Journal of Advanced Research in Technology and Management Sciences
Volume: 03,Issue: 02, ISSN: 2582-3078 March-April 2021
Available online at: http://www.jartms.org

DNN-RBF & AHHO for Speaker Recognition using MFCC

P S Subhashini Pedalanka[@] Dr M. Satya Sai Ram[$] Dr Duggirala Sreenivasa Rao[*]
@ Associate Professor, Department of E.C.E, R.V.R& JC College of Engineering, Guntur.
@ Research scholar, Department of E.C.E, JNTUH, Hyderabad.
$ Professor, Department of E.C.E, R.V.R& JC College of Engineering, Guntur.
* Professor, Department of E.C.E, JNTUH, Hyderabad.

Abstract: Speaker Recognition is essential in the field of authentication, and surveillance to validate the user's identity using extracted feature characteristics of audio speech signal. In this work, the speaker recognitions performed by deep neural network-Radial Basis Function (DNN-RBF). Initially, the available speech signals are preprocessed to remove the noise from the input signal. The noise removal in the input signal is performed by wiener filter. From this pre-processed signal Mel frequency cepstral coefficients (MFCC) features are extracted. The i-vector is estimated from the Gaussian Mixture Model (GMM) super vector in which the dimensionality of extracted features is reduced.Extracted i-vector features are then injected within classifier for recognizing the specific speaker. Based on these extracted features, the speakers are recognized by Adaptive Harris Hawk Optimization (AHHO) based DNN-RBFin an appropriate manner. The performance of this speaker recognition process is evaluated with TIMIT (Texas Instruments/Massachusetts Institute of Technology)dataset. Some of the performance metrics like precision, accuracy, and recall are evaluated to evaluate the effectiveness of this proposed technique.The proposed speaker recognition technique is evaluated with various performance measures such as EER, precision, recall, and accuracy. The accuracy, precision, and recall values attained by proposed AHHO based DNN-RBF is 94.92%, 89.87 and 94.67 respectively. The presence of adaptive optimization approach improves the performance of DNN-RBF in speaker recognition. The implementation process is performed in Mat lab platform.

Key words: Speaker Recognition, Adaptive Harris Hawk optimization, DNN-RBF, MFCC.

1. SPEAKER RECOGNITION

Speaker Recognition plays a major role in communication and surveillance area. This recognition process is evaluated by matching the training data with the test data. In recent years number of experiments was done to improve this recognition approach. The existing approaches outcomes include some trouble causing factors they are linear channel distortion, reverberation, and additive noise. The classifier modeling and feature extraction of audio are the two important components in speaker recognition. The features that are used for speaker recognition system are fundamental and spectrum frequency histograms, linear prediction cepstral coefficients (LPCC), instantaneous spectra covariance matrix, averaged auto-correlation, and MFCC. Among all these, MFCC has a major significant in speaker recognition process.

2. FEATURE EXTRACTION

The most commonly applied feature for speaker recognition is MFCC. The obtained MFCC 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. The MFCC assumes the signal as stationary therefore it fails to accurately analyze the localized events. To avoid such issue the DWT feature extraction process is also included in this proposed work.The GMM introduced a number of methods for speaker recognition they are support vector machine (SVM), Joint Factor Analysis (JFA), GMM-Universal background model (GMM-UBM), and i-vector models.The data gathered by GMM is used to enhance i-vector performance in speaker recognition. The GMM used in speaker recognition includes speech signals probability density function and Gaussian components. The entire framework of i-vector is employed to obtain a better performance from the low dimensional speech sounds. It is identified as a predominant approach due to its extraordinary performance, condensed representation, and less computational complexity. I-vector models the speech sounds from an inconsistency subspace. This i-vector has two important demerits in real time application they are; first the fewer amounts of data rapidly reduce the robustness of i-vector. Secondly, an unwanted latency is introduced after ending the computation process.

Deep learning provides an enormous success in neural networks. Two feed-forward architectures most effectively used for speaker recognition are Convolutional Neural Networks (CNN), and DNNs. The suitable information from the raw data is extracted by DNN-RBF. Initially, the layer-by-layer unsupervised learning is proposed to train the DNN-RBF and it is finely tuned by the supervised learning algorithm. Finally, the DNN-RBF is trained to remove the unwanted features from the audio signal. The DNN-RBF output defines the arrangement to gather considerable amount of statistics for i-vector extraction. An AHHO algorithm is introduced to optimize the weight parameter of DNN-RBF, whereas this optimization may improve the recognition activity of DNN-RBF. It is a gradient-free and population-based optimization technique so it is applied for various optimization issues. The major tactic of this HHO algorithm is “seven kills” strategy, which is normally defined as “surprise pounce”. It is mainly inspired by the cooperative behavior that is exhibited by the most intelligent Harris’ Hawks birds while hunting the escaping preys.

3. PROPOSED METHODOLOGY

The trained speech of particular person's voice is translated by this recognition method to identify the particular speaker. It is essential to obtain the speaker identity for some security purpose, so this method also authenticate or validate the speaker's identity. The speaker recognition approach mainly depends on both feature extraction and speaker classification process. The overall architecture of the proposed work is shown in Figure. 1.

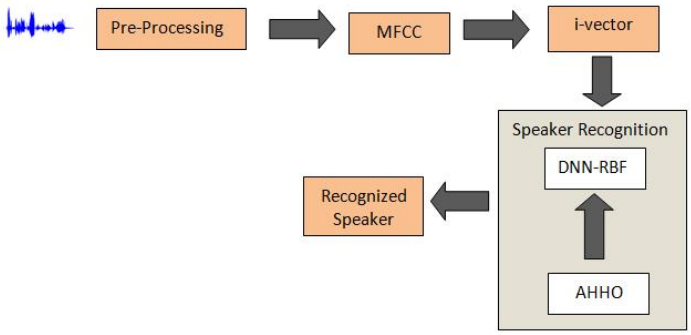

Figure 1. Workflow of DNN-RBF-AHHO-MFC Speaker Recognition System

The entire workflow of this method is demonstrated in Figure.1. Here, initially the audio signal is preprocessed to remove the noise and also for feature extraction. The weiner filter is introduced here for noise removal. The MFCC features are extracted during the pre-processing stage. The i-vector is extracted from the MFCC coefficients. The i-vector feature is then given to DNN-RBF based AHHO for speaker recognition.

3.1 MFCC (Mel frequency cepstral coefficients) feature:

Mel Frequency Cepstral Coefficients (MFCCs) are audio representation coefficients. The MFCCs analysis is derived from cepsfal representation of audio data.The difference between mel frequency cepstrum and standard cepstrum is that the MFCC frequency bands axe positioned logarithmically. The coefficients generated by algorithm are fine representation of signal spectra with great data compression (test application generates 13 elements vectors of MFCCs).

Speech feature extraction is to reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. Here MFCC feature extraction method was implemented to extract features from speech signal, which consist of the following step.

- Pre-Emphasizing
- Framing and Windowing
- FFT
- Mel-Scaled Filter Bank
- Logatithm
- DCT

The extracted MFCC features are applied as input for DNN-RBF and weights are updated using Adaptive Harris Hawk Optimization (AHHO) for classification of speaker.

4. RESULTS AND DISCUSSION

The proposed method for speaker recognition is evaluated with TIMIT dataset. The implementation is carried out in Matlab platform. The performance measures such as EER, DCF, Precision, recall and accuracy are evaluated and the experimental outcomes are compared with prevailing methods.

4.1 Dataset description

TIMIT Corpus

The TIMIT corpus of read speech is used for the proposed speaker recognition task. Total of 6300 sentences are included in this dataset, this 6300 sentences are spoken by 630 speakers (10 sentences each). These speakers were selected from the 8 important dialect regions of US (United States). The TIMIT corpus contains a 16-bit, 16 kHz speech waveform file for each utterance. This whole database is classified as training (70%) and testing (30%) files. From each speaker 7 audio files are used for training and 3 files are used for testing.

4.2 Evaluation metrics

A. Accuracy

It determines the system ability for the accurate detection of speaker.

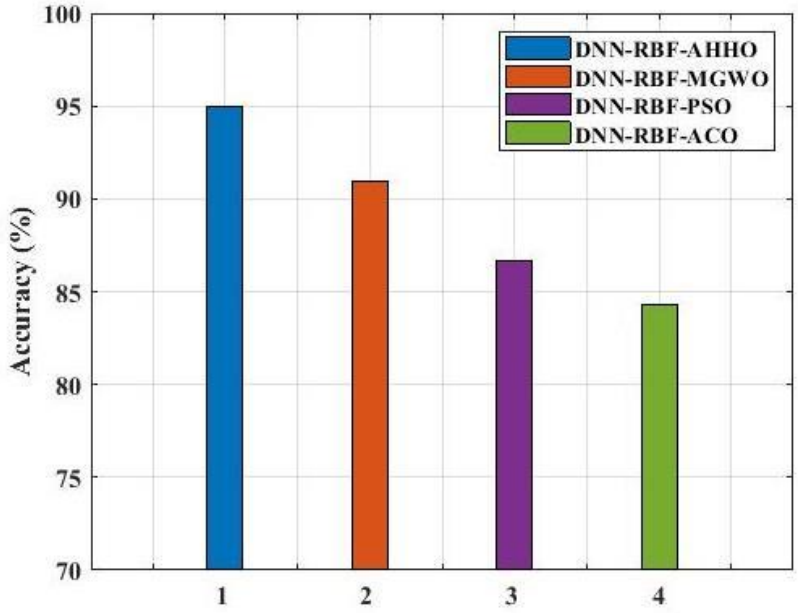

Figure 2. Graphical representation for accuracy Recognition based result

The recognition based accuracy shown in figure 2. The accuracy of this proposed method DNN-RBF-AHHO-MFCC is found to be much better than the other three existing algorithms. Due to this, reduced execution time is obtained for this proposed DNN-RBF-AHHO based method.

B. Precision

The fraction of recognized features which are more relevant at TP rates provide the precision value.

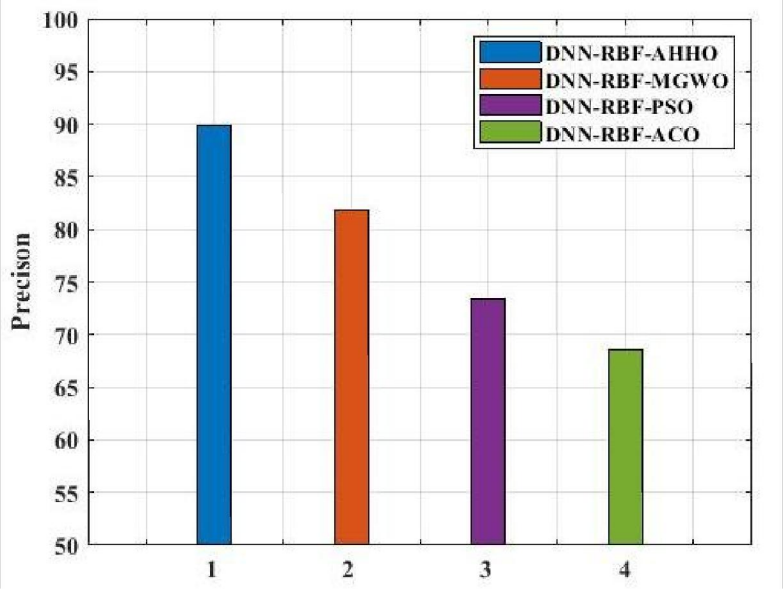

Figure 3. Graphical representation for precision Recognition based result

C. Recall

Recall is determined in-terms of feature classification recognized at both FN and TP predictions.

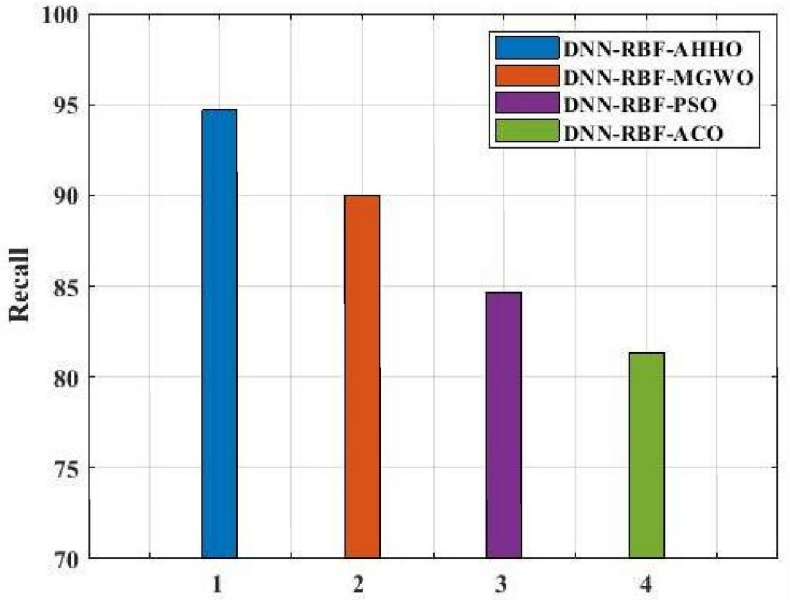

Figure 4. Recall of proposed algorithm Recognition based result

The graphical representations for the recall of proposed and existing techniques are shown in Figure. (4). the recall results of proposed is found to be much better than the other existing methods.

The accuracy, precision, and recall outcomes of this proposed and other three existing algorithms are shown in table. 3 and its comparison results are depicted in Figures. (2, 3, & 4).

Table3. Accuracy, Precision, recall of proposed and existing approaches

Method	Accuracy (%)	Precision (%)	Recall (%)
DNN-RBF-AHHO-MFCC (proposed)	94.92	89.87	94.67
DNN-RBF-MGWO-MFCC	90.93	81.82	90
DNN-RBF-PSO-MFCC	86.71	73.41	84.67
DNN-RBF-ACO-MFCC	84.27	68.54	81.33

The performance metrics of proposed AHHO is compared with other met heuristic optimization based deep earning approach and its results are given in table 3. The accuracy of proposed optimization based neural network is compared with various other optimization techniques. However, the accuracy attained by proposed adaptive algorithm is found higher than the other optimization algorithms. This is because the fuzzy logic combined with HHO

attains a crisp output during optima weight parameter selection. Formation of crisp output further improves the recognition performance deep learning approach.

D. Equal error rate (EER)

During simulation, the EER is measured based on the accuracy. EER is the measure of false acceptance rate (FAR) to the false rejection rate (FRR).This parameter evaluates the number of incorrect acceptance and number of incorrect rejection. Here, the EER comparison is performed with various other optimization techniques. The comparison results of EER for different optimization algorithms.This outcome indicates that MFCC with DNN-RBF-AHHO provides better recognition.

5. CONCLUSION

Highly analyzed area in speech processing field is speaker recognition. It has various applications like intelligent voice-identification applications like answering machines, telephone banking, and forensic science. In this work, DNN-RBF based AHHO approach is proposed for speaker recognition which highly depends on i-vector extraction and DNN-RBF with AHHO. The speech utterance from the TIMIT dataset is preprocessed to obtain MFCC feature vectors. The i-vector is extracted. DNN-RBF is used for the purpose of classifying the speaker and the feature vectors in the output layers are optimized with AHHO. In this method, theTIMITdatasets are taken into consideration for speaker recognition.The proposed approach attains 94.92% accuracy which is found 3.99% higher than the existing optimization based deep learning approach. The evaluation metrics of this proposed method is found to be higher than the other existing methods.

REFERENCES

[1] Kim, C and Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 24(7), 1315-1329 (2016).
[2] Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J. and Marxer, R.: An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Computer Speech & Language. 46, 535-557 (2017).
[3] Mannepalli, K., Sastry, P.N. and Suman, M.: MFCC-GMM based accent recognition system for Telugu speech signals. International Journal of Speech Technology. 19(1), 87-93 (2016).
[4] Wang, K., An, N., Li, B.N., Zhang, Y and Li, L.: Speech emotion recognition using fourier parameters. IEEE Transactions on Affective Computing. 6(1), 69-75 (2015).
[5] Borde, P., Varpe, A., Manza, R and Yannawar, P.: Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. International journal of speech technology. 18(2), 167-175 (2015).
[6] Singer, E and Reynolds, D.A.: Domain mismatch compensation for SR using a library of whiteners. IEEE Signal Processing Letters. 22(11), 2000-2003 (2015).
[7] Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P.J. and Gonzalez-Rodriguez, J.: Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks. 64, 49-58 (2015).
[8] Cumani, S., Laface, P., Cumani, S. and Laface, P.: Nonlinear i-vector transformations for PLDA-based SR. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 25(4), 908-919 (2017).

[9] Miao, Y., Zhang, H. and Metze, F.: Speaker adaptive training of deep neural network acoustic models using i-vectors. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 23(11), 1938-1949 (2015).
[10] Liu, Z., Wu, Z., Li, T., Li, J. and Shen, C.: GMM and CNN hybrid method for short utterance SR. IEEE Transactions on Industrial Informatics. 14(7), 3244-3252 (2018).
[11] Heidari, A.A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M. and Chen, H.: Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems. 97, 849-872 (2019).
[12] Du, P., Wang, J., Hao, Y., Niu, T. and Yang, W.: A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily PM2. 5 and PM10 forecasting. arXiv preprint arXiv:1905.13550,2019.
[13] Fayek, H.M., Lech, M and Cavedon, L.: Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks. 92, 60-68 (2017).
[14] Jia, F., Lei, Y., Lin, J., Zhou, X and 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. 72, 303-315 (2016).
[15] Zeinali, H., Sameti, H. and Burget, L.: Text-dependent speaker verification based on i-vectors, Neural Networks and Hidden Markov Models. Computer Speech & Language. 46, 53-71 (2017).
[16] Wang, J.C., Wang, C.Y., Chin, Y.H., Liu, Y.T., Chen, E.T and Chang, P.C. Spectral-temporal receptive fields and MFCC balanced feature extraction for robust SR. Multimedia Tools and Applications. 76(3), 4055-4068 (2017).
[17] Visalakshi, R., Dhanalakshmi, P and Palanivel, S.: Analysis of throat microphone using MFCC features for SR. In Computational Intelligence, Cyber Security and Computational Models, Springer, Singapore.35-41 (2016).
[18] Yu, C., Ogawa, A., Delcroix, M., Yoshioka, T., Nakatani, T and Hansen, J.H.: Robust i-vector extraction for neural network adaptation in noisy environment.2016