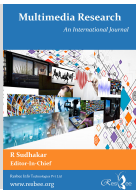
Vol.2 No.4 Oct.2024

Multimedia Research

Received 28 August, Revised 24 September, Accepted 14 October



Emotion Recognition from Speech Signals Using DCNN with Hybrid GA-GWO Algorithm

Raviraj Vishwambhar Darekar
A. G. Patil Institute of Technology,
Solapur,Maharashtra, India
ravirajvishwambhardarekar@gmail.com

Ashwinikumar Panjabrao Dhande
Pune Institute of Computer Technology,
Pune, Maharashtra, India

Abstract: In recent days, from the speech signal the recognition of emotion is considered as an extensive advanced investigation subject because the speech signal is considered as the rapid and natural method to communicate with humans. Numerous examinations have been progressed related to this topic. This paper develops the emotions recognition from the speech signal in an accurate way, with the knowledge of numerous examined models. Therefore, to study the multimodal fusion of speech features, a Deep Convolutional Neural Network model is proposed. Moreover, the hybrid Genetic Algorithm (GA)-Grey Wolf Optimization (GWO) algorithm is presented that is the combination of both the GA and GWO technique features towards training the network. Finally, the developed recognition model is verified and compared with the existing techniques in correlation with diverse performance measures such as Accuracy, Sensitivity, Precision, Specificity, False Positive Rate (FPR), False Discovery Rate (FDR), False Negative Rate (FNR), F1Score, Negative Predictive Value (NPV), and Matthews correlation coefficient (MCC).

Keywords: Speech; Recognition Model; Emotion; Neural Network; Optimization

Nomenclature

Abbreviations	Descriptions
DAE	Denoising Auto Encoder
GWO	Grey Wolf Optimization
IBM	Ideal Binary Mask
GA	Genetic Algorithm
PSD	Power Spectral Density
LSTM	Long Short-Term Memory
FT	Fourier transform
IRM	Ideal Ratio Mask
DNN	Deep Neural Networks
UDS	Ultrasonic Doppler Signal
DFT	Discrete Fourier transform
MMSE	Minimum Mean-Square Error
STFT	Short-time Fourier Transform
SNR	Signal-to-Noise Ratio
ResLSTM	Residual Long Short-Term Memory
ZNCC	Zero-mean Normalized Correlation Coefficient
PASTFT	Pitch-Adaptive STFT
MLSE	Machine-Learning Spectral Envelope
WF	Wiener filter
DCNN	Deep Convolutional Neural Network
PSO	Particle Swarm Optimization
IR	Impulse Response

1. Introduction

In several day-to-day life speech communication technologies, speech enhancement is considered as specific attention, whereas noise environments challenges urge for a consistent speech processing device performance [1]. For speech enhancement when important attempts have been committed within the past decades, in order to find innovative ideas to push the restriction of the attainable noise reduction.

https://doi.org/10.46253/j.mr.v2i4.a2

For instance, in a cocktail party case, attracts a huge amount of attention with many possible applications in hearing technologies and mobile phones.

By humans or enhance decoding by systems the noisy speech signals processing is to enhance their perception is how speech enhancement deals with. A mode of speech enhancement techniques is to enhance the system performance while the input of the speech is harmed by noise [2]. Usually, it is difficult to keep speech undistorted when minimizing noise and hence, restrictions on the performance of the speech enhancement systems, the compromise among noise reduction and speech distortion [3]. To high SNR, for distorted speech with medium, the objective will be able to create subjectively natural signal by minimizing noise level and for those with minimum SNR [6]. Moreover, when preserving the intelligibility, the purpose of this technique could be to minimize the noise level. The general factor which creates the speech's quality degradation and intelligibility is environment noise that can be non-stationary or stationary and is understood to be additive and uncorrelated to the speech signal [4]. An extensive categorization of speech enhancement techniques can be specified as temporal processing and spectral processing techniques. In frequency domain, the degraded speech leaves during processing in the spectral processing techniques. However, for temporal processing technique, the processing will be in time domain [5].

The research concentrates on the speech enhancement community, which turns into deep learning techniques. Recently for the speech enhancements, the deep learning techniques have been utilized, which has shown outstanding performance [8]. The NN was exploited for clean speech spectra from noisy speech spectra [18] [19] as non-linear maps. By utilizing clean speech and noisy pairs, a DAE was pretrained for this task [9]. A non-causal NN clean speech spectrum estimator was presented which generated improved speech with maximum objective quality scores [10], that later integrated multi-objective IBM-based post-processing [10]. In order to calculate the time-frequency masks, the neural networks were exploited. Recently, an LSTM network was utilized to calculate the IRM [11]. It is significant to notice that both noise and speech signals are extremely non-stationary, that is, the signal energy distribution over frequency, which is not constant in time. Moreover, the smoothing operations, which does not consider, might produce biased approximates.

To ease this restriction, a supervised learning technique was exploited whereas some hours of the training data were frequently exploited to calculate the association among noisy signals as well as equivalent clean signals [12]. By real data, the signal estimation rules are constructed, also discrepancy among the expected and actual model can be minimized. It tends to enhance speech enhancement performance were adequate training data state is present. Lately, in order to accurately model the clean speech spectrum, a deep autoencoder framework was used [13].

The main objective of this paper is to develop a model for recognize the emotions from the speech signals in an accurate manner. Hence, a DCNN model is presented in order to study the multimodal fusion of speech features. Here, the speech features are prosodic features, cepstral, and NMF. Multimodal features can represent each speech signal at unique margin that is extremely suggested by numerous researchers.

2. Literature Review

In 2019, Ki-Seung Lee et al [1], examined the utilization of the ultrasonic doppler frequency shifts, which occurred using movements in facial. It was used for improving the audio speech adulterated using high levels of acoustic noise. At first, the conventional signals were demodulated and it was transformed into a parameter of the spectral feature. From the UDS, the spectral feature was derived, and from noisy speech, it was concatenated with spectral features that were subsequently exploited to estimate the spectrum magnitude of clean speech. In this estimation, a nonlinear regression technique was exploited whereas the connection among audio-UDS features as well as the equivalent clean speech was indicated using DNN.

In 2019, Johannes Stahl and Pejman Mowlaee [2] presented a PASTFT model in order to attain a signal-dependent time-frequency depiction for the input signal. The association of inter-frame for the consecutive speech was analyzed, the DFT bins occurring from the harmonic signal modeling and the PASTFT. If the phase advanced development using the harmonic nature of the speech signal was considered and this analysis shows notable correlation. Therefore, consecutive speech DFT bins represented as composite-valued autoregressive procedures were modeled and to integrate the harmonic stage development into a state-transition designed.

In 2019, Aaron Nicolson and Kuldip K. Paliwal [3] investigated deep learning algorithm for the MMSE method, with the aim of processing comprehensible improved speech with maximum quality. As the speech enhancement performance for an MMSE technique enhances with the precise of the exploited a priori SNR estimator, a ResLSTM network was exploited here to precisely calculate a priori SNR.

MMSE algorithms utilizing the ResLSTM a priori SNR estimator were estimated by objective and subjective measures of intelligibility and speech quality.

In 2017, Qi He, Feng Bao and Changchun Bao [4], proposed a novel wiener filtering speech enhancement technique in order to estimate the short period of time linear predictive parameters for noise and speech in the codebook-driven. A pre-trained spectral form codebook for the speech was exploited in order to model a priori instruction regarding coefficients of linear predictive for speech. In the presented technique, a multiplicative update rule was exploited to calculate the linear predictive gains high precisely were used.

In 2017, Ji Ming and Danny Crookes [5], developed a novel technique that aspires to minimize or efficiently remove requirements. By exploiting the ZNCC, it was demonstrated that was in contrast to the measure, and by expanding the efficient speech segment length corresponding to sentence long speech statement. From the noise, it was probable to attain an accurate speech calculation without necessitating the particular knowledge regarding the noise. A new realization which incorporates the full-sentence speech correlation with uncontaminated recognition of speech, established as a constrained maximization issue, to control the data sparsity issue.

In 2018, Robert Rehr and Timo Gerkmann [6], presented a theoretical and experimental estimate of the MLSE-based methods. The super-Gaussian prior's permits for a minimization of noise among speech spectral harmonics that was not attainable by Gaussian estimators like the WF. A deep neural network on the basis of the low-rank nonnegative matrix factorization and a phoneme classifier model was exploited for the evaluation as instances of MLSE based techniques.

In 2018, Johannes Stahl and Pejman Mowlaee [7], presented a new outlook producing three major contributions. Initially, a pitch-synchronous signal indication was contemplated and exhibited to be dominant for the calculation of the harmonic parameters model. Subsequently, the harmonic amplitudes were modeled in voiced speech as arbitrary variables with frequency bin determined by Gamma distributions. At last, for the different methods, the distinct estimators of unvoiced speech, voiced speech, as well as speech non-presence was derived.

3 Emotional Speech Signal Analysis

3.1 Analysis of NMF

Let us assume the non-negative data vector of $\mathrm{D}(\mathrm{n})$ as $\mathrm{D} \in \mathrm{T}^{\mathrm{l}\times\mathrm{j}};1\times\mathrm{j}$, which represents the 1-dimensional samples, the main contribution of NMF [14] is to recognize two non-negative matrices that are stated in eq.(1) and (2), Here v indicates the dimensional space.

$$\mathrm{X} \in \mathrm{T}^{\mathrm{l}\times\mathrm{v}} \tag{1}$$

$$\mathrm{Y} \in \mathrm{T}^{\mathrm{v}\times\mathrm{j}} \tag{2}$$

By the non-negative constraints, the parts-based indication is achieved as they permit one and only addition and not either combinations or subtractive. It is necessary to describe the function of the cost to recognize the approximate factorization that enumerates the approximation quality. The most well-known cost models are:

(a)The squared Euclidean distance among YZ as well as S is described in eq. (3), here $|\,|\cdot|\,|_{\mathrm{Fb}}$ indicates the matrix Frobenius norm.

$$|\,|\mathrm{D}-\mathrm{XY}|\,|_{\mathrm{Fb}}=\sum_{\mathrm{m}=\mathrm{l}}^{\mathrm{y}}|\,|\mathrm{S}_{\mathrm{m}}-(\mathrm{XY})_{\mathrm{m}}|\,|^{2} \tag{3}$$

(b)In eq. (4), the general Kullback-Leibler divergence K between R and XY is described.

$$\mathrm{K}(\mathrm{R}\,|\,|\,\mathrm{XY})=\sum_{\mathrm{m}=\mathrm{l}}^{\mathrm{y}}\left(\mathrm{R}_{\mathrm{m}}\log\frac{\mathrm{R}_{\mathrm{m}}}{(\mathrm{XY})_{\mathrm{m}}}-\mathrm{R}_{\mathrm{m}}+(\mathrm{XY})_{\mathrm{m}}\right) \tag{4}$$

By multiplicative technique, both the functions of cost are resolved [15]. For the square Euclidean distance, the update instruction for non-negative matrices Y_{m} and X_{m} are stated in eq. (5) and (6) as well as the updating rule for the divergence of Y_{m} and X_{m} are stated in eq. (7) and (8), whereas $\mathrm{m}=1....\mathrm{y}$. In eq. (9), the feature of NMF $NM_{t}(n)$ is stated.

$$\mathrm{Y}_{\mathrm{m}}\leftarrow\frac{(\mathrm{X}^{\mathrm{T}}\mathrm{R})_{\mathrm{m}}}{(\mathrm{X}^{\mathrm{T}}\mathrm{XY})_{\mathrm{m}}}\mathrm{Y}_{\mathrm{m}} \tag{5}$$

$$X_m \leftarrow \frac{\left(RY^T\right)_m}{\left(XYY^T\right)} Y_m \tag{6}$$

$$Y_m \leftarrow \frac{\sum_l X_l R_{lm} / (XY)_{lm}}{\sum_l Y_l} Y_m \tag{7}$$

$$Z_m \leftarrow \frac{\sum_l Z_{ml} R_m / (XY)_{lm}}{\sum_l Y_{ml}} X_m \tag{8}$$

$$NM_t(n) = (X_m, Y_m) \tag{9}$$

3.2 Cepstral Analysis

Consider the speech signal as $S_s(n)$, that is attained from the complexity of two signals $h(n)$ and $j(n)$ as the summation of two signals as well as it is stated in eq. (10). Here $\widehat{S}_s(n)$ represents the complex cepstrum and $\widehat{S}_s(n)$ is indicated in eq. (17).

$$S_s(n) = h(n) + j(n) \rightarrow \widehat{S}_s(n) = \hat{h}(n) + \hat{j}(n) \tag{10}$$

In eq. (10), $h(n)$ represents the speech portion and $j(n)$ represents the noiseless piece of the speech recording correspondingly and eq. (11) states the analysis of the cepstral, and eq. (12) states the log of the signal $SL(z)$.

$$S_s(n) = S_{s_1}(n) * S_{s_2}(n) \Leftrightarrow SL(z) = SL_1(z) SL_2(z) \tag{11}$$

$$\log\{SL(z)\} = \log\{SL_1(z)\} + \log\{SL_2(z)\} = \hat{S}L(z) \tag{12}$$

If the transform Z is reasonable and the composite log is characteristic, subsequently the two convolved signals $\hat{S}_{s_1}(n)$ and $\hat{S}_{s_2}(n)$ are added, as well as it is demonstrated in Eq. (13).

$$\widehat{S}_s(n) = \hat{S}_{s_1}(n) + \hat{S}_{s_2}(n) \tag{13}$$

In eq. (14), the signal $S_s(n)$ is limited to have poles and zeros into the unit circle, here $\log\{SL(u)\}$ is the complex logarithm of $SL(u)$.

$$\log\{SL(U)\} = \log\{|SL(u)|\} + 1 \lhd SL(u) \tag{14}$$

Eq. (15) states if $SL(u) = SL_1(u) SL_2(u)$ subsequently $\log\{|SL(u)\}$ is indicated. In eq. (16), the real cepstrum $RS_y(n)$ is stated where the magnitude of $RS_y(n)$ is non-negative and real. The $\hat{S}_s(n)$ complex cepstrum is stated in eq. (17), whereas the stage is indicated as $\arg(\ )$, $\log\{|SL(u)\}$ and $\log|SL(e^{lu})$ are the log spectrum of the signal. This is intricate due to it uses the complex logarithm. Additionally, the compound cepstrum of the real series is real.

$$\log\{|SL(u)|\} = \log\{|SL_1(u) SL_2(u)\} = \log\{|SL_1(u)|\} + \log\{|SL_2(u)|\} \tag{15}$$

$$RS_y(n) = \frac{1}{2\Pi} \int_{-\Pi}^{\Pi} \log|SL(e^{lu})| e^{lun} du \tag{16}$$

$$\hat{S}_s(n) = \frac{1}{2\Pi} \left[\left[\log|SL(e^{lu})|\right] + l \arg\left(SL(e^{lu})\right)\right] e^{lun} du \tag{17}$$

In reality, the real cepstrum is the even part of $\hat{S}_s(n)$, as well as it is shown in eq. (18). In general, speech processing $RS_y(n)$ is exploited that is attained by using an inverse FT of the log spectrum of the signal.

$$RS_y(n) = \frac{\hat{S}_s(n) + \hat{S}_s(n)}{2} \tag{18}$$

Eq. (19) represents the FT is replaced by DFT in digital signals. Here, $\hat{S}L_t(g)$ is indicated as the sampled version of $\hat{Y}(e^{kw})$ and hence $\hat{S}_{s_t}(n)$ is stated as demonstrated in Eq. (20), and N indicates the period. Likewise, the misidentification of a signal frequency, by repeating the cepstrum with N, the $RS_t(n)$ cepstral feature of $S_s(n)$ is stated in Eq. (21).

$$SL_t(g) = \sum_{n=0}^{N-1} S_s(n) e^{-1\frac{2\Pi}{N}gn} \qquad 0 \le g \le N-1$$

$$\hat{S}L_t(g) = \log\{SL_t(g)\} \qquad 0 \le g \le N-1 \tag{19}$$

$$\hat{S}_{s_t}(n) = \frac{1}{N}\sum_{g=0}^{N-1} \hat{S}_t(g) e^{1\frac{2\Pi}{N}gn} \quad 0 \le n \le N-1$$

$$\hat{S}_{s_t}(n) = \sum_{p=-\infty}^{\infty} \hat{S}_s(n+mN) \tag{20}$$

$$RS_t(n) = \frac{1}{N}\sum_{g=0}^{N-1} \log|SL_t(g)| e^{1\frac{2\Pi}{N}gn} \quad 0 \le n \le N-1 \tag{21}$$

3.3 Analysis of Pitch

The entire estimation of pitch filter consists of some steps, and it is stated as below:

(a) At first to transform the $t_f(n)$ to the time-frequency domain with the help of STFT)= $TF_t(f)$ that is stated in eq. (22). Here, $V_t(f)$ indicates the PSD of the exasperating noise as well as $t_f(n)$, the power of the P^{th} harmonic, fr_0 as well as *t* indicates the time and frequency for the ideal periodic source.

$$TF_t(f) = \sum_{p=1}^{P} t_f(n)\Delta(f - pfr_0) + U_t(f) \tag{22}$$

(b) In this process, the PSD of each frame against a Log spaced frequency grid $TF_t(l)$ that is stated in eq. (23), here $l = \log f$. Moreover, the harmonic spacing is unconventional to fr_0 using convolving $TF_t(f)$ with IR $I(l)$ the energy should be integrated that is stated in eq. (24).

$$TF_t(l) = \sum_{p=1}^{P} t_f(n)\Delta(l - \log p - \log fr_0) + V_t(l) \tag{23}$$

$$I(l) = \sum_{p=1}^{P} \Delta(l - \log P) \tag{24}$$

(c) As stated in eq. (25), formulate $\beta_t(l)$. To decide the compression exponent $\beta_t(l)$, in both the log-frequency and time $L_F(l)$, the first procedure is to construct the smoothened spectrum $\overline{F}_t(l)$ using low pass filtering $F_t(l)$. Both $F_t(l)$ and $\overline{F}_t(l)$ are normalized to the power of $L_F(l)$ as well as sets the compression exponent $\beta_t(l)$. So, the normalized smoothened spectrum $\overline{F}_t(l)$ equals $L_F(l)$ and the compressed PSD formulation $F'_t(l)$ is performed as stated in eq. (26)

$$\beta_t(l) = \frac{\log L_F(l)}{\log \overline{F}_t(l)} \tag{25}$$

$$F'_t(l) = F_t(l)^{\beta_t(l)} \tag{26}$$

(d) Here, $F'_t(l)$ is convolved with $I(l)$ and selects the maximum peak in the sensible range as $P_t(n)$, the assessed pitch.

$$RR_t(n) = \{F'_t(l) * I(l)\} \tag{27}$$

The initial process is feature extraction, in that from the input $t_f(n)$, the features are extracted. The resultant features such as NMF feature $NM_t(n)$, Cepstral feature $RS_t(n)$, and pitch feature $RR_t(n)$ are stated as the input to the dimensionality minimized phase. The produced dimensional minimized features are integrated and stated as the input to the classification procedure that outcome the classified output. Fig 1 demonstrates the architectural diagram of the proposed model.

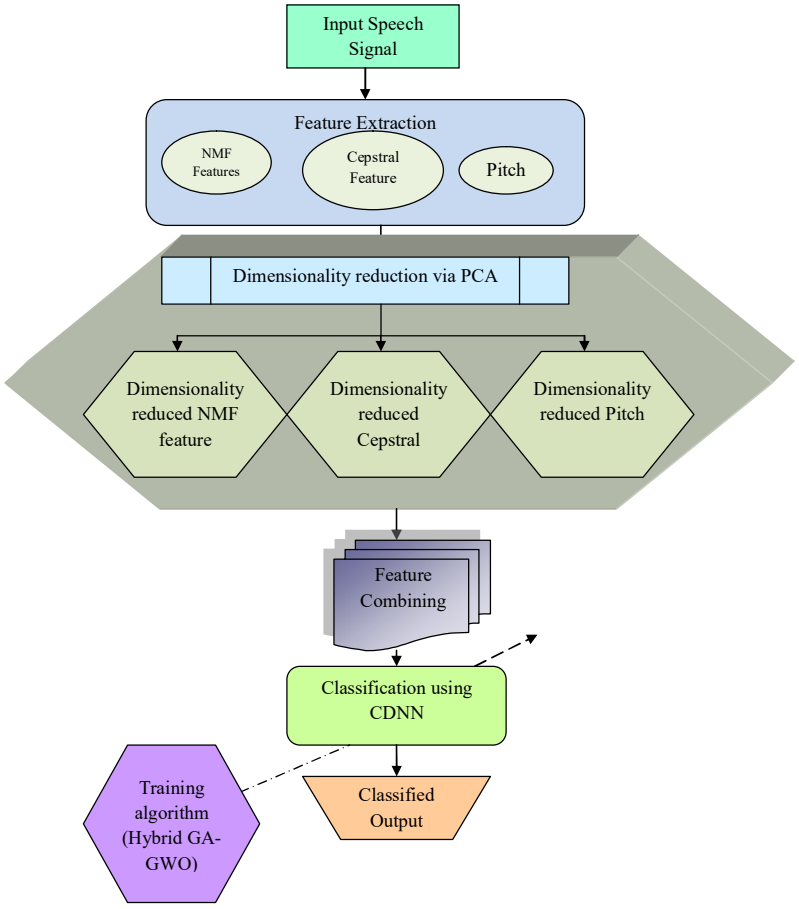


Fig. 1. *Architectural diagram of the proposed model.*

4. Detailed Analysis of the Proposed Algorithm

4.1 Feature Extraction

The proposed method consists of three features such as $\mathrm{RS_t(n)}$, $\mathrm{NM_t(n)}$ and $\mathrm{RR_t(n)}$ are extracted that is demonstrated in Eq.(9), (21) and (27). The feature extraction is stated to the subsequent procedure in order to minimize the dimensions of the features.

4.2 Reduction of Dimensionality

In this phase, the extracted $\mathrm{CF_t(n)}, \mathrm{NM_t(n)}$ and $\mathrm{RR_t(n)}$ are given as the input. Here, the well-known process called PCA is exploited to reduce the dimensions of the features. A collection of dv data vectors $\mathrm{g_1},.....,\mathrm{g_{dv}}$ are selected from the extracted feature and g_o indicates the well-defined group analysis of variables u. Subsequently, the empirical mean formulation is performed with all column $\mathrm{cl}=1,.....,\mathrm{u}$, and the ensuing mean value is positioned in the $e_{(u)}$ with $u\times 1$ dimensions, that is stated in eq. (28), the computation of mean deviation is done as per eq. (29). Here, F indicates the $d_v \times u$ matrix and k represents the column vector $\mathrm{d_v}\times 1$ of all 1s: $\mathrm{k[i]}=1$.

$$\mathrm{e_{(u)}[cl]}=\frac{1}{\mathrm{d_v}}\sum_{\mathrm{i}=1}^{\mathrm{d_v}}\mathrm{RS[i,cl]} \tag{28}$$

$$\mathrm{F}=\mathrm{S}-\mathrm{kcl}^{\mathrm{T}} \tag{29}$$

As demonstrated in eq. (30), the covariance matrix CM is stated. Moreover, both the formulation of eigenvalues and eigenvector is processed using valuing the M matrix, that diagonalizes CM and the formulation is stated in eq. (31)

$$CM = \frac{1}{d_v - 1} F^{*}.F \tag{30}$$

$$P^{-1}CM(P) = D^{(P)} \tag{31}$$

In eq. (31), $D^{(P)}$ indicates the diagonal matrix of eigenvalues of CM. In eigenvector matrix N the column sorting is performed, the eigenvalue matrix minimization of D^{P} is attained. The computation of cumulative energy content ce is stated in eq. (32) that embraces the sum of the energy content of eigenvalues from 1 through cl.

$$ce[cl] = \sum_{d_v=1}^{cl} D^{(p)}[d_v, d_v], \text{ for } cl = 1,.....u \tag{32}$$

As *B* matrix by storing the *cl* column of P, the eigenvectors subset is chosen. To select the cost of cl, the l vector is exploited. The column of $\hat{P}$ the matrix $\hat{P} = Q.B = KKL\{R\}$ is the vector, as well as in the row matrix R, it decides the Kosambi-Karhunen-Loeve transform. In eq. (33), the Q matrix is stated, here $t = t(c) = \left\{\sqrt{CM \mid cl, cl \mid}\right\}$, $cl = 1....v$ as well as this is the generated dimensionally minimized speech signal $S_d(n)$.

$$Q = \frac{F}{j.t^{\hat{P}}} \tag{33}$$

For all the extracted features individually the PCA process is exploited, as well as the dimensional minimized signal $S_d(n)$is stated as stated in eq. (34), here $CF_t^{dv}(n), NM_t^{dv}(n), m_t^{dv}(n)$indicates the dimensional minimized NMF, cepstral, and pitch features in the proposed method.

$$S_{dv}(n) = \left\{CF_t^{dv}(n), NM_t^{dv}(n), p_t^{dv}(n)\right\} \tag{34}$$

4.3 Emotion Recognition using the Proposed Method

(a) DCNN: In this paper, the deep neural network models, the DCNN is exploited. In recent years, it has been acquired huge popularity in classification. An archetypal DCNN comprises 5 convolutional layers, like three fully-connected layers and three pooling layers [16]. Using competent alternating and pooling layers and stacking convolutional layers, feature learning is attained. The one –dimensional feature vector obtained from plotting two-dimension feature vectors using the polling layers as well as the fully-connected layers follow the convolution layers.

The mapping of output feature maps for the m^{th} layer y_j^m can be computed on the basis of the Eq. (35) if there are P feature maps as the input and Q filters for convolutional layers.

$$y_j^m = f\left(\sum_{i=1}^{P} y_j^{m-1} * r_{ij}^m + a_j^m\right), j = 1,....., Q \tag{35}$$

In eq. (35), r_{ij}^m indicates the kernel of the j^{th} filter associated with the i^{th} input map, y_j^{m-1} indicates the i^{th} input map, a_j^m is the bias equivalent to j^{th} filter, $*$ indicates the convolutional operation, and $f(\ )$ indicates the activation function. Hence, Q feature maps are attained as the output. Using eq. (36), the number of all the parameters of a convolutional layer is computed.

$$PR = N \times (k \times k \times P + 1) \tag{36}$$

A convolutional layer is gone after by the pooling layer directly holds the activations within the little spatial region. The nature of the operations for pooling is characterized into two types: average and maximum pooling whereas the maximum-pooling unit calculate the utmost for a local patch of unites in a feature map, as well as the average pooling calculates average. The computation procedure of output feature maps of the m^{th} layer is same as with eq. (35) in a pooling layer.

$$y_j^m = f\left(\alpha_j^m down\left(y_j^{m-1}\right)\right), j = 1,....., Q \tag{37}$$

In eq. (37), $down(\ )$ indicates the sub-sampling function and α_j^m represents the multiplicative bias equivalent to the j^{th} filter. Subsequent to the pooling layers and the convolutional layers; from the raw data, the fully connected layer is exploited to classify the features extracted. Into the one-dimensional vector, the learned feature vectors are demolished that is an input of the fully connected layers. For the input vector each value the is linked for every value of the output vector by one neuron in a fully

connected layer. If the lengths of the output and input vectors are Q and P correspondingly; as eq. (38) the output vector of the m^{th} layer is computed.

$$y_j^m = f\sum_{i=1}^{P} y_j^{m-1} \times z_{ij}^m + a_j^m,\ j = 1,.....Q \tag{38}$$

In eq. (38) z_{ij}^m indicates the weight of the j^{th} output value linked with the i^{th} input value. The calculation for the number of the entire parameters of a fully connected layer is stated in eq. (39).

$$M = P \times Q + 1 \tag{39}$$

(b) Hybrid GWO-GA:

In this paper, the GWO is integrated GA to enhance the effectiveness of the method [17]. Assume that the present population as $P_p(Y_1, Y_2, ...Y_D)$, the fitness of the individual Y_i is f_i. Compute the fitness $f_i(i = 1,2,...., D)$ of each individual and organize them in descending order. Choose the individuals with the maximum fitness to copy directly into the subsequent generation of the population. Compute the total fitness T of the residual individuals and the probability P_{p_i} that each individual is chosen.

$$T = \sum_{i=1}^{D-1} f_i (i = 1,2,....D-1) \tag{40}$$

$$P_{p_i} = \frac{f_i}{\sum_{i=1}^{D-1} f_i} (i = 1,2,....D-1) \tag{41}$$

For each individual, compute the cumulative fitness value c_i, and subsequently, the selection operation is done in the way of the stake roulette until the number of individuals in the children population is reliable with the parent population.

$$c_i = \frac{\sum_{i=1}^{i} f_i}{T} (i = 1,2,....D-1) \tag{42}$$

In the subpopulation, each individual is cross-operated in a linear crossover way in the proposed method. For each individual y_i in the subpopulation, generate a corresponding arbitrary number $r_i \in (0,1)$. While the random number r_i is lesser than the C_P crossover probability, the equivalent individual y_i is paired for cross-operation. Here, the crossover operators are (c_1, c_2). Here, generate a random number $\eta \in (0,1)$ and it consists of two children $cl^1(cl^1,....,cl_D^1)$, $cl^2(cl^2,....,cl_D^2)$, which are produced by two parents $pr^1(pr^1,....,pr_D^1)$ and $pr^2(pr^2,....,pr_D^2)$.

$$cl_i^1 = \eta pr_i^1 + (1-\eta) pr_i^2,\ i = 1,2,.....D \tag{43}$$

$$cl_i^2 = \eta pr_i^2 + (1-\eta) pr_i^1,\ i = 1,2,.....D \tag{44}$$

The best individual is $y_i(y_1, y_2, ..., y_d)$ with the mutation probability M_p, the mutation operation is done on y_i namely, choose a gene y_m from the best individual with probability M_p, in place of the gene y_m with an arbitrary number among lower and upper bounds to create a new individual $y_i' = (y_1', y_2', ..., y_d')$. Eq. (45), the precise operation is described.

$$y_i' = \begin{cases} 1 + \eta * (v - 1) & i = m \\ y_i & i \neq m \end{cases} \tag{45}$$

In eq. (45), η represents a random number in [0, 1], and v and l represents the upper and lower bounds of the individual y_i, correspondingly.

5. Results and Discussions

5.1 Experimental Procedure

In this paper, two databases are exploited such as benchmark and Marathi database. Here, both the databases are obtained from http://neuron.arts.ryerson.ca/ravdess/index.php. The benchmark database comprises of numerous emotions such as 96 happy, angry, surprise, fear, neutral, and sad emotions. Likewise, the Marathi database includes 40 neutral, surprise, angry, sad, fear, and happy emotions.

5.2 Analysis of the Proposed Model

This section exhibits the analysis of the proposed Hybrid GWO-GA technique and conventional techniques such as GWO-NN, GA-NN and PSO-NN models for both the Marathi and benchmark databases. In fig 2, the performance analysis of the proposed technique over the existing techniques for the benchmark databases is shown. Here, the result analysis demonstrates the accuracy of the proposed approach 15% is better than the GWO-NN, 22% better than the GA-NN, and 28% better than the PSO-NN methods. Fig 3 shows the performance analysis of the proposed approach over the existing techniques for the Marathi databases. Here, the result analysis demonstrates the accuracy of the proposed approach 23% is better than the GWO-NN, 27% better than the GA-NN, and 31% better than the PSO-NN methods. The analysis outcome reveals that the proposed technique is better than the existing techniques.

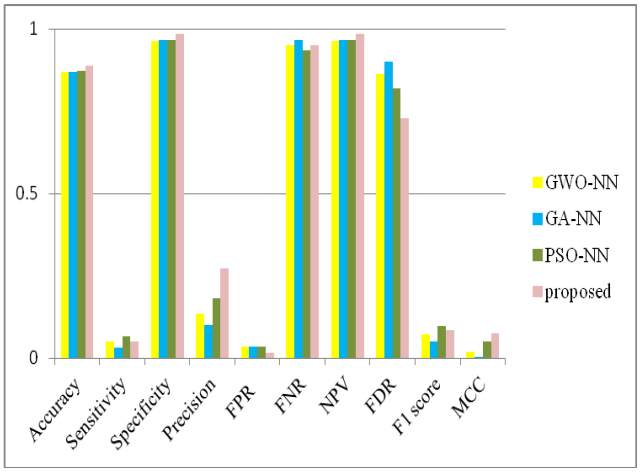


Fig. 2. Analysis of the proposed method for benchmark database

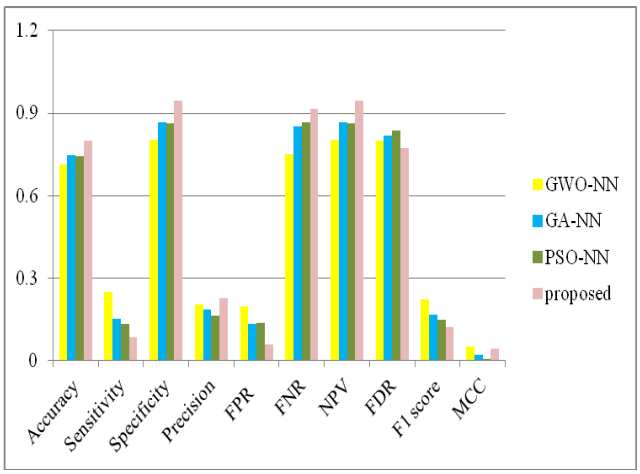


Fig. 3. Analysis of the proposed method for Marathi database

In both fig 4 and 5, the analysis of the proposed feature techniques such as an amalgamation of NMF, cepstra, and Pitch over the performance of NMF feature individually, cepstra feature individually and Pitch feature individually for both the Marathi databases and benchmark database. In fig 4, the analysis of the proposed feature technique with the traditional models for the benchmark databases is demonstrated. Here, the result analysis showed that the superiority of the proposed technique over the existing techniques. Here, the result analysis demonstrates the accuracy of the proposed approach 14% is better than the GWO-NN, 16% better than the GA-NN, and 19% better than the PSO-NN methods. Fig 5 demonstrates the performance analysis of the proposed method with the traditional techniques for the Marathi databases. The analysis outcome reveals that the proposed technique is better than traditional techniques. Here, the result analysis demonstrates the accuracy of the proposed approach 31% is better than the GWO-NN, 35% better than the GA-NN, and 37% better than the PSO-NN methods.

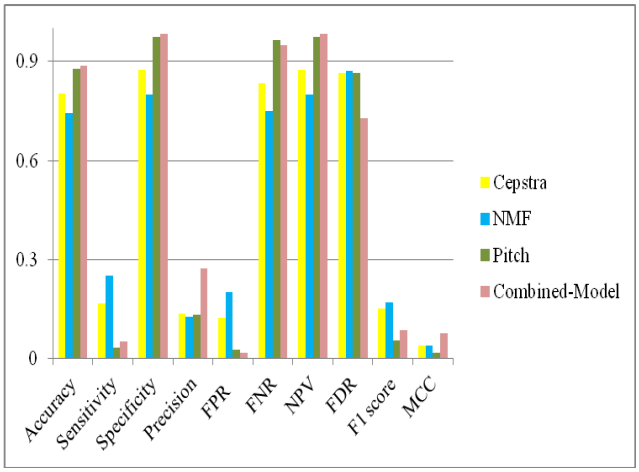


Fig. 4. *Analysis of proposed feature model for benchmark database*

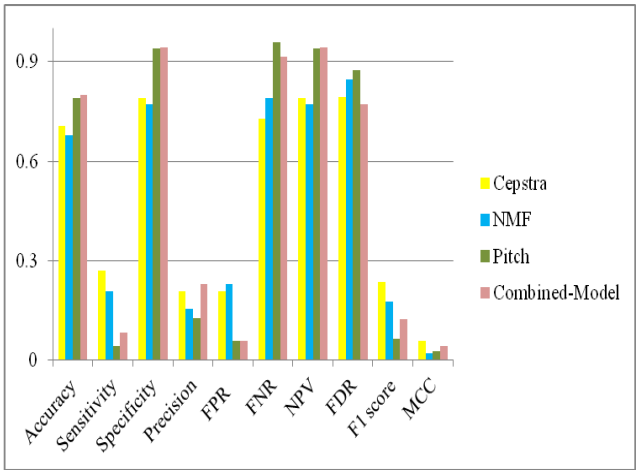


Fig. 5. *Analysis of the proposed feature model for Marathi database*

6. Conclusion

Generally, speech analysis has become an important element in minimizing the gap among physical and digital world with the increase in man to machine communication. A significant subpart within this domain is the emotion recognition in speech signals that was conventionally examined in linguistics and psychology. Moreover, speech emotion recognition is a field having diverse applications. The main objective of this paper is to present a novel emotion recognition model. Hence, a DCNN model was developed to study the multimodal fusion of speech features. The features such as cepstral, NMF, and prosodic features. Moreover, a hybrid GA-GWO technique was also attained to train the classifier. The proposed recognition method performance was evaluated and compared it with the existing techniques. The proposed method has shown accurate emotions recognition from the speech signal.

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] Ki-Seung Lee,"Speech enhancement using ultrasonic doppler sonar", Speech Communication, Volume 110, July 2019, Pages 21-32.

[2] Johannes Stahl, Pejman Mowlaee,"Exploiting temporal correlation in pitch-adaptive speech enhancement",Speech Communication, Volume 111, August 2019, Pages 1-13.

[3] Aaron Nicolson, Kuldip K. Paliwal,"Deep learning for minimum mean-square error approaches to speech enhancement", Speech Communication, Volume 111, August 2019, Pages 44-55.
[4] Q. He, F. Bao and C. Bao, "Multiplicative Update of Auto-Regressive Gains for Codebook-Based Speech Enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 457-468, March 2017.
[5] J. Ming and D. Crookes, "Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 531-543, March 2017.
[6] R. Rehr and T. Gerkmann, "On the Importance of Super-Gaussian Speech Priors for Machine-Learning Based Speech Enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 2, pp. 357-366, Feb. 2018.
[7] J. Stahl and P. Mowlaee, "A Pitch-Synchronous Simultaneous Detection-Estimation Framework for Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 2, pp. 436-450, Feb. 2018.
[8] Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.-D., Jin, W., Schuller, B., 2018. Deep learning for environmentally robust speech recognition: an overview of recent devel- opments. ACM Trans. Intell. Syst. Technol. 9 (5), 1–28.
[9] Lu, X. , Tsao, Y. , Matsuda, S. , Hori, C. , 2013. Speech enhancement based on deep denoising autoencoder. In: Proceedings Interspeech 2013, pp. 436–440 .
[10] Xia, Y., Stern, R., 2018. A priori SNR estimation based on a recurrent neural net- work for robust speech enhancement. In: Proc. Interspeech 2018, pp. 3274–3278. doi: 10.21437/Interspeech.2018-2423 .
[11] Chen, J., Wang, D., 2017. Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. 141 (6), 4705–4714.
[12] Xu, Y. , et al. , 2015. A regression approach to speech enhancement based on deep neu- ral networks. IEEE/ACM Trans. Audio Speech Lang. Process (TASLP) 23 (1), 7–19 .
[13] Li, J.-j. , et al. , 2014. Whisper-to-speech conversion using restricted boltzmann machine arrays. Electron Lett. 50 (24), 1781–1782.
[14] J. Deng, X. Xu, Z. Zhang, S. Frühholz and B. Schuller, "Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition,"IEEE Access, vol. 4, no. , pp. 4299-4309, 2016.
[15] M. El Ayadi, M.S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, vol.44, pp.572-587, 2011.
[16] D. D. Pukale, S. G. Bhirud and V. D. Katkar, "Content-based Image Retrieval using Deep Convolution Neural Network," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-5.
[17] E. Daniel, "Optimum Wavelet-Based Homomorphic Medical Image Fusion Using Hybrid Genetic–Grey Wolf Optimization Algorithm," in IEEE Sensors Journal, vol. 18, no. 16, pp. 6804-6811, 15 Aug.15, 2018.
[18] Renjith Thomas and MJS. Rangachar,"Hybrid Optimization based DBN for Face Recognition using Low-Resolution Images",Multimedia Research, Volume 1, Issue 1, October 2018.
[19] J.S. Anita and J.S. Abinaya,"Impact of Supervised Classifier on Speech Emotion Recognition",Multimedia Research, Volume 2, Issue 2, January 2019.

.