Acoustic Feature Extraction and Optimized Neural Network Based Classification for Speaker Recognition

P. P. S. Subhashini, M. Satya Sai Ram, D. Sreenivasa Rao

Abstract*: *Identifying the person from his or her voice characteristics is an essential trait for human interaction. Automatic speaker recognition (ASR) systems are developed to find the identity of the speaker in the field of forensics, business interactions and law enforcement. It can be achieved by extracting prosodic, linguistic, and acoustic speech characteristics. Furthermore optimized neural network based approaches are reviewed to classify the extracted features. In this paper, literatures are surveyed on recognition of speaker through the neural network using an optimization algorithm that has developed from the previous years for ASR systems. We deliberate different characteristics of ASR arrangements, containing features, neural network based classification, performance metrics and standard evaluation data sets. ASR system is discussed in two parts. The first part illustrates different feature extraction techniques and the second part involves the classification approaches which identify the speaker. We accomplish this evaluation through a comparative analysis of various recognition of speaker approaches and compare the results of the same.

***Index Terms*: *Speaker recognition, feature extraction, classification, Deep neural network, optimization algorithm*.**

I. INTRODUCTION

Automatic recognition related to the speakers are the process of identifying speech signal with the corresponding speaker. It be able to be used as a biometric tool for personal authentication and recently it has become a high security concern [1]. The biometric approaches such as voice, face, and fingerprints can able to handle mismatched or noisy comparison. Transmission or sensor attacks are serious issues for this kind of biometric recognition system [2]. The identification system is divided into two types: speaker verification and identification. Speaker identification system detects the unknown speaker from the database whereas the verification system validates the identity [3]. In both systems, the speech signal of the speaker is taken into account for speaker recognition [4]. The speaker identification system can be commonly implemented by noticing the unvoiced or voiced components or it can be obtained by evaluating the energy distribution of speech. For that, the system is divided into two steps namely feature extraction and speaker classification [5]. It is essential to obtain sufficient performance in terms of accuracy [6]. For recognition process, probabilistic LDA method is used in several kinds of literature. The researchers have focused the speaker verification task with DNN and contribute there result to the classification process [7]. Comparison of standard UBM/i-vector framework with speaker recognition models like universal background model (UBM) and Gaussian mixture model (GMM) gives suitable result for speaker recognition [8]. Hence, some categories like speech frames are aligned softly with DNN/i-vector framework and it results in accurate Automatic speaker recognition (ASR). Speaker recognition can be stable while performing the DNN with frame by frame classification [9]. The posterior probabilities of different Signal to noise ratio (SNR) levels from i-vectors can compute by using the DNN [10]. The trained CNN are used to compute the frame posteriors for automatic speaker recognition [11]. The i-vector trained by two indirect methods: Bottle neck features and DNN posteriors. The BNF extracted frame-level features from DNN with special BN layer. The second one extracts the posterior from DNN to accumulate the multi-model statistics [12]. The utterance occurred from i-vectors offer closed representation of the corresponding speaker specific attributes through confining them to a low-dimensional space [13]. The most popular method for feature extraction is classical Mel-frequency cepstral co-efficients (MFCC) method. Another one method linear predictive coding (LPC) which is suitable for immensely aided text-dependent identification tasks. These two methods are globally used for speech analysis [14]. MFCC is also used for different applications like speech recognition, noise classification and speaker identification [15]. But recently, joint factor analysis is the basis for the Eigen voice (i-vector). Eigen voice adaptation is the operation towards evaluate i-vector this addresses a latent factor with low dimensional aimed at all classes in a corpus [16]. For improving the performance of fundamental recognition, the researcher focuses on the optimized neural network based approaches with combined feature extraction. Different approaches are used to characterize the speaker at the feature level [17]. Those approaches are voice source features, short-term spectral, spectro temporal features (segmented information, pitch and rhythm) and prosodic, high level features (lexical,, idiolect and phonetic) [18]. The phonetic and acoustic

Revised Manuscript Received on July 06, 2019.

P.P.S. Subhashini, Associate Professor, Dept. of ECE, RVR & JC College of Engg., Chowdavaram, Guntur-522019.

Dr. M. Satya Sai Ram, Associate Professor, Dept. of ECE, RVR & JC College of Engg., Chowdavaram, Guntur - 522019.

Dr. D. Sreenivasa Rao**,** Professor, Dept. of ECE, JNTUH, Hyderabad.

methods are results the corresponding data for speaker recognition. To overcome the limitations in speaker recognition performance, two methods are used such as the model based and feature based method. The model based method is used for severe noise and channel distortion. For noise robust features uses feature based method [19]. In automatic speaker recognition, errors can occur such as different speaking styles, feature variation; translational invariance may cause the result [20]. The main objective of this work is to analyze different feature extraction techniques and comparing the performance of optimized DNNs such as DE optimized DNN, ABC optimized DNN and ACO optimized DNN. The DNN is commonly used for speaker recognition which provides the best classification. The classification accuracy of the speaker is improved by optimizing the DNN parameters. Weight updation is the main strategy of the neural network and it is required to select the optimal weight for an accurate solution. It can be accomplished through an optimization algorithm which selects the global optimal solution for speaker recognition. The outline of the work is described as follows. Section 2 describes different kind of features and their extraction procedure. Section 3 discusses the DNN approaches and optimization approaches for improving the DNN performance. The experimental results of speaker recognition tasks are presented in section 4. Section 5 describes the significant aspects of our work and concludes.

II. FEATURE EXTRACTION

The feature extraction is used to extract the speaker based attributes from the speech signal. It consists of two parts namely preprocessing and cepstral feature vector extraction. The overall flow for feature extraction is shown in figure 1.

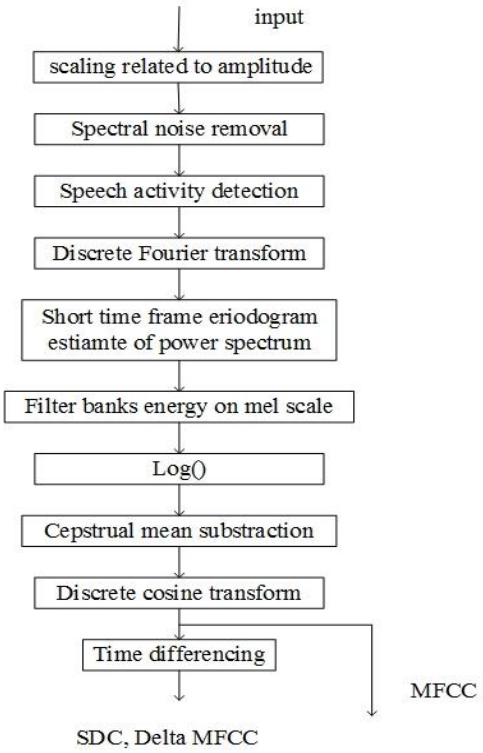

Figure 1: Overall process of feature extraction

Spectral Noise Removal

In the spectral domain, two pass effect is used to eliminate the background noise. An initial pass is implemented through the noise and this is obtained via signal via the speech. The FFT is computed for all frequency band in each windowed sample. The highest level is obtained by considering at least n sampling windows and the value of n is varied based on the signal. When removing the noise, the control related to gain for the frequency range is set in which the sound exceeded the previously obtained threshold. The control gain is used for the frequency is set to 0 and this gain is reduced with noise suppression. The method of smoothing frequency remains smeared in which the unique rate is not destroyed and it is enhanced through boosting. The unique frequency is not affected by the isolation procedure of frequency smoothing when applying time based smoothing. For the complex FFT signal, the gain controls are captured and the inverse is taken following the Hanning window. This kind of noise removal is applicable to noise reduction from the voice signal.

Speech activity detection based on energy

The activity detection can be accomplished based on the non-speech frames and the silent speech signal. The performance of the system is affected by the availability of non-speech frames. Energy based activity detection is mainly useful for the detection of Speech and speaker specific applications. The phoneme based method provides a better result for this detection but it can be avoided with the computational complexity of the algorithm. The energy base detection can be accomplished in this system is straightforward. Initially, from the given speech utterance, the energy of speech frames are computed and from each frame energies, the specific threshold is selected. The reference threshold is chosen with the factors $0.06 \times E_{avg}$, where E_{avg} denotes the frames with average energy in each utterance of speech. The defined threshold is based on the maximum available energy of each frame.

Cepstral Mean Subtraction

In order to achieve CMS, the mean M is computed for the feature vectors and the mean is subtracted from the feature vector. It is an efficient technique due to its simplicity. It is accomplished in the cepstrum domain and log spectrum domain due to the linear transformation of the log spectrum. The convolution for the utterance of the speech signal $X(N,K)$ is computed in the time domain. The speaker based properties $B(N,K)$ and a constant channel noise $C(N,K)$ are computed in the domain of DFT, then the model related to channel which is calculated as:

$$Y(N,K) = C(N,K).B(N,K).X(N,K) + G(N,K) \qquad (1)$$

This approach, the frame index is denoted as n and the frequency is indicated as k . The additive noise $G(N,K)$ is considered in log cepstral area otherwise spectral domain which could be computed for both non-speech and speech frames.

$$\log Y(N,K) = \log(C(N,K).B(N,K).X(N,K) + G(N,K)) \quad (2)$$

$$\log Y = \log(C.B.X + G) = \log(C.B.X) + \log(1 + G/C.B.X) \quad (3)$$

Thee log domain notations are expressed as follows.

$$Y = C + B + X + R;\ R = \log(1 + G/B.C..X) \quad (4)$$

For foremost signals, $G << X.B.C$

$$Y = C + B + X \quad (5)$$

In case of frames with non-speech, $X, B = 0$. Thus, additive noise N is $\log Y = \log G(N,K) => Y = N$. Currently received signal's mean 'Y' of the identical model could be described in both expressions subjective through the quantity of non-speech and speech frames.

$$M = \alpha * Yspeech + \beta * Ypause \quad (6)$$

$$M = \alpha * (Xavg + B + C + Ravg) + \beta * N \quad (7)$$

For speech utterances with long time signals, maximum SNR, we can eliminate *xavg* and *ravg* by using the estimate $M = \alpha(B + C) + \beta * N$. Nowadays, performing the CMS.

$$Z = Y - M \quad (8)$$

For the speech signal, the value is computed as $Y = X + \beta B - \beta C$

For the non-speech signal, it can be computed as: $Y = \alpha N - \alpha B - \alpha C$

Several frames with non-speech signals are avoided in SAD stage, the range of $\beta << \alpha$ then in frames with speech alone and *x* ruins as the substantial measure as predictable. After performing CMS, constant channel component of the speech frames is computed. In the case of frames with non-speech signals, here is a shift relevant to each medium. The greater variance in terms of β proportion remains applied for the conversational speech. The relevant word error rate is reduced around 4% for the speaker based CMS when compared with the conversational speech based CMS. The performance of CMS is higher than RASTA when applying channel normalization techniques. After applying RASTA filtering, the shape of the speech signal is avoided in the time domain by preserving CMS. It leads to the introduction of RASTA filter. Several kinds of features are extracted for the purpose of getting efficient feature extraction. The details of this feature extraction procedure are described as follows.

A. *Linear predictive coefficient (LPC) features*

There are three concepts involved in the extraction of LPC features. They are mainly based on an equal-loudness curve, critical band spectral resolution and power law based on intensity. The Mel filters which are used to extract MFCCs with critical bands. The variance is that Mel scale is replaced with Bark scale and trapezoidal like masking curves are used with triangular filters. The sensitivities related to non-linear characteristics of hearing level of human at varying frequencies are modelled with equal loudness curve models. In between the supposed intensity of sound, loudness and the intensity power law describes the nonlinear relationship. This nearly equals the logarithm of MFCC computation. The compression of the cubic root is utilized for PLPs. Based on these procedures; the autoregressive all-pole method is used to compute auditory spectrum.

B. *Delta and delta*

The feature extraction techniques like MFCC provides better estimates for finding the local spectra of statistical features but it fails to detect the specific aspects of human speech which is necessary for differentiating speakers. In addition to the static parameters, the time derivatives are added to improve the efficiency of the speech processing system. The local temporal derivatives are estimated from the speech cepstrum by using delta and delta-delta cepstra with a local slope and least square approximation are computed over several frames. The delta coefficients denote the first order derivatives and it can be computed based on the following regression procedure.

$$\Delta C_i(N) = \frac{\sum_{k=-G}^{G} kC_i(N+K)}{\sum_{k=-G}^{G} K^2} \quad (9)$$

Where, $\Delta C_i(N)$ denotes the coefficient of Delta are computed at the i^{th} cepstral stream C_i with nth frame and N is utilized to find the number of frames across which the delta cepstrum is computed. It can be varied based on the received voided signal but it is commonly set to 2 or 4. Similar to that, the delta-delta coefficients of second order derivatives such as are estimated by an identical formula for delta coefficients without using original cepstral coefficients. The past and future cepstral coefficients are computed with the cepstrum features with changing requirement of starting and the ending of the speech signspeech signal's starting and ending. The issue associated with the end effect of signals can be resolved with the help of simplified first order variations and the initialization of speech. The number of T frames can be implemented with

$$\Delta C_i(N) = C_i(N+1) - C_i(N), N < G \quad (10)$$

$$\Delta C_i(N) = C_i(N) - C_i(N-1), N \geq T - G \quad (11)$$

Similar to that, the relevant delta-delta coefficients are computed accordingly. These features are reliable for artifacts of channel with the usage of delta-delta and delta cepstral. Through assuming the value $N = 1$, the delta and delta-delta coefficients are constructed. The denominator used in the equation is eliminated with the following formula,

$$\Delta C_i(N) = C_i(N+1) - C_i(N-1) \quad (12)$$

The delta-delta and delta cepstrum are usually joined through the static cepstrum for formulating the unique vector regarding the features consisting of together the dynamic and static features from the signal.

C. *Bottleneck feature extraction*

The bottleneck features used are computed based on the specific DNN structure. It consists of a narrow amount of BN layer in its position. Because of the amount of BN layer with hidden nodes is commonly very lesser while comparing to another layers. Training of DNN strengthen the initiation of the data, low dimensional compact representation of the signal is formed with the specially trained generation of less classification error for the output signal. The training procedure for BN DNN is started from a general supervised pre-training process with the limited

Boltzmann machines are utilized for the initialization of matrices of weight for efficient starting functions. Next to enable the procedure of training process, the bottleneck deep neural network used are finely altered with supervised manner through the help of standard error back-propagation procedure for optimizing the specific target function. Assumed the training set $X = \{x_t, l_t\}_{t=1}^{T}$ X = { } where l_i is the phone label off x_i, the target task could be denoted by the way of training data with entire negative log posterior probability:

$$D = -\sum_{t=1}^{T} \log P(l_t \mid x_t) \quad (13)$$

After training the voice signals with BN-DNN, the linear output from the hidden layers are obviously utilized as features of BN. If V Indicates the outputs with hidden layer sigmoid function is just beforehand the BN layer, with BN topographies Y could be get from $Y = W^T.V + b$, in which W is the weight matrix of BN layer and b denotes the bias vector [33].

D. I-vector feature extraction

The speaker and channel variability are introduced in two sub spaces which are based on the JFA training procedure. There are some available information on the channel space which differentiates the speaker. Due to this reason, the single space model is utilized with variabilities of two function. It is assumed that the channel and speaker reliant GMM super vector M is to be modelled as,

$$m = M + tW \quad (14)$$

In which M , represents a channel and super vector for speaker independent (UBM super vector with the worthy estimation of M), here low rank matrix is represented as T , that signifies a source of the condensed entire inconsistency space then w represents the usual standard vector. t Remains called as the entire matrix variability; the modules of w in which the entire features with the signifies the coordinates with in the minimum overall predictability gap. Distinctive vectors are mentioned or i-vectors for short. The vectors are related with a given recording is the MAP estimate of W , whose intention is described. We represent with this through $\hat{W}$.

E. Spectral Centroid

For the given frame, the weighted average frequency is computed with the spectral centroid. The weights required to be computed are based on the energy normalization of each frequency component in the particular frame.

$$SC_i = \frac{\sum_{<f>} f \mid S_i[f]\mid^2}{\sum_{<f>} \mid S_i[f]\mid^2} \quad (15)$$

Wherever, frame number is denoted as i , indicates the frame frequency range is mentioned as $< f >$ and Fourier transform is mentioned as $S_i[f]$ represents the. It is noted that the centroid of spectral and the SC feature set for each image frame is destroyed through the pitch frequencies and structure of harmonic which are induced by vocal source.

F. Renyi Entropy

An information theoretic measure is mentioned as Renyi Entropy that classifies the frames in a random manner. Here, the standardized energy of each frame is computed as a probability distribution for computing the entropy and is given by G.

$$RE_i = \frac{1}{1-\alpha} \log_2 \left(\sum_{<f>} \left| \frac{S_i[f]}{\sum_{<f>} S_i[f]} \right|^{\alpha} \right) \quad (13)$$

This kind of entropy measure is beneficial to detect both the unvoiced and voiced speech components due to the capacity of finding the degree of randomness for all speech signal. The less structured speech contains higher entropy when compared with the high structure speech signal. The less structure speech is based on the unvoiced speech and it has a higher value of entropy.

G. Shannon Entropy

Shannon Entropy is one of the information theoretic computations which classify the frames in a random manner. For entropy determination, the normalized energy of several frame is recognized as a probability distribution

$$SE_i = -\sum_{<f>} \left| \frac{S_i[f]}{\sum_{<f>} S_i[f]} \right| \log_2 \left| \frac{S_i[f]}{\sum_{<f>} S_i[f]} \right| \quad (14)$$

The SE trend is used similarly to the RE trend that is required for finding the unvoiced and voiced components of speech. For speaker identification, the features used are innovative and it can be used in the field of biometric recognition.

III. SPEAKER RECOGNITION BASED NN WITH OPTIMIZATION

Efficient and improved performance of ASR framework is based on the proper system training with the help of machine learning techniques. With the help of extracted feature vectors and trained data, network models are developed in which the testing data is correctly classified. There are three different kinds of neural network based approached are involved and they are worked based on the extracted features from the spoken words. The commonly used approaches for speaker classification and DNN based optimized approaches are described as follows. The extracted features from the voice signal are given as an input for classification approaches.

A. Multi-layer perceptron neural network

In MLP, each neuron computes the weighted sum for input values and this obtained input values is multiplied with the coefficient value. Finally, the outcomes of this multiplications are added together. Each node is considered as a linear classifier and the complex nonlinear classifiers are developed by combining all nodes. MLP is a kind of feed forward artificial neural network which maps the set of input values into its corresponding output set. It contains several layers which form the directed graph with each node of the next layer. Each node is having the activation function except the input node. This MLP is identified as an improved form of linear perceptron network which differentiates data that are linearly separable.

Activation function

The MLP has the activation function in the linear form in each nodes which maps the weighted input with the output of the node. It is possible to vary the number of nodes and layers used in the MLP based

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

on the complexity. The frequency of action in MLP is modelled with the nonlinear activation function of each node. It can be modelled in various methods. The sigmoid is the commonly used activation functions which are described in two specific ways. They are represented as follows.

$$Y(B_i) = \tanh(B_i) \quad (15)$$

$$Y(B_i) = (1+e^{-B_i})^{-1} \quad (16)$$

The first one represents the function of the hyperbolic tangent which is in the range of -1 to 1. Another one is the logistic function which is similar to the previous one but the ranges are varied from 0 to 1. The expression y_i represents the i^{th} nodes output and the weighted sums of input sequences are indicated with v_i. Instead of using these activation functions, alternate functions such as soft plus or rectifiers can be used. The radial basis function is one common kind of activation function which is based on supervised DNN models.

Layers

MLP consists of three or more layers. They are one input, one output and any number of hidden layers. Due to the presence of enormous hidden layers it is also referred as DNN. It is a fully connected network and each node of one layer is connected with a certain weight to every node in the successive layers. The weight factor w_{ij} indicates that the weight from the i^{th} node to the j^{th} node. The out layer act as a classifier which provides the classification result for the given speech signal based on the voice features extracted. It identifies the correct speaker by processing the hidden layers of DNN.

B. DNN with back propagation

The back-propagation model is based on the learning approach of perceptron by modifying the weights of the connection after processing each set of data. The amount of error produced is compared with the error occurred in the process and the error is reduced in the further iterations. It is a kind of supervised learning which is accomplished through the concept of back propagation. The generalizations of linear perceptron is applied with the least mean square. The error occurred in the j^{th} output node of n^{th} data point which is denoted by $e_j(n) = d_j(n) - y_j(n)$. Here, the target value is represented as, d, and y represents the value obtained by the perceptron. Then the weight of the node is modified based on the correction which reduces the error from the obtained output which is denoted as

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \quad (17)$$

By utilizing the gradient descent, the weight updation factor is denoted as

$$\Delta W_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \quad (18)$$

Where y_i indicates the previous nodes output and learning rate is represented as η, which is chosen to provide sufficient weight convergence without generating oscillations. Based on the obtained local values v_j, the derivatives are computed. For the output node, the simplification of this factor can be described as,

$$-\frac{\partial \varepsilon(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n)) \quad (19)$$

Where ϕ' indicates the derivative of the mentioned activation function, and it does not vary by itself. The weight updation of the hidden nodes is difficult to analyze but the derivatives are described as

$$-\frac{\partial \varepsilon(n)}{\partial v_j(n)} = \phi'(v_j(n))\sum_k -\frac{\partial \varepsilon(n)}{\partial v_k(n)} w_{kj}(n) \quad (20)$$

This is based on the weight updation of the k^{th} nodes, which indicates the output layer. In order to change the weight of the hidden node, the output weights are updated based on the activation function derivatives. Hence it is called the back propagation DNN. The output produced by this DNN is improved in terms of accuracy by reducing the error produced.

C. DNN with Differential evolution algorithm

The mutation, selection, and crossover operation are included in this DE algorithm based on the variation randomly generated pairs in the population procedure. It functions as a search procedure when the crossover operators are combined with the parent vectors and mutated vectors. All solution has the equal probability of selecting the parent vector. The step by step procedure is mentioned as follows. Initially, the populations are mentioned with parameter vectors of D dimensions and the size of the population is indicated as NP. $X_{i,G}(i = 1,2,...NP)$, indicates the generation of each target vector G and $X_{i,G} = \{x_{i1,G}, x_{i2,G}, ...x_{iD,G}\}$ represents the overall parameters that needs to be optimized. From the randomly chosen vectors, the donor vector for the subsequent generation $V_{i,G+1}$ is created based on $X_{r1,G}, X_{r2,G}$ and $X_{r3,G}$ as follows.

$$V_{i,G+1} = X_{r1,G} + F(X_{r2,G} - X_{r3,G}) \quad (21)$$

Where, F denotes the mutation factor of the random number which is distributed uniformly with the range of $[0,2]$. The randomly chosen indexes $r1, r2, r3 \in \{1,2,...NP\}$ are varied from one another and also from the entire running index i as well. The trial vector $U_{i,G+1} = \{u_{i1}, G+1, u_{i2,G+1}, ..., u_{iD,G+1}\}$ can be described by the crossover vector which can be described as follows.

$$u_{ij,G+1} = \begin{cases} v_{ij,G+1} & if\ rand \le CR\ or\ j = I_{rand} \\ x_{ij,G} & if\ rand > CR\ and\ j \ne I_{rand} \end{cases} \quad (22)$$

Where, $j = 1,2,...,D$, the crossover constant is represented as CR within the range $[0,1]$; $rand_{ij}$ represents a random number in the range of zero and one; and I_{rand} denotes a arbitrary index from $[1,2,...,D]$, projects which in the sample course one element remains selected from $V_{i,G}$. In the entire population, the diversity change is maintained by the parameters of CR. The vector required for the successive iterations are selected with the following minimization problem.

$$X_{i,G+1} = \begin{cases} U_{i,G+1} & if\ f(U_{i,G+1}) \le f(X_{i,G}) \\ X_{i,G} & otherwise \end{cases} \quad (23)$$

Where, the objective function related with $X_{i,G}$ is indicated as $f(X_{i,G})$. The target value is compared with the trail vectors and the lower function provides the value for the upcoming iterations. It is commonly used as a simple algorithm due to its control parameters because its control parameters such as NP, F and CR.

D. DNN with Artificial bee colony optimization

The foraging behavior performed by honey bee swarms is highly taken into consideration by this ABC algorithm to perform the optimization process. In this optimization approach, the obtained entire possible solutions are identified as a food source for this artificial bee. While performing the execution process, the overall bees may interact among themselves to increase the total quantity of gathered food source. The onlooker, scout and employed bees are grouped, then the duties and distribution of this grouped bees have been described. In this approach, food source position is $\{X_1, X_2, ...X_{NP}\}$ of the population is randomly adjusted. Both the population quantity and food source number are represented as NP . Each food source $X_i = \{x_{i1}, x_{i2}, ...x_{iD}\}$ represents the D-dimensional vector, consists of D variables for obtaining an optimal solution to the problem. During the phase of employed bees, the nearest solution v_{ij} obtained from the original value x_{ij} by using the equation mentioned below

$$v_{ij} = x_{ij} + rand(-1,1) \times (x_{ij} - x_{kj}) \quad (24)$$

Where, the subscript $j \in [1, D]$ indicates arbitrarily has selected files and $k \in [1, NP]$ is an arbitrary adjacent file which is varied from i. Afterwards performing the nearest neighbor then the unique answer based on the fitness solution of the best problem, the improved explanation from the population is obtained. The information based on the source is transmitted from the employed bees and the selection is based on the obtained probability value as follows.

$$P_i = \frac{f(X_i)}{\sum_{k=1}^{NP} f(X_k)} \quad (25)$$

For solution X_i, its fitness function is represented as $f(X_i)$. For the selected solution, fitness value based on likelihood rate p_i is made. If p_i is greater than an arbitrary amount solutions obtained within the range of [0, 1]. The nearest solution is obtained based on the local search of onlooker bees. During the phase of the onlooker and employed, the food source is exhausted and it is destroyed. The original solution is obtained by replacing the nearest neighbor which is not higher than the required number of cycles $Limit$. It is considered as an exhaustive search by replacing it with the random solution. The newest food sources obtained by employing the subsequent equation.

$$x_{ij} = x_j^{\min} + rand(0,1) \times (x_j^{\max} - x_j^{\min}) \quad (26)$$

Where, $x_j^{\max}$ and $x_j^{\min}$ denotes the maximum and minimum value of the appropriate solution, respectively. Similar to that, the correct solution is obtained by evaluating and comparing with the existing one. The solution is selected based on the optimal fitness value and it can be accomplished with two parameters NP and $Limit$.

E. DNN with Ant colony optimization

The ACO algorithm is based on the behavior of ants based on the objective of finding minimum distance between food and the nest. The communication mechanism used between ants is a pheromone laid. Based on the intensity of pheromone, the paths are selected independently. The solution for the component is obtained by selecting the higher value of pheromone. This value is reinforced with the pheromone value based on the suitable solution. The working concept of this presented algorithm is defined as shown below In the search space, the number of parameters required to be optimized is set to be D. The parameters used are divided into a group of discrete points. For parameter $x_j (j = 1,2,..D)$ within a range of $[a_{j1}, a_{jN}]$, there are totally N discrete points where the possible range is divided into $N-1$ shares uniformly. Each point based on the candidate value of the parameter, namely the solution component $a_{jk} (k = 1,2,...,N)$. When searching the parameters D by an ant $i (i = 1,2,...,NP)$, it selects the value for its parameters from the candidate points and stores it in the relevant tag. For each candidate point, the intensity value is represented as τ_{jk} for tag k. When it reaches the parameter x_j, the following probability is required to be selected for each component.

$$P_{jk} = \frac{\tau_{jk}}{\sum_{1 \le m \le N} \tau_{jm}} \quad (27)$$

When the probability value is greater than the arbitrary range in the range one and zero, the relevant point a_{jk} is chosen related with this parameter. Next to selecting all the required parameters $X_i = \{x_1, x_2, ...x_D\}$ by the ant, it goes towards the location and the intensity of pheromone is updated based on the following equation.

$$\tau_{jk}(i+1) = \rho \tau_{jk}(i) + \Delta \tau_{jk} \quad (28)$$

$$\Delta \tau_{jk} = \begin{cases} Q / f(X_i) & if\ a_{jk} is selected as\ x_j\ and\ belongs to\ X_i \\ 0 & else \end{cases} \quad (29)$$

The process of evaporation is indicated in the first term by which $\rho \in [0,1]$ indicates the pheromone duration coefficient. The reinforcement is represented by the second term and the amount of pheromone available for the solution is considered as apportion of X_i. Within the identified best solution, the component a_{jk} is not taken into consideration so the reinforcement parameter becomes $\Delta \tau_{jk} = 0$. The user-defined parameter obtained in the equation is normally referred as pheromone constant Q which is found to be similar for the entire ants. The, NP, N, ρ and Q are the parameters that are normally included in this ACO algorithm.

IV. EVALUATION OF DIFFERENT APPROACHES

There are three optimization algorithms are evaluated with the DNN and the speaker specific features are extracted along with i-vector features. Rather than employing, the algorithm of back propagation for parameter updation, it is accomplished with the help of optimization algorithms. The results produced with different algorithms and different datasets used for the speaker recognition tasks are discussed in this section.

A. Dataset

The performances of different neural network based machine learning approaches are evaluated for automatic speaker recognition.

Table 1: The comparison of different speaker recognition approaches

Authors	Language	Feature extraction	Classifier	Dataset	Accuracy
Themos Stafylakis et al. [21]	English	i-vector	DNN	NIST-SRE 2010,	96%
David Snyder et al. [22]	English	i-vector with probabilistic linear discriminant (PLDA)	DNN	US English telephone speech	89.8%
Waad Ben Kheder et al. [23]	English	short utterance i-vectors	GMM	SITW database	91.2%
Michel Vacher et al. [24]	French	Features from Subspace GMM	Semantic Classification Trees	7-channel raw audio stream	80%
Waad Ben Kheder et al. [25]	English	I-vector +MAP	GMM	NIST SRE 2004, 2005, 2006 and Switchboard data	83.73%
Zhiyuan Tang et al. [26]	English	recurrent information	unified neural network	WSJ database,	96.87
Lantian Li et al. [27]	Chinese	i-vector	DNN, GMM-UBM	SUD12	82.57
MdSahidullah et al. [28]	English	frequency domain linear prediction (FDLP), mean Hilbert envelope coefficients (MHECs) and power-normalized cepstral coefficients (PNCCs)	GMM–UBM	NIST SRE and text-dependent (RSR2015) speech corpora	89%
Ahilan Kanagasundaram et al. [29]	English	i-vector	GPLDA scoring	NIST 2008 SRE	96.13
Omid Ghahabi et al. [30]	English	i-Vector and PLDA	Deep Belief Networks (DBN) and Deep Neural Networks (DNN)	NIST SRE 2006	95.89
Suwon Shon et al. [31]	english	LDA and Heteroscedastic-LDA	Maximum likelihood detector	NIST SRE 2010	97.57
Sharada V. Chougule et al. [32]	Hindi	Normalized Dynamic Spectral Features	GMM	MVSR	89%

(i) Benchmark TEDLIUM dataset

It is developed for ASR and TED recognition tasks. This is one of the publicly available datasets which consists of 774 TED spoken utterances with the amount of 118 hours of speech data. Each talk of TED is considered as a speaker for ASR task. Both the tst2010 and dev2010 test sets are described to perform the decoding process by the ASR path of the already existing IWSLT estimation process. The IWSLT evaluation process mainly focuses on developing end-to-end speaker recognition systems, where this ASR is considered as typical component [13].

(ii) TIMIT

The TIMIT dataset is utilized with the standard training set having 462-speaker and the SA records are eliminated. Each speaker in the data base has identical speak. The set having 50 speakers are developed separately to tune the entire meta parameters. Results are developed by employing non-overlapped 462-speaker training set along with the development set. In the test set, totally eight utterances are included for each speaker [34].

(iii) SUD12

It is a recorded dataset developed for SUSR tasks and it is suitable for the speaker recognition tasks. This dataset includes 28 male and 28 female speakers, and their utterances are obtained in standard Chinese. In this dataset, each speaker contains 100 Chinese sentences and 15 ~ 30 Chinese characters are included in this each sentence. The voice signal is sampled with 16 kHz and 16-bits precision. It contains 56 speakers, whereas, 62-63 short speech utterances are included in this each speakers and the total duration of each speech is about 35 seconds. The utterance length is 0.5-2 seconds [27].

(iv) NIST 2014

The whole database of i-vector based speaker recognition challenge dataset is made available in the NIST 2014 dataset and this dataset is applied in this section to perform the experimentation process. Instead of applying speech signals, the NIST provide i-vectors to overcome the challenges of training and testing process while developing the

speaker recognition system. The system comparison is more readily enabled along with the consistency in the amount, front-end and also for the type of the background data.

B. Experimental results

The experiments are implemented with the DNN by setting the following parameters. The DNN is used with 6 hidden layers and sigmoid activation function. 1024 units are used with each layer. The remaining experimental parameters are based on the DNN parameters used in [13]. The experimental results are analyzed for various speakers' recognition techniques such as DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO. The speaker recognition techniques are compared for finding the efficient algorithm for updating the parameters in speaker recognition using DNN. Figure 2, 3 and 4 shows the comparison of speaker recognition approaches used with DNN classifier. The speaker recognition results are compared with the metrics used for speaker recognition such as EER, DCF, and C-average. The c-average comparison for the work is shown in figure 1.

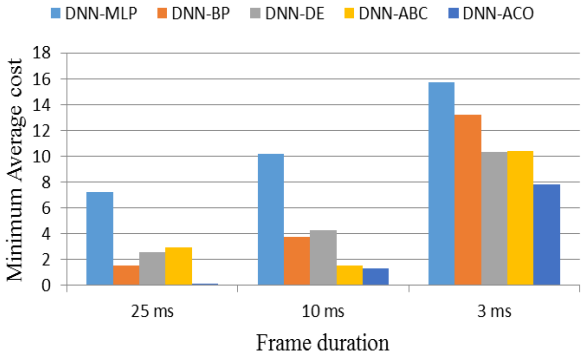

Figure 2: C-average comparison for speaker recognition approaches

Figure 2 shows the performance comparison of speaker recognition approaches by varying the frame duration. The frame durations are taken as 25ms, 10ms and 3ms. These values are computed for all techniques. The better value is obtained for the DNN-ACO technique. The higher value of minimum average cost shows the inefficient performance and the lower value of c-average indicates the highest performance in terms of efficiency. At 25 ms, the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO produces values such as 6.5, 0.5, 2.5 and 2.7 respectively. At 10ms, the c-average values produced by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 10, 2.9, 4.1, 0.7 and 0.5. At 3ms, the c-average values produced by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 15, 13, 10, 9 and 7.5. When the numbers of frames are increased then the c-average performance is degraded and the better performance can be obtained with increasing the frame duration.

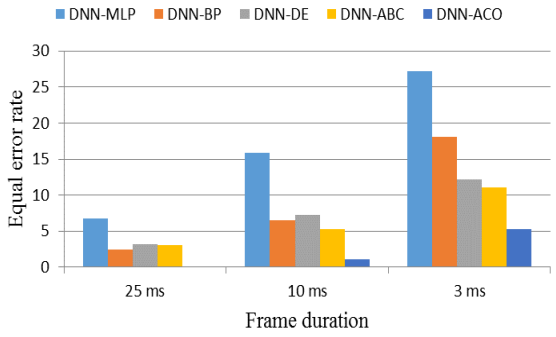

Figure 3: EER comparison for speaker recognition approaches

The EER performance comparison is shown in figure 3 which compares three optimization approaches which are suitable for the process of speaker recognition. The EER performance of the recognition task is compared by varying the frame duration with several levels like 25ms, 10ms, and 3ms. At 25ms duration, the EER values produced by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 7, 2.5, 2.7, 3 and 1. At 10ms duration, the EER values produced by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 15, 5.2, 5.5, 5 and 3. At 3ms duration, the EER values produced by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 25.5, 15.5, 12, 11 and 5. When reducing the frame duration the numbers of frames used are increased and it leads to the degradation in performance. When the frame duration is reduced, the numbers of frames used are reduced in which the performance is enhanced.

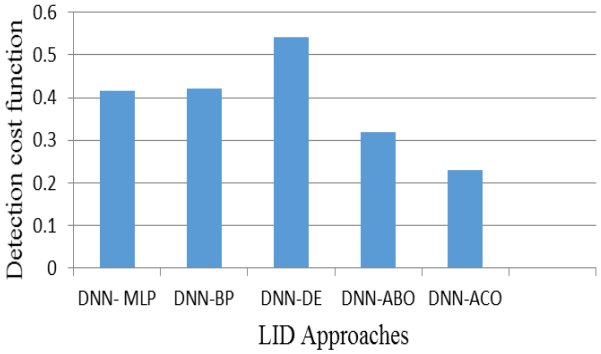

Figure 4: DCF comparison for speaker recognition approaches

In speaker recognition procedure, the DCF is compared with the other DNN based speaker recognition approaches. Its performance is found high when the parameters used in the DNN are updated with the optimization algorithms. The optimized DNN structure is shown in the figure2, figure3, and figure4. The DCF value obtained by the DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 0.4, 0.45, 0.55, 0.35 and 0.25 respectively.

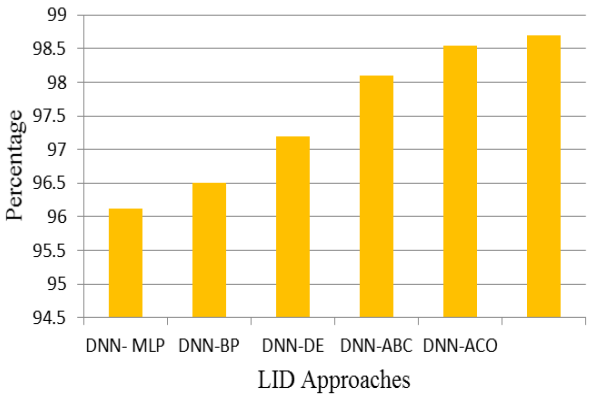

Figure 5: Accuracy comparison for speaker recognition approaches

The accuracy comparison for several speaker recognition technique is shown in figure 5. To vary the proposed optimized DNN methods performance, various classifiers are compared along with the speaker values. The accuracy of the DNN based approaches such as DNN-MLP, DNN-BP, DNN-ABC and DNN-ACO are 96.12, 96.15, 97.2, 98.1, 98.54 and 98.7. These values high for the optimization based DNN approaches and for other approaches this value is relatively low. This kind of DNN based system is suitable for the speaker recognition task when processing it with i-vector features. The performance measures such as DCF, EER, c-average and accuracy are presented in figures 1, 2, 3, 4 and 5 shows the performance of optimized DNN based approaches. The performance of both DNN-MLP and DNN-BP produces less performance than that of DNN-DE, DNN-ABC and DNN-ACO. These approaches are compared in this review. Several feature extraction and speaker recognition approaches are reviewed in this work for analyzing the performance of the system in various DNN environments.

V. CONCLUSION

In this review, different feature extraction approaches are analyzed for finding the speaker specific attributes. The features are efficiently extracted with the i-vector extraction and bottleneck features. Nonetheless, from what we know, it very well may be said that automatic speaker-recognition systems should focus mainly on the high-level features to achieve performance. The combined feature extraction provides more accurate results for the speaker recognition task. Neural network based classification algorithms and other states of the art machine learning algorithms are evaluated, then the simulation outcomes are compared. The DNN with better optimization provides better results for the classification tasks when compares with other classification mechanisms. The comparative study of different speaker recognition approaches and their performance is presented and it helps the researchers to develop an efficient system for better recognition.

REFERENCES

1. Hanilçi, Cemal. 2018. Data selection for i-vector based automatic speaker verification anti-spoofing. Digital Signal Processing, Elsevier. 72: 171-180.
2. Sizov, Aleksandr, Khoury E, Kinnunen T, Wu Z and Marcel S. 2015. Joint Speaker Verification and Anti spoofing in the $ i $-Vector Space. IEEE Transactions on Information Forensics and Security. 10(4): 821-832.
3. Wang, Jia-Ching, Li-XunLian, Yan-Yu and Jia-Hao. 2015. VLSI design for the SVM-based speaker verification system. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 23(7): 1355-1359.
4. Poignant, Johann, Besacier L and Quénot G. 2015. Unsupervised speaker identification in TV broadcast based on written names. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 23(1): 57-68.
5. Daqrouq, Khaled and Tutunji TA. 2015. Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. Applied Soft Computing, Elsevier. 27: 231-239.
6. Karthick, S. "TDP: A Novel Secure and Energy Aware Routing Protocol for Wireless Sensor Networks." In International Journal of Intelligent Engineering and Systems, vol. 11, no. 2, pp. 76-84. 2018.
7. Moro-Velázquez, Laureano, Gómez-García JA, Godino-Llorente JI, Villalba J, Orozco-Arroyave JR and Dehak N. 2018. Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. Applied Soft Computing, Elsevier. 62: 649-666.
8. Wu, Zhizheng, Evans N, Kinnunen T, Yamagishi J, Alegre F and Li H. 2015. Spoofing and countermeasures for speaker verification: a survey. Speech Communication, Elsevier. 66: 130-153.
9. Safavi, Saeid, Russell M and Jančovič P. 2018. Automatic Speaker, Age-group and Gender Identification from Children's Speech. Computer Speech & Language, Elsevier.
10. Ranjan, Shivesh, et al. 2018. Curriculum Learning Based Approaches for Noise Robust Speaker Recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 26(1): 197-210.
11. Sainath, Tara N, et al. 2015. Deep convolutional neural networks for large-scale speech tasks. Neural Networks. 64: 39-48.
12. Richardson, Fred, Reynolds D and Dehak N. 2015. Deep neural network approaches to speaker and language recognition. IEEE Signal Processing Letters. 22(10): 1671-1675.
13. Miao, Yajie, Zhang H and Metze F. 2015. 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.
14. Almaadeed, Noor, Aggoun and Amira A. 2015. Speaker identification using multimodal neural networks and wavelet analysis. IET Biometrics. 4(1): 18-28.
15. Qawaqneh, Zakariya, Mallouh AA and Barkana BD. 2017. Deep neural network framework and transformed MFCCs for speaker's age and gender classification. Knowledge-Based Systems. 115: 5-14.
16. Franco-Pedroso, Javier and Gonzalez-Rodriguez J. 2016. Linguistically-constrained formant-based i-vectors for automatic speaker recognition. Speech Communication, Elsevier. 76: 61-81.
17. Misra, Abhinav and Hansen JHL. 2018. Modeling and compensation for language mismatch in speaker verification. Speech Communication, Elsevier. 96: 58-66.
18. Saon, George and Soltau H. 2014.A comparison of two optimization techniques for sequence discriminative training of deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. 5567-5571. IEEE.
19. Ferrer, Luciana, Lei Y, McLaren M and Scheffer N. 2016. Study of senone-based deep neural network approaches for spoken language recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). 24 (1): 105-116.
20. Garcia-Romero, Daniel, Zhang X, McCree A and Povey D. 2014. Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. In Spoken Language Technology Workshop (SLT), 2014 IEEE. 378-383. IEEE.
21. Stafylakis T, Kenny P, Gupta V, Alam J and Kockmann M. 2016.Compensation for phonetic nuisance variability in speaker recognition using

DNNs. In Odyssey: The Speaker and Language Recognition Workshop. 340-345.
22. Snyder D, Ghahremani P, Povey D, Garcia-Romero D, Carmiel Y and Khudanpur S. 2016. Deep neural network-based speaker embeddings for end-to-end speaker verification. In Spoken Language Technology Workshop (SLT), 2016 IEEE. 165-170. IEEE.
23. BenKheder W, Matrouf D, Ajili M and Jean-François. 2016. Probabilistic Approach Using Joint Long and Short Session i-Vectors Modeling to Deal with Short Utterances for Speaker Recognition. In Interspeech. 1830-1834.
24. Vacher M, Lecouteux B, Romero JS, Ajili M, Portet F and Rossato S. 2015. Speech and speaker recognition for home automation: Preliminary results. In Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. 1-10. IEEE.
25. BenKheder W, Matrouf D, Jean-François, Ajili M and Pierre-Michel. 2015. Additive noise compensation in the i-vector space for speaker recognition. In Acoustics, Speech and Signal Processing (ICASSP). 2015 IEEE International Conference on. 4190-4194. IEEE.
26. Tang Z, Li L and Wang D. 2016. Multi-task recurrent model for speech and speaker recognition. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. 1-4. IEEE.
27. Li L, Wang D, Zhang X, Zheng TF and Jin P. 2016. System combination for short utterance speaker recognition. In Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2016 Asia-Pacific. 1-5. IEEE.
28. Md Sahidullah and Kinnunen T. 2016. Local spectral variability features for speaker verification. Digital Signal Processing. 50: 1-11.
29. Kanagasundaram A, Dean D and Sridharan S. 2015. Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. 4654-4658. IEEE.
30. Ghahabi O and Hernando J. 2017. Deep learning backend for single and multisession i-vector speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 25(4): 807-817.
31. Shon S, Mun S, Han DK and Ko H. 2015. Maximum likelihood Linear Dimension Reduction of heteroscedastic feature for robust Speaker Recognition. In Advanced Video and Signal Based Surveillance (AVSS). 2015 12th IEEE International Conference on, 1-5. IEEE.
32. Chougule SV and Chavan MS. 2015. Robust spectral features for automatic speaker recognition in mismatch condition. Procedia Computer Science 58: 272-279.
33. Song, Yan, Jiang B, Bao YB, Wei S and Li-Rong. 2013. I-vector representation based on bottleneck features for language identification. Electronics Letters. IET. 49(24): 1569-1570.
34. Xue, Shaofei, Jiang H, Dai L and Liu Q. 2016. Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition. Journal of Signal Processing Systems. 82(2): 175-185.