International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 7 Issue 3, March-2018, Impact Factor: 4.059

Optimization Based Speech Authentication System to Web Content for Disabled Users

Anand R Mehta

ABSTRACT

The objective of this research is to apply artificial neural network (ANN) to authenticate the disable users with their speech using MFCC features. To access the web content, a speech authentication system is proposed for disable users. We have used Genetic algorithm (GA) to create unique feature sets according to the objective function for the training of ANN for the acceptance of classification accuracy. This research work has dealt in the improvement of speech authentication system. To extract the feature sets from the speech signals MFCC feature extraction technique is used with GA. In this work, GA is used as a feature optimization techniques and ANN as classifier to train the speech authentication system. The evaluation and validation is being conducted on MATLAB 2016a simulator. To check the performance, parameters, like, precision, recall, f-measure, execution time and accuracy have been calculated.

Keywords: Artificial neural network, MFCC features, Genetic algorithm, and speech authentication system

I. INTRODUCTION

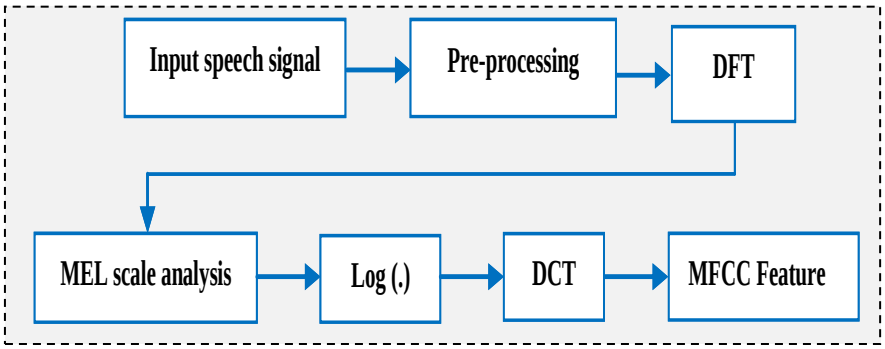
Speech authentication system has huge number of applications in the security systems, like in, healthcare, telephone military, security systems and the equipments being considered for disable users [1]. In this work, a speech authentication system is presented to the authentication of users for web content. This speech authentication work is important for disabled users due to the several reasons. By using speech authentication system, first of all, the disabled users can access the web content just as easily as the sighted users.

Speech signal is considered as an analogue signal that varies according to time and the silent part of signal may affect the system and therefore, the recognition accuracy of the system gets affected. To minimize this problem, pre-processing step is used in the proposed work to remove the silent signal by means of speech signal. Accurate algorithm for the digital processing has to be chosen for automatic speech authentication system. The speech authentication system has been categorized in speaker dependent with speaker independent system.

The speaker dependent system focuses on the development of system for the recognition of accurate individual's voiceprint whereas the speaker independent system identifies the words spoken by the speaker. In this research work, speech dependent system has been used for the classification of disabled users for accessing the web content. The primary phase of any speech authentication system is the extraction of audio signal features i.e. to classify the components and patterns of the audio signal that are superior for identification of the linguistic data and discarding another stuff that carries unwanted information like background noise, emotion etc [2].

The significant point is the understanding of speech produced by human being filtered with vocal area shape such as teeth and tongue etc [3]. The shape finds the kind of sound that produced. With the precise finding of shape, precise illustration of expression would be produced [4].

The vocal area shape marked itself in the envelope of the less time power spectrum, and the main task of MFCC (Mel Frequency Cepstral Coefficient) algorithm is to precisely symbolize this envelope. In this work we have used MFCC as a feature extraction technique that finds out the feature sets of audio signal [5]. The structure of MFCCs feature extraction technique is given below as figure 1:



MFCC is the feature widely used in speech authentication system but to achieve better accuracy, we need to improve the quality of MFCCs by using the optimization techniques. The basic steps of MFCCs implementation in the speech authentication system are given as [6];

Step 1: Firstly, the conversion of audio signals into short frames has been done with the easy operation.
Step 2: For each frame, the periodogram estimation for power spectrum has been calculated.
Step 3: After, the Mel filter bank toward the power spectra is being applied with the calculation of the sum of the energy in every filter.
Step 4: Consider the logarithm of each 'filter bank energy'
Step 5: Apply the log filter bank energies for DCT.
Step 6: Consider 12 to 13 DCT lower coefficients as features and discard the rest coefficients.

An audio signal changes constantly, so, to simplify things it is being assumed that on less time scales, the audio signal doesn't modify a lot (While the variation is not consider, than it mean it is stationary statically, therefore, the variation of samples are static on less span of time). That is why, we have framed the signal ranges of 20-40ms frames and in the proposed work, the frame size if 25ms. But when the frame is very less, then the achievement of enough samples for the reliable spectral estimate cannot takes place, on the other side, if the frame is longer than required than the variation of signals takes place throughout the frame.

The subsequent step is the calculation of the power spectrum of every frame with the finding of periodogram estimation of power spectrum. The periodogram estimate helps to identify the frequencies which are there in the frame [7].

The periodogram spectral estimate has a lot of data in which some information is required and some is not for speech authentication system. This effect turn into extra pronounced with the change of frequencies. For this reason, cluster of periodogram bins are taken with the addition to know the amount of energy exists in diverse frequency regions. This is executed by Mel filter bank in which the primary filter is very thin and provides a sign of the amount of energy exists around 0 Hertz. As the frequencies get higher, the filters get wider with the less consideration towards variations. The interest is in the occurrence of energy at every spot. The Mel scale signifies accurately how to space the filter banks and how broad to create them. The calculation of spacing has been explained below. The method of conversion from frequency to Mel scale is:

$$\mathrm{Mel}(f) = 1127 \ln(1 + f/700) \quad (1)$$

To exit from Mels back to frequency:

$$\mathrm{Mel}^{-1}(m) = 700(\exp(m/1127) - 1) \quad (2)$$

Once the filter bank energies are achieved, the logarithm of them is taken. This is also motivated by human hearing process because we don't hear loudness on a linear scale. At the final step, the log filter bank energies DCT has been calculated The DCT de-correlates the energies which mean diagonal covariance matrices could be utilized to produce the features. Except only 12 from 26 DCT coefficients are useful so we considered only 12 coefficients. This is because the higher DCT coefficients represents fast changes in the filter bank energies and it turns out that these fast modify essentially degrade the performance of speech authentication system, so, the removal of coefficients is taken place resulted in the accurate classification. There is a problem in MFCC feature extraction method. MFCC feature extraction technique uses the pattern of signal and if the silent signal is present in the speech signal then the uniqueness of feature is less and recognition

accuracy decreases. So, to resolve this difficulty, pre-processing step is involved in the proposed work with an optimization technique to make sure that the input feature of training system is unique or not.

MATERIALS AND METHODS

In this subdivision, the description of methods being utilized for the speech authentication system on the basis of MFCC feature-extraction algorithm based on GA (Genetic Algorithm) along with ANN (Artificial neural network) are defined below:

1.1 MFCC (Mel Frequency Cepstral Coefficient) Algorithm

There are several steps of MFCC feature extraction method as defined below:

i. Upload an audio signal with sample frequency at 16 kHz and after that frame the audio signal into 25 ms frames. We use 25 ms because it is standard. This means the frame length for a 16 kHz signal is 0.025X16000 = 400 samples. Frame step is typically something similar to 10ms (160 samples), that allows a few overlap to the frames. The initial 400 sample frame establishes at sample 0, the subsequent 400 sample frame begin at sample 160 and so on till the ending of the speech file is achieved. If the speech file does not separate in an even number of frames, include it with zeros so that it do [8].
The subsequent steps are useful to each single frame; for each frame, 12 MFCC coefficients are extracted as set of features. The time domain signal of proposed work is considered as Signal(n) and once it is framed, we have total $\text{Signal}_i(n)$ signals in the outline of frames, where, n ranges over 1-400 (if the frames are of 400 samples) and i vary from number of frames. When we compute the composite DFT of audio signal, we obtain $\text{Signal}_i(k)$ – where i indicate the frame numeral corresponding to the time-domain frame. $\text{Power}_i(k)$ is then the power spectrum of frame i [9].

ii. To apply the Discrete –Fourier- Transform on the each frame, we have performed the following equation:

$$\text{Signal}_i(k) = \sum_{n=1}^{m} \text{signal}_i(n) h(n) e^{-\frac{j2\pi kn}{m}} \qquad 1 \le k \le K \qquad (3)$$

Where, h(n)hamming window and K is is the extent of the DFT. The periodogram-based power-spectral - estimate for the speech frame $\text{Signal}_i(n)$ is specified by:

$$\text{Power}_i(k) = \frac{1}{m} |\text{Signal}_i(k)|^2 \qquad (4)$$

This is termed as the Periodogram-estimate-of-the power spectrum. We take complete value of the Complex Fourier Transform, and later square the outcome [10].

iii. After the step 2, we calculate the Mel-spaced-filter-bank. This is a group of 20 to 40 triangular filters that we have developed for periodogram power spectral estimate since step two. Our filter bank comes in the structure of 26 vectors of 257 lengths. Every vector is typically zeros; however is non-zero for a definite section of the spectrum. To compute the filter bank energies, we have multiplied every filter bank by the power spectrum, and after that insert the coefficients.

iv. Log on to 26 energies from step three has been excuted. This has given 26 log filter bank energies.

v. At the last, we have applied Discrete- Cosine-Transform (DCT) of the 26 log filter bank energies to give 26 cepstral coefficients. For the proposed speech authentication system, we have considered only the lower 12-13 of the 26 coefficients as a set of features and that are called as Mel-Frequency -Cepstral -Coefficients.

Algorithm of MFCC as a Feature Extraction

1. Load speech signal for feature extraction
2. Describe the parameters - Sample frequency = 16 kHz
   - Frame length = 25 ms
3. Amount of frames = $\frac{(\text{Frame Length X Sample frequency})}{1000}$
4. **If** Amount of frames ~= **Even number**
   Amount of frames = Pad (Zero)

Else

   Amount of frames = Amount of frames

End

5. Apply the Discrete-Fourier-Transform on the each frame using equation number 3

$\text{Signal}_i(k) = \sum_{n=1}^{m} \text{signal}_i(n)h(n)e^{-\frac{j2\pi kn}{m}} \qquad 1 \le k \le K$

6. Calculate the periodogram estimate of the power spectrum using equation number 4

$\text{Power}_i(k) = \frac{1}{m}|\text{Signal}_i(k)|^2$

7. Apply Mel-spaced filter bank
8. Filter bank energies = Energy Vector
9. Log Energy Vector = log (Filter bank energies)
10. Cepstral coefficients = DCT (Log Energy Vector)
11. MFCCs = Lesser 12-13 of the 26 cepstral coefficients
12. Return; MFCCs

End

1.1 GA (Genetic Algorithm) with ANN (Artificial neural network)

After the feature-extraction as of the audio signal, we need to optimize MFCC uses the feature optimization technique. In the proposed work, Genetic algorithm is used for feature optimization using the novel fitness function [11]. The fitness function could be described according to the requirement of speech authentication system. In the proposed work, fs is the current selected feature and ft is the threshold value of feature points. As per the given condition, we check the fit value which exists in new feature set.

$$f(\text{fit}) = \begin{cases} 1, & fs < ft \\ 0, & fs \ge ft \end{cases} \qquad (5)$$

Where, f(fit) optimizes the value as per fitness function and 1 represents the true value and 0 represents the false value. If condition is true, then genetic algorithm develops an enhanced solution of the feature sets. Integration of genetic algorithm with MFCCs is described in the below section with the used parameters in optimization to create a unique feature sets from the extracted feature sets. The used genetic parameters and operators are population size, crossover function, mutation function and selection function. To create a unique and organized feature sets according to the requirement, selection of individuals feature is used by selection function. The selection of individual features is done according to its fitness value and denoted asfs. The selection procedure is given is the equation 6 which is depends on the extracted feature by means of speech signal [12].

$$fs = \sum_{i=1}^{\text{Popsize}} f(i) \qquad (6)$$

Where, F(i) is the individually selected feature according to i and Popsize shows population-size of genetic algorithm. The fitness-function is defined in terms of a distance measure between selected value and threshold value of features based on the crossover function. Crossovers and mutations function are the operators and it is a general operator type to establish the relationship between selected feature value fs and threshold feature value ft. The crossover function is dependent on an individual feature (parents) and a new individual feature (children) while mutation changes the genes of one individual to produce a new one feature (mutant) as per fitness function. New optimized feature sets are transferred to ANN as an input or training set to create a trained ANN structure for classification [13]. The methodology of proposed genetic algorithm with ANN is described as an algorithm in the below section.

Algorithm of an optimization technique for the training using ANN

1. Load speech feature sets
2. Calculate the length of feature [r, c]
3. Define genetic parameters and operators to initialized the genetic algorithm
4. Population size = 50 (When number of variables less than 50 then Popsize 50 is sufficient for optimization)
5. Selection function = Manage to the function that chooses crossover parents from feature sets
6. Crossover function = Manage to the function in which genetic algorithm utilizes for optimal solution
7. Mutation function = Manage to the function that develops mutation children which are called optimized feature

8. Define fitness function using the equation 5
9. **for each components of feature according to rows**
10. **for each components of feature according to columns**

$$fs = \sum_{i=1}^{Popsize} f(i)$$

$$ft = \frac{\sum_{i=1}^{Popsize} f(i)}{\text{Length of feature}}$$

$$f(fit) = \begin{cases} 1, & fs < ft \\ 0, & fs \geq ft \end{cases}$$

$$\text{No. of variables} = 1$$

$$O_{value} = GA(f(fit), \text{No. of variables}, \text{Initialized parameters})$$

End

End

$$\text{Training data} = O_{value}$$

11. **for each set of** Training data

$$\text{Group} = \text{Training data}(i)$$

End

12. Initialized the ANN using Training data and Group
13. Net = Newff (Training data , Group, Neurons)
14. Set the training parameters according to the requirements and train the system
15. **Net = Train (Net, Training data , Group)**
16. **Return; Net as ANN trained structure**

Save the ANN training structure and used in the classification of speech authentication system.

After the training of ANN, we hae evaluated the performance parameters of training as MSE, gradient, mutation and validation checks [14]. MSE is known as a function of network performance and it measures the neural network's performance as per categories of training data. The MSE graph of proposed work is depicted in illustration 2(a) by means of epochs. In this graph, epochs denotes the iterations amount which is used by the ANN during the training of speech MFCC feature sets.The training perfomance of the ANN is given below with regression plot of training structure [15].

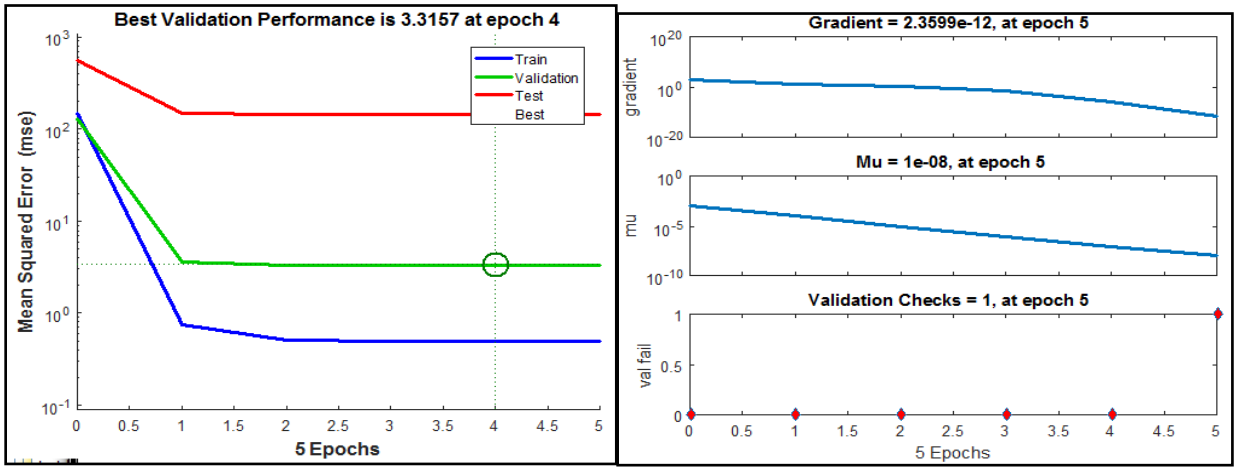


Figure 2: (a) Performance of ANN and (b) Training state of ANN

Figure 2(a) shows the graph of performance parameters of training of ANN and figure 2(b) is the training state of ANN. In the above figure, the green circle denotes the best performance in terms of least mean square error value of 3.3157 at iteration number 4.

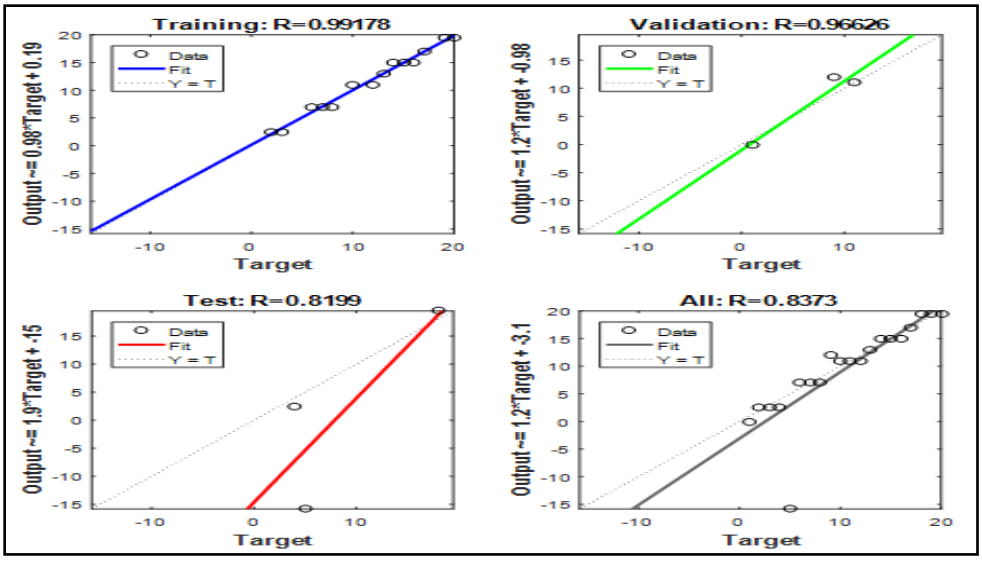


Figure 3: Regression plot of ANN

Figure 3 shows the description of datasets which are utilized for the training purpose. In the above figure, there are total of four graphs: first graph is for training data, second graph is for validation, third graph is for test data which are automatically taken from the training dataset and last graph is for output of training. The black solid line shows the finest fit linear decay line between outputs and targets. The value of regression is denoted by the R and if R is near to 1 then the training will be better. If R is close to zero, then there is no direct relationship between outputs and targets and we can say that training is not proper for proposed speech authentication system. According to ANN training, we have simulated the proposed work with several audio files and the results are described in the below section.

I. RESULT AND ANALYSIS

In this part, the results and analysis of proposed speech authentication system is described. In this work, MFCC feature extraction method based on the genetic algorithm along with the ANN is used for speech authentication system. It is hard to distinguish the speech in presence of background noise, emotion etc. The proposed work is tested on several audio signals to calculate the performance of proposed work according to classification parameters. Due to distortion, speech authentication becomes difficult, so, we have used genetic algorithm as feature optimization technique to optimize the MFCC feature sets for the training. There are two sections in the proposed work, first is training section and second one is the testing section. The step involves in the training and testing is same like feature extraction as of the speech signal by utilizing the MFCC feature extraction, feature optimization using the genetic algorithm and for the classification, artificial neural network is used. The simulator of proposed speech authentication system is shown in figure 3.

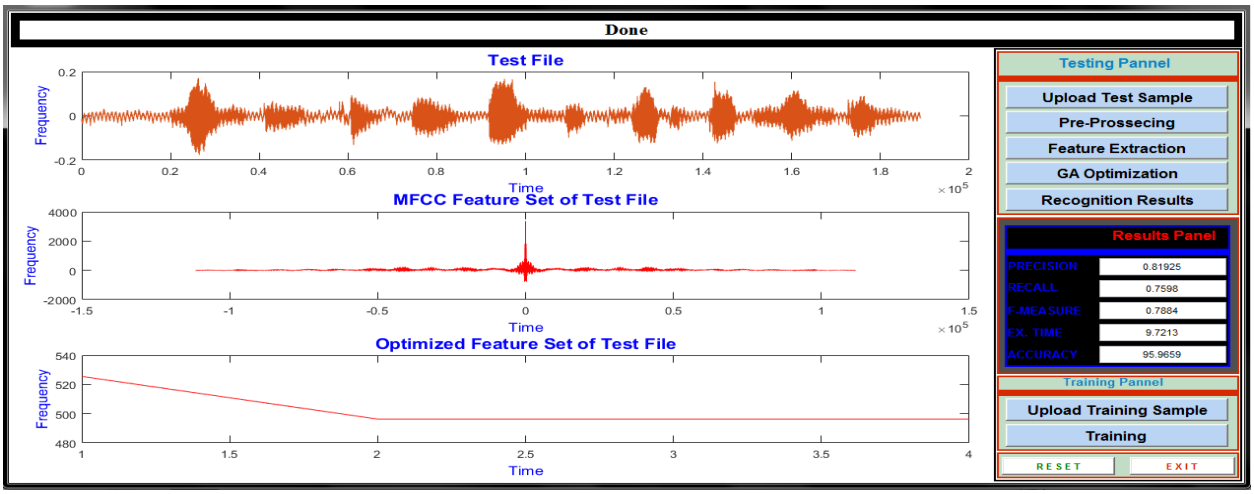


Figure 3: Simulator of proposed speech authentication system

Above figure represents the simulator of proposed work. In the simulator, two sections are categorized; first section is training and second is the classification. In the training segment, we have used ANN (Artificial neural network) with tan sigmoid transfer function to train the input feature sets. After the training of the proposed work, we have classified a test audio signal and the performance metrics like precision, recall, f-measure, execution time and accuracy have been calculated. The used performance metrics are defined below:

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp}+\text{Fp}} \quad (7)$$

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp}+\text{Fn}} \quad (8)$$

$$\text{F} - \text{mesure} = \frac{2\text{X}(\text{Precision} \quad \text{X Recall})}{(\text{Precision} \quad +\text{Recall})} \quad (9)$$

$$\text{Execution Time} = \text{Total time taken by simulator} \quad (10)$$

$$\text{Accuracy} = \frac{\text{Tp}+\text{Tn}}{\text{Tp}+\text{Tn}+\text{Fp}+\text{Fn}} \quad (11)$$

As shown in equations (7), (8), (9) and (11), Tp is true positive, Fp is false positive, Tn is true negative and Fn is false negative. In this section, we have analyzed the obtained results of proposed work for speech authentication system based on the artificial intelligence technique. The reliability of classification depends on feature extraction and training process. The experimental results have confirmed our expectations by giving good values in terms of measuring metrics like precision, recall, f-measure, execution time and accuracy.

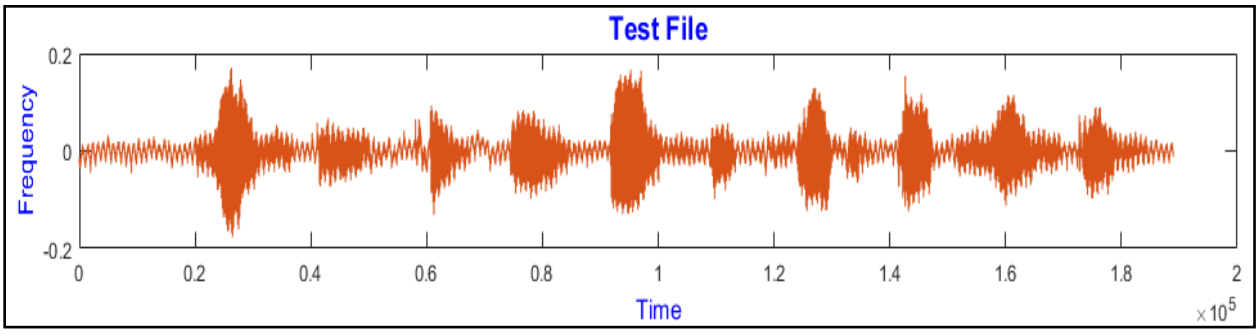


Figure 4: Audio File

Above figure represents the audio signal which used in the speech authentication system to authenticate the users. The probablity of silence present in the audio signal is more which is represented in above figure by linear signal. So, in the speech authentication system, removal of silence signal plays an essential role to achieve better accuracy. When the audio signals are uploaded, the pre-processing steps on the audio signal are applied.

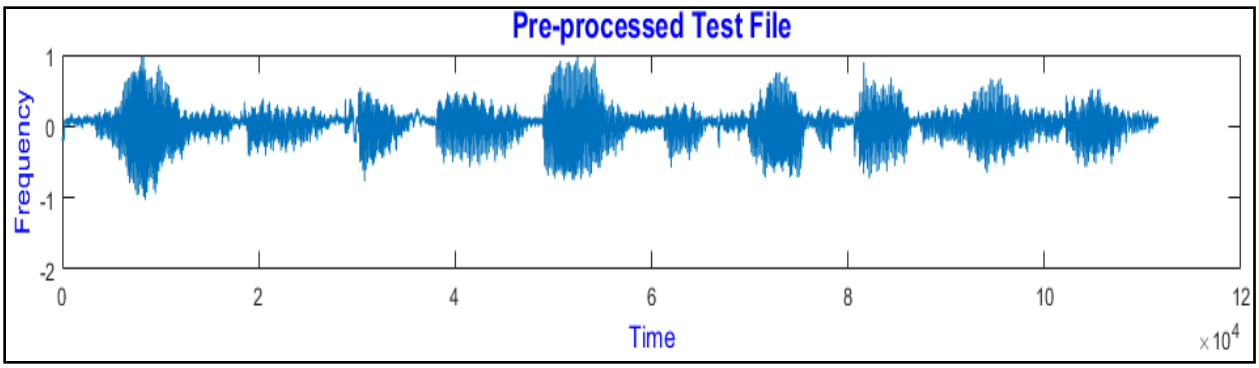


Figure 5: Pre-processed Audio File

Above figure represents the pre-processed audio signal which is used in the speech authentication system to authenticate the users with more accuracy. From the above figure, we have observed that almost silent signal is removed and silence free audio signal for the further processing is achieved. After the pre-processing step, we have calculated MFCCs features for the pre-processed audio signal.

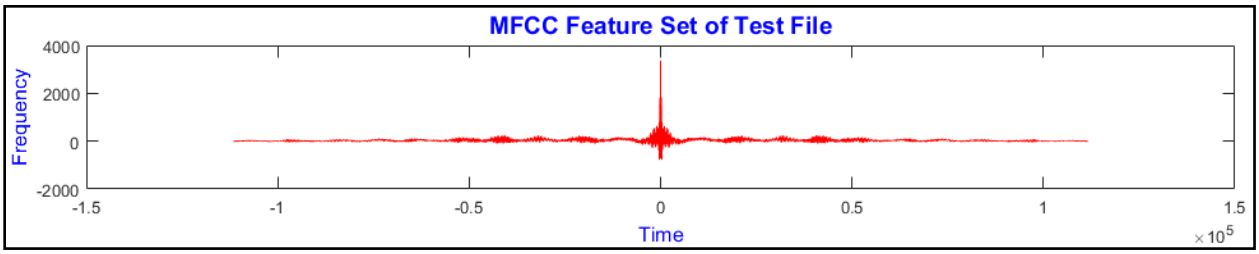


Figure 6: MFCC of Pre-processed Audio File

After the MFCC feature extraction process, we have obtained the above signal which represents the MFCC of audio signal. To optimize MFCC of audio signal, we have used Genetic algorithm with an optimal objective function which is described in the equation 5 and for the training of proposed speech authentication system, ANN as a classifier. By using the artificial neural network, the accuracy rate of proposed work came out to be better than the previous work. The outcome of proposed work is described in the below section with the graphical representation.

Table I: Performance parameters

Sample	Precision	Recall	F-measure	Execution time (s)	Accuracy (%)
1	0.81925	0.7598	0.7884	9.73	95.97
2	0.74462	0.7654	0.7549	7.69	96.74
3	0.73976	0.8659	0.7979	12.34	98.55
4	0.68269	0.6679	0.6752	9.47	95.77
5	0.56784	0.7384	0.6420	8.78	98.69
6	0.78558	0.7364	0.7602	13.74	97.56
7	0.56889	0.6484	0.6060	9.34	98.57
8	0.78586	0.8476	0.8156	8.95	98.66
9	0.85789	0.9377	0.8960	12.86	95.77
10	0.89553	0.7893	0.8391	13.25	98.54
Average	**0.74479**	**0.7757**	**0.7578**	**10.62**	**97.48**

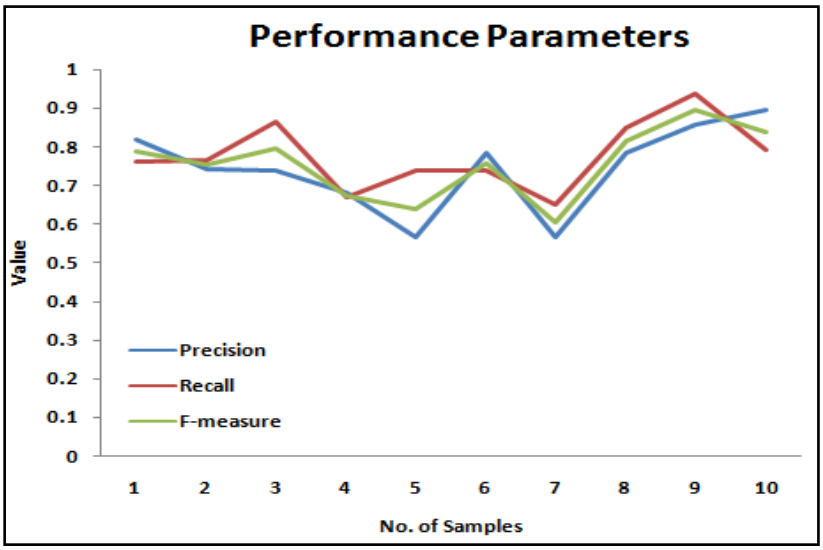


Figure 7: Performance parameters of proposed work

Above figure is the graphical illustration of the performance parameters of proposed work in terms of precision, recall and f-measure. These parameters depict the performance of proposed work. Form the above figure; it is clear that the performance parameters are acceptable for speech authentication system and the average value of precision is 0.74479, recall is 0.7757 and for f-measure, it is 0.7578.

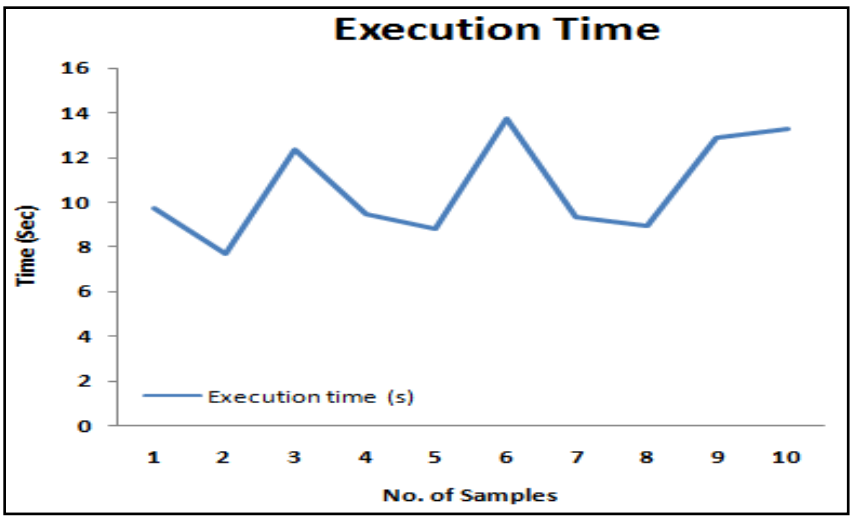


Figure 8: Execution time of proposed work

Above illustration shows the execution time of the proposed work and it should be less for more accuracy of the work. By using artificial neural network, it becomes possible to minimize the execution time of speech authentication system. The execution time is directly proportional to the training of system, if the training of system is better than the execution time will be less.

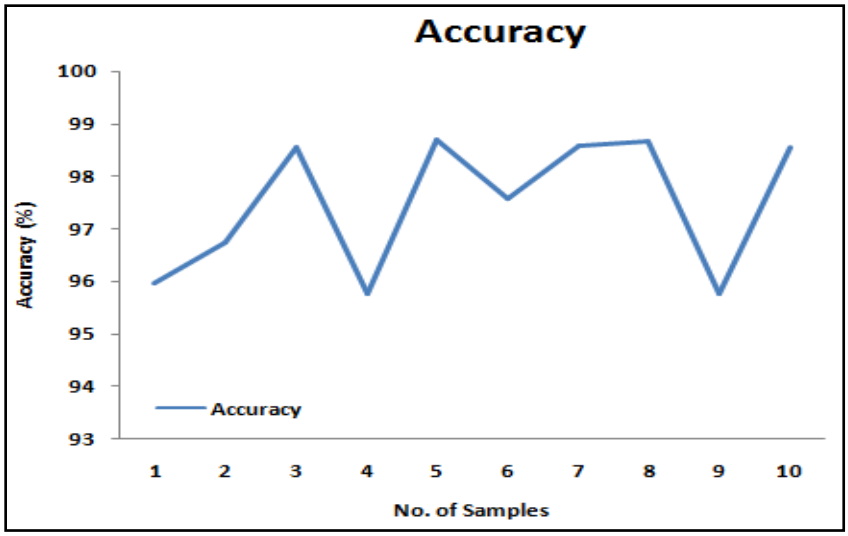


Figure 9: Accuracy of proposed work

Figure 9 represents the recognition accuracy for 10 different speech samples. For the figure, it is being observed that the average recognition accuracy is more that 97% and it shows improvement in the speech authentication system.

Table II: Comparison of proposed work with existing work

Method	Accuracy (%)
MFCC	79.74
PNCC	81.88
M.GSD + NR/PNCC	83.00
MFCC+GA+ANN	97.48

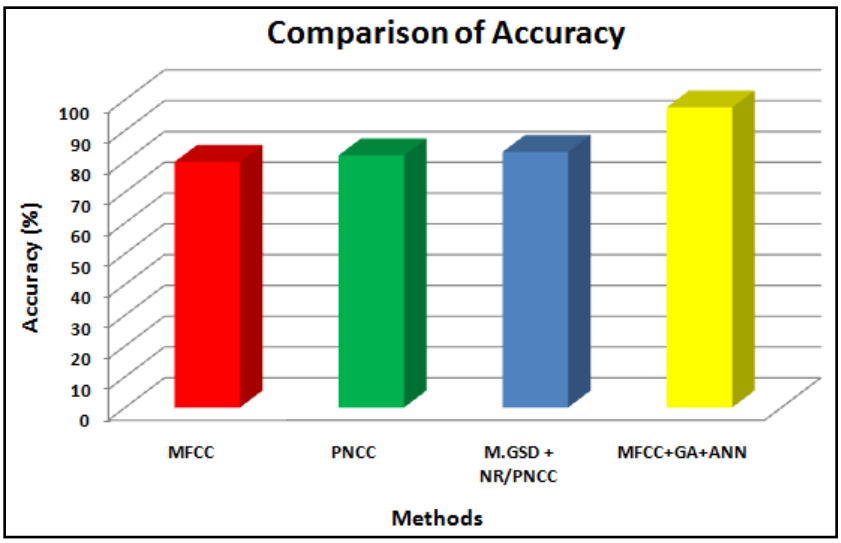


Figure 10: Comparative analysis of proposed work with existing work

Above table and figure represents the comparative analysis of proposed work with existing work for recognition accuracy. In the above figure, the recognition accuracy of existing work with MFCC, PNCC and M.GSD + NR/PNCC has been compared to the proposed work. In the proposed work, MFCC is utilized with GA and ANN. Form the figure, we have observed that the recognition accuracy of proposed work is improved as compared to existing work and the recognition accuracy is improved by 14.85%.

CONCLUSION

In this research, the comparison of speech authentication accuracy with the use of MFCC feature sets for the authentication of disabled users to execute the web content has been considered. By using ANN as a classifier along with the genetic algorithm, the accuracy of proposed speech authentication system is up to 97%. To resolve the issue of feature uniqueness, genetic algorithm is adopted with ANN to improve feature sets of audio signals. With the usage of GA, the system recognition rate increases up to 97% with minimum execution time. Under the circumstances of adopting only MFCC feature sets, speech recognition rate is not acceptable so we have used genetic algorithm with an optimal objective function. This feature optimization technique has provided substantially better recognition accuracy than only MFCC features for speech. The experimental results has analyzed that proposed optimization based speech authentication system using MFCC with ANN and provides enhanced results with precision as 0.74479, recall as 0.7757, f-measure as 0.7578, execution time as 10.62 seconds and accuracy as 97.48%.

In the future, deep learning techniques has been utilized instead of ANN for the classification of speech authentication system to authenticate the disabled users. The deep Learning methods may capture composite relations between extracted feature sets and speech words pattern to improve the classification accuracy.

REFERENCES

[1] Voice, C., Smith, M., McCulligh, M. and Zuccherato, R., Entrust Ltd, 2016. Offline methods for authentication in a client/server authentication system. U.S. Patent 9,281,945.

[2] Kartik, P., Prasanna, S.M. and Prasad, R.V., 2008, November. Multimodal biometric person authentication system using speech and signature features. In TENCON 2008-2008 IEEE Region 10 Conference (pp. 1-6). IEEE.

[3] Eshwarappa, M.N. and Latte, M.V., 2010. Bimodal biometric person authentication system using speech and signature features. International Journal of Biometrics and Bioinformatics, 4(4), p.147.

[4] Slaney, M., Interval Research Corp, 2001. System and method for automatic classification of speech based upon affective content. U.S. Patent 6,173,260.

[5] Han, W., Chan, C.F., Choy, C.S. and Pun, K.P., 2006, May. An efficient MFCC extraction method in speech recognition. In Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on (pp. 4-pp). IEEE.

[6] Hirsch, H.G. and Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW).

[7] Muda, L., Begam, M. and Elamvazuthi, I., 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.
[8] Zhen, B., Wu, X., Liu, Z. and Chi, H., 2000. On the Importance of Components of the MFCC in Speech and Speaker Recognition. In Sixth International Conference on Spoken Language Processing.
[9] Tiwari, V., 2010. MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), pp.19-22.
[10] Dave, N., 2013. Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), pp.1-4.4
[11] Benkhellat, Z. and Belmehdi, A., 2012, July. Genetic algorithms in speech recognition systems. In Proceedings of the International Conference on Industrial Engineering and Operations Management (pp. 853-858).
[12] Karaali, O. and Mackie, A.W., Motorola Solutions Inc, 2001. Method, device and system for part-of-speech disambiguation. U.S. Patent 6,182,028.
[13] Quixtiano-Xicohtencatl, R., Flores-Pulido, L. and Reyes-Galaviz, O.F., 2006, November. Feature Selection for a Fast Speaker Detection System with Neural Networks and Genetic Algorithms. In Computing, 2006. CIC'06. 15th International Conference on (pp. 126-134). IEEE.
[14] Lim, C.P., Woo, S.C., Loh, A.S. and Osman, R., 2000. Speech recognition using artificial neural networks. In Web Information Systems Engineering, 2000. Proceedings of the First International Conference on (Vol. 1, pp. 419-423). IEEE.
[15] Sanderson, C., Bengio, S., Bourlard, H., Mariéthoz, J., Collobert, R., BenZeghiba, M.F., Cardinaux, F. and Marcel, S., 2003, July. Speech & face based biometric authentication at idiap. In Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on (Vol. 3, pp. III-1). IEEE.