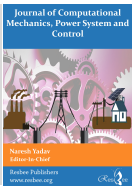
Resbee Publishers

Journal of Computational Mechanics, Power System and Control

Received 3 May, Revised 28 June, Accepted 14 July



Hybrid Particle Swarm Optimization-Gravitational Search Algorithm based Deep Belief Network: Speech Emotion Recognition

Dr. J Rajeshwar
Professor in CSE,
Guru Nanak Intitutions Technical Campus
Ibrahimpatnam, Hyderabad, India
rajeshwarj.2013@gmail.com

Abstract: One of the most important research areas is the Speech Emotion Recognition (SER) technique, which is applied in many fields such as speech processing and human-computer interaction. In general, it is mainly concentrated on using the techniques of machine learning in order to predict the precise emotional category from speech. In affective computing as well as the human-computer interaction, the developed applications of SER are very effective that are considered as the important module of the computer's next-generation system. It is due to the automatic service provisions are granted by the natural human-machine interface that requires an improved approval of user emotional conditions. Hence, this work proposes a novel SER model which integrated both emotion and gender recognition. Various features are extracted and that is fed for the emotions classifications. In this work, the Deep Belief Network (DBN) is exploited. At last, performance analysis of the developed technique is seen that better accuracy rate while comparing with the conventional models. This work proposes a novel technique for the SER model which helps both emotion as well as gender recognition. Here, the proposed Hybrid Particle Swarm Optimization (PSO) and Gravitational Search Algorithm (GSA) algorithm are introduced to identify the optimal weight of the DBN technique.

Keywords: DBN, Emotion, Gender, Human-Computer Interface, SER.

Nomenclature

Abbreviations	Descriptions
EEG	Electroencephalography
Bi-GRU	Bi-direction Gated Recurrent Unit
KLT	Kosambi-Karhunen-Loeve transform
LR	Logistic Regression
HCI	Human–Computer Interaction
STFT	Short time Fourier Transform
INCA	Iterative Neighborhood Component Analysis
MLP	Multilayer Perceptron
PSD	Power Spectral Density
MSE	Mean Squared Error
UAR	unweighted average recall
NMF	Non-Negative Matrix Factorization
ECFW	Emotional-Category Based Feature Weighting
SVM	Support Machine Vector
CRNN	Convolutional Recurrent Neural Network
CM	Confusion Matrix
MSDA	Multi-Scale Discrepancy Adversarial
RBM	Restricted Boltzman Machine
PSD	Power Spectral Density
MFCCs	Mel-Frequency Cepstral Coefficients
PSO	Particle Swarm Optimization
LPF	Low Pass Filtering
MLP	Multi-Layer Perceptron
GSA	Gravitational Search Algorithm
UAR	Unweighted Average Recall
IR	Impulse Response
TQWT	Tunable Q wavelet transform
WAR	Weighted Average Recall

https://doi.org/10.46253/jcmps.v4i3.a4

1. Introduction

In speech emotion recognition research, a high-quality emotional corpus is a vital prerequisite. Nevertheless, it's found that numerous emotional corpora are more or less unsatisfactory in the course of practice, and there are basically issues of inconsistent duration of samples and unbalanced sample classifications. Speech signal comprises a lot of information which increases the written message content involving metrics like speaker identity, emotional state, the status of information, and intentional patterns [1]. Usually, by exploiting the algorithms, the speech recognition system involves the speech features behind the speech signal and recognizes their content. Various researches were performed on speech recognition since the late 1950s which attained important improvement. Similarly, to attain a superior HCI experience, authorizing machines with emotional expressive capability and creating them able to recognize the human emotional state has received huge popularity in the HCI fields [2]. Therefore, an SER research field is introduced, which is represented as a significant research field in the last decade. By the automated analysis of human affective behavior, several researchers are concentrated. Nevertheless, despite important efforts done using speech recognition, to attain huge natural communications among machines and humans, the SER needs extensive work.

In humans, to detect the emotional changes the emotion recognition fields are exploited by employing up-to-date smart systems. By exploiting the facial images, EEG signals, and speech signals, automatically emotions can be recognized. Moreover, to ascertain the emotional state, the SER is exploited which is the employ of effectual features of speech signals. To express emotions as well as human mood, speech is representing as a physiological signal. During the speech, because of the environmental factors, by mental states, the people are affected. In humans, this communication leads to diverse emotional states. People react to diverse circumstances during communication with feelings namely happiness, neutral, sadness, and anger. In decision-making mentally, emotions play an important role and these circumstances lead the speech to alter [3].

Besides the signals, the speech signal possesses distinctive series patterns as well as emotional information which are unevenly distributed. In order to detain the temporal clues of speech in numerous existing researches, the SER is indicating a static or dynamic classification in proportion to two timescales, such as the complete and sub utterance. In aforesaid two timescales, based on the domain alignment it can be individually implemented. Nevertheless, this method might not be adequately strong to handle the analyzed maximum variance [4].

SER is considered as a machine learning area that set up the technique for the human emotions classification on the basis of the speech itself. Emotions possess numerous modalities which can be apparent using the facial expression, psychological signals, and speech so on. Speech signal exhibits huge versatility while comparing with the other modalities and it is simple to obtain

The most important aim of this research is to present a novel SER technique that involves two kinds of recognition such as emotion and gender recognition. From input speech signal, pitch feature is extracted for gender recognition, in order to classify the gender, and extracted features are fed to NN based classifier. Here, both pitch features and Non-Negative Matrix Factorization (NMF) are extracted for emotion recognition, and this work exploits the DBN classifier to classify emotion. The corresponding emotions are given by the classifier. In addition, exploiting a novel Hybrid PSO-GSA technique, DBN weight is optimally chosen, so that the recognition accuracy rate is higher.

2. Literature Review

In 2020, Zijiang Zhu et l [1], developed an SER technique on the basis of the Bi-GRU and Focal Loss. The technique was enhanced based on the learning CRNN intensely. In order to lengthen the speech samples efficiently, the Bi-GRU was exploited with minimum time as well as to tackle the classification issues the focal loss function was exploited which occur using the imbalance of emotional samples classifications. Using various techniques for evaluation analysis, CM, UAR, and WAR, UAR was exploited as an evaluation index of the technique.

In 2020, Wanlu ZHENG et al [2], introduced a new MSDA network to conduct multiple timescales domain version for cross-corpus SER, that was the integration of domain discriminators of hierarchical levels into emotion recognition model to alleviate the breach among source as well as target domains. Particularly, two types of speech features were extracted such as deep features, as well as handcraft features from three timescales of local, global, as well as hybrid levels.

In 2020, Turker Tuncer et al [3], developed a nonlinear multi-level feature generation technique exploiting a cryptographic technique. To choose the features, the innovation of this paper exploited a cryptographic structure named shuffle box for the iterative neighborhood and feature generation module analysis. The developed technique possesses three important phases such as multi-level feature

generation exploiting TQWT, for feature generation twine shuffle pattern (twine-shuf-pat) and by exploiting the INCA, the discriminative features were chosen and classified.

In 2020, Dongdong Li et al [4], developed an ECFW technique that aspires to find the importance of every feature in diverse emotions and using this importance as a priori knowledge. Here, diverse integration of models was explained, and features consequence in large differences in the SER performance that were estimated using various simulations. To extract the most precious information features should be designed with suitable techniques for emotional demonstration.

In 2021, Reinert Yosua Rumagit et al [5], examined the dissimilarity in model performance by exploiting the MLP, SVM, and LR with MFCCs on the Indonesian language. In the dataset, several people's voices were recorded that were gathered by exploiting a peer-to-peer method.

3. Proposed SER Model

The architecture model of the proposed SER model is illustrated in Fig 1. Here, it involves two objectives such as emotion recognition and gender recognition. In a precise manner, input speech signal $s(n)$, the recognition of both the emotion and gender has been done. From the input signal, pitch features $pt_t(n)$ are extracted, for gender recognition. Furthermore, by exploiting the NN classifier, the classification is performed and presents the ouput1 represents male, 2 represents female. At first, the features such as NMF $nmf_t(n)$ as well as pitch $pt_t(n)$ are extracted for the emotion recognition, which is subjected to the PCA in order to reduce their dimension. Moreover, the dimension reduction $nmf_t^d(n), pt_t^d(n)$ is subjected to the classification.

In this work, the DBN classifier is exploited to classify emotions, as well as the classified output may be Calm, Angry, Disgust, Neutral, Happy, Fear, Sad as well as Surprise.

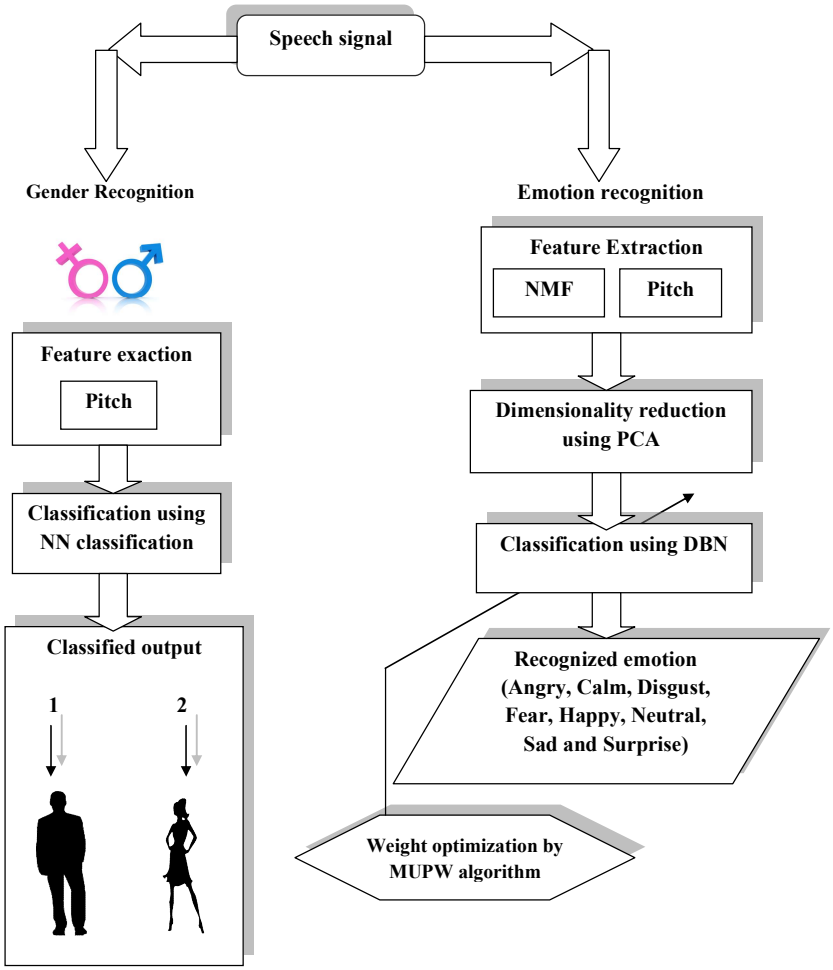


Fig 1: *Architecture of developed speech recognition technique*

4. Summary of Gender Recognition Model

In this SER model, the gender recognition technique is considered as the initial objective. Here, 2 major stages namely feature extraction and classification is performed. Initially, the $pt_t(n)$ feature will be extracted from the input signal, $s(n)$ to the NN classifier, which is subjected as the input, whereas it presents the classified output as female or male.

4.1 Pitch Feature Extraction [9]

In the evaluation of the pitch feature, some steps are presented, which are stated as follows:

a) Initially, by exploiting the STFT $FT_t(f)$, $s(n)$ is altered to a time-frequency domain as well as it is stated in Eq.(1), in that $PW_t(f)$ states PSD of aggravating noise as well as $s(n)$, that is H^{th} harmonic power t and f_0 states the time and frequency for accurate periodic source.

$$FS_t(f) = \sum_{h=1}^{H} P(n)\delta(f - hf_0) + PO_t(f) \tag{1}$$

b) Eq.(2) represents the PSD of the total frame on Log spaced frequency grid, $FT_t(g)$, where, $g = \log f$. The spacing of harmonic is released to f_0 as well as the convolution of $FT_t(f)$ with IR $M(g)$ energy is connected, as well as it is stated in Eq.(3).

$$FT_t(g) = \sum_{h=1}^{H} s(n)\delta(g - \log h - \log f_0) + PW_t(g) \tag{2}$$

$$M(g) = \sum_{h=1}^{H} \delta(g - \log H) \tag{3}$$

c) Evaluate $\alpha_t(g)$ as stated Eq. (4). To state compression exponent $\alpha_t(g)$, the initial procedure is to estimate the $\overline{F}T_t(g)$ smoothened spectrum using the LPF $FT_t(g)$ in log-frequency as well as time $TM(g)$. Both $\overline{F}T_t(g)$ as well as $FT_t(g)$ normalize to the power of $TM(g)$ as well as sets $\alpha_t(g)$. Hence, $\overline{F}T_t(g)$ equivalents $TM(g)$ as well as evaluation of compressed PSD $FT_t'(g)$ is stated in Eq.(5)

$$\alpha_t(g) = \frac{\log TM(g)}{\log \overline{F}T_t(g)} \tag{4}$$

$$FT_t'(g) = FT_t(g)^{\alpha_t(g)} \tag{5}$$

d) Finally, $FT_t'(g)$ is convolved by means of $M(g)$ as well as chooses uppermost peak in rational range as estimated pitch $pt_t(n)$.

$$pt_t(n) = \{FT_t'(g) * M(g)\} \tag{6}$$

4.2 Classification using Neural Network

To classify the gender as aforesaid, this work exploits NN [10] classifier. For this classifier, the input is extracted $pt_t(n)$. In Eq. (7), (8), and (9), the network model is stated the hidden neuron is stated as l, $w_{(\overline{Bl})}^{(HI)}$ signifies bias weight to l^{th} hidden neuron, $w_{(kl)}^{(HI)}$ signifies k^{th} input to l^{th} hidden neuron, N_I states count of input neurons, N_h states count of hidden neurons, $w_{(lj)}^{(o)}$ states the output weight from l^{th} hidden neuron to j^{th} layer, $w_{(\overline{Bj})}^{(o)}$ states output bias weight to j^{th} layer, and NF states he activation function. In eq. (8), network output $\hat{O}_j$ is given, whereas H_j indicates actual output, $\left|H_j - \hat{H}_j\right|$ indicates error amid actual as well as predicted output.

$$IN^{(HI)} = NF\left(w_{(\hat{Bl})}^{(HI)} + \sum_{k=1}^{N_j} w_{(kl)}^{(HI)} pt_t(n) \right) \tag{7}$$

$$\hat{O}_j = NF\left(w_{(\overline{Bj})}^{(o)} + \sum_{l=1}^{N_h} w_{(lj)}^{(o)} IN_l^{(HI)} \right) \tag{8}$$

$$w^{*} = \underset{\left\{w_{(Bl)}^{(HI)}, w_{(kl)}^{(HI)}, w_{(Bj)}^{(o)}, w_{(lj)}^{(o)}\right\}}{\arg\min} \sum_{j=1}^{N_o} \left| O_j - \hat{O}_j \right| \tag{9}$$

The classifier outputs particular gender, if it is female or male.

5. Summary of Emotion Recognition Technique

Emotion recognition is considered as the second contribution. The model extracts the features such as pitch feature and NMF feature to identify the emotion. By exploiting the PCA model, features extracted are given to the dimensionality reduction. Subsequently, the dimension reduction $pt_t^d(n)$ and $nmf_t^d(n)$ is subjected to input to the DBN classifier, whereas it classifies relevant emotion.

5.1 Feature Extraction

NMF feature [8]: Consider non-negative data vector of $s(n)$ be $NV \in R^{1\times sa}$; $1\times sa$ states 1-D samples. Furthermore, the most important intention of NMF [6] is to decide 2 matrices (non-negative), which is stated in Eq. (10) and Eq. (11), whereas di shows the dimensional space.

$$X \in R^{1\times di} \tag{10}$$

$$Y \in R^{di\times sa} \tag{11}$$

Moreover, using the non-negative constrictions the parts-based revelation is increased as they agree only preservative other than not either combinations or subtractive. The renowned stated cost functions are represented as follows that is required discovering the approximate factorization.

a) In Eq. (12), the square Euclidean distance amid NV and XY is stated, whereas $||\cdot||_E$ signifies matrix Frobenius norm.

$$||NV - XY||_E = \sum_{l=1}^{sa} ||NV_l - (XY)_l||^2 \tag{12}$$

b) Eq. (13) represents the Kullback-Leibler deviation DI between NV and XY.

$$DI(NV||XY) = \sum_{l=1}^{sa}\left(NV_l \log\frac{NV_l}{(XY)_l} - NV_l + (XY)_l\right) \tag{13}$$

Further, by multiplicative technique, cost functions are solved. Eq. (14) as well as (15) exhibits square Euclidean distance updating of Y_l as well as X_l (non-negative matrices). Likewise, Eq. (16) and (17) represent the update rule for deviation of Y_l as well as X_l, whereas $l = 1....s$. Eq. (18) indicates the ensuing NMF feature $nmf_t(n)$.

$$Y_l \leftarrow \frac{\left(X^T NV\right)_k}{\left(X^T XY\right)_l} Y_l \tag{14}$$

$$X_l \leftarrow \frac{\left(NVY^T\right)_l}{\left(XYY^T\right)} Y_l \tag{15}$$

$$Y_l \leftarrow \frac{\sum_m X_m NV_{ml}/(XY)_{ml}}{\sum_m X_m} Y_l \tag{16}$$

$$X_l \leftarrow \frac{\sum_m X_{lm} NV_l/(XY)_{ml}}{\sum_m Y_{lm}} X_l \tag{17}$$

$$nmf_t(n) = (X_l, Y_l) \tag{18}$$

Pitch Feature: The clear interpretation of the pitch feature $pt_t(n)$ is stated in the aforesaid section.

5.2 Dimensionality reduction

This developed technique involves PCA [7] technique to reduce the extracted dimensions $nmf_t(n)$ and $pi_t(n)$. A cluster of DA data vectors $a_1,.....,a_{DA}$ is chosen from the mined features, and a_o exhibits the unique group observation of *va* variables. Then, assessment of empirical mean is carried out with all column $cl = 1,.....,va$, the relevant mean value is stated in $e_{(va)}$ with $va\times 1$ dimensions,

which is ascertained in Eq.(19). Moreover, as Eq. (20), the evaluation of mean difference is carried out, in that G denotes $DA \times va$ matrix and ja denotes column vector $DA \times 1$ of all 1s: $ja[i] = 1$.

$$e_{(va)}[cl] = \frac{1}{DA}\sum_{i=1}^{DA} R[i, cl] \tag{19}$$

$$G = R - jacl^{T} \tag{20}$$

Eq. (21) exhibits covariance matrix evaluation $CV^{(MA)}$. Both eigenvalues, as well as eigenvector evaluation, are handled using valuing the matrix MA as well as diagonalizes $CV^{(MA)}$, which is exhibited in Eq.(22).

$$CV^{(MA)} = \frac{1}{DA-1} G^{*}.G \tag{21}$$

$$MA^{-1} CV^{(MA)} MA = DM^{(MA)} \tag{22}$$

$DM^{(MA)}$ shows eigenvalues diagonal matrix of $CV^{(MA)}$. By sorting the eigenvector matrix column, eigenvalue matrix minimization of $DM^{(MA)}$ is attained. Furthermore, as per Eq. (23), the cumulative energy content formulation $cm^{(en)}$ is stated that holds the energy content augmented of eigenvalues from one through cl.

$$cm^{(en)}[cl] = \sum_{DA=1}^{cl} DM^{(MA)}[DA, DA], \text{ for } cl = 1,.....va \tag{23}$$

In addition, the subset selection is occurred by storing cl column of MA as C matrix. Moreover, using *ve* vector the cl cost is chosen. The $\overline{MA}$ column, $\overline{MA} = Q.C = KLT\{RW\}$ is a vector, that ascertains KLT in row matrix, RW . Eq. (24), states the determination of Q matrix, whereas $ta = ta(cl) = \left\{\sqrt{CV^{(MA)} \mid cl, cl \mid}\right\}$; $cl = 1....va$ and that is represented as an output signal, dimensionally reduced speech signal $s_d(n)$.

$$Q = \frac{G}{ja.ta^{\overline{MA}}} \tag{24}$$

For both the NMF and pitch feature the PCA model is independently used, and as stated in eq. (25) $s_d(n)$ is determined, in that the dimensional minimization of NMF is stated as $mnf_t^d(n)$, and the dimensional minimized pitch feature is stated as $pt_t^d(n)$.

$$s_d(n) = \left\{mnf_t^d(n), pt_t^d(n)\right\} \tag{25}$$

Hence, the output $s_d(n)$ will be input to the DBN classification procedure, whereas speech emotions are routinely identified.

5.3 DBN based classification

This paper exploits the DBN classifier, for emotion recognition which is the second contribution. Generally, the DBN [6] is a well-known intelligent model, the model consists of multilayers, as well as each layer involves visible neurons which use both input layer and hidden neurons. For the input, this particular neuron model ascertains a precise output. Eq. (26) states the output as well as Eq. (27) states feasibility in sigmoid-shaped function, in that t^P states pseudo-temperature. Eq. (28) states the deterministic method of stochastic technique.

$$\overline{OP}_q(\zeta) = \frac{1}{1 + e^{\frac{-\zeta}{t^P}}} \tag{26}$$

$$\overline{PY} = \begin{Bmatrix} 1 & \text{with } 1 - \overline{OP}_q(\zeta) \\ 0 & \text{with } \overline{OP}_q(\zeta) \end{Bmatrix} \tag{27}$$

$$\lim_{t^P \to 0^+} \overline{OP}_q(\zeta) = \lim_{t^P \to 0^+} \frac{1}{1 + e^{\frac{-\zeta}{t^P}}} = \begin{cases} 0 & \text{for } \zeta < 0 \\ \frac{1}{2} & \text{for } \zeta = 0 \\ 1 & \text{for } \zeta > 0 \end{cases} \tag{28}$$

In the DBN model, using a set of RBM layers, the feature extraction procedure occurs, and using MLP classification procedure is performed. The arithmetic model exhibits the Boltzmann machine energy for

neuron formation or binary state bi that is stated in Eq. (29), whereas $w_{a,l}$ indicates weights among neurons, θ_a indicates biases.

$$\Delta EY(bi_a) = \sum_l bi_a w_{a,l} + \theta_a \tag{29}$$

Eq. (30), (31), and (32) states the procedure of energy regarding the joint composition of visible and hidden neurons (x,y). Moreover, x_a indicates either binary or neuron state of a visible unit, BI_l indicates the binary state of l hidden unit, as well as k_a.

$$EY(x,y) = \sum_{(a,l)} w_{a,l} x_a y_l - \sum_a k_a x_a - \sum_l BI_l y_a \tag{30}$$

$$\Delta EY(x_a, \bar{y}) = \sum_l w_{al} y_l + k_a \tag{31}$$

$$\Delta EY(\vec{x}, y_a) = \sum_l w_{al} x_a + BI_l \tag{32}$$

$$\hat{w}(\hat{M}) = \max_{\hat{w}} \prod_{\vec{x} \in N} c(\vec{x}) \tag{33}$$

Eq. (34) represents the possibility of allocated RBM model for hidden as well as visible vectors pair $(\vec{x}, \overrightarrow{hi})$, whereas PA^F represents partition function, as well as it is stated in Eq. (35).

$$c(\vec{x}, \overrightarrow{hi}) = \frac{1}{PA^F} e^{-EY(\vec{x},\vec{y})} \tag{34}$$

$$PA^F = \sum_{\vec{x},\vec{y}} e^{-EY(\vec{x},\vec{y})} \tag{35}$$

$$w'_{a,l} = \Delta w_{a,l} + w_{a,l} \tag{36}$$

Ahead of the happening learning procedure of MLP technique, let $\left(K^{\hat{C}}, L^{\hat{C}}\right)$ as training patterns, in that $\hat{C}$ indicates the count of training patterns, $1 \le \hat{C} \le \overline{OP}$.

$$e_l^{\hat{C}} = K^{\hat{C}} - L^{\hat{C}} \tag{37}$$

Hence, Eq. (40) ascertains squared error of $\hat{C}$ pattern pursued using MSE, as well as it is stated as Eq. (41).

$$ME_{\hat{C}}^{mean} = \frac{1}{\tilde{o}_y} \sum_{l=1}^{\tilde{o}_y} \left(e_l^{\hat{C}}\right)^2 = \frac{1}{\tilde{o}_y} \sum_{l=1}^{\tilde{o}_y} \left(K^{\hat{C}} - L^{\hat{C}}\right)^2 \tag{38}$$

$$ME_{avg} = \frac{1}{\tilde{P}} ME_{\hat{C}}^{mean} \tag{39}$$

6. Proposed Hybrid Model

6.1 Objective Function

In the DBN model, the weight is optimally chosen, hence, this work develops a novel hybrid optimization technique called the PSO-GSA algorithm which discovers the optimal weight helping in increasing recognition accuracy. The input to the proposed method is the weight (arbitrary weights), from that optimal weight will be chosen hence recognition accurateness obtains maximized. Eq. (40), indicates the most important objective of the developed model.

$$OB = \max(\text{recognition accuracy}) \tag{40}$$

6.2 Proposed PSO-GSA

In this paper, PSO and GSA are hybridized to avoiding premature convergence to local optima and present appropriate tradeoff among exploitation as well as exploration capabilities and can implement the Speech Emotion Recognition.

Generally, PSO algorithm is inspired from birds' and fish's flocking model. A general model of PSO algorithm consist of particles (Search Agents) indicating the solutions of an optimization issue.

GSA is a meta- heuristic approach of optimization on the basis of the laws of gravity. According to Newton's basic laws of gravity whereas every object of mass exerts a force on every other object of mass called as a gravitational force.

In local search, the GSA is better; however, PSO needs to obtain near to optimum values efficiently after that some optimal solution issues are faced by the GSA.

The developed Hybrid PSO-GSA initiates with separating the whole population into two sub-categories [11].In these two categories, one category is followed by the PSO and the other one is by GSA. In both categories, the populations are arbitrarily initialized. In the proposed approach strategy of the migrant is used in that the optimal solution in PSO restores GSA poor solution, as well as the optimal solution of GSA, restores poor solution of PSO. This substitution is performed subsequent to the evaluation of the optimal and poor solution of each category. Ahead of the migration, the location equivalent to the optimal solution from each category that is both the sub-populations experience crossover. It is used amid PSO's and GSA optimal solution locations and that are described in eq. (41).

$$C_{i,n}^{c}(t) = i \times g_{best_{i,n}}(t) + (1-i)x_{b}^{n}(t) \tag{41}$$

$C_{i,n}^{c}(t)$ indicates the search agents' location post crossover, $i \in (0,1)$. The optimal and crossover location experiences mutation, Cauchy mutation is used in Eq (42)–(44):

$$C_{i,n}^{m}(t) = C_{i,n}^{c}(t) + (ub_n(t) - lb_n(t)) \times \text{Cauchy}(0,S) \tag{42}$$

$$g^{m}{}_{best_{i,n}}(t) = g_{best_{i,n}}(t) + (ub_n(t) - lb_n(t)) \times \text{Cauchy}(0,S) \tag{43}$$

$$x_{b}^{n,m}(t) = x_{b}^{n}(t) + (ub_n(t) - lb_n(t)) \times \text{Cauchy}(0,S) \tag{44}$$

where, $lb_n(t)$ and $ub_n(t)$ indicates lower and upper bounds of decision variables in n^{th} dimension and at the t^{th} iteration. $\text{Cauchy}(0,S)$ indicates the Cauchy distribution with mean 0 as well as scaling parameter S. The scaling parameter minimizes linearly with iteration as $S(t+1) = S(t) - \frac{1}{t_{max}}$ and $S(1)$ is equal to 2. For each location, the Crossover location, and optimal location of PSO and GSA are performed. In the GSA environment, the PSO search agents obtain mass value allocated on the basis of their fitness as stated in Eq. (45) and (46).

By exploiting the Mass value in (45), the force, velocity, acceleration, as well as locations are computed.

$$m_{P(i,n)}^{mig}(t) = \frac{fit_{P(i,n)}(t) - w'(t)}{b(t) - w'(t)} \tag{45}$$

$$M_{P(i,n)}^{mig}(t) = \frac{m_{P(i,n)}^{mig}(t)}{\sum_{l=1}^{N} m_l(t)} \tag{46}$$

On basis of the Eq. (47) and (48) the migrated particle's position, and velocity, are updated.

$$v_{(i,n)}^{mig}(t+1) = W * v_{(i,n)}^{mig}(t+1) + c_1' r_1 \begin{pmatrix} p_{best(i,n)}^{mig}(t) - p_{(i,n)}^{mig}(t+1) \\ + c_2' r_2 \left(g_{best(i,n)}^{mig}(t) - g_{(i,n)}^{mig}(t) \right) \end{pmatrix} \tag{47}$$

$$p_{(i,n)}^{mig}(t) = p_{(i,n)}^{mig}(t) + v_{(i,n)}^{mig}(t) \tag{48}$$

7. Result and Discussion

The proposed emotion recognition model was developed and the employed dataset consists of 8 emotions namely Fear, Angry, Disgust, Calm, Neutral, Happy, Sad, and Surprise. The performance of the proposed method was analyzed with the existing techniques such as Genetic Algorithm (GA), Particl)e Swarm Optimization (PSO), Firefly (FF), and Whale Optimization Algorithm (WOA regarding certain measures namely Accuracy, Precision, Sensitivity, False Negative Rate (FNR), Specificity, False Discovery Rate (FDR), False Positive Rate (FPR), NPV, Matthews correlation coefficient (MCC) as well as F_1Score to corroborate the performance of the proposed model.

Fig 2 exhibits the performance evaluation of the developed technique against existing techniques. In this figure, the proposed method is 20% better than the GA, 20% better than the PSO, 15% better than the FF, 13% better than the WOA in terms of accuracy. Here, the performance outcomes demonstrate the superiority of the developed method over the conventional techniques.

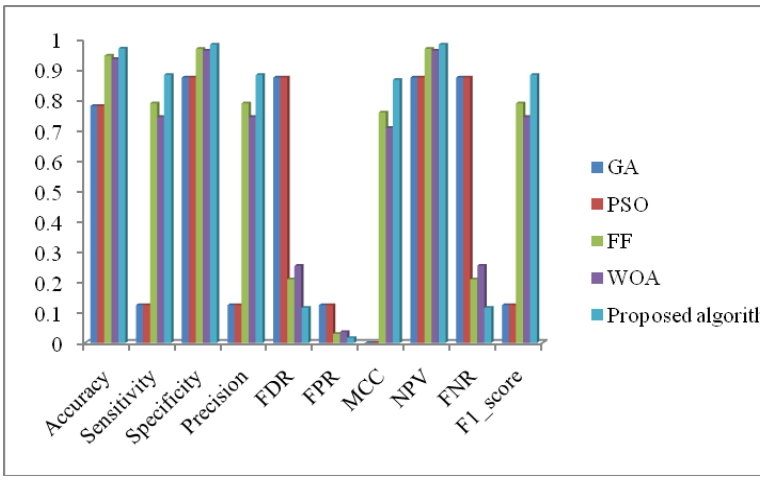


Fig 2: *Analysis of developed technique with existing techniques*

8. Conclusion

A new SER recognition model was proposed with two important objectives such as emotion recognition and gender recognition. The pitch features were extracted in order to classify the gender and that was fed to the NN classifier. From the given input signal, the corresponding gender has been classified by exploiting the classifier. Initially, the features such as pitch and NMF were extracted for the emotion recognition and the PCA model was exploited for extracted features for dimension reduction. In order to recognize the corresponding emotions, the dimensional minimized features were subjected to the DBN technique. Moreover, the DBN model weight was optimally chosen using a hybrid Particle Swarm Optimization - Gravitational Search Algorithm optimization approach named Hybrid PSO-GSA technique. At last, the performance of the proposed method was analyzed over the existing techniques as well as superior consequences were obtained using a developed technique with the maximum accuracy rate.

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] Zijiang ZhuWeihuang DaiJunshan Li,"Speech emotion recognition model based on Bi-GRU and Focal Loss", Pattern Recognition Letters,11 November 2020.
[2] Wanlu ZhengWenming ZhengYuan Zong,"Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition",Virtual Reality & Intelligent Hardware24, February 2021.
[3] Turker TuncerSengul DoganU. Rajendra Acharya,"Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques", Knowledge-Based Systems,31 October 2020.
[4] Dongdong LiYijun ZhouDaqi Gao," Exploiting the potentialities of features for speech emotion recognition", Information Sciences,8 October 2020.
[5] Reinert Yosua RumagitGlenn AlexanderIrfan Fahmi Saputra,"Model Comparison in Speech Emotion Recognition for Indonesian Language",Procedia Computer Science,19 February 2021.
[6] Kasiprasad Mannepalli, Panyam Narahari Sastry and Maloji Suman, "A novel Adaptive Fractional Deep Belief Networks for speaker emotion recognition", Alexandria Engineering Journal, October 2016.
[7] S. C. Ng, " Principle component analysis to reduce dimension on digital image", Procedia Computer Science, vol. 111, pp. 113–119, 2017.
[8] Lijun ZHANG, Zhengguang CHEN, Miao ZHENG and Xiaofei HE, " Robust non-negative matrix factorization", Front. Electr. Electron., vol.. 6, no.. 2, pp. 192–200, 2011.
[9] Sira Gonzalez and Mike Brookes, " A Pitch Estimation Filter Robust To High Levels Of Noise (PEFAC)", 19th European Signal Processing Conference, 2011.

[10] Yogeswaran Mohan, Sia Seng Chee, Donica Kan Pei Xin and Lee Poh Foong, " Artificial Neural Network for Classification of Depressive and Normal in EEG", IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2016.

[11] S. Mirjalili, S.Z.M. Hashim, A new hybrid PSO-GSA algorithm for function optimization, in: International Conference on Computer and Information Application, IEEE, 2010, pp. 374–377.

[12] Cristin R,Gladiss Merlin N.R,Ramanathan L,Vimala S,"Image Forgery Detection Using Back Propagation Neural Network Model and Particle Swarm Optimization Algorithm", Multimedia Research, vol 3, no. 1, January 2020.

[13] R. Cristin,Dr.V.Cyril Raj and Ramalatha Marimuthu,"Face Image Forgery Detection by Weight Optimized Neural Network Model",Multimedia Research, vol 2, no. 2, April 2019.