International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)
ISSN: 2305-0543 (Online), ISSN: 2411-6173 (Print)
Vol. 11 (42), Oct. 2021, pp. 5059-5064

Recognition of Persian Digits from Zero to Nine using Acoustic Images based on Mel Capstrom Coefficients and Neural Network

Seyed Mehdi Hoseini
Department of Computer Science, University of Mazandaran, Babolsar, Iran
mehdihoseini.cs@gmail.com

Abstract—in this article, first, the database of zero to nine Persian digits have been recorded and collected using the voices of 50 men and women in the environment .In the proposed method, we first frame the preprocessed signal and then go through the improved window; in the next step, it enters the Fourier transform block .Now the Fourier transform spectrum is given to the Gaussian filter bank, and then the output power spectrum of the Gaussian bank filter is passed through Root Function, and then by applying cosine transform to compress the components, Mel-Capstrom coefficients are obtained .Finally, the acoustic image is formed as a matrix containing the temporal and frequency features of the speech signal using a two-dimensional inverse Fourier transform of the Mel Capstrom coefficient matrix .To classify and test the data, the features obtained are trained using an improved algorithm in the perceptron neural network with two hidden layers, and the recognition rate is reported at the end .The test results for the signal to different noises show the improvement of the noise signal detection rate by the proposed method, so that the recognition rate of the proposed algorithm without noise is 98.85.

Keywords— digits recognition, acoustic image, perceptron neural network, Mel-Capstrom coefficients, Gaussian bank filter

I. INTRODUCTION

One of the subfields of signal processing is speech processing. Speech processing includes the three main branches of text-to-speech, speech recognition, and speech improvement .Every speech recognition system needs to extract a feature that can accurately detect input speech .The reinforcement of these features, and in particular the feature of Mel-Capstrom frequency coefficients as the most common of them against noise is very important in speech recognition .

If we limit the range of words to be recognized to numbers, we call that recognition, the recognition of numbers .A number here is a set of one or more digits that are arbitrarily expressed .Recognition of numbers is very important today because of its many uses .Applications of number recognition include automatic identification of personal identification number, national code, bank account identification or membership number of users of a service system, connection to a remote database, automatic telephone dialing, computer password logging, and student registration from Mentioned by phone or internet after announcing the student number or the code of the desired courses and the like .Many researches have been done on recognizing Persian numbers, some of which we will mention.

In the system that was implemented in 1998, the word model from zero to nine was made and each word was modeled with six modes .In teaching this system, 200 samples of each word were uttered by an equal number of male and female speakers .The coefficients used in this system were the Capstral coefficients of linear prediction analysis [1].

Another study to identify discrete digits over the telephone was the multilayer perceptron network, in which the neural network estimated the next feature vector .The training of this system was done using a combination of dynamic programming algorithm and network training algorithm .This system was tested on a telephone database consisting of zero to nine Persian numbers .The coefficients used in this system were Capstral coefficients at the Mel scale .The symbol MFCC[1] represents this type of coefficient . Recognition results in this study for experimental data were 81% [2].

In the article [3], which attempted to recognize Persian continuous speech by a combined system of Markov model and neural network, they achieved 75% efficiency of word recognition. In 2012, an algorithm was proposed using the MFCC method that has a good performance against noise in the environment. In this method, real Capstrom

[1] Mel Frequency Cepstrum Coefficients (MFCC)

Article History:
Received Date: May. 23, 2021, Accepted Date: Aug. 10, 2021, Available Online: Oct. 01, 2021

(logarithmization), Gaussian filter bank and an improved window have been used to improve the results [4].

In this paper, first the collected database is introduced and then the basic method for extracting Mel-Capstrom coefficients is stated and in the next step, Capstral at the Mel scale is used to extract the feature from the speech signal using the proposed method. In section three, we will introduce the neural network and how to recognize it. In the final part, the recognition rate is reported and compared with feature extraction methods such as Mel-Capstrom coefficients in the basic method, Gaussian Mel-Capstrom coefficients and Gaussian root Mel-Capstrom coefficients for speech recognition in signal to different noises. In this article, instead of the phrase (Mel-Capstrom coefficients), (Mel coefficients) will be used briefly.

The collected database has the sound of zero to nine Persian numbers, which is recorded and prepared by 50 people, including men and women in the age range of 20 to 30 years, by the authors of this article in the environment, and It contains a total of 500 number sounds, which are cut evenly with Maker Move software, and the sounds of each number are stored in a separate file.

The basic method for extracting Mel coefficients is shown in Figure 1. In this method, first the speech signal frame is passed through the Hamming window and then past the discrete Fourier transform block and the result is applied to the triangular filter bank. In the next step, the output of the triangular filter bank passes through the logarithm block and the inverse Fourier transform and forms the Mel coefficients.

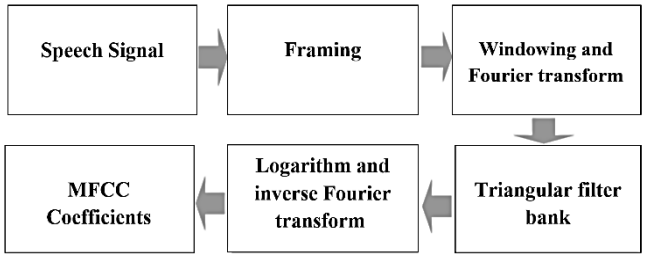

Fig. 1. Extraction steps of Mel-Capstrom coefficients in the basic method

In 2009, Gaussian functions were used instead of the triangular bank filter, which improved the detection rate due to the greater correlation between the frames. The Latin symbol GMFCC[2] represents this type of coefficient [5].

In general, three different areas can be used to improve this feature extraction method; the first is the improvement in the algorithm and base blocks. The second field, considering the importance of hardware implementation of this algorithm, is the improvement in the hardware part, and finally, the third field is the improvement or creation of new and complementary blocks in the basic algorithm. The method of improvement in this article is of the first and third types, ie the improvement of the basic algorithm and its complement.

II. Proposed Method to Improve Mel Coefficients

The block diagram of the proposed method for improving the basic method of Mel-Capstrom coefficients is shown in Figure 2.

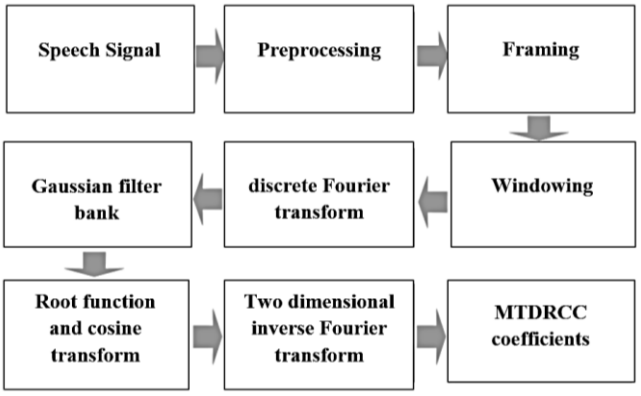

Fig. 2. Block diagram of the proposed feature extraction method

Since the proposed method uses root Capstrom and acoustic image, the coefficients obtained from this method are represented by the abbreviated symbol MTDRCC[3]. In the following, the steps of the proposed method are described.

A. *Preprocessing*

First, we delete the silence at the beginning and end of the speech file and then we delete the DC value of the signal using the speech signal difference and its average value to improve the next operation. In the next step, the speech signal is sent to the pre-emphasis high-pass filter. One of the reasons for using the pre-emphasis filter is that this filter effectively eliminates the adverse effects of the larynx and lips and the sudden changes in the signal caused by ambient noise and makes the speech signal uniform. If S (n) is the speech signal and P(n) is the output of the pre-emphasis filter, relation (1) defines the conversion function of this filter [6].

$$P(n) = S(n) - \alpha . S(n-1) \quad (1)$$

The parameter α is a pre-emphasis coefficient and its range is usually 0.9 and 1.

Figure 3 shows the speech signal before and after applying the pre-emphasis filter with a pre-emphasis coefficient of 0.95.

[2] Gaussian Mel Frequency Cepstrum Coefficients (GMFCC)

[3] Mel-scale Tow Dimension Root Cepstrum Coefficients (MTDRCC)

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

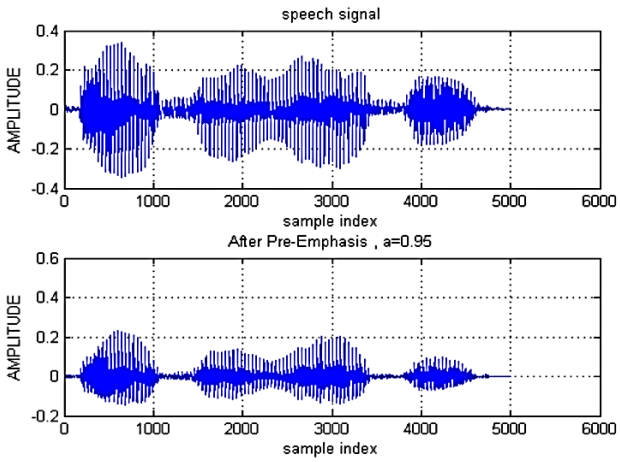

Fig. 3. Speech signal before and after applying the pre-emphasis filter

B. Framing

After the preprocessing operation, it is time to frame the speech signal. Speech signal is non-static, but because the organs of speech cannot change faster than a certain limit, it can be assumed to be static at short intervals. Thus, the speech signal is usually divided into 20 or 25 millisecond frames with overlapping 1.2 or 1.3 frame length [7]. If the frame length is shorter, the speech signal is covered by more frames, so the length of the extracted feature vector increases and the computational volume increases, but more speech information is provided. Here we consider each frame with a length of 500 Samples and an overlap of 150 Samples.

C. Windowing

In this step, each frame is multiplied separately in a window to reduce the effect of signal discontinuity at the beginning and end of the signal. The choice of window is very important, because the margins of a frame have an effect on increasing or decreasing the error signal, the basic algorithm uses Hemming windowing W(n) in relation (2) [6]. But the window we use here is according to Equation (3), the enhanced window $W_{New}(n)$. If the frame is displayed with X(n) and the windowed frame is displayed with $\overline{X}(n)$, the window operations will be the same as in Equation (4). N is the number of samples in a frame and K is the number of frames.

$$W(n) = 0.54 - 0.46\cos\frac{2n\pi}{N-1} \qquad 0 \le n \le N-1 \qquad (2)$$

$$W_{New}(n) = nW(n) \qquad 0 \le n \le N-1 \qquad (3)$$

$$\overline{X}_k(n) = X_k(n)W_{New}(n) \qquad 0 \le n \le N-1 \qquad (4)$$

In the improved window, three important indicators of dispersion, convergence of lateral parts and width of the main part of the window have been considered. In this window, compared to a simple Hamming window, the spectral scattering factor increases as well as the width of the main aspect and the convergence factor of the lateral aspects decreases, the first two of which are desirable changes and the last of which is an undesirable effect [8]. Due to the improved results, the adverse effect can be ignored. Figure 4 shows the difference between these two windows.

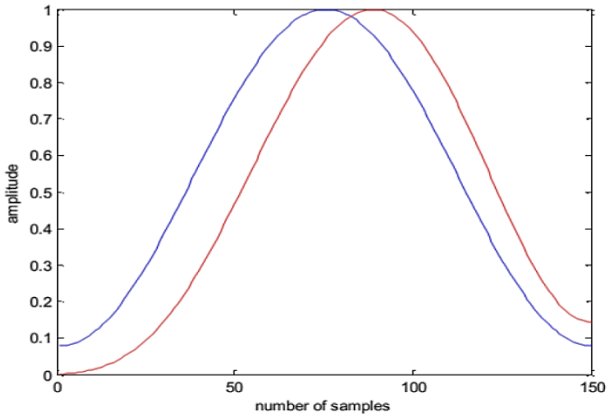

Fig. 4. Hamming window (blue) and improved window (red)

D. Discrete Fourier Transform

Because the type of sound is related to the distribution of signal energy in the frequency domain, we need its information in the frequency domain. Using Equation (5), we take the signal x(n) inside each frame to the frequency domain. Since most of the audio information is in its Fourier transform spectrum, we obtain the Fourier transform spectrum of each frame. The Ms parameter is the number of points in the Fourier transform.

$$X(k) = \sum_{n=1}^{Ms} x(n)e^{-j\frac{2\pi nk}{Ms}} \qquad 1 \le k \le Ms \qquad (5)$$

E. Gaussian filter bank

The desired coefficients are obtained by applying a set of filter banks that cover the entire frequency spectrum. In the bandpass filter bank structure, which covers the entire signal bandwidth, the Fourier transform spectrum of the frames passes through the filters. Different types of filter banks are used in speech recognition. These filters simulate the frequency separation of the human ear perception system. In the basic algorithm for calculating MFCC coefficients, a triangular filter bank is usually used.

In this type of bank filter, if the overlap of the frames is not enough, the information of the parts of the frame that are located at the beginning and end points and outside the subsections, is lost, because the triangles have no weight outside the subbands. But if we use a Gaussian bank filter instead of this filter, due to the weight outside its subbands, it prevents the loss of information in these sections.

Also, in the Gaussian filter bank, there is more correlation between adjacent filters and can be Increased or decreased this correlation using the parameter α in relation (8). To create a filter bank, we must first transfer the frequencies to the Mel domain, which is the auditory unit of the human ear. In fact, the relation (6) is a mapping from the actual frequency f to the Mel frequency f_{mel}.

$$f_{Mel} = 2595\text{Log}_{10}\left(1 + \frac{f}{700}\right) \qquad (6)$$

Next, we obtain the parameter k_{bi}, which specified the boundary points in the triangular filter bank, in relation

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

(7). k_{bi} is also used to determine the scattering parameter σ_i. After calculating the variance of each subset of the filter bank in relation (8), we finally obtain the Gaussian bank filter $\Psi_i(k)$ with the final equation in relation (9) [5].

$$k_{bi} = \left(\frac{Ms}{Fs}\right).f^{-1}mel[f_{mel,min} + \frac{i.(f_{mel,max}-f_{mel,min})}{Q+1}] \quad (7)$$

$$\sigma_i = \frac{k_{bi+1}-k_{bi}}{a} \quad (8)$$

$$\Psi_i(k) = e^{-\frac{(k-k_{bi})^2}{2\sigma_i^2}} \quad (9)$$

The parameter Q is the number of bank filters and $1 \le i \le Q$. also Ms is the number of points in the discrete Fourier transform that leads to the energy spectrum. Fs is also the sampling frequency. The parameter α regulates the variance, so that the larger the alpha parameter, the smaller the standard deviation of Gaussian filters, and vice versa. In this paper, the alpha value is equal to 2, because in this case a better correlation is created between adjacent subbands in the Gaussian filter bank [5]. Gaussian filter bank can now be designed to extract GMFCC coefficients. Figures 5 and 6 show the diagrams of the triangular filter bank and the Gaussian filter bank, respectively.

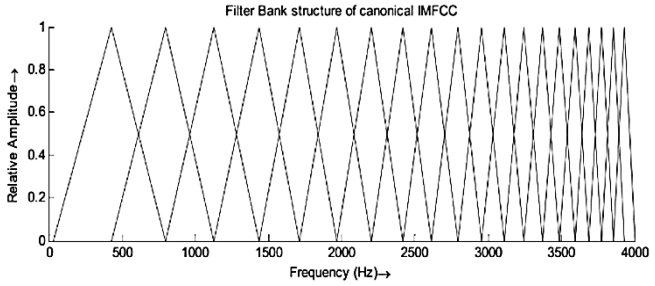

Fig. 5. Triangular filter bank

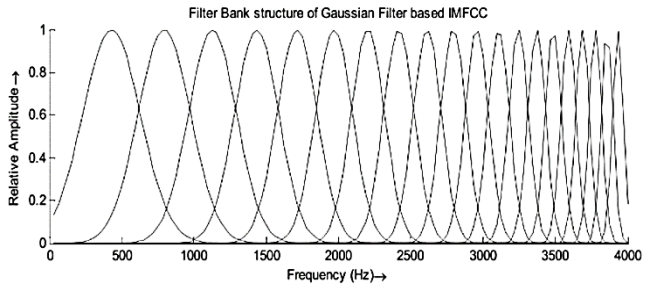

Fig. 6. Gaussian filter bank

F. *Root function and cosine transform*

In the basic MFCC algorithm, after the output of the filter bank, the size logarithm is performed, which is known as the real Capstrom method. In the proposed method, we pass the output of the Gaussian filter bank through the root function. This method, called root Capstrom, will cause the obtained Mel coefficients to move more smoothly, and therefore it will be useful in removing the noise mounted on the speech signal [9].

The gamma parameter in the root function can be a number between -1 and 1, so that the closer the gamma is to one, the more the obtained Mel coefficients will change [10]. In this paper, to achieve better results, the value of the gamma parameter is considered to be 0.3. Figure 7 shows the changes in the logarithm function and the root function for an increase in noise or a decrease in signal to noise.

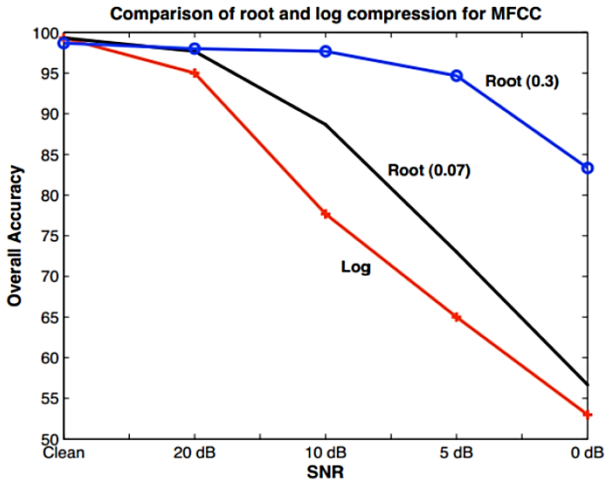

Fig. 7. Comparison of root function and logarithm function in different signal to noise

As mentioned, the root function is more resistant to the increase in noise in the environment than the logarithm function [11]. Root Capstrom operations are performed using equation (10). To reduce the amount of components and compression, a discrete cosine transform according to equation (11) is applied to the output of the root function.

$$Y_{\gamma,j}(m) = |Y_j(k)|^{\gamma} \qquad -1 \le \gamma \le 1 \quad (10)$$

$$C_j(n) = \sum_{m=1}^{F} Y_{\gamma,j}(m) \cos\left(\frac{n(m-0.5)\pi}{F}\right) \qquad 1 \le n \le P \quad (11)$$

In Equations (10) and (11). $Y_j(k)$ is Gaussian filter bank output number k for frame j, F is Number of Gaussian filters, $Y_{\gamma,j}(m)$ is output from root function, $P \le N$ is length of Mel coefficient vector, N is length of a speech frame, γ is Capstrom root parameter and Cj is frequency coefficient vector obtained for frame j. Since the coefficients considered in relation (11) are obtained using the Gaussian filter bank and the root function, we represent these coefficients with the symbol GMFRCC[4] [12].

G. *Two-dimensional inverse Fourier transform*

Usually the features extracted from the speech signal are one-dimensional and from the time or frequency domain, while there are useful features in both time and frequency domains. Therefore, first by Putting together the GMFRCC coefficient vector of each frame, a matrix Y(i,j) containing

[4] Gaussian Mel Frequency Root Cepstrum Coefficients (GMFRCC)

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

the total GMFRCC coefficients of the speech signal is formed. Then, using Equation (12), we obtain a two-dimensional inverse Fourier transform from the matrix Y [10]. The result is a matrix called an acoustic image that has two dimensions of time and frequency [13], meaning it includes the features of both time and frequency domains [9]. The acoustic image matrix is also used to create a widely used spectrogram diagram.

$$\hat{x}(u,v) = \frac{1}{MP}\sum_{m=1}^{M}\sum_{p=1}^{P} Y(i,j) e^{\frac{j2pv\pi}{P}} e^{\frac{j2mu\pi}{M}} \quad 1 \le u \le P$$

$$1 \le v \le M \qquad (12)$$

Where M, the number of frames used to calculate the matrix Y and P, represents the vector length of the Mel coefficients in a frame. The matrix of final coefficients is obtained after calculating the absolute value in relation (13).

$$C(i,j) = |\hat{x}(u,v)| \qquad (13)$$

Figure 8 shows the coefficients of the first column and the first row of the matrix C(i,j). According to Figure 8, most of the important and useful features of the speech signal are in the lower dimensions of the matrix C(i,j), so we separate the low-dimensional data from the matrix A as a sub-matrix containing the important features of the speech signal [9]. In this paper, the sub-matrix has dimensions equal to 20*5, i.e. 100 temporal and frequency features are extracted from each speech signal.

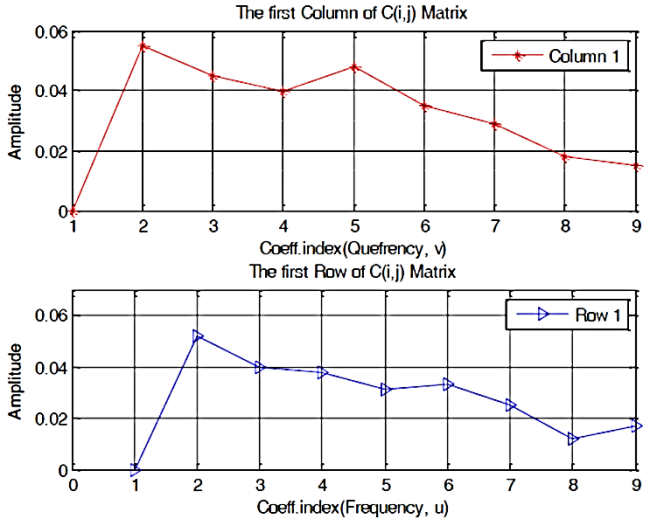

Fig. 8. The coefficients of the first column and the first row of the matrix C(i,j)

III. Classification with Artificial Neural Network

The study of artificial neural networks is largely inspired by natural learning systems in which a complex set of interconnected neurons is involved in learning. A type of artificial neural network is built based on a computational unit called a perceptron. A perceptron takes vectors of inputs with real values and calculates a linear combination of these inputs. In this paper, a perceptron network with an input layer, two hidden layers and an output layer is used. Here we have 10 classes. The number of input layer neurons is determined by the size of the feature vector and the number of hidden layer neurons is customized by the user. The output layer also has a neuron. The connection between each layer is made by connections that connect to neurons. These connections are called network weights [14].

It should be noted that in addition to weights, each neuron is affected by another independent quantity called the bias. The influence of the values of weights and biases on the neurons of each layer to the next layer is determined by a function called the transfer function. Weights related to each neuron are an important factor in the network, so we have tried to select the appropriate values of these weights that have the least network error. This is called network training [15]. Figure 9 shows an overview of the neural network structure used.

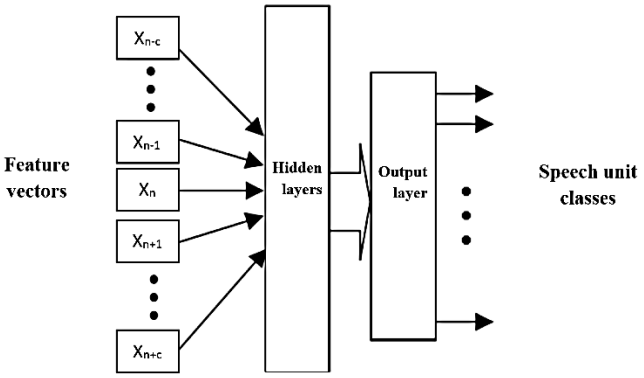

Fig. 9. Structure of a perceptron neural network with hidden layers

In artificial neural network classification, Delta's law is used to reduce errors in weight gain. The main idea of this law is to use a descending gradient to search the space of possible weight hypotheses. The descending gradient algorithm in the weight space seeks vectors that minimize error. This algorithm starts from an arbitrary value for the weight vector and changes the weights at each step to reduce the error. This is the basic rule of the backward propagation method used to train a network with several hidden layers [16,17].

IV. Simulation Results

In order to test the proposed method and evaluate the performance of MTDRCC coefficients at the speech recognition rate that is infected with noise, a series of tests have been performed. The experiments in this paper are performed on a database of Persian numbers between zero and nine described. 70% of these data were used for network training and 30% for network testing. First, four features with the symbols MTDRCC, GMFRCC, GMFCC and MFCC were extracted from each speech signal. For noise-free speech signal, we test the effect of each feature on the recognition rate, then gradually reduce the signal-to-noise ratio. Gaussian white noise is used to generate noise on the signal. The results are shown in Table 1. Comparing the network output for the proposed method and the other three methods in different the signal-to-noise ratio, it is observed

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

that the proposed algorithm is more resistant to applied noise and also has a higher recognition rate than other methods.

TABLE I. COMPARISON OF RECOGNITION RATES FOR DIFFERENT FEATURES AND SIGNAL TO NOISE RATIO

SNR/Feature	MFCC	GMFCC	GMRFCC	MTDRCC
No Noise	85.5	89.55	92.79	**98.85**
25 dB	83.4	87.56	88.22	**97.38**
10 dB	78	82.4	85.45	**96.26**
5 dB	70.45	75.6	76.7	**89.19**
0 dB	61.52	65.15	69.05	**80.09**
-5 dB	54.1	64.18	66.4	**79.54**

V. CONCLUSION

In this paper, an improved algorithm based on Mel-Capstrom coefficients is proposed to increase the recognition rate of noise-infected speech. Using an improved window has a significant impact on improving results. Due to the weight outside the Gaussian bank filter subbands, it prevents the loss of information in these sections and has a significant effect on improving the algorithm. Also, the use of root capstrom makes the mel coefficients more resistant to the noise applied to the speech signal. Extracting the final features from the acoustic image matrix makes it possible to use the important and useful features available in both time and frequency domains of the speech signal. The neural network is designed in such a way that we have the least error in training the network and then the most appropriate weights. According to Table 1, the proposed algorithm is more resistant to noise increase than other methods and has a higher recognition rate than other methods.

REFERENCES

[1] A. Rostamzadeh, "Recognition of discontinuous Persian speech, independently of the speaker with the help of hidden Markov models with continuous density", 6th Iranian Conference on Electrical Engineering, Tehran, 1998.

[2] M. M. Homayounpour and A. Najjari, "Recognition of Persian numbers independent of the speaker using neural predictor model", 7th Iranian Conference on Electrical Engineering, Tehran, pp. 75-81, 1999.

[3] A. Taheri, "Recognition of Persian continuous speech in medium vocabulary by combining neural networks and hidden Markov models", 10th Iranian Conference on Electrical Engineering, Tabriz, 2002.

[4] H. Marvi and D. Darabian, "Resistant Persian Speech Recognition Using Improved Mel-Capstrom Coefficients and Neural Network", 11th Iranian Intelligent Systems Conference, Tehran, 2013.

[5] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter", International Journal of Signal Processing, 5(1), 2009, pp. 11-19.

[6] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques", *arXiv preprint arXiv:1003.4083*, 2010.

[7] V. Skorpil and J. Stastny, "Back-propagation and k-means algorithms comparison", IEEE 8th international Conference on Signal Processing, Vol. 3, 2006.

[8] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition", IEEE signal processing letters, 20(2), 2012, pp. 149-152.

[9] H. Marvi, "Efficient feature extraction based on two-dimensional cepstrum analysis for speech recognition", University of Surrey (United Kingdom), 2004.

[10] X. Wang and Z. Han, "A novel acoustic feature extraction algorithm based on root cepstrum coefficients and CCBC for robust speech recognition", In 2008 Second International Symposium on Intelligent Information Technology Application, Vol. 1, December 2008, pp. 643-647.

[11] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments", Speech communication, 12(3), 1993, pp. 277-288.

[12] M. Babaei and A. Soleimani, "Acoustic images for numbers recognition", 8th International Conference on Information Technology and Knowledge, Hamadan, 2016.

[13] H. Marvi and E. Chilton, "Acoustic classification using time-frequency distributions", In Proceeding of the Institute of Acoustics, Vol. 2, 2004, pp. 612-620.

[14] J. Esmaily, R. Moradinezhad, and J. Ghasemi, "Intrusion detection system based on multi-layer perceptron neural networks and decision tree", IEEE 7th Conference on Information and Knowledge Technology (IKT), may 2015, pp. 1-5.

[15] A. M. Othman and M. H. Riadh, "Speech recognition using scaly neural networks", World Academy of Science, Engineering and Technology, 2008, pp. 253-258.

[16] V. Skorpil and J. Stastny, "Back-propagation and k-means algorithms comparison", IEEE 8th international Conference on Signal Processing, Vol. 3, November 2006.

[17] K. Gurney, "An introduction to neural networks", CRC press, 1997.

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173