Triumvirate Features and Adaptive Harris Hawk Optimization Based Speaker Recognition using RBFNN

P S Subhashini Pedalanka[1] Dr. M. Satya Sairam[2] Dr. Duggirala Sreenivasa Rao[3]

[1]*Associate Professor, Department of E.C.E, R.V.R& JC College of Engineering, Guntur.*

[1] *Research scholar, Department of E.C.E, JNTUH, Hyderabad.*

[2] *Professor, Department of E.C.E,R.V.R& JC College of Engineering, Guntur.*

[3] *Professor, Department of E.C.E, JNTUH, Hyderabad.*

[1]pssubhashini.pedalanka@gmail.com, [2] msatyasairam1981@gmail.com,[3]dsraoece@gmail.com,

Abstract: Speaker Recognition is necessary in the field of authentication, and surveillance to validate the user's identity utilizing extracted feature characteristics of audio speech signal. In this work, the speaker recognition is 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, the wavelet, Mel frequency cepstral coefficients (MFCC), and Gammatone frequency cepstral coefficients (GFCC) features are extracted. The Triumvirate features are estimated from the Gaussian Mixture Model (GMM) super vector in which the dimensionality of extracted features is reduced. Extracted Triumvirate 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-RBF in an appropriate manner. The performance of this speaker recognition process is evaluated with Voxceleb and TIMIT dataset. Some of the performance metrics like precision, accuracy, and recall are evaluated to evaluate the effectiveness of this proposed technique. Meanwhile, Decision Cost Function (DCF) and Equal Error Rate (EER) for Voxceleb dataset is also evaluated in this method. The EER and DCF of Voxceleb is compared with EER and DCF of TIMIT speaker recognition dataset. The proposed speaker recognition technique is evaluated with different performance measures like EER, DCF, accuracy, recall and precision. The accuracy, EER, DCF, precision, and recall values attained by proposed AHHO based DNN-RBF is 97.4%, 0.8 (EER-Voxceleb), 1.25 (EER-TIMIT), 0.006 (DCF-Voxceleb), 0.0135 (DCF-TIMIT), 94.8%, and 96.3% respectively. The presence of adaptive optimization approach improves the performance of DNN-RBF in speaker recognition. The performance metrics of presented approach (DNN-RBF-AHHO) is correlated with some existing algorithms. The implementation process is performed in Matlab platform.

Key words: Speaker Recognition, Adaptive Harris Hawk optimization, DNN-RBF, MFCC, GFCC, and discrete wavelet transform (DWT).

1. Introduction

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 were done to improve this recognition approach. The existing approaches outcomes include some trouble causing factors they are additive noise, linear channel distortion and reverberation [1, 2]. The classifier modeling and feature extraction of audio are the two important components in speaker recognition [23]. The features that are used for speaker recognition system is LPCC (linear prediction cepstral coefficients), fundamental & spectrum frequency histograms, averaged auto-correlation, MFCC and instantaneous spectra covariance matrix [3]. Among all these MFCC has a major significant in speaker recognition process [4, 23].

More basically applied feature for speaker recognition is MFCC [24]. 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 [5, 22]. 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 SVM (support vector machine), JFA (Joint Factor Analysis), GMM-Universal background model (GMM-UBM), and i-vector models. [9, 10].The data gathered by GMM is used to enhance i-vector performance in speaker recognition [8, 21]. The GMM used in speaker recognition includes speech signals probability density function and Gaussian components [23]. 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 [6, 20]. This i-vector has two important demerits in real time application they are, first the less amount of data rapidly reduces the robustness of i-vector. Secondly, an unwanted latency is introduced after ending the computation process [7, 25].

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 [13]. 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 [14]. The DNN-RBF output defines the arrangement to gather considerable amount of statistics for Triumvirate feature extraction [21]. An AHHO algorithm is introduced for optimizing a DNN-RBF weight parameter, whereas this optimization may improve DNN-RBF recognition activity. 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 that is normally defined as "surprise pounce". It is mostly inspired by the cooperative behavior that is exhibited by the more intelligent Harris' Hawks birds while hunting the escaping preys (most cases rabbits are considered as prey) [11, 12].

The major contribution of the proposed deep learning based speaker recognitionarchitecture is:

A novel optimization algorithm is introduced to enhance the performance of neural network.This optimization enhance the performance of this entire speaker recognition process. It performs the optimization process by utilizing seven kills strategy, so that the best solution can be attained within less time, however the complexity also gets reduced.

The selection of random energy value in HHO delays the performance of HHO algorithm. Therefore, for improving HHO performance the fuzzy logic concept is hybrid with HHO for improving the optimal weight selection of HHO algorithm.

In recent works the amount of input data is reduced to improve the accuracy and perform number of iterations to identify the appropriate speaker. But in this approach, the issues regarding the accuracy is removed by injecting Triumvirate features based DNN-RBF-AHHO. The entire unpredictable covariance matrix for extracted features is determined by this Triumvirate feature vector in single iteration. This entire process does not require any word hypotheses or transcriptions.

The proposed methods outline is as follows. The related work for speaker recognition is discussed in Section 2. Section 3 designates the presented scheme. The evaluation for results and performance of proposed speaker recognition approach is discussed in Section 4. Finally, the conclusion for the presented work is discussed in Section 5.

2. Literature Survey

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. Sub-field of this speaker recognition is speaker identification, the problems identified in this speaker identification field is removed by introducing most powerful MFCC feature of audio signal in Sengupta et al, 2019 [15]. The co-occurrence matrices are applied, and from this matrices the statistical measures were derived. The derived measures were then incorporated within the feature vector. Finally, the classifier was applied to recognize the speaker accurately by utilizing the sample speech. It shows higher recognition output, and it was the only method that includes the co-occurrence matrices to derive the statistical measure.

Omid Ghahabi, et al. 2017, [16], proposed a Deep Learning 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 as a backend. Two methods, adaptation process for universal model and impostor selection algorithm are included in DNN and DBN (Deep Belief Networks) 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. This defect decreases the speaker recognition performance using i-vector.

Weighted-Correlation PCA (WCR-PCA) was introduced by Ahmed et al, 2019 [17], to perform the speech feature transformation in an effective manner in speaker recognition field. The RNN (recurrent NN) was introduced in this technique to accomplish the weighted RNN process. Here, the log likelihood values (i.e. weights) were selected from the fitted SGBM (Single Gaussian-Background Model). Huge difference is obtained among the feature variances of speech features, as it allows the covariance based PCA on less optimal solution. The comparative study was carried out for speaker recognition by employing the weighted and un-weighted correlation & covariance based PCA. The MFCC and LPCC feature extraction process were enhanced by introducing the extensions. The NIST2010 and VoxCeleb1 dataset were applied in this work for performance testing.

Rohdin et al 2018 [18], introduced DNN based end-to-end speaker verification technique. These techniques shows efficient outcome for the short utterances of both text- independent and dependent tasks. Recently, the machine learning based speaker recognition techniques attains a huge demand. In this, end-to-end speaker verification system is introduced. The training process for this system is performed in end-to-end but in regularized form, thus it won't deviate much long from the primary system. Due to this, the over-fitting can be minimized as it retards the effectiveness of these end-to-end methods. It provides better outcomes for both short and long duration utterances than i-vector + PLDA baseline.

Themos Stafylakis, et.al, 2016, [19], 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. This database includes 300 speakers between the age groups 17-42 (i.e., 143 females and 157 males). 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.

Improving the ASR performance was considered as the most challenging and significant issue. Therefore, Devi and Thongam, 2019 [36] developed an efficient ASR technique. Before going for recognition, initially the speech signal was pre-processed using the LMS adaptive filter. From, the pre-processed signal some basic and essential features like MFCC, zero-crossing rate, energy, and auto-correlation function were extracted. Then, the dimensions of extracted features were minimized by a swallow swarm optimization algorithm. At last the recognition is performed by ensemble classification methods they are SVM, improved CNN and LSTM (Long-short term memory).

Presence of additive noise in input signal degrades recognition process accuracy. To remove such adaptive noise Alabbasi et al, 2020 [37] developed an algorithm by combining the extracted features. The input signal was initially pre-processed using a wavelet thresholding approach. Then, the features like GFCC and PNCC (power normalized cepstral coefficients) were extracted. Next, wrap the extracted features for improving the robustness of classifier. The features between the actual and claim speakers were matched using UBM-GMM model.

Devi et al, 2020 [38] introduced an ANN (artificial neural network) for ASR. It was developed for improving the recognition accuracy and to resolve the ASR issues. The MFCC features from the speech signal were highly exploited by this method for speaker recognition. Then, the dimensions of extracted features were minimized by a SOFM (Self-Organizing Feature Map) scheme. The speech recognition is accomplished finally by MLP (Multi-layer perceptron) with Bayesian regularization scheme. Real speech dataset was used in this method for training purpose.

Problem statement: There are some issues found in this speaker recognition process. In the existing techniques number of flaws are identified. Among them, one of the major issue identified in this field is accurate recognition. Existing techniques failed to recognize the speaker in accurate manner. Furthermore, maximum error rate is obtained during recognition process. Further, the existing deep learning approaches used for speaker recognition process fails to attain a desired accuracy rate which is mainly due to the random generation of weight metric. The learning process with random metric degrades the accuracy rate by increasing the system complexity. Therefore, an efficient and modified approach that reduces the system complexity by increasing the recognition accuracy is required. To avoid such defects, a novel optimization algorithm is introduced in this method which improves the recognition performance by minimizing the computational complexity. Large number of existing works includes MFCC alone for speaker recognition process, but here two additional features are included along with this MFCC. Based on these features, the speaker recognition process is performed in an effective manner.

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 presented work is illustrated in Figure. 1.

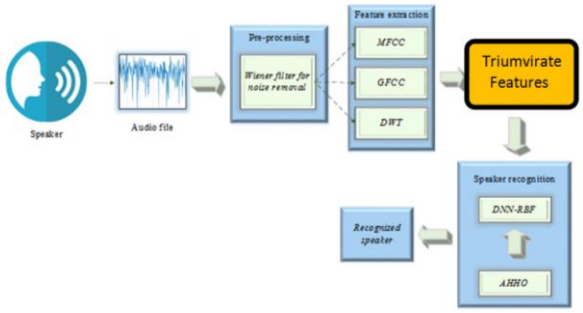


Figure 1. Workflow of proposed Speaker Recognition

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, GFCC, and DWT features are extracted during the pre-processing stage. These three features are extracted from speech signal because these three extraction techniques provide valuable features for speaker recognition. DWT extracts the signal features over the spectrum without considering about the dominant frequency band. The extracted features are then stacked into an array by introducing the Triumvirate Feature vector extraction process. The stacked feature is then given to DNN-RBF based AHHO for speaker recognition.

3.1 Pre-processing

3.1.1 Weiner filter: The unwanted noise found in the audio signal is removed by this Weiner filter [26, 27]. Here, the additive noise power spectrum is represented as

$S_n(a,b)$ and the power spectrum for speech signal is represented as $S_x(a,b)$. Normally, in the high frequency range the, $S_n(a,b)$ has a foremost effect over $S_x(a,b)$. Because, this $S_x(a,b)$ highly concentrates in low frequency spectrum range. The expression for weiner filter in Fourier domain is given in equation. (1),

$$W(a,b)=\frac{H^*(a,b)}{|H(a,b)|^2+S_{n,x}(u,v)} \quad (1)$$

Where, the noise-to-signal ratio is represented as $S_{n,x}(a,b)=S_n(a,b)/S_x(a,b)$. This ratio value is found larger in high frequency region, i.e. $S_{nx}(a,b)>>|H(a,b)|$. Meanwhile, the high frequency response of restoration filter gets suppressed. It is noteworthy, that if the noise is absent (i.e. $S_n(a,b)\to 0$) then the weiner filter will act as inverse filter.

$$W(a,b)\big|_{S_n(a,b)\to 0}=\begin{cases}\dfrac{1}{H(a,b)}, & H(a,b)\neq 0\\ 0 & H(a,b)=0\end{cases} \quad (2)$$

3.1.2 MFCC (Mel frequency cepstral coefficients) feature: This feature is normally applied for both speaker identification and speech recognition. The major aim of this MFCC method is to obtain the significant information from speech waveform by eliminating the redundant information. In frame-by-frame way this process is achieved, next each frame is transformed into a single N-dimensional feature vector [28]. In each frame the number of samples is greater than the taken N value. It provides the data needed for processing by back end system where it reduces the data quantity. The input of audio is transformed into a vector sequence in feature extraction $X=[x_1,x_2,...x_k]$, here k denotes frame index, x_k indicates N-dimensional vector.

The MFCCs can be determined using these following steps:

(i)First one is speech signal Pre-emphasizing.

(ii) The speech signal is separated into a larger number of frames with 20ms size & shift 10ms, after that hamming window is applied over these each frames.

$$P_k(a)=\frac{1}{n}|DFT_k(a)|^2 \qquad 1\le a\le A \quad (3)$$

Here, A indicates the length of DFT, the power spectrum of k^{th} hamming window is given as $P_k(a)$. n denotes sample number.

(iii) The magnetic spectrum is computed for each windowed frame by applying DFT. The k^{th} hamming window DFT is estimated as:

$$DFT_k(a)=\sum_{x=0}^{n-1}s_k(x)e^{\frac{-j2\pi ax}{n}} \quad (4)$$

Where, $s_k(x)$ indicates k^{th} hamming window time domain signal.

(iv) Mel spectrum computing is obtained through passing the DFT signal via Mel filter bank.

(v) DCT is applied to the coefficients of log Mel frequency for obtaining required MFCCs.

3.1.3 GFCC (Gammatone frequencycepstral coefficients) Feature: Recently, evolved GFCC shows promising recognition performance in the applications of speaker recognition, most particularly in noisy acoustical environment. In this method, the GFCC is extracted along with MFCC and DWT for enhancing the speaker recognition performance. The auditory features based on Gammatone Filter banks are defined as GFCC. It performs its process with cube root for more robustness, whereas these MFCCs may include log for processing. The process flow of this GFCC is same as that of MFCC, but it introduce Gammatone filter banks instead of Mel filter banks [28, 29].

$$g(f,t)=\begin{cases}pt^{n-1}e^{2\pi qt}\cos(2\pi f_{c_i}t+\phi) & t\ge 0\\ 0 & else\end{cases} \quad (5)$$

Where, the central frequency of Gammatone filter is indicated as f_c, time is indicated as t, the order and gain of the filter is represented as n and p respectively. Between the filter bank boundaries, the f_{c_i}, are equally spaced on the ERB (Equivalent Rectangular Bandwidth). Then the Gammatone features are obtained by performing the cubic root operation over the decimated output. Next, the DCT (Discrete Cosine Transform) is employed to obtain the GFCC features.

3.1.4 DWT (Discrete Wavelet Transform): DWT is a special form of WT which represents the signal in frequency and time domain in a compact manner to achieve effective computation. DWT has time-frequency localization property therefore it effectively analyze the sound signal. Wavelet based feature extraction techniques reduce the size of feature vectors and also it reduces the computational cost of several folds. DWT effectively analyze the non-stationary signals like speech. In practice, the DWT is figuredout bysuccessivelypassing the audio signal $x(n)$ via a low-pass and high-pass filter having impulse response $l(n)$ and $h(n)$ respectively [29]. The signal is convolved with filter's impulse response to perform the signal filtering process. In each decomposition level, the approximation coefficient a and detailed

coefficient d are produced by low and high-pass filters. The attained filter outputs are then down-sampled by 2. It contains single decomposition level and it is expressed as,

$$a_1(n) = \sum_{m=-\infty}^{\infty} x[m].l[2n-m] \quad (6)$$

$$d_1(n) = \sum_{m=-\infty}^{\infty} x[m].h[2n-m] \quad (7)$$

Here, the variables m and n indicates the discrete time coefficients. The a is then split into a_2 and d_2 respectively. This process gets repeated, until getting the desirable result.These three features are taken as the basic features for speaker recognition in this method. These features identified from the speech signal is stacked into an array by this Triumvirate Feature vector process. Among these three features, MFCC provide higher accuracy results in this speaker recognition field.

3.2 Triumvirate Feature vector: The identified features are then stacked into array for speaker recognition. For that, here we are applying the GMM based Triumvirate Feature vector extraction process. In speaker recognition field, Triumvirate Feature vector based on GMM is gaining a huge significance. Here, the speaker and channel information are compressed within the low-dimensional space basically referred as total variability space (TVS), and then each evaluated GMM super-vector is projected towards Triumvirate Feature vector (i.e.total factor feature vector). To perform the inter-session compensation, both Probabilistic Linear Discriminant Analysis (PLDA) and LDA are employed on Triumvirate Feature vector. The Triumvirate Feature vectors are applied to compress the large-dimensional feature within the small-dimension by retaining maximum amount of relevant information. The sub space is considered because the channel space that may not be utilized to differentiate between the speakers [30]. The total variability sub space training assumes the representation of utterance by the GMM super vector is given by,

$$M = m + Tw \quad (8)$$

Where M contains the session and speaker independent mean super vector m from the UBM model, T is the total variability space that is a low rank matrix indicates the variation of primary direction across development data collection, w represents the Triumvirate Feature vector representation which is normally distributed with the parameters $N(0,1)$. The extraction of Triumvirate Feature vector is depends on Baum Welch zero order, N & centralized first order F, statistics. For the given utterance, the statistics is computed with respect to C UBM components and F dimension of extracted features (MFCC, GFCC, and DWT). For the given utterance the Triumvirate Feature vector is extracted as follows.

$$w = (I + T^T \sum^{-1} NT)^{-1} T^T \sum^{-1} F \quad (9)$$

Where, N is a diagonal matrix with $F \times F$ blocks $N_c I(c = 1,2,...C)$, I denote the $CF \times CF$ identity matrix, F represents the super vector which is obtained by concatenating the centralized statistics of first order. The residual variability which is not captured by total variability T is represented as a covariance matrix $\sum$. The process of training the total variability T is similar to the training of JFA eigen-voice except one variation. JFA considers all session of a speaker to be the same person. For capturing the total variation, the total variability training considers all the speakers as the different person. Hence the total variability is utilized for capturing both channel and speaker variation.

3.3 DNN-RBF and Adaptive Harris Hawk Optimization (AHHO) based classification: DNN-RBF is a popular former method which achieves better recognition performance in speaker recognition [33, 34]. A large number of hidden layers which must be linear or non-linear are included in this DNN-RBF this hidden layers represents the 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 enormous hidden layers. The relationship between both the input and the first hidden layer is given by,

$$a_1 = F(w_1 x + b_1) \quad (10)$$

Where W_1 and b_1 are the weight and bias matrix of the first layer and the activation function gaussian is denoted as F(.). It is a special case of logistics function and it can be defined as,

$$F(x) = e^{-x^2} \quad (11)$$

Here, x represents activation input. The mapping between the present and next hidden layer after getting the first hidden layer is given as,

$$a_1 = F(w_l a_{l-1} + b_l), l = 2,..., L \quad (12)$$

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

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

$$\hat{y} = G(a_L) \quad (13)$$

Where, $G(x) = \dfrac{1}{1 + e^{-\phi x}}$ indicates the softmax function and for the label y, the parameters of DNN-RBF is defined as shown in Equation. (14):

$$\theta^* = \arg\min\{C(y, \hat{y}; x, \theta) + \gamma R(w) + \eta\rho(A)\} \quad (14)$$

Where, $\theta = \{w_l, b_l, l = 1, 2, ..., L\}$ 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 $X = [x_1, ...x_i, ...x_n]$ and the output labels are $Y = [y_1, ...y_i, ...y_N]$ where, N indicates the total number of training samples. The cost function is denoted as,

$$C(Y, \hat{Y}; X, \theta) = \frac{-1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} [y_{i,j} \log \hat{y}_{i,j}] \quad (15)$$

Where, $\hat{Y} = [\hat{y}_1, ...\hat{y}_i, ...\hat{y}_N]$ 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,

$$R(w) = \sum_l \|w_l\|_F^2 \quad (16)$$

Where, $\rho(A)$ represents the penalty sparsity of the hidden layer output, $\|.\|_F^2$ is the Frobenius norm, η and γ denotes the controlling coefficients. Finally, the complexity of this recognition process is reduced by AHHO by identifying the optimal weight. This optimal identification process enhances the performance of entire process by enhancing accuracy.

3.3.1 **Harris Hawk Optimization (HHO):** An HHO algorithm is introduced for optimizing the DNN-RBF, weight parameter, 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. It is mainly inspired by the cooperative behavior that is exhibited by the most intelligent Harris' Hawk birds while hunting the absconding preys (rabbits are considered as prey). The prey exploration, surprise pounce and various attacking tricks inspire HHO algorithm to introduce both the exploration (searching) and exploitation (Hunting) phases [11].

a. Exploration phase: The Harris Hawk is taken as candidate solution (DNN-RBF weight parameter), in the same way the candidate which is found closer to local or prey minima is taken as better candidate solution. Based on these two conditions, it perch over few locations in random way to prey identification. If q equal chance is provided to the strategy of several perching then hawks can perch based on both prey and family members position that is depicted in equation (17), in addition this strategy of perching for $q < 0.5$ is modeled in following equation,

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t)| & q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3 (L_b + r_4 (U_b - L_b)) & q < 0.5 \end{cases} \quad (17)$$

For subsequent iteration, the hawk position vector is indicated as $X(t+1)$, $X_{rabbit}(t)$, indicates the location of prey (rabbit), the hawk current location is indicated as $X(t)$, the upper and lower bounds are represented as U_b, and L_b, the random numbers $r_1, r_2, r_3, r_4, and\ q$ are updated in each iteration process, $X_{rand}(t)$ represents the hawks which are chosen randomly from current population and X_m indicates current hawks population average position. The $X_m(t)$ is attained by applying the subsequent equation,

$$X_m(t) = \frac{1}{n} \sum_{i=1}^{N} X_i(t) \quad (18)$$

Where, the location of several hawk at iteration t is represented as $X_i(t)$, and total hawks are taken as n. Different ways are available for average location identification, but here, only a simplest rule is applied.

b. Switch from exploration to exploitation

Exploration phase is transferred into exploitation phase, based on the preys escaping energy. However performing the process of escape, the prey might lose its energy significantly. Due to this fact, the energy is modeled as:

$$E = 2E_0 \left(1 - \frac{t}{T}\right) \quad (19)$$

Where, energy of prey's escaping is indicated as E, E_0 represents the initial energy, maximum number of iterations is represented as T. Within a interval (-1, 1) the E_0 represents the variations in each iteration in a random way. If E_0 gets minimized from 0 to -1 then the rabbit is founded physically flagging. If the E_0 gets increased from 0 to 1 then the rabbit might accomplish more strength. The exploration phase would occur when |E| ≥1, and exploitation phase will happened when |E| <1.

c. Exploitation phase

The surprise pounce might take place here this phase is executed by performing attack over the intended prey which is obtained from previous phase. Basing on chasing and escaping behaviors four different strategies are introduced in HHO for illustrating attacking phase.

i. Soft besiege

The prey sustains its energy till |E| ≥ 0.5 and r ≥ 0.5, during this interval the prey performs some misleading jumps in random manner to escape from that place, but it cannot. At that time, these Harris hawks slowly encircles the prey and makes the prey exhausted, after that it

performs the surprise pounce. The equation for encircling behavior is shown below:

$$X(t+1) = \Delta X(t) - E\left|IX_{rabbit}(t) - X(t)\right| \quad (20)$$

$$\Delta X(t) = X_{rabbit}(t) - X(t) \quad (21)$$

Here, the difference among preys position vector & present location in iteration t provides $X(t)$c, the random number within (0, 1) is represented as r_5 and $I = 2(1-r_5)$ represents the strength of the prey to perform the random jump throughout the escaping process. The J value randomly alters in each iteration for simulating rabbit motions nature.

ii. Hard besiege

The prey is found exhausted and tired during r ≥0.5 and |E| <0.5. At that time, the hawk encircle the prey and performing surprise pounce. During this hard besiege, the equation. (22), is applied to update its present location.

$$X(t+1) = X_{rabbit}(t) - E\left|\Delta X(t)\right| \quad (22)$$

iii. Soft besiege with progressive rapid dives

Enough energy is retained in prey at |E| ≥0.5 but r <0.5, during this situation the prey tries to escape, therefore still the soft besiege is performing before starting the surprise pounce. This process is found most intelligent than the former condition.

The concept of levy flight (LF) is included in this HHO to model the mathematical expression for the escaping patterns of leapfrog and prey movements. The hawks can perform its best pounce during the right time to catch the prey. The expression that is applied to evaluate this desired pounce of hawk in soft besiege is shown below

$$x = X_{rabbit}(t) - E\left|IX_{rabbit}(t) - X(t)\right| \quad (23)$$

The possible results of this particular movement are compared with the previous dives to identify whether it is best or worst dive. If the best candidate is performing the worst dive, then the remaining also perform rapid, inappropriate and abrupt dives when reaching near to the prey.

iv. Hard besiege with progressive rapid dives

The energy gets reduced for prey when both $|E|, r$ are less than 0.5. Then the hard besiege is developed before performing the surprise ponce for prey catching and killing. At this step, the prey energy is same as that the energy maintained in the soft besiege phase, but here the hawk may minimize its average location to catch the prey. The hard besiege is carried out based on the below condition:

$$X(t+1) = \begin{cases} x & if\ F(x) < F(X(t)) \\ y & if\ F(y) < F(X(t)) \end{cases} \quad (24)$$

Here, new rule is then included to obtain x and y

$$x = X_{rabbit}(t) - E\left|IX_{rabbit}(t) - X_m(t)\right| \quad (25)$$

$$y = x + S \times LF(D) \quad (26)$$

This entire optimization process takes place based on two parameters they are $|E|$ and, r.

3.3.2 Training DNN-RBF using AHHO

In this section, the DNN-RBF training using an adaptive HHO is elaborated. The major goal of DNN-RBF-AHHO is speaker recognition, for such it uses the features that are extracted from speech signal. Training of DNN-RBF is achieved using an AHHO algorithm, which is developed by incorporating the fuzzy logic within the HHO to enhance HHO algorithm performance. The steps that are used for training the DNN-RBF is discussed subsequently,

Initialization: At first, initialize the weights of DNN-RBF in random manner, which is represented in equation (27)

$$y = \{y_1 y_2, .., y_d, .., y_\alpha\};\ 1 < d \le \alpha \quad (27)$$

Where, α represents the total weight.

Error estimation: Apply, the extracted features F_s and weight y to DNN-RBF for recognizing the speaker from the output layer. The squares of present output obtained from the network and the training label output used for network training are summed for obtaining the final output error. The output error is represented in equation (28),

$$Er^{e+1} = \frac{1}{d_s}\sum_{x=1}^{d_s}\left[O_x^e - X_x^e\right] \quad (28)$$

Where, the predicted output is represented as, X_x^e, the total number of available data samples are represented as d_s, and the output that is estimated at the current iteration is represented as O_x^e.

Incremental learning: After identifying the best weight for one instance then input the new instance, for that instance compute the error and update the weight value. This is the major role of incremental learning process.

Update weight using AHHO: Determine the updated weight using AHHO algorithm, which is determined based on the equation (29),

$$X(t+1) = \begin{cases} X_k(t) - r_1\left|X_k(t) - 2r_2 X(t)\right| & q \ge 0.5 \\ \left(X_p(t) - X_m(t)\right) - r_3(L_b + r_4(U_b - L_b)) & q < 0.5 \end{cases} \quad (29)$$

While performing the process of escape, the prey might lose its energy considerably. Due to this fact, the energy is modeled as:

(30)

Where, energy of prey's escaping is indicated as E, E_0 represents the initial energy, maximum number of

iterations is represented as T . Within a interval (-1, 1) the E_0 represents the variations in each iteration in a random way. If E_0 gets minimized from 0 to -1 then the rabbit is founded physically flagging. If the E_0 gets increased from 0 to 1 then the rabbit might accomplish more strength. The exploration phase will occur when |E| ≥1, and exploitation phase will happened when |E| <1.

Based on the, r, the fitness (prey) will perform the escaping process, i.e. if (r <0.5), the prey would escape, or else if (r ≥0.5), the prey won't escape before performing the surprise pounce. Whatever the situation, the weight parameter can execute the hard or soft besiege for reaching the required fitness. Hard besiege happens, when |E| <0.5, and at |E| ≥0.5, then the soft besiege will occur.

This entire optimization process takes place based on two parameters they are $|E|$ and, r. But, the random selection of these two parameters may reduce the effectiveness of optimal weight selection process. Therefore here the Fuzzy logic is utilized to adopt the parameters dynamically for getting HHO better performance. An optimal $|E|$ and, r are identified by applying fuzzy logic which enhances the optimization process of HHO. The adaptive parameters $|E|$ and, r using fuzzy logic is defined as follows:

$$|E| = \frac{\sum_{k=1}^{R_{|E|}} \mu_k^{|E|}\left(|E_k|\right)}{\sum_{k=1}^{R_{|E|}} \mu_k^{|E|}} \quad (31)$$

$$r = \frac{\sum_{k=1}^{R_r} \mu_k^{r}\left(r_k\right)}{\sum_{k=1}^{R_r} \mu_k^{r}} \quad (32)$$

Where, the entire rules of fuzzy system is represented in $R_{|E|}$ and R_r, the outcomes of rule k is represented as $|E|_k$ and r_k respectively [31]. The membership function in association with the rule k is represented.

$$E = 2E_0\left(1 - \frac{t}{T}\right)$$

The fuzzy rules are set based on the iteration of algorithm. By using this AHHO, an optimized weight parameter is determined which is very much useful for DNN-RBF based speaker recognition process.

4. Experimental Results and Discussion

The presented method for speaker recognition is evaluated with Voxceleb dataset. Some familiar information regarding Voxceleb datasets are provided in this section. But, the EER, and DCF of Voxceleb dataset is compared with TIMIT corpus dataset to show the effectiveness of Voxceleb dataset. The implementation is carried out in Matlab platform. The performance measures like 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 consist a 16-bit, 16 kHz speech waveform file for every 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.

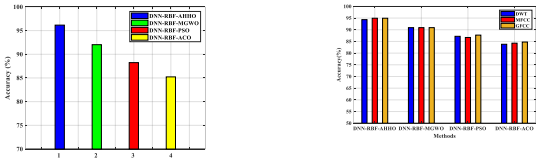
Voxceleb corpus: The Voxceleb dataset achieves high accuracy in speaker recognition process. This Voxceleb dataset is developed with large amount of 'real world' utterances for over 1000 celebrities. This utterances for Voxceleb dataset is collected from YouTube. Total of 100,000 utterances are included in this Voxceleb dataset, this 100,000 utterances for 1,251 celebrities are obtained from YouTube. In this dataset, total 55% of the speakers are male. This speakers included in this dataset are from various professions, accents, ages, and ethnicities. The development and test set does not include any overlapping identities.

The TIMIT is a noise-free dataset but Voxceleb is a noisy dataset, however we have included some noise in both TIMIT and Voxceleb datasets.

4.2 Evaluation metrics

A. Accuracy: It determines the system ability for the accurate detection of speaker. The expression for accuracy is shown in Equation. (33),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (33)$$



a) b)

Figure 2. Graphical representation for accuracy results, a) Recognition based result, and b) feature based result

The recognition based accuracy and feature extraction techniques based accuracy results are shown in figure 2 (a & b). The accuracy of this presented technique is determined 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. Similarly the feature based accuracy results are also shown in figure 2 (b). The accuracy, precision, and recall outcomes of this proposed and other three existing algorithms are shown in table. 1 and its comparison results are depicted in Figures. (2, 3 & 4).

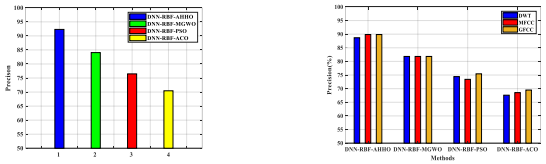
Table1. Accuracy, Precision, recall of proposed and existing approaches

Method	Accuracy (%)	Precision (%)	Recall (%)
DNN-RBF-AHHO (proposed)	97.4	94.8	96.3
DNN-RBF-MGWO	92.5	84.8	91.5
DNN-RBF-PSO	88	77	86.75
DNN-RBF-ACO	85.15	70.2	83.5

The performance metrics of proposed AHHO is compared with other metaheuristic optimization based deep earning approach and its results are given in table 3. The accuracy of presented optimization based neural network is compared with various other optimization techniques. However, the accuracy attained by presented 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. The three optimization algorithms taken for comparison are MGWO (Modified grey wolf optimization), PSO (Particle swarm optimization), and ACO (Ant colony optimization).

B. Precision: The fractions of recognized features that are more relevant at TP rates provide the precision value. The precision evaluation is performed by utilizing the equation (34),

$$\Pr ecision = \frac{TP}{TP + FP} \quad (34)$$



a) b)

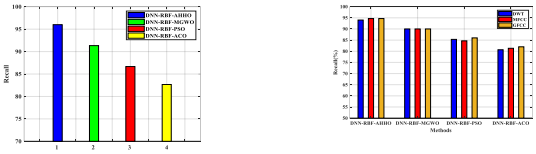
Figure 3. Graphical representation for precision results, a) Recognition based result, and b) feature based result

The effectiveness of precision due to the presence of feature extraction techniques and the precision of overall recognition performance is shown in figure 3 (a, & b). The graphical representation for the precision of prevailing and presented speaker recognition based algorithms are provided in Figure. 3 (a). The existing methods that are compared with this proposed DNN-RBF-AHHO are DNN-RBF-MGWO, DNN-RBF-PSO, and DNN-RBF-ACO. In this method the comparison is performed with different metaheuristic optimization algorithms. With the newly developed optimization algorithm high accuracy is achieved during speaker recognition.

- 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.

C. Recall: Recall is determined in-terms of feature classification recognized at both FN and TP predictions. The expression that is introduced to measure this recall value is given in equation. (35).

$$recall = \frac{TP}{TP + FN} \quad (35)$$



a) b)

Figure 4. Recall of proposed and existing algorithmsa) Recognition based result, and b) feature based result

The graphical representation for the recall of presented and existing techniques are shown in Figure. (4). The recall performance shown by feature extraction techniques and overall recognition process is separately shown in figure 4 (a, & b). The recall results of proposed is found to be much better than the other existing methods.

D. Decision Cost function (DCF)

Number of trials used to detect the system performance of recognition task is known as DCF. The DCF is defined as

$$C_{DET} = C_{miss} P_{tar} P_{miss} + C_{fa} (1 - P_{tar}) P_{fa} \quad (36)$$

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.

$$P_{miss} = \frac{FN}{TP + FN} \quad (37)$$

$$P_{fa} = \frac{FP}{FP + TN} \quad (38)$$

Table 2. DCF comparison

Methods	DCF		
	DWT	MFCC	GFCC
DNN-RBF-ACO	0.3005	0.3629	0.4565
DNN-RBF-PSO	0.2225	0.3005	0.3629
DNN-RBF-MGWO	0.1758	0.1965	0.2225
DNN-RBF-AHHO	0.1247	0.1588	0.1758

The DCF comparison is shown in table 4 and it contains the minimum value of 0.1247 for DWT. This value is high for the traditional approaches such as DNN-RBF-ACO, DNN-RBF-PSO, and DNN-RBF-MGWO. The obtained DCF values for DNN-RBF-AHHO are 0.1247, 0.1588, and 0.1758. This obtained values are found to be much better than the other three existing methods.

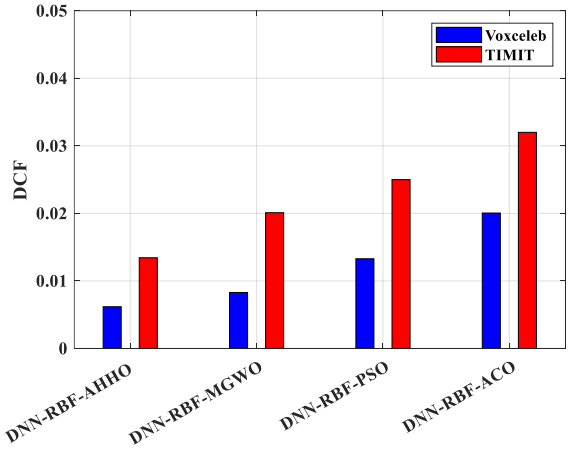


Figure 5: DCF of Voxceleb and TIMIT dataset

The DCF of proposed and existing algorithms using two different datasets (Voxceleb and TIMIT) are shown in figure 5. The DCF of proposed AHHO provides better result than other existing MGWO, PSO, and ACO algorithms.

E. Equal error rate (EER)

During simulation, the EER is measured based on the accuracy. EER is the measure of FAR (false acceptance rate) to the FRR (false rejection rate). Here, the parameters like FAR & FRR is evaluated as follows.

$$FAR = \frac{Number\ of\ false accep\tan ce}{Number of\ identificationattempt} \quad (42)$$

$$FRR = \frac{Number of\ false rejection}{Number of\ identificationattempt} \quad (43)$$

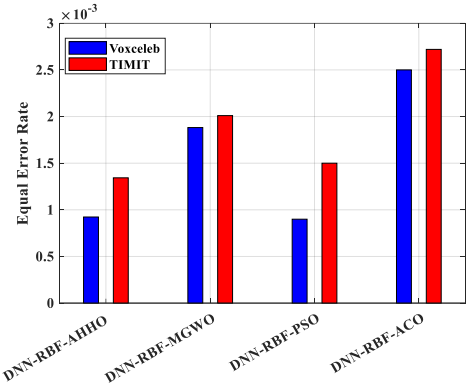


Figure 6: EER of Voxceleb and TIMIT dataset

Both these parameters evaluate the number of incorrect acceptance & number of incorrect rejection. Here, the EER, and DCF comparison are performed among Voxceleb and TIMIT dataset. The EER and DCF values of both TIMIT and Voxceleb are provided in above table. The comparison results of EER and DCF for different optimization algorithms are provided in Figure. (5 & 6).

Table 3. EER and DCF of TIMIT and Voxceleb dataset

Methods	DCF		EER	
	Voxceleb	TIMIT	Voxceleb	TIMIT
DNN-RBF-AHHO	0.006	0.0135	0.8	1.25
DNN-RBF-MGWO	0.0087	0.02	1.86	2
DNN-RBF-PSO	0.014	0.025	0.7	1.5
DNN-RBF-ACO	0.02	0.032	2.5	2.75

The DCF and EER value for these proposed and existing algorithms are tabulated in Table. 3. This outcome indicates that this Voxceleb dataset along with DNN-RBF-AHHO provides better recognition than the TIMIT based recognition.

Table 4. EER comparison for speaker identification approaches

Speaker identification approaches		30 ms	20ms	10 ms	3 ms
DWT	DNN-RBF-ACO	0.055	0.107	0.265	1
	DNN-RBF-PSO	0.0335	0.075	0.2007	0.7857
	DNN-RBF-MGWO	0.020	0.056	0.1625	0.6583
	DNN-RBF-AHHO	0.012	0.0436	0.1372	0.5743
MFCC	DNN-RBF-ACO	0.0722	0.1333	0.3166	1.1722
	DNN-RBF-PSO	0.055	0.1075	0.265	1
	DNN-RBF-MGWO	0.026	0.064	0.1794	0.7148
	DNN-RBF-AHHO	0.0162	0.0493	0.1487	0.6123
GFCC	DNN-RBF-ACO	0.0981	0.1721	0.394	1.4312
	DNN-RBF-PSO	0.0722	0.133	0.3166	1.1722
	DNN-RBF-MGWO	0.033	0.075	0.2007	0.7857
	DNN-RBF-AHHO	0.0208	0.0565	0.1625	0.65833

The EER rate of MFCC, GFCC, and DWT for 30ms, 20ms, 10ms, and 3ms are shown in table 6.Each features from different sizes with various optimization algorithms such as PSO, ACO, and MGWO, and AHHO. Among several feature extraction techniques, our proposed DWT provides better EER performance. The EER of this presented DNN-RBF-AHHO is found less than other optimization algorithms.

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 scheme is presented for speaker recognition which highly depends on Triumvirate Feature vector extraction and DNN-RBF with AHHO. The speech utterance from the Voxceleb dataset is preprocessed to obtain MFCC, GFCC, and DWT feature vectors and also for noise removal. The GMM super vector and Baum Welch statistics are calculated to extract the Triumvirate Feature vector. DNN-RBF is utilized for classifying feature vectors and speaker in output layers are optimized with AHHO. In this method, the Voxceleb datasets are taken into consideration for speaker recognition. The EER and DCF of both Voxceleb and TIMIT speaker recognition datasets are related. The EER attained for Voxceleb (0.8) is found less than the TIMIT (1.25) dataset. The proposed approach attains 97.4% accuracy which is found 4.9% higher than the existing optimization based deep learning approach. The presented DNN-RBF based AHHO performance is compared with three different deep learning based optimization algorithms they are MGWO, PSO and ACO. The evaluation metrics of this presented scheme is determined to be higher than the other schemes. The experimental outcomes are compared with some existing metaheuristic optimization based deep learning and it exhibits the efficiency of this presented scheme.

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] Sengupta, S., Yasmin, G. and Ghosal, A.: Speaker Recognition Using Occurrence Pattern of Speech Signal. In Recent Trends in Signal and Image Processing, Springer, Singap.207-216 (2019).

[16] Ghahabi, O and Hernando, J.: Deep learning backend for single and multisession i-vector speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 25(4), 807-817 (2017).

[17] Ahmed, A.I., Chiverton, J.P., Ndzi, D.L. and Becerra, V.M.: Speaker recognition using PCA-based feature

transformation. Speech Communication. 110, 33-46 (2019).

[18] Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P. and Burget, L.: End-to-end DNN Based Speaker Recognition Inspired by i-vector and PLDA. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.pp. 4874-4878 (2018, April)

[19] Stafylakis, T., Alam, M.J. and Kenny, P.: Text-dependent SR with random digit strings. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 24(7), 1194-1203 (2016).

[20] Zhang, C., Koishida, K and Hansen, J.H.: Text-independent speaker verification based on triplet convolutional neural network embeddings. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 26(9), 1633-1644 (2018).

[21] 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).

[22] Sahidullah, M and Kinnunen, T.: Local spectral variability features for speaker verification. Digital Signal Processing. 50, 1-11 (2016).

[23] 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).

[24] 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).

[25] Chougule, S.V. and Chavan, M.S.: Robust spectral features for automatic SR in mismatch condition. Procedia Computer Science. 58, 272-279 (2015).

[26] Makandar, A and Patrot, A.: Computation pre-processing techniques for image restoration. International Journal of Computer Applications. 113(4), 11-17 (2015).

[27] Siam, A.I., El-khobby, H.A., Elnaby, M.M.A., Abdelkader, H.S and El-Samie, F.E.A.: A Novel Speech Enhancement Method Using Fourier Series Decomposition and Spectral Subtraction for Robust Speaker Identification. Wireless Personal Communications. 1-14 (2019).

[28] Nguyen, S.T., Lai, V.D., Dam-Ba, Q., Nguyen-Xuan, A. and Pham, C.: Vietnamese Speaker Authentication Using Deep Models. In Proceedings of the Ninth International Symposium on Information and Communication Technology, ACM.177-184 (2018, December).

[29] Sekkate, S., Khalil, M and Adib, A.: A feature level fusion scheme for robust speaker identification. In International Conference on Big Data, Cloud and Applications, Springer, Cham. (289-300, 2018, April)

[30] 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

[31] Wu, D., Warwick, K., Ma, Z., Gasson, M.N., Burgess, J.G., Pan, S and Aziz, T.Z.: Prediction of Parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization. International journal of neural systems. 20(02), 109-116 (2010).

[32] Sreedharan, S and Eswaran, C.: Optimized Variable Size Windowing Based Speaker Verification. In Proceedings of the 2018 International Conference on Electronics and Electrical Engineering Technology, ACM.202-206 (2018, September)

[33] Richardson, F., Reynolds, D and Dehak, N.: A unified deep neural network for speaker and language recognition. arXiv preprint arXiv:1504.00923.2015.

[34] Han, H.G., Guo, Y.N and Qiao, J.F.: Nonlinear system modeling using a self-organizing recurrent radial basis function neural network. Applied Soft Computing. 71, 1105-1116 (2018).

[35] Saranya, M.S., Padmanabhan, R and Murthy, H.A.: Feature-switching: Dynamic feature selection for an i-vector based speaker verification system. Speech Communication. 93, 53-62 (2017).

[36] Devi, K.J. and Thongam, K., 2019. Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals. Journal of Ambient Intelligence and Humanized Computing, pp.1-14.

[37] Alabbasi, H.A., Jalil, A.M. and Hasan, F.S., 2020. Adaptive wavelet thresholding with robust hybrid features for text-independent speaker identification system. International Journal of Electrical & Computer Engineering (2088-8708), 10(5).

[38] Devi, K., Singh, N. and Thongam, K., 2020. Automatic speaker recognition from speech signals using self organizing featuremapand hybrid neural network. Microprocessors and Microsystems, p.103264.