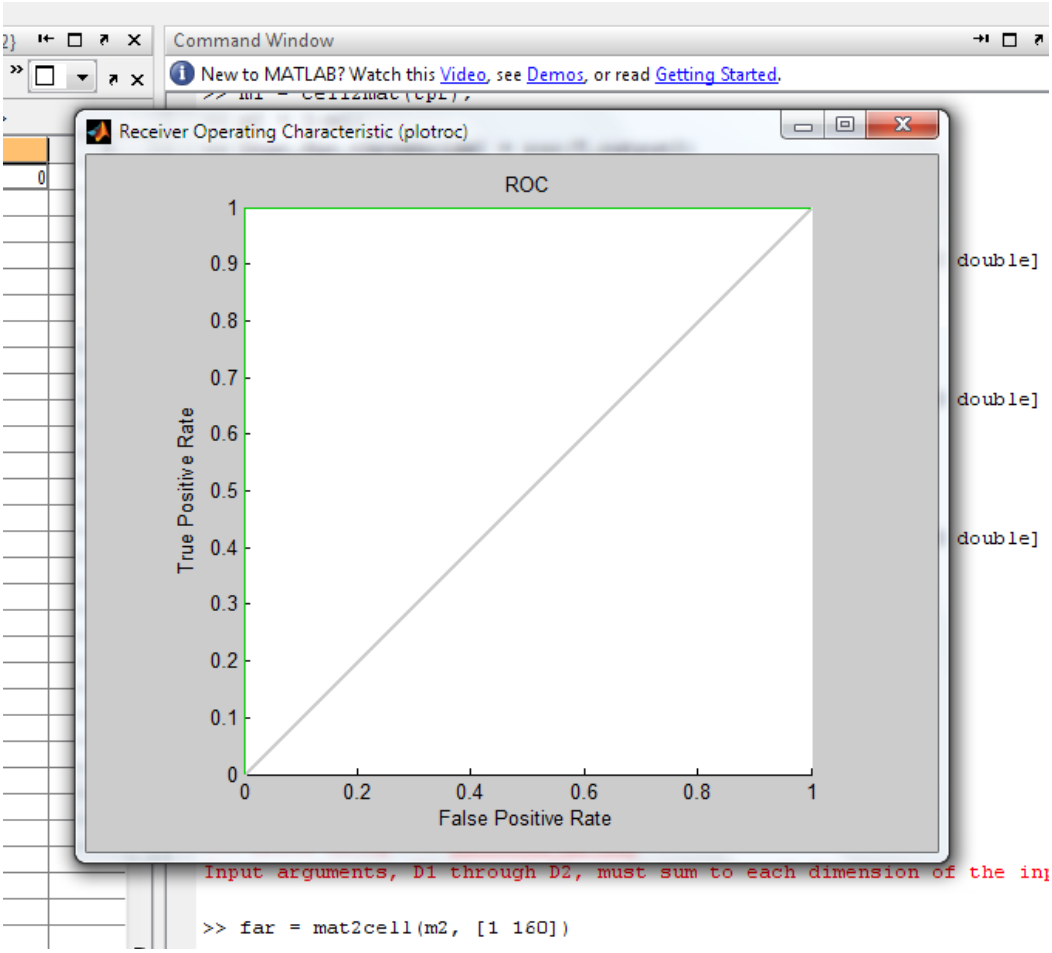

Fig 4.8 ROC curve

Table 4.2 below shows the result of speaker verification. We trained the ANN with speech signals which says “National Institute of Technology, Rourkela” and at the time of verification speech signal saying same line is done and the following results are obtained.

Name of the speaking person	No. of signal tested	No. of signal got matched
Arghya	**3**	**3**
Debesh	**3**	**2**
Jatindra	**3**	**3**
Sujit	**3**	**2**
Sunil	**3**	**2**
Others	**5**	**0**

Table 4.2 Result verification of 20 speech signals

We have taken 15 speech signals, 3 signals each from the persons whose voices are enrolled and 5 speech signals from others. We got 80% of matching among the enrolled persons and none of the speech signal matched from the persons who were not enrolled.

CHAPTER 5

CONCLUSION

The results obtained in this project using MFCC and Artificial Neural Network. We have computed MFCCs of all speech signals. We have used MFCCs because these coefficients follow the human ear's response. We have taken "**National Institute of Technology, Rourkela"** as input speech signal.

The performance analysis of neural network method says that, neural networks perform better for varying speech signals. As, we are dealing with speaker verification, hence the neural network should be trained by taking enough no of samples so that it remains unaffected by the deviations from standard. More the no of samples, more the compression is achieved. But while testing, it is possible that we are left with very small no of samples to test, which may not yield good result. In case of multi-layer networks, the Learning Coefficient η determines the size of the weight changes. A small value of η results in a very slow learning process. The large weight changes may cause the desired minimum to be missed if the learning coefficient is too large. Depending on the problem statement, 'eta' should lie between 0.005 to 0.1. The multilayer feed forward networks trained with the Back propagation method are probably the most practically used networks for real world applications.

In the simulation we got an ideal ROC curve. It is so because of the low number of speech samples we have collected. We have taken 10 speech samples of each person for the experiment. To get a better ROC curve we need speech samples more than 100.

REFERENCES

[1] Campbell, J.P., Jr.; **"Speaker recognition: a tutorial"** Proceedings of the IEEE Volume85, Issue 9, Sept. 1997 Page(s):1437 – 1462.

[2] Reynolds, Douglas A., Rose, Richard C.; **"Robust Text-Independent Identification Using Gaussian Mixture Speaker Models",** IEEE Transection on Speech and Audio Processing, Volume 3, Number 1, January 1995, Page(s): 72-83

[3] Childers, D.G.; Skinner, D.P.; Kemerait, R.C.; **"The cepstrum: A guide to processing"**Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.

[4] Seddik, H.; Rahmouni, A.; Sayadi, M.; **"Text independent speaker recognition usingthe Mel frequency cepstral coefficients and a neural network classifier"** FirstInternational Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s):631 – 634.

[5] S. Furui, **"Speaker independent isolated word recognition using dynamic features ofspeech spectrum"**, IEEE Transactions on Acoustic, Speech, Signal Processing, Vol.34, issue 1, Feb 1986, pp. 52-59.

[6] John G. Proakis, Dimitris G. Manolakis, "Digital Signal Processing".

Stephen P. Banks, "Signal Processing, Image Processing and Pattern Recognition".

[7] Soft Computing Course Lecture, notes, slides, R C Chakraborty. www.myreaders.info/html/soft_computing.html.

[8] Speech processing tool box of MATLAB R2009a, www.mathworks.com

[9] Reddy, N. P.,Buch Ojas; "**Speaker verification using Committee Neural Network."** Computer methods and programs in biomedicine, Volume 72, Issue II, Oct 2003, Page(s): 109-115

[10] M. Zaki, A. Ghalwash, A. Elkouny;"**Speaker recognition system using a cascade neural network**", Int. J. Neural Syst., 7 (1996), pp. 203–212

[11] C.A. Norton, S.A. Zahorian; "**Speakerverification based on speaker position in a multidimensional speakeridentification space"**, Intelligent Engineering Systems Trough Artificial Neural Networks, 5ASME Press, New York (1995), pp. 739–744

[12] M.W. Mak, S.Y. Kung; "**Estimation of eliptical basis function parameters by EM algorithm with application tospeaker recognition"**, IEEE Trans. Neural Networks, 11 (2000), pp. 961–969