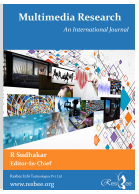
Vol.3 No.1 Jan 2020

Multimedia Research

Received 4 September, Revised 18 November, Accepted 22 December



Hybrid Particle Swarm Optimization-Deep Neural Network Model for Speaker Recognition

Vasamsetti Srinivas
Department of Electronics and Communication Engineering, Swarnandhra Institute of Engineering and Technology, Narsapur, Andhra Pradesh, India
srinivas.vasiet@gmail.com

Santhirani Ch
DMS&SVH Engineering college, Machilipatnam, Andhra Pradesh, India

Abstract: Nowadays, speaker recognition is considered as a current research topic. Moreover, the voice biometrics which is attained from the speaker's behavior or physical related features offers a pattern of data that contains sensitive information regarding the speaker. The efficiency of speaker recognition systems is observed to minimize expeditiously because of the mismatch incidence, such as noise and channel degradations. With the aspire to promise security and effectual recognition, a Hybrid Particle Swarm Optimization–based Deep Neural Network (Hybrid PSO-based DNN classifier) is used to identify a speaker for that the frequency-dependent features, like MKMFCC, autocorrelation, and spectral skewness, are exploited. The classification is done by exploiting the DNN classifier based on feature extraction and classifier is performed optimally by exploiting the proposed Particle Swarm Optimization. Finally, the simulation analysis of the proposed technique is compared with the LM, SVM, GMM, and BSW. It shows the performance of the proposed technique outperforms the conventional techniques concerning the accuracy, FAR FRR.

Keywords: Speaker Recognition; Feature Extraction; Classification; DNN; PSO

Nomenclature

Abbreviations	Descriptions
MKMFCC	Multiple spectral kurtosis, Kernel Weighted Mel Frequency Cepstral Coefficient
SI	Speaker Identification
SV	Speaker Verification
ELSDSR	English Language Speech Database for Speaker Recognition
SVM	Support Vector Machine
UBM	Background Model
PLDA	Probabilistic Discriminant Analysis
VQ	Vector Quantization
GMM	Gaussian Mixture Model
ResNet	Residual Network
DFT	Discrete Fourier Transform
SITW	Speakers in the Wild estimate
LDV	Laser Doppler Vibrometer
OMLSA	Optimally Modified Log-Spectral Amplitude
SGD	Stochastic Gradient Descent
SID	Speaker Identification
FFT	Fast Fourier transform
BSW	Blind Spectral Weighting
DTW	Dynamic Time Warping
DNNs	Deep Neural Networks
GMM	Gaussian Mixture Model
ML	Machine Learning
MSE	Mean Square Error
MKMFCC	Multiple Kernel Weighted Mel Frequency Cepstral Coefficient
DCT	Discrete Cosine Transform
LM	Levenberg Marquardt
BP	Back Propagation
AI	Artificial Intelligence
SVM	Support Vector Machine

© Resbee Info Technologies Pvt Ltd
https://doi.org/10.46253/j.mr.v3i1.a1

1. Introduction

The human voice is a phenomenon that is mainly based upon the speaker who creates it. Various research exhibits that no two individuals sound precisely similar [1]; the acoustic features differentiate the dissimilarity among the speaker's voices that are indecisive and complicated to separate from signal features, which reflect segment recognition. In [2], the sources of difference between speakers are threefold, such as the difference in speaking styles (pronunciation incorporated), the difference in vocal cords and vocal tract shapes, and how speakers articulate themselves to express an exacting message (phrases or words exploited). On the other hand, like a speaker's tendency to exploit definite words, phrases, and syntactic structures (indicating to third source) is not simple to enumerate or manage in experimentation, the first two sources are used by the automatic speaker recognition systems by means of searching the minimum-level acoustic features of a speech signal. In signal processing, speaker recognition is a significant subject particularly in security systems that encompass a diversity of applications [3]. On speaker recognition, voice-controlled devices and systems rely plays a great role. In speaker recognition, few applications are security control for secret information, forensics and remote access to computers and confirmation customers for bank transactions [2].

Generally, speaker recognition represents the procedure of robotically identifying a speaker by her or his audio samples. It possesses to turn out to be a gradually more significant means of identifying identities in numerous e-commerce technologies within business interactions, common forensics, and law enforcement [1]. In addition to SI and SV, speaker recognition is subcategorized into text-independent and text-dependent recognition tasks. Speaker substantiation aspires to confirm if a word fits into an exact registered speaker when speaker recognition aspires to categorize the recognition of an unidentified word between an exact set of registered speakers. Text-dependent speaker identification tasks need speakers to complete a definite phrase when there are no phonetic constraints in the text-independent speaker identification.

A study in automatic speaker recognition [2] has concentrated gradually more on improving sturdiness in adverse circumstances persuaded by reverberation, background noise, and minimum-quality recordings. Numerous methods were examined to undertake these concerns, the main flourishing being i-vector technology [3] exploited together with the PLDA back-end [4]. Besides, the novel word-level features back-ends and (i-vectors), enhancements have been attained in the first segment of the speech processing chain by emergent robust acoustic features [6]. In both topics, current advances have brought the speaker recognition systems performance nearer to the level predictable in applications like surveillance, forensics, and authentication.

The main contribution of this work is to identify a speaker from the speech signal on the analysis with other speech signals using an automated speaker recognition system on the basis of the frequency-dependent features. The proposed speaker recognition procedure includes two major steps, such as feature extraction and classification. At first, the frequency-based features, like MKMFCC, spectral skewness, spectral kurtosis, and autocorrelation are extracted. Moreover, the weight values of the classifier are attained by exploiting the proposed PSO approach optimally, by exploiting the classifier which classifies the speech signal and recognizes the speaker. Moreover, the extracted features are subjected to the classifier that is exploited to classify the features to recognize the speaker.

2. Literature Review

In 2019, Tengyue Bian et al [1], presented to integrate the ResNet with the self-notice method to attain enhanced analysis of speaker recognition in text-independent with minimum computational cost and fewer parameters. Moreover, the Cluster-Range Loss was on the basis of an ingenious online exemplar mining that was proposed to unswervingly shrink the intra-class difference and to expand inter-class distance.

In 2019, Jes´us Villalba et al [2], worked on the pooling layers, network architectures, back-end adaptation, training objectives approaches, and calibration methods. Moreover, the VAST audio quality was relentlessly degraded evaluated to SITW, although they both comprise of Internet videos. This degradation reasons a strong domain mismatch among VAST and training data. Because of this disparity, great networks carried out just somewhat superior to lesser networks. In addition, this complicated VAST calibration. Nevertheless, to standardize VAST using acclimatize SITW scores distribution to VAST was managed, by exploiting a small number of in-domain improvement data.

In 2018, Shuping Peng et al [3], addressed the concerns in remote speaker identification. Hence, an LDV was exploited to recognize the remote speaker. Here, three LDV speech corpuses, each comprises of 50 speakers, which were gathered from the vibrations of a mineral water bottle, a plastic bag, and a computer screen, by exploiting the LDV. To enhance LDV-captured speech quality, speech enhancement

applications on the basis of the OMLSA approach was exploited. A GMM-UBM technique was build to differentiate remote speakers based on improved LDV-captured speech.

In 2019, Emma Jokinen et al [4], worked on two compensation techniques that were proposed to undertake them was equivalent in a normal vs. shouted speaker recognition mission. These techniques were used in the extraction of the feature phase for the recognition of a speaker. In estimation by exploiting the conventional ivector based recognition system, the proposed methods offered substantial enhancements in recognition rates evaluated to the case while shouted speech spectra were not processed.

In 2019, Michael Jessen et al [5], worked on two automatic speaker recognition systems proposed using the company Phonexia. The primary called SID -XL3 was i-vector PLDA system that operates with two streams of features, one of them exploiting MFCCs in a traditional sense, second one exploiting features of DNN-Stacked Bottle-Neck.

In 2018, Ville Vestman, et al [6], worked on a disadvantageous source of acoustic deviation, although; occurs from incompatible speaking styles persuaded by speaker, most important to a considerable performance drop in recognition precision. This was the main crisis particularly in forensics whereas perpetrators might with determination disguise their uniqueness by varying their speaking style. Moreover, they had focused on the majority usually exploited ways of disguising one's speaker identity, such as whispering. From the perspective of robust feature extraction, the issue of normal-whisper acoustic mismatch compensation was approached.

In 2017, Ing-JrDing and Jia-YiShi [7], presented a Kinect microphone array-based technique for the voice-based control of humanoid robot shows by speaker and speech recognition. For speaker identification and verification, and speech recognition, the GMM, SVM, and DTW were exploited they were efficiently integrated for understanding proposed voice-based control of humanoid robot demonstrations.

In 2019, JohanRohdin et al [8], presented DNNs, which was based on the numerous end-to-end speaker corroboration systems. For text-dependent errands with small utterances, these systems were confirmed to be competitive. Here, overfitting mitigates that usually restricts the end-to-end systems performance. On both short and long duration utterances, the proposed model was superior to the i-vector+PLDA baseline.

3. Hybrid PSO- Based DNN Model for Speaker Recognition

The major aim of the work is to model and propose an automatic speaker recognition system by exploiting the frequency-based features. Fig. 1 exhibits a schematic model of proposed Hybrid PSO – based DNN. The speaker recognition is performed by exploiting two steps, such as feature classification and extraction. Initially, to extract features, the subjected speaker signal is processed. From the speaker signal, frequency-based features, like spectral kurtosis, spectral skewness, MKMFCC, and autocorrelation are extracted. Subsequently, the extracted features are subjected to DNN. The high-frequency marginal features are produced. By exploiting the theory of Particle Swarm Optimization the texture details are enhanced. The optimization parameters of the Hybrid PSO-based DNN classifier are adjusted. By exploiting the PSO method weights of DNN are considered. Consider X as the input database comprising of each speaker P_ω and λ speakers comprises n a number of training speech signals.

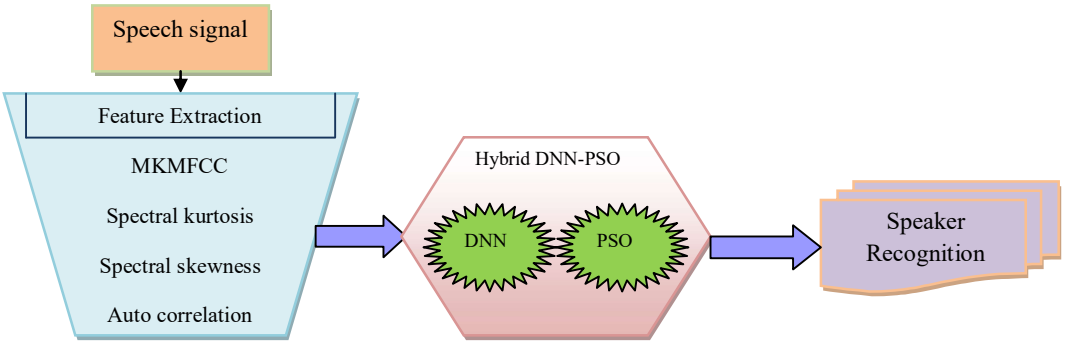


Fig. 1. *Schematic illustration of the proposed algorithm*

$$X \in P_\omega \;\; ; \;\; \{1 \le \omega \le \lambda\} \tag{1}$$

$$P_\omega = \{H_\delta \;\; ; \;\; 1 \le \delta \le n \;\; \} \tag{2}$$

Each speech signal associated with H_δ denotes as a signal vector $s(\omega)$ which is exploited to identify the speaker. Here, a major concern to be described is the identification of λ speakers independently on the basis of their speech signal.

3.1 Feature Extraction

In feature extraction, from the subjected speech signal, the extraction of frequency-based features is done for the precise speaker recognition. For speech signal, frequency-based features like spectral skewness, MKMFCC, spectral kurtosis, and autocorrelation are used.

3.1.1 MKMFCC

Here, MKMFCC [9] is selected as a feature parameter which is Mel frequency scale, which extracts entire phonetic component features and it is not able to differentiate frequencies above 1 kHz as same as the human. In this approach, coefficients of MFCC are attained by exploiting 2 diverse kernel functions. An improved method to incorporate and combine diverse kinds of data is to employ kernel-based weightage. Moreover, the integrated function of necessary kernel methods and models tends to a coherent and competent class of algorithms. The statistical and computational properties of the kernel method and model are obviously specified by exploiting a variety of ML techniques [10]. The tangential and exponential functions of the multiple kernel weighted model are exploited in Mel filter bank energy utilization. Each phonetic part is given without neglect on taking into consideration the maximum and minimum energy frame of the audio signal that is the chief advantage of exploiting the multiple kernels. The Mel scale is estimated by exploiting the weighted sum of spectral models regarding the exponential and tangential function that tends to enhanced diarization performance. The function of MKMFCC feature parameterization and the mathematical model is described and it is stated as follows:

i) Pre-emphasis: In this procedure, filter is subjected to input audio signal that denotes the superior frequencies available in acoustic signal. The procedure for extraction of feature is performed, while the superior frequency intensity raises the signal energy of superior frequencies and it is stated in eq. (3).

$$O(n) = H(n) - Y * H(n-1) \tag{3}$$

In eq. (3), H indicates the input signal, E indicates the output signal, $n = 1$ to N indicate the samples of the audio signal, and Y represents a constant value which creates audio sample to produce from the preceding sample.

ii) Framing: By exploiting the framing procedure, the sample speech signal is partitioned into a number of frames. The audio frame length lies among 20 and 40 ms and therefore partitioned into 1frames of N samples. F indicates a factor, exploited to divide the neighboring frames of the stream, whereas $F < N$.

iii) Hamming windowing: In audio streams, close frequency lines are combined by exploiting the Hamming window process which is helpful in the procedure for the extraction of feature. The Hamming window model is stated in eq. (4).

$$Z(n): 1 \le n \le N-1 \tag{4}$$

After windowing attained signal is stated in eq. (5) as,

$$O(n) = H(n) \times Z(n) \tag{5}$$

In eq. (6), hamming window value of $Z(n)$ stated as below,

$$Z(n) = 0.56 - 0.46\left(\frac{2\pi n}{N-1}\right) \tag{6}$$

Whereas, $0 \le n \le N-1$.

iv) FFT: In the time domain, the audio signal is transformed into the frequency domain of the sample frame in Fast Fourier Transform. Using the eq. (7), the power spectrum is calculated.

$$M_1(d) = \frac{1}{N}\left|H_1(d)\right|^2 \tag{7}$$

The power spectrum is exploited to the summation of the domain conversion, in that the periodogram spectral estimate is attained. The Discrete Fourier Transform (DFT) of the equivalent framed signal is stated in eq. (8).

$$H_1(d) = \sum_{n=1}^{N} O(n) \bullet e^{-j2\pi dn} \tag{8}$$

In eq. (8), $1 \le d \le D$, k represents DFT length, $A(m) = B(m) \times Z(m)$ comprising of R sample long window analysis.

v) Mel filter bank processing: In the FFT spectrum, the wider range of frequency is not appropriate in a linear scale of the audio stream. Hence, the unwanted signals in the spectral estimates

are filtered with the bank of the filter by exploiting the Mel scale processing. The signal frequencies are filtered by exploiting the triangular filter to decide the weighted summation of filter spectral modules and a result, Mel scale is calculated by exploiting the triangular filter. As stated in eq. (9), filters are set to variety equal spacing among each other which presents linear frequency in the Mel frequency.

$$\text{Mel}(t) = 1125 * \text{In}\left(1 + \frac{t}{700}\right) \tag{9}$$

In eq. (9), $\text{Mel}(t)$ indicates the linear frequency.

vi) Filter bank energy: For MKMFCC computation, indecisive band total energy attained is used. At first, the power spectrum and filter bank are multiplied, and at last, summed up to few coefficients to attain filter band energy. In the feature parameterization process, the energy computation needs multiple kernel weightage function. In all filters, the log energy is attained with the multiplication of the triangular filter and the magnitude frequency response.

vii) DCT: From the values of the log Mel spectrum the time domain values are attained with the DCT. The outcomes attained from the DCT are indicated as coefficients of MKMFC and the set of MKMFC coefficients offers the acoustic feature vectors of every input utterance and is stated in eq. (10).

$$G(a) = \bar{G}(d) \tag{10}$$

In eq. (10), $G(a) = \begin{cases} G(a), & d = d_1 \\ 0, & \text{otherwise} \end{cases}$

In eq. (11), the cepstral coefficient attained from the calculated.

$$YJ_s(N) = \frac{1}{N'} \sum_{d=0}^{N'-1} \left[\bar{G}(D) e^{id\left(2\pi / N'\right)n} \right] \tag{11}$$

In eq. (11), $YJ_s(N)$ indicates the MKMFCC.

viii) Delta spectrum and energy: In the acoustic feature vector, the energy features are augmented with cepstral features to raise the accurateness in the speech recognition process and therefore, reduces the complexity of echo and noise.

ix) Cepstral normalization: In acoustic signals, the normalization process minimizes the remaining mismatch of the feature vector. The MKMFCC feature is indicated as, f_1.

3.1.2. Spectral Kurtosis

In [11], kurtosis of a complex arbitrary variable $y(r)$ at each frequency bin r is indicated as spectral kurtosis $y(n)$ and stated in eq. (12).

$$f_2 = D_x(n) = \frac{D_4\left\{Z^+(n), Z^+(n), Z^+(n), Z^+(n)\right\}}{\left[D_2\left\{Z^+(n), Z^+(n)\right\}\right]^2} \tag{12}$$

In eq. (12), $Z^+(n) \in \left\{Z(n), Z^*(n)\right\}$, $Z^*(n)$ represents complex conjugate of $Z(n)$.

3.1.3. Spectral Skewness

In [12], skewness of a spectrum coefficient is calculated exploiting spectral skewness. By exploiting skewness, distribution symmetry is attained that is besides called a third central moment. Distribution can also be negatively or positively skewed. The distribution is given with a longer tail to the right in positive skew, and in left, the distribution is provided with a long tail for the negative skew. For symmetrical distribution, the skewness is '0'. The skewness ratio to std dev for the third power is called coefficient of skewness which is stated in eq. (13).

$$f_3 = \text{Skewness} = \sum_{j=1}^{R} \frac{\left[(t_i - \varphi)^3 \bullet y_j\right]}{\alpha^3} \tag{13}$$

3.1.4. Autocorrelation Coefficients

To discover the correlative behavior, for real-time signal processing, the autocorrelation coefficients [13] are used at different time intervals that provide few exclusive of speaker characteristics which are exploited to recognize the speaker efficiently. Eq. (14) states coefficients of autocorrelation extraction $rr_s(q)$ from $s(j)$.

$$f_4 = rr_s(p) = \frac{1}{R} \sum_{j=p}^{R-1} s(j)s(j-p) \quad (0 \le p \le R) \tag{14}$$

In eq. (14), N denotes the total count of samples in the speech signal and p states the sample delay.

3.1.5. Feature Concatenation

The Spectral skewness, MKMFCC features, Autocorrelation coefficients, and Spectral kurtosis, are concatenated. The feature vector is stated as f which is indicated in eq. (15).

$$f' = \{f_1', f_2', f_3, f_4'\} \tag{15}$$

In eq. (15), f_1' indicates the MKMFCC feature, f_2' is the spectral kurtosis feature, f_3' indicates the spectral skewness feature, and f_4' indicates the coefficients of autocorrelation feature. The feature vector f' indicates and it is considered as input to the classifier which tries to identify the speaker.

4. Hybrid PSO-DNN Model

Basically, the speaker recognition is considered as a biometric system that possesses the capacity to recognize an exacting individual by comparing and analyzing the speech signal features. At first, the speech signal is given to the extraction of feature procedure, whereas features are extracted. Here, Particle swarm optimization is exploited to train the DNN classifier. Subsequently, the extracted features are subjected to the proposed classifier that classifies the speech signal to identify the speaker.

In this section, the hybrid model of PSO [14] and the DNN model are proposed. Here, to enhance the performance of the convolutional neural network, and also to optimize parameters of deep learning technique for speaker recognition.

4.1. Conventional PSO Algorithm

In [16], the PSO approach as well called the flock foraging approach. At first, an arbitrary solution is created and subsequently discovers the best solution by means of optimal fitness value iteratively. This category of technique is extensively used to the BP NN since of the benefits of simple performance, maximum accuracy, and rapid convergence. In addition, it possesses shown an advantage in resolving realistic issues and at first used in the area of DL.

The fundamental model of the PSO technique comprises a collection of particles that correspond with each other to attain the optimal place frequently. The technique updates the velocity, location, and each particle fitness value which are decided using the mathematical model in order to optimize the issue. Particles' location states the candidate solution to the issue required and stored as the individual optimal solution p_{best_i}.

The altering of location, which is prejudiced by its individual optimal value of the fitness as $p_{fitness}$ that is the least attained value in preceding iterations and directed toward global optimal location g_{best_i} equivalent to the global fitness value $g_{fitness}$ between each and every outcomes in complete space.

The eq. (16) and (17) represents the mathematical model of the PSO technique approach,

$$U_i[j+1] = wU_i[j] + c_1 r_1 (p_{best_i} - P_i(j)) + c_2 r_2 (g_{best} - P_i(j)) \tag{16}$$

$$P_i(j+1) = P_i(j) + U_i(j+1) \tag{17}$$

In eq. (16), w indicates the inertia weight, which aids the particles to shift during the interior to an improved location, c_1 indicate the constants and r_1 indicate the uniform arbitrary value, velocity and location vector of i^{th} particle are $P_i(j)$ and $U_i(j)$ correspondingly at the j^{th} iteration and the $P_i(j)$ is updated by $U_i(j+1)$.

4.2. DNN

Deep Learning is a base of numerous contemporary AI applications that comprises of multiple neural nodes and hidden layers. Presently, it has been outrageously used in voice processing, image recognition, and so on, in addition to a few developments in communication has been attained.

Speaker recognition has slowly changed from the conventional approach to the DNN approach [16] with the fast growth of Deep Learning. Also, this novel technique can be observed as a learning idea from the learning system model to signal features. Actually, the procedure of extravagance DNN as a classifier is observed as an amalgamation of features and ML. Also, DNN input is called the training instance, which is a multidimensional data vector obtainable invisible layer. Subsequently, every hidden layer achieves a series of non-linear transformations which is explained as below:

$$X = sig(EW * I + v) \tag{18}$$

In eq. (18), I indicates each neural node input, EW and sig represented as sigmoid activation function, v equivalent the bias vector and encoding weight matrices correspondingly, that is $1/(1+e^{-y})$.

Here, the SGD approach is adopted for training hidden layers and exploited the MSE model to compute the output error which is stated as below:

$$\text{Error}_{\text{out}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{K}\left(I'_{ki} - I_{ki}\right)^{2} + \lambda * \Omega_{\text{weight}} \tag{19}$$

In eq. (19), λ maps the coefficient for the L_2 regularization, I'_{ki} represents the actual output subsequent to a series of operations and Ω_{weight} represents used for the cost function weights and it is stated as eq. (20).

$$\Omega_{\text{weight}} = \frac{1}{2}\sum_{j}^{L}\sum_{k}^{Z}\sum_{i}^{M}\left(W_{ki}^{(j)^2}\right) \tag{20}$$

In eq. (20) M and L indicates a number of input variables and hidden layers correspondingly. Subsequent to the training DNN model, the test data can be fed subjected to this technique for prediction and transformed into identification precision at the output.

However, the DNN model is a black box that cannot be seen how it works and the features it learns cannot be seen neither that tends to a series of issues. Hence, the PSO approach is used as the solution to the issue in which the DNN is simple to fall into the local minimum value and the number of hidden layer nodes is not fixed, thus enhancing the precision of speaker recognition.

4.3. Speaker Recognition On The Basis Of the Hybrid PSO-DNN Model

Subsequent to DNN training, speech recognition is done by exploiting an amalgamation of DNN architecture and the PSO approach as the proposed model, and it is demonstrated in fig. 2. Here, the input of the network model is considered a normalized feature vector. Subsequently, the count of double hidden layer nodes of DNN is routinely adjusted exploiting global optimization ability of PSO method to attain the best number of nodes and to develop recognition precision, on the basis of examining the Mean Square Error attained. To encompass an improved examination of identification outcomes, to normalize the output layer to acquire the ultimate recognition rate $\hat{x}$ of every output node, which is stated as eq. (21).

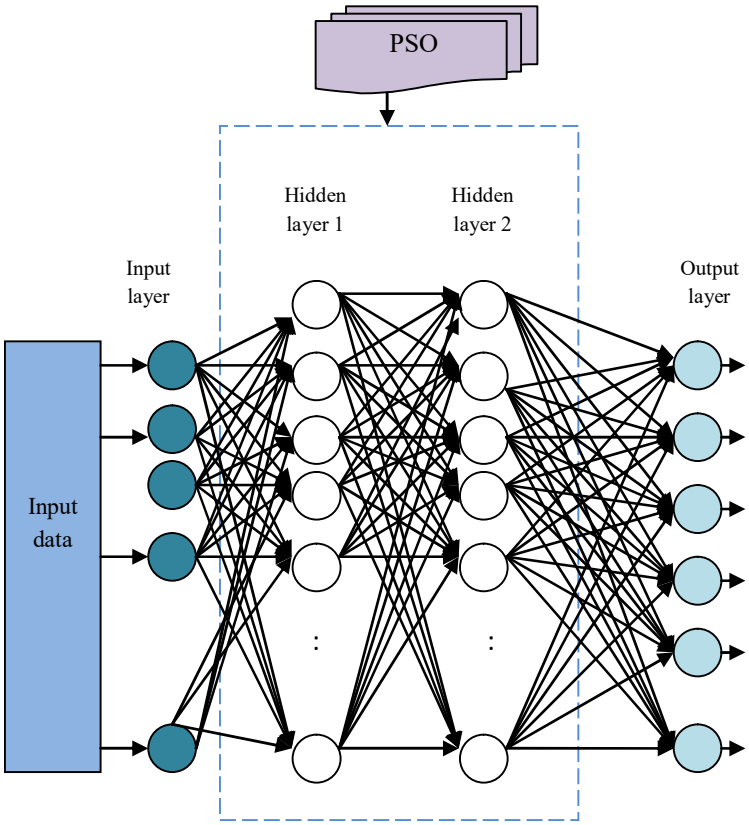


Fig. 2. *Diagrammatic representation of hybrid PSO-DNN model*

$$\hat{x}^{i} = P(x = i \mid out) = \frac{e^{out^{i}}}{\sum_{i=1}^{6} e^{out^{i}}} \tag{21}$$

In eq. (21), out^{i} indicates output vector out for the i^{th} element. At last, recognition modes are classified that are equivalent to the utmost $\hat{x}$, on the basis of the difference of features by exploiting the PSO-DNN approach.

Moreover, α sets of data are produced, which is the number of particles in a swarm. The individual fitness value of optimal p_{fit} and global optimal fitness value g_{fit} indicates the least recognition error equivalent to the best number of nodes in i^{th} the group and complete group, correspondingly. Likewise, the individual value of the optimal p_{best_i}, and global optimal value g_{best} indicate in i^{th} group for an optimal number of nodes and the complete group, correspondingly.

Algorithm 1: Pseudo code of Hybrid PSO-DNN
Number of the particle in a swarm α, cognitive coefficients c, Inertia weight w
Initial nodes of double hidden layers p, initial velocity U
Initial individual optimal fitness value $p_{fitness}$, Individual optimal value p_{best_i}
Using DNN, initial global optimal fitness value $g_{fitness}$ is computed; g_{best} global optimal value
While $g_{fitness} > 0$ do
For each $i = 1; \alpha$ do
Compute update $U_i \leftarrow wU_i[j] + c_1 r_1 (p_{best_i} - P_i(j)) + c_2 r_2 (g_{best} - P_i(j))$
Compute update $P_i \leftarrow P_{i-1} + U_i$
Using P_i, compute $p_{fit}(i)$ by DNN
If $p_{fit}(i) < p_{fit}(i-1)$ do
Update global optimal value; $p_{best_i} \leftarrow P_i$
If $p_{fit}(i) < g_{fit}$ do
Global optimal fitness value $g_{fit}(i) < p_{fit}(i)$ is updated
Global optimal value $g_{best} \leftarrow p_{best_i}$ is updated
End if
End if
End for
End while
Return g_{fit}, g_{best}

5. Results and Discussions

5.1. Experimental Procedure

In this section, the outcomes of the proposed PSO-based DNN classifier was discussed. Moreover, the proposed classifier performance was evaluated by means of the traditional classifiers with respect to the accuracy, FAR and FRR were shown.

Here, the data set description is based upon the ELSDSR which is considered as a database and it is selected for the simulation procedure that is obtained from [17]. For evaluation, the speech data is exploited and production of the automatic speaker identification system subsequent to the consideration of the ELSDSR speech corpus.

English is the text language, and 1 Icelander, 1 Canadian, and 20 Danes are done to read it. Because of the nonattendance of the recognized rehearsal, the prospect of attaining the ideal pronunciation is small. Nevertheless, it is not so significant to get the unique and precise characteristics of the individuals. '.wav' the file type is exploiting to record the voice message. PCM technique is chosen at a bit rate of 16 with the sampling frequency of 16 kHz.

ELSDSR boards "the voice messages of 22 speakers comprising of 12 males and 10 females of age variation from 24 to 63. Between them, many of the speakers are faculties and the Ph. D students, who are working at IMM, and four of them are the master students, and the other one is the international master student. The gender distribution is not regular at the site of the experiment, and thus, the average age of male speakers is lower than the average age of female speakers".

5.2. Performance Analysis

In this section, several traditional approaches, like LM, GMM, SVM, and BSW are evaluated with the proposed Hybrid PSO-based DNN classifier to express the efficiency of the proposed model in speaker recognition. Fig 3 exhibits the performance evaluation of the proposed model without and with noise in the signal with respect to the accuracy. Here, the proposed method is 15% better than the LM, 17% better than the GMM, 21% better than the SVM, 23% better than the BSW conventional models.

Fig 4 states the performance evaluation of the proposed model without noise in the signal with respect to the FAR and FRR. In Fig 5, the performance evaluation of the proposed model with noise in the signal with respect to the FAR and FRR is demonstrated. Here, the overall analysis states the proposed method is better than the conventional models.

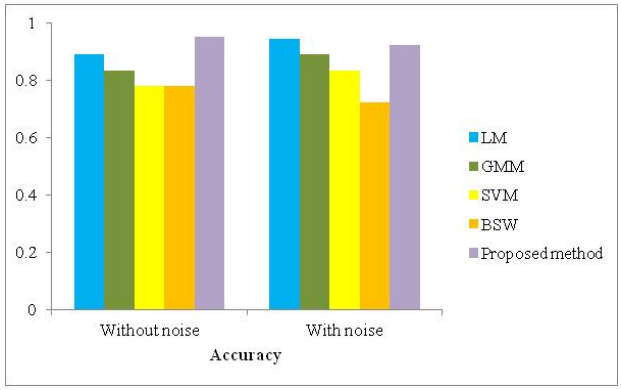


Fig. 3. *Performance evaluation of the proposed technique with and without noise in the signal with respect to the accuracy*

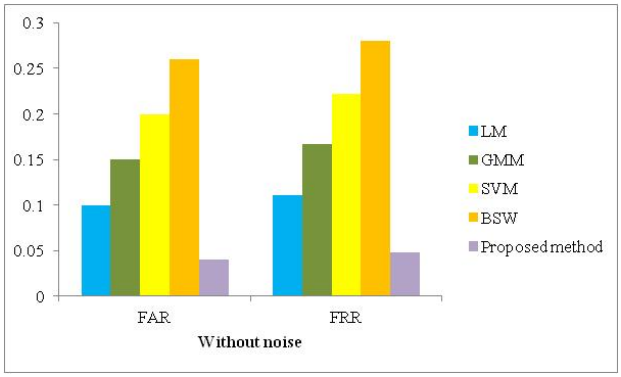


Fig. 4. *Performance evaluation of the proposed technique without noise in the signal with respect to the FAR and FRR*

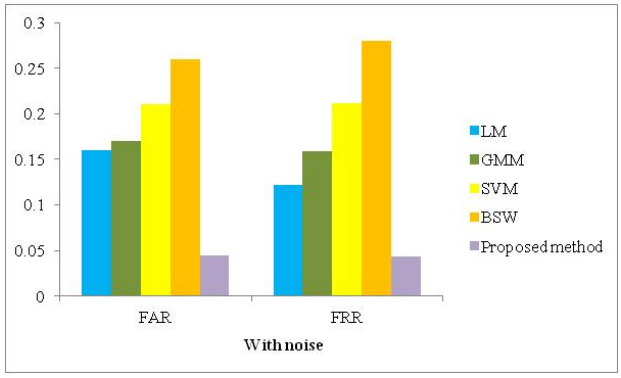


Fig. 5. *Performance evaluation of the proposed technique with noise in the signal with respect to the FAR and FRR*

6. Conclusion

In this work, Hybrid PSO- based DNN classifier was proposed, for recognition of the speaker by exploiting speech signals for that an automated speaker recognition method was proposed on the basis of the frequency-dependent features. For the proposed algorithm the speaker recognition procedure was

performed by exploiting 2 main steps, such as feature classification and extraction. At first, from the input speech signal, frequency-based features, like MKMFCC, spectral skewness, spectral kurtosis, and autocorrelation, were extracted. Moreover, the DNN classifier was exploited to classify the speech signal and recognizes the speaker. The classifier weight values were attained optimally by exploiting the proposed PSO method. Finally, the outcome attained exhibits that the proposed Hybrid PSO- based DNN classifier offers superior accuracy and minimum error value as evaluated to conventional techniques.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Tengyue Bian, Fangzhou Chen, Li Xu,"Self-attention based speaker recognition using Cluster-Range Loss", Neurocomputing, vol. 368, pp.59-68, , 27 Nov. 2019.
[2] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Najim Dehak,"State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations", Computer Speech & Language, vol. 60, March 2020.
[3] Shuping Peng, Tao Lv, Xiyu Han, Shisong Wu, Heyong Zhang,"Remote speaker recognition based on the enhanced LDV-captured speech", Applied Acoustics, vol. 143, pp. 165-170, 1 January 2019.
[4] Emma Jokinen, Rahim Saeidi, Tomi Kinnunen, Paavo Alku,"Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task",Computer Speech & Language, vol. 53, pp. 1-11, January 2019.
[5] Michael Jessen, Jakub Bortlík, Petr Schwarz, Yosef A. Solewicz,"Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)", Speech Communication, vol. 111, pp 22-28, August 2019.
[6] Ville Vestman, Dhananjaya Gowda, Md Sahidullah, Paavo Alku, Tomi Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction", Speech Communication, vol 99, pp. 62-79, May 2018.
[7] Ing-Jr Ding, Jia-Yi Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots", Computers & Electrical Engineering, vol. 62, pp. 719-729, August 2017.
[8] Johan Rohdin, Anna Silnova, Mireia Diez, Oldřich Plchot, Ondřej Glembek, "End-to-end DNN based text-independent speaker recognition for long and short utterances", Computer Speech & Language, vol. 59, pp. 22-35, January 2020.
[9] Ramaiah, V.S. Rao, R.R, "Speaker diarization system using MKMFCC parameterization and WLI-fuzzy clustering," International Journal of Speech Technology, vol. 19, no. 4, pp. 945–963.
[10] Valsalan, P. Manimegalai. S.O and Augustine, S," Non invasive estimation of blood pressure using a linear regression model from the photoplethysmogram (PPG) signal," Perspectivas em Ciencia da Informacao, vol. 22, no. 4, 2017.
[11] Vrabie, V. Granjon, P. and Serviere, C" Spectral kurtosis: from definition to application," In the proceedings on 6th IEEE International Workshop on Nonlinear Signal and Image Processing (NSIP), 2003.
[12] Auto content analysis, https://www.audiocontentanalysis.org/code/, 2018.
[13] Morales-Cordovilla, J.A. Peinado, A.M. Sánchez, V. González, J.A," Feature Extraction Based on Pitch-Synchronous Averaging for Robust Speech Recognition," IEEE transactions on audio, speech, and language processing, vol. 19, no. 3, pp. 640-651, 2011.
[14] Manoela Kohler, Marley M. B. R. Vellasco, Ricardo Tanscheit,"PSO+: A new particle swarm optimization algorithm for constrained problems", Applied Soft Computing, 21 Oct. 2019.
[15] J. Ren and S. Yang, ``An improved PSO-BP network model," Proc. 3rd Int. Symp. Inf. Sci. Eng. (ISISE), Shanghai, China, pp. 426-429, Dec. 2010.
[16] Ravi Kumar Vuddagiri, Hari Krishna Vydana, Anil Kumar Vuppala,"Curriculum learning based approach for noise robust language identification using DNN with attention", Expert Systems with Applications, Volume 110, 15 November 2018, Pages 290-297.
[17] English Language Speech Database for Speaker Recognition (ELSDSR) from http://www2.imm.dtu.dk/~lfen/elsdsr/