INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Mel Frequency Cepstral Coefficients based Bacterial Foraging Optimization with DNN-RBF for Speaker Recognition

P S Subhashini Pedalanka[1,2]*, M SatyaSai Ram[3], Duggirala Sreenivasa Rao[4]

1 Associate Professor, Department of E.C.E, R.V.R & JC College of Engineering, Chowdavaram, Guntur, India
2 Research scholar, Department of E.C.E, JNTUH, Hyderabad, India
3 Professor, Department of E.C.E, R.V.R & JC College of Engineering, Chowdavaram, Guntur, India
4 Professor, Department of E.C.E, JNTUH, Hyderabad, India

OPEN ACCESS

Received: 04.10.2021
Accepted: 18.11.2021
Published: 04.12.2021

Citation: Subhashini Pedalanka PS, SatyaSai Ram M, Rao DS (2021) Mel Frequency Cepstral Coefficients based Bacterial Foraging Optimization with DNN-RBF for Speaker Recognition. Indian Journal of Science and Technology 14(41): 3082-3092. https://doi.org/10.17485/IJST/v14i41.1858

* **Corresponding author**.

pssubhashini.pedalanka@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To improve the accuracy and to reduce the time complexity of the Speaker Recognition system using Mel-Frequency Cepstral Coefficients (MFCCs) and Bacterial Foraging optimization (BFO) with DNN –RBF. **Method:** The MFCCs of each speech sample are derived by pre-processing the audio speech signal. The features are optimized with BFO algorithm. Finally, the probability score for each speaker is generated to identify the speaker. Then the features are classified towards the target speaker using DNN-RBF. For the proposed MBFOB speaker recognition function, the TIMIT read corpus is used. It contains a total of 6300 phrases, 10 phrases each. **Findings:** the identity of user is validated in the fields of authentication and surveillance for recognition of speaker. By using the audio speech signal, features are extracted. This paper suggests an MBFOB solution based on Mel-frequency Cepstral Coefficients and DNN-RBF with BFO, for the identification of speakers. The speech utterance from the TIMIT data corpus is preprocessed to obtain MFCC feature vectors DNN-RBF is used for the purpose of classifying the speaker and the feature vectors in the output layers are optimized with Bacterial Foraging optimization. Finally, the scores for each speaker are calculated to identify the speaker. Different output metrics like EER, DCF, Cavg and accuracy are used to test the proposed speaker recognition technique. The execution time of this proposed method is found to be lesser than the other existing methods. The experimental findings are contrasted with other current methods and it shows the efficiency of our approach. **Novelty:** A novel MFCC-based Bacterial Foraging Optimization with Deep Neural Network-Radial Basis Function (DNN-RBF) for identification of exact speaker is proposed in this study.

Keywords: BFO; DNN; RBF; Speech processing; speaker recognition; MFCC extraction; deep neural network; and Bacterial foraging optimization; scoring

1 Introduction

Speaker recognition plays a major role in communication and surveillance area. This recognition process was evaluated by matching the training data with the test data. In recent years number of experiments was done to improve this recognition approach. The best generally used feature in recognition of speakers was MFCC. The MFCC of speech signal provides the local spectral properties for frame analysis. The obtained MFCC features are calculated for the entire training set samples, and they are stored for speaker recognition. MFCC feature computes both the training and testing set samples.

In [(1)], described a complete Audio-Visual Speech Recognition system which include face and lip detection, features extraction and recognition. This article carried out using our own database, which was designed for evaluation purpose. We used a new technique for visual speech recognition that is Zernike Moment.

In [(2)], approaches to be more robust than cepstral features like MFCCs and LPCCs. The case of sensor mismatch was investigated with promising results. In [(3)], proposed approach which gave results that compare favorably with inter-dataset variability compensation and whitening using matched domain data.

In [(4)], presented a new method for speaker recognition (SR) and language recognition (LR) using DNN. The improved performance obtained by DNN in automatic speech recognition (ASR) encourages it to employ its application in speaker and language recognition. The important benchmarks provided for both SR and LR was DAC13 and LRE11. In this method, both the integrated DNN and baseline contains the i-vector classifier. The tandem features or score fusion was used by DAC13 SR to reduce the error rates. The tandem features provide better performance in SR, but in LRE11 performance, the score fusion does not provide any significant improvements.

In [(5)], introduced a new method to recognize the speaker using random digit strings. In this method, the Joint Factor Analysis (JFA) method was explored with arbitrary digit strings for speaker recognition. The database RSR2015 (part III) was utilized by this method.

Zhang, et al, introduced the DNN with de-noising auto encoder (DAE) based reverberation and for recognition of speaker, bottleneck features are considered. In this, the distance-talking set was selected for speaker identification. This method introduces the multichannel least mean squares (MCLMS) method to reduce the error rates. The corresponding error rate obtained by this method for bottleneck feature is 21.4% and for the auto encoder feature the error rate was 47.0%. Furthermore, the performance of this method was enhanced by fusing the DAE-based de-reverberation and the DNN-based bottleneck feature. The two features used by DNN train Gaussian mixture model (GMM) for recognition of speakers. The likelihood features obtained from both the DAE and bottleneck enhances the speaker recognition performance [(6,7)].

In this method, GMM-UBM benchmark contains 60-dimensional, variance and mean normalized PLP instead of MFCC. Two diverse techniques were applied in score normalization to improve the performance. Among that data string technique improves the performance result. At last, the perfectly matched data string was obtained between both test utterance and impostor cohorts [(8)].

The suitable information from the raw data was extracted by DNN. Initially, the layer-by-layer unsupervised learning was proposed to train the DNN and it was finely tuned by the supervised learning algorithm. Finally, the DNN was trained to remove the unwanted features from the audio signal [(9)].

It seems that we need more efforts on the BN-based method, especially for solving its weakness in the open phrase-set scenario. Although the DNN-based method in some cases outperformed the HMM-based one, we believe that the HMM method reflects the very nature of the text dependent task and we should be able to improve its performance. Experimenting with triphone models to improve the context modeling will be the first natural step of our future work [(10)].

The proposed STRF based feature to perform speaker recognition and also presents a feature set that combines the proposed STRF feature with conventional Mel frequency cepstral coefficients (MFCCs). The support vector machines (SVMs) are adopted to be the speaker classifiers. To evaluate the performance of the proposed speaker recognition system, experiments on 36-speaker recognition were conducted. Comparing with the MFCC baseline, the proposed feature set increases the speaker recognition rates by 3.85 % and 18.49 % on clean and noisy speeches, respectively [(11)].

In this article, exception of minimum variance distortion less response (MVDR) beam forming, most algorithms perform consistently on real and simulated data and can benefit from training on simulated data. We also find that training on different noise environments and different microphones barely affects the ASR performance, especially when several environments are present in the training data: only the number of microphones has a significant impact. Based on these results, we introduce the CHiME-4 Speech Separation and Recognition Challenge, which revisits the CHiME-3 dataset and makes it more challenging by reducing the number of microphones available for testing. [(12)]

DL provides an enormous success in neural networks. Two feed-forward architectures most effectively used for speaker recognition are Convolutional Neural Networks, and Deep Neural Networks. In this Deep Learning (DL) technique for speaker recognition. In this method, speaker recognition was achieved by filling the gap between i-vector cosine and oracle scoring

system. In this method, the process was done by choosing the Deep Learning (DL) as a backend [13].

2 Related Work

The existing methods introduce a number of speaker recognition algorithms. None of these existing approaches provides an algorithm to increase the precision of the input signals. The speaker recognition technique has been widely attractive for its extensive application in many fields. As a critical method, the Gaussian Mixture Model (GMM) makes it possible to achieve the recognition capability that is close to the hearing ability of human in a long speech. In this paper, we propose a novel model to enhance the recognition accuracy of the short utterance speaker recognition system. Different from traditional models based on the GMM, we design a method to train a Convolutional Neural Network (CNN) to process spectrograms, which can describe speakers better. Thus, the recognition system gains the considerable accuracy as well as the reasonable convergence speed. The experiment results show that our model can help to decrease the equal error rate of the recognition from 4.9% to 2.5%.The experimental results illustrates that BFO model is better than other traditional algorithms. This BFO system has therefore been integrated with the DNN-RBF for optimum classification performance [14,15].

The effectiveness of introducing deep neural networks into conventional speaker recognition pipelines has been broadly shown to benefit system performance. A Euclidean distance similarity metric is applied in both network training and SV testing, which ensures the SV system to follow an end-to-end fashion. For datasets with more severe training/test condition mismatches, the probabilistic linear discriminant analysis (PLDA) back end is further introduced to replace the distance based scoring for the proposed speaker embedding system. Comparison with the state-of-the art SV frameworks on three public datasets justifies the effectiveness of our proposed speaker embedding system [16,17].

This article proposes a BFO algorithm based on MFCC to improve speaker recognition accuracy. In recent works the amount of input data was reduced to improve the accuracy and perform number of iterations to identify the appropriate speaker. But in this approach, the issues regarding the accuracy was removed by injecting MFCC based BFO. The entire unpredictable covariance matrix was determined by this MFCC in single iteration. Finally, the perfectly optimized result was obtained by the BFO algorithm.

Two methods, adaptation process for universal model and impostor selection algorithm are included in Deep Neural Networks (DNN)and Deep Belief Networks (DBN) based hybrid system for performance improvement. The performance gap of about 46% was filled by this method. The explicit session model was not included in this method so it fails to outperform the PLDA.

3 Methodology: Speaker Recognition with MBFOB

The identification of speakers is based on extraction and classification of speakers. Figure 1 shows the complete design of work.

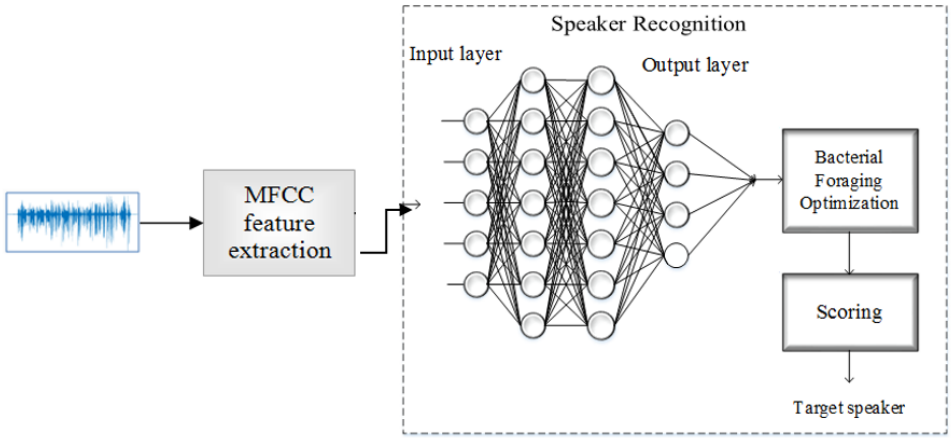

Fig 1. Proposed Speaker Recognition system Block diagram

The proposed architecture divides the speaker recognition task into two modules. They are MFCC feature extraction and speaker identification. Initially, the feature vector from the speech signal is obtained as MFCC feature. The input for the neural

network is MFCC feature vector which optimizes the features at the output layer and they are scored to identify the specific speaker.

A. Preprocessing

The main purpose of this parameterization is to remove appropriate information by eliminating redundant information from speech waveforms. Frame by frame entire process is performed. Single N-dimensional feature vector is converted in each frame. Value of N is greater than number of samples within each frame. Backend device processing will provide the requisite data based on the reduction of data volume. Hence for extraction the sequence of vector is converted based on input audio,

$$X = [x_1, x_2, \ldots x_k \ldots]$$

Here frame index is represented by using k, N-dimensional vector is represented by using x_k

- The determination of MFCCs is done by using the following steps
- Pre-emphasizing of voice signal is the first step.
- Frame sequence is separated from the speech signal by using 20ms frame and shift is obtained is 10ms. each frame is applied in the hamming window.
- The magnetic spectrum for each window frame is determined by applying DFT.
- The spectrum of Mel calculated obtained via the Mel filter bank by passing the DFT signal.
- DCT is used to derive the desired MFCCs for the log Mel frequency coefficients (log Mel spectrum).

B. DNN-RBF and Bacterial Foraging Optimization (BFO) based classification

DNN-RBF is a state-of-the-art technique which achieves better recognition performance in speaker recognition. DNN-RBF is considered with multiple non-linear or linear hidden layers that represents data in encoded form. The main concept of DNN-RBF is activating the current output layer to the input of next hidden layer. Identification capability is enhanced by using large number of hidden layers. The relationship between the input layers and the first hidden layer is given by,

$$a_1 = F(W_1 x + b_1) \tag{3}$$

Where w_1 and b_1 are weight, bias matrix of first layer and the activation function Gaussian is denoted as F(.). The definition of special case for logistics function is given as shown in below,

$$F(x) = e^{-x^2} \tag{4}$$

Here, input of activation is denoted as x. First hidden layer is obtained by providing mapping between current and next hidden layer and this relation is given as shown in below

$$a_1 = F(W_1 a_{l-1} + b_1), l = 2 \ldots, L \tag{5}$$

Where L signifies total quantity of layers, a_{l-1} represent the first layer, F(.) represent the activation function.

For speaker recognition, G (.) is applied in output layer. Hence the output of DNN-RBF is represented by,

$$\hat{y} = G(a_L) \tag{6}$$

Where, G (.) is the soft max function and for the label y, the parameters of DNN-RBF can be defined as follows:

$$\theta' = \text{argmin} \{ \digamma(y, \hat{y}, x, \theta) + \gamma R(W) + \eta \phi A) \} \tag{7}$$

Where, $\theta = \{W_l, b_l, l = 1, 2, \ldots, L\}$ denotes the parameter set and C(.) is the cost function. Cross entropy is considered as the cost function. For DNN-RBF training, the training data $X = [x_1, \ldots x_i, \ldots x_n]$ and the output labels are $Y = [y_1, \ldots y_i, ..y_N]$ where, N denotes the total number of training samples. The cost function is denoted as,

$$C(Y, \hat{Y}; X, \theta) = \frac{-1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} [y_{i,j} \log \hat{y}_{i,j}] \tag{8}$$

Where $\hat{Y} = [\hat{y}_1, \ldots \hat{y}_i, \ldots \hat{y}_N]$ denotes the DNN-RBF output, $\hat{y}_{i,j}$ and $y_{i,j}$ is the j^{th} element of $\hat{y}_i$ and y_i respectively. The value of R(W) is calculated as,

$$R(W) = \sum_l \|W_l\|_F^2 \tag{9}$$

Where, $\rho(A)$ represents the penalty sparsity of the hidden layer output, $\|\cdot\|_F^2$ is the Frobenius norm, $r_{\text{and}}\,\gamma$ denotes the controlling coefficients. In the output layer, the BFO is used to recognize the target speaker. The required speaker dependent features are obtained from the bottleneck layer which has low dimensionality and the dimension is further reduced with principal component analysis (PCA). Thus the abstract level of feature is obtained with reduced dimension.

i) Bottleneck feature extraction and dimensionality reduction

The bottleneck features (BFs) from the hidden layer of neural network are extracted which contains fewer units than other hidden layers. This layer creates the abstract features which is suitable for speaker recognition in DNN-RBF. BFs are generally used in auto encoders and to predict input features, a neural network is trained. The output of hidden layers can be used as the features for further processing. Nonlinear transformations are represented for preprocessing the speech signal. The dimensionalities of bottleneck features are required to be reduced due to its large size. Principal component analysis (PCA) is a technique which reduces the original variable p by q number of derivative variables. Principal components are denoted as the linear combinations of original variables with reduced size than the original variables. Due to its simplicity, it is used in dimensionality reduction of features.

When using large set of features, measurements of these features are required to be processed. It is possible to lessen the features when much of them are available. Consider the vector x of p variables and n measurements. For reducing the dimension from p to q, PCA detects the linear combinations

$$a_1'x, a_2'x, \ldots, a_q'x$$

This is known as principal components. It has maximum variance and they are being correlated with other $a_k'xs$ The Eigen vectors $a_1, a_2, \ldots a_q$ of the covariance matrix S correspond to q largest Eigen value are used to solve the given maximization problem. This value provides variance for their principal components and the proportion of total variance. The first q principal components are denoted as the ratio between q Eigen values and p original variables.

ii) Bacterial Foraging optimization algorithm

A BFO algorithm consists of multiple bacteria that consists optimization solutions and consist of processes in three stages: elimination dispersal, chemo taxis and replication. The below Table 1 shows the BFO algorithm.

Table 1. BFO Algorithm

Input: M unaligned set of sequence
// parameter initialization
P, , , , , , ,
Step 1. Removal and dispersal loop considered as
Step 2. Duplicate loop considered as
Step 3. Chemotaxis phase
i. For , proceed the chemotaxis steps on behalf of bacterium 'i' as specified below:
ii. Perform calculation of fitness function of
iii. Additional cells have been added to the influence of the attractant cell on the nutrient concentration and its fitness function has been obtained as a nutrient concentration calculated as
iv. Let collect this value subsequently the better cost can be found through a iteration run. If the fitness values are in increased order, continue the process for all samples. Otherwise perform Tumbling.
Swim
a. Initialize (swim length counter)
b. While
Let
If (if doing better)
Let and
Let to compute the new as in step 3(e) above.
Else let , end the while loop

Continued on next page

Table 1 continued

Repeat the same procedure for next bacterium , go to step (b) if
v. Tumble: Create a random vector for each element a random number on (-1,1)
vi. Move using
It will result in a step of size in the way of the tumble for bacterium i
Calculate the fitness function
Let
Step 4. If , then perform this step. In this case, continue chemotaxis because the life of the bacteria is not over. Else end chemotaxis process.
Step 5. Duplicate loop phase
a. For considered k and for each
Let sort the health values in descending order (higher cost means lower health)
b. Bacterial with lower relative fitness die when bacteria with higher relative fitness divide to replicate and are put in the same spatial position. So the population remains stable.
Step 6. If , go to step 2, increment the reproduction step by starting the next generation of chemotactic step else end the process and go to step 7.
Step 7. Elimination and disperse Stage: Individually bacterium with probability . If the bacterium is removed, separate another one at an arbitrary location in the optimization search space.
If, then got to step 2 or else end the process.

Direction of bacterium is represented by using the chemistaxis based on the random-directed bacterium. Next stored is bacteria and reproduction cycle is survived during the half of the fittest bacteria. Two similar bacteriaare spitted from the remaining bacteria to form new bacteria. Different random locations are moved probabilistically by using thebacterium. This mainly performs elimination / dispersion procedure. While this action preserves the diversity of the output, the optimization cycle may be interrupted and thus carried out after many stages of the reproduction process.

C. Scoring

The work proposed is implemented as a framework level so that each frame is fed to the network and for each target speaker a class posterior likelihood is achieved. This makes MBFOB suitable for real-time applications as the decision concerning the target speaker is made in each frame. In the optimization process, the input from the past system is combined with the new frame. The probability and the definition of target speakers are defined as

$$P(y) = \frac{e^{y}}{\sum_{k=1}^{K} e^{y}} \tag{10}$$

If the output layer is indicated, K is the total class number and the rear of the target speaker is described as P(y). Using the output layer, you can make flexible decisions on the target speaker. The target speaker is highly scored based on the performance criterion by using this probability measure.

4 Results and Discussion

The TIMIT data corpus assesses the proposed MBFOB method for the recognition of speakers. Information about the data corpus and data from the extraction and description of the function are presented in this section. Evaluation of performance measures such as EER, DCF, Cavg, precision, recall, and F-measure and comparison of test results with the current methods.

A. Corpus

For the proposed MBFOB speaker recognition function, the TIMIT read corpus is used. It contains a total of 6300 phrases, 10 phrases, spoken by 630 speakers from eight major United States dialect regions. For each utterance, the TIMIT corpus includes a 16-bit 16 kHz waveform file. The entire archive is marked as files for training and research. Seven audio files for training are used for each speaker and three files for testing.

B. Parameter selection

The speech is analyzed with a 20-ms Hamming window with a fixed frame rate of 10-ms to extract features. The vector is generated by the Fourier filter-based transformer bases that consist of 13 Mel-sized coefficients and their energy along with

its first and second time derivatives. The MFCC feature vector dimension is says the number of units in input layer for DNN-RBF+BFO, and the output layers with 64 units are used. There are 6 hidden layers and the bottleneck layer has been used with 64 units. For the purpose of comparison, the work is simulated with other approaches like DNN-multilayer perceptron, DNN-back propagation, and DNN-Radial basis function. Similar parameter settings are used for all approaches and the LPC, delta and delta features extracted for comparing with the performance of MFCC feature vector.

C. Equal Error Rate (EER)

The ratio of false acceptance rate (FAR) is equal to the false rejection rate (FRR) is nothing but EER. Speaker recognition accuracy prediction measurement is considered. FAR and FRR parameters tested which are given below.

$$FAR = \frac{\text{No.of falseacceptance}}{\text{No.of identificdionattempt}} \tag{11}$$

$$FRR = \frac{\text{No.of falserejection}}{\text{No.of identificdionattempt}} \tag{12}$$

Both the number of incorrect rejection and number of incorrect acceptance parameters are evaluated. By using the existed and proposed approaches EER value is featured and shown in Tables 2, 3 and 4.

Table 2. Speaker identification approaches with EER comparison

Feature extraction methods	EER comparison for Existing Speaker Recognition approaches		
	DNN-RBF	DNN-BP	DNN-MLP
MFCC	0.0208	0.0336	0.055
LPC	0.0265	0.055	0.0722
Delta	0.0336	0.0722	0.0981
Delta - Delta	0.055	0.0981	0.1414

The EER rate of LPC, delta and delta-delta, MFCC for 30ms is shown in Figure 2. Each features with various neural network models such as multi-layer perceptron (MLP), back propagation, and RBF. Among several feature extraction techniques, MFCC feature extraction provides better performance of EER. The EER of this proposed DNN-RBF with MFCC provide minimum error rate value.

Table 3. EER comparison for Proposed DNN-RBF+BFO Speaker Recognition Approach

Feature extraction methods	EER comparison for Proposed DNN-RBF+BFO Speaker Recognition
MFCC	0.0124
LPC	0.0162
Delta	0.0208
Delta Delta	0.0265

Table 4. EER comparison for Existing and **Proposed** Speaker Recognition

Feature extraction methods	EER comparison for Existing and Proposed Speaker Recognition approaches			
	DNN-RBF+BFO	DNN-RBF	DNN-BP	DNN-MLP
MFCC	0.0124	0.0208	0.0336	0.055
LPC	0.0162	0.0265	0.055	0.0722
Delta	0.0208	0.0336	0.0722	0.0981
Delta Delta	0.0265	0.055	0.0981	0.1414

The comparison results for the EER with LPC, MFCC, delta, delta-delta is shown in Figure 2 and the obtained results of this proposed approach DNN-RBF+BFO is found to be superior to the other three existing approaches.

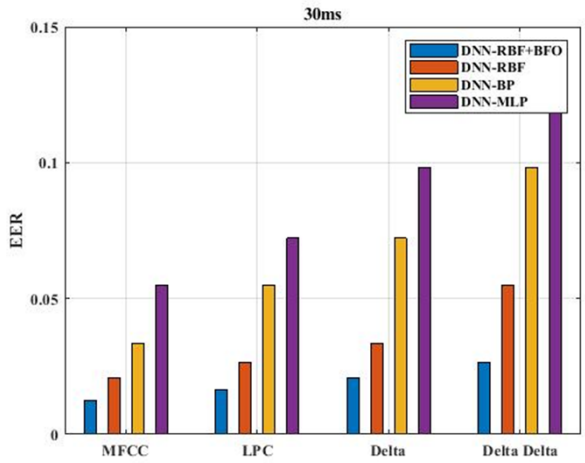

Fig 2. Comparison graph for the EER values of various features (MFCC, LPC, Delta, Delta-Delta)

D. Decision Cost function (DCF)

Number of trials used to detect the system performance of recognition task is known as DCF. The accuracy of the task is based on four conditional parameters. They are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The DCF is defined as

$$C_{DET} = C_{mi\ ss} P_{tar} P_{mi\ ss} + C_{fa}\left(1 - P_{tar}\right) P_{fa} \tag{13}$$

Where, C_{miss} represents the cost of miss detection, C_{fa} represents the false alarm cost, P_{tar} represents the target speakers probability, P_{miss} represents the miss probability and the value of P_{miss} and P_{fa} is calculated as follows.

$$P_{\text{miss}} = \frac{FN}{TP+FN} \tag{14}$$

$$P_{fa} = \frac{FP}{FP+TN} \tag{15}$$

True positive (TP) - number of samples that are correctly labeled as positive.

False positive (FP) - number of samples that are incorrectly represented as positive.

True negative (TN) - samples that are correctly labeled as negative.

False negative (FN) - number of samples that are incorrectly represented as negative.

The values shown in Figure 3 represent the DCF for the proposed DNN-RBF+BFO and some other existing methods. The methods that are taken for comparison are DNN-MLP, DNN-BP, and DNN-RBF. Among all these existing methods, our proposed method results better output than others.

Table 5. DCF comparison

Methods	DCF			
	LPC	Delta	Delta-Delta	MFCC
DNN-Multi layer perceptron	0.3629	0.4565	0.5127	0.3005
DNN-Back propagation	0.3005	0.3629	0.4565	0.2225
DNN- Radial basis function	0.1965	0.2225	0.3005	0.1758
DNN-RBF+BFO optimization	0.1588	0.1758	0.1965	0.1247

The DCF comparison is shown in Table 5 and Figure 3 and it contains the minimum value of 0.1247 for proposed method. This value is high for the traditional approaches such as DNN-MLP, DNN-BP, and DNN-RBF. The obtained DCF values for DNN-RBF+BFO are 0.1588, 0.1758, 0.1965, and 0.1247. This obtained value are found to be much better than the other three existing methods

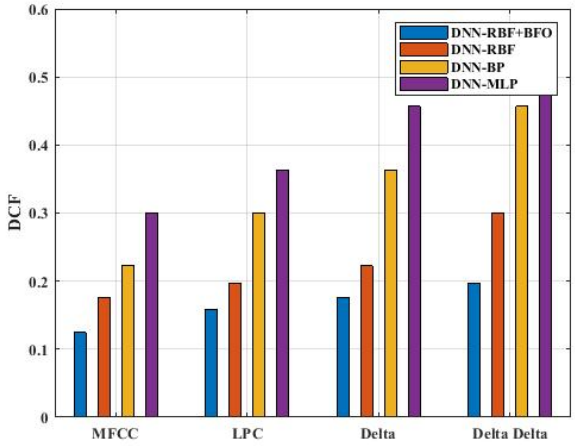

Fig 3. Comparison results for DCF vs. various features

E. Minimum average cost (C_{avg})

The minimum average cost is calculated as

$$C_{\text{avg}} = \frac{1}{N_{\text{tar}}} \sum_{l \in L_{w_w}} 0.5 P_{\text{miss}}(l) + 0.5 + \frac{1}{N_{\text{tar}} - 1} \sum_{r \in L_{\text{ren}}} P_{FA}(l, l') \tag{16}$$

Where, L_{tar} denotes the set of N target speakers, N_{tar} denotes the target speakers, P_{FA} denotes the probability of false alarm for target and non-target speaker pairs and L_{non} denotes the non-target speakers. This probability values are calculated for each trail of recognition task in order to detect the accuracy.

Table 6. Comparisonof C_{avg} values

Methods	C			
	LPC	Delta	Delta-Delta	MFCC
DNN-Multi layer perceptron	0.1173	0.1422	0.1837	0.1008
DNN-Back propagation	0.1008	0.11733	0.1422	0.8023
DNN- Radial basis function	0.0734	0.0802	0.1008	0.068
DNN-RBF+BFO optimization	0.0635	0.068	0.07342	0.0499

The C_{avg} comparison is presented in table 6 and Figure 4. The minimum average costs for the proposed approach with various features are 0.0635, 0.068, 0.07342 and 0.0499. The C_{avg} values for various existing methods like DNN-BP, DNN-RBF, DNN-MLP, and also for the proposed method DNN-RBF+BFO is specified here. The comparison results shows that better C_{avg} should be obtained by our proposed method.

The proposed method produces the lower value of C_{avg} for all approaches. The lower value of EER, DCF and C_{avg} shows the improved speaker recognition performance than the other existing methods. The results of this proposed performance factors provide better results for speaker recognition. The proposed DNN-RBF+BFO based MFCC achieve high accuracy and better performance than other methods.

F. Accuracy

It determines the system ability for the accurate detection of speaker. The expression for accuracy is shown in Equation. (17),

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

The accuracy of the methodology proposed was much higher than the three other approaches currently in use. As a result, the time for this proposed approach is popular. The precision and time-consuming results of this proposal and of the three other

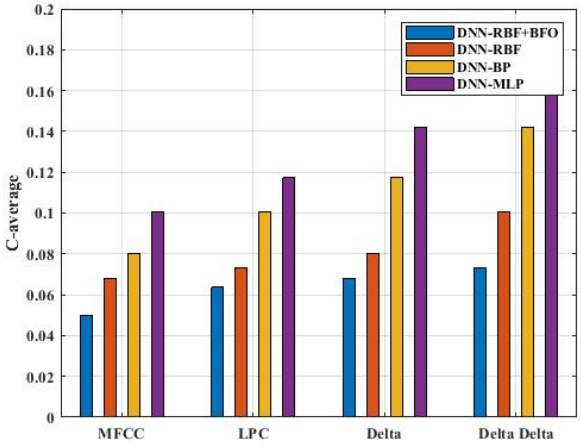

Fig 4. Comparative result for the C-average values of various features

existing methods are shown below and are shown in Table 7.

Table 7. Comparison results for the accuracy and execution time of proposed and existing methods

Methods	Accuracy (%)	Execution time
DNN-RBF+BFO	95.5	0.6165
DNN-RBF	93.205	2.6561
DNN-BP	92.10	4.3544
DNN-MLP	90.9	8.2882

Table 7 Comparison results for the accuracy and execution time of proposed and existing methods]

The below table 7 shows comparison results of execution time and accuracy. Compared with existing method, performance of proposed method is 2.295 times better. Figure 5 shows the comparison graph of execution time and accuracy of proposed method.

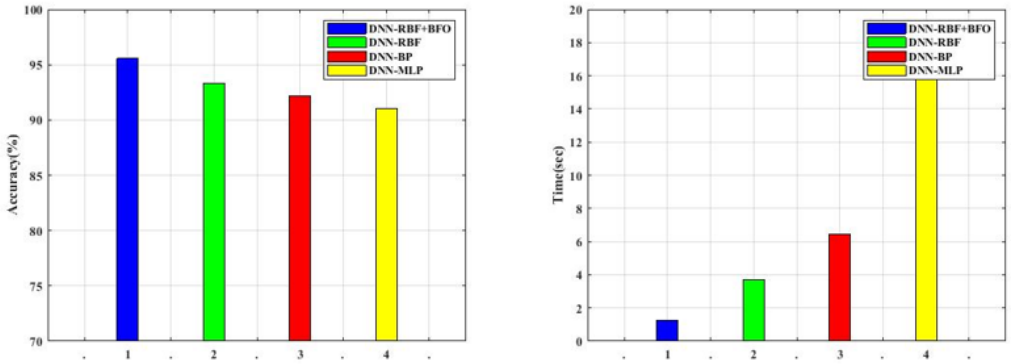

Fig 5. comparison results for both accuracy and execution time

The accuracy value of this proposed DNN-RBF+BFO for speaker recognition is found to be better than the other methods. The obtained values indicate that our proposed method higher performance due to the involvement of novel BFO algorithm.

5 Conclusion

This study suggests an MBFOB solution based on Mel-frequency Cepstral Coefficients and DNN-RBF with BFO, for the identification of speakers. The speech utterance from the TIMIT data corpus is preprocessed to obtain MFCC feature vectors Deep Neural Network-Radial Basis Function is used for the purpose of classifying the speaker and the feature vectors in the output layers are optimized with BFC. Finally, the scores for each speaker are calculated to identify the speaker. The execution time of this proposed method is found to be lesser than the other existing methods. The experimental findings are contrasted A novel MFCC-based Bacterial Foraging Optimization with DNN-RBF 0.1008 for identification of exact speaker is proposed in this article of paper Different output metrics like EER, DCF, Cavg and accuracy are used to test the proposed speaker recognition technique.

References

1) Borde P, Varpe A, Manza R, Yannawar P. Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. *International Journal of Speech Technology*. 2015;18(2):167–175. Available from: https://dx.doi.org/10.1007/s10772-014-9257-1.
2) Chougule SV, Chavan MS. Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition. *Procedia Computer Science*. 2015;58:272–279. Available from: https://dx.doi.org/10.1016/j.procs.2015.08.021.
3) Singer E, Reynolds DA. Domain Mismatch Compensation for Speaker Recognition Using a Library of Whiteners. *IEEE Signal Processing Letters*. 2015;22(11):2000–2003. Available from: https://dx.doi.org/10.1109/lsp.2015.2451591.
4) Richardson F, Reynolds D, Dehak N. Deep Neural Network Approaches to Speaker and Language Recognition. *IEEE Signal Processing Letters*. 2015;22(10):1671–1675. Available from: https://dx.doi.org/10.1109/lsp.2015.2420092.
5) Stafylakis T, Alam MJ, Kenny P. Text-Dependent Speaker Recognition With Random Digit Strings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2016;24(7):1194–1203. Available from: https://dx.doi.org/10.1109/taslp.2016.2546458.
6) Visalakshi R, Dhanalakshmi P, Palanivel S. Analysis of Throat Microphone Using MFCC Features for Speaker Recognition. In: Computational Intelligence, Cyber Security and Computational Models. Springer Singapore. 2016;p. 35–41. Available from: https://doi.org/10.1007/978-981-10-0251-9_5.
7) Kim C, Stern RM. Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012;24(7):1315–1329. Available from: https://doi.org/10.1109/TASLP.2016.2545928.
8) Mannepalli K, Sastry PN, Suman M. MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology*. 2016;19(1):87–93. Available from: https://dx.doi.org/10.1007/s10772-015-9328-y.
9) Jia F, Lei Y, Lin J, Zhou X, Lu N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mechanical Systems and Signal Processing*. 2016;72-73:303–315. Available from: https://dx.doi.org/10.1016/j.ymssp.2015.10.025.
10) Zeinali H, Sameti H, Burget L. Text-dependent speaker verification based on i-vectors, Neural Networks and Hidden Markov Models. *Computer Speech & Language*. 2017;46:53–71. Available from: https://dx.doi.org/10.1016/j.csl.2017.04.005.
11) Wang JC, Wang CY, Chin YH, Liu YT, Chen ET, Chang PC. Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. *Multimedia Tools and Applications*. 2017;76(3):4055–4068. Available from: https://dx.doi.org/10.1007/s11042-016-3335-0.
12) Vincent E, Watanabe S, Nugraha AA, Barker J, Marxer R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*. 2017;46:535–557. Available from: https://dx.doi.org/10.1016/j.csl.2016.11.005.
13) Ghahabi O, Hernando J. Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2017;25(4):807–817. Available from: https://dx.doi.org/10.1109/taslp.2017.2661705.
14) Liu Z, Wu Z, Li T, Li J, Shen C. GMM and CNN Hybrid Method for Short Utterance Speaker Recognition. *IEEE Transactions on Industrial Informatics*. 2018;14(7):3244–3252. Available from: https://dx.doi.org/10.1109/tii.2018.2799928.
15) Cai Z, Gu J, Wen C, Zhao D, Huang C, Huang H, et al. An Intelligent Parkinson's Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach. *Computational and Mathematical Methods in Medicine*. 2018;2018:1–24. Available from: https://dx.doi.org/10.1155/2018/2396952.
16) Zhang C, Koishida K, Hansen JHL. Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2018;26(9):1633–1644. Available from: https://dx.doi.org/10.1109/taslp.2018.2831456.
17) Subhashin PSP, Ram MSS, Rao DS. Bacterial Foraging Optimized Parameters for ANN using Adaptive Harris Hawks Weight Optimization. 2021.