Accurate Speech Emotion Recognition by using Brain-Inspired Decision-Making Spiking Neural Network

Dr. Madhu Jain[1]
Electronics and Communication Engineering Department,
Jaypee Institute of Information Technology, A-10,
Sector 62, Noida

Ms. Shilpi Shukla[2*]
Mahatma Gandhi Mission's College of Engineering and Technology, A-9
Sector 62, Noida

***Abstract*—A portion of speech recognition is taken away by emotion recognition which is a smart update and it is necessary for its gain massively. Feature selection is an indispensable stage among the furtherance of various schemes in order to implement the classification of sentiments in speaking. The communication among features prompted from the alike audio origin has been rarely deliberated at present, which might yield terminated features and cause an upswing in the computational costs. To resolve these defects the deep learning-based feature extraction technique is used. An incredible modernization in speech recognition in recent years incorporates machine learning techniques with a deep structure for feature extraction. In this paper, the speech signal obtained from the SAVEE database is used as an input for a deep belief network. In order to perform pre-training in the network, the layer-wise rapacious feature extraction tactic is implemented and by using systematic samples, the smearing back-propagation method is accomplished for attaining fine-tuning. Brain-inspired decision-making spiking neural network (SNNs) is used to recognize different emotions but training by deep SNNs remains a challenge, but it improves the determination of the result. In order to enhance the parameters of SNNs, a social ski-driver (SSD) evolutionary optimization algorithm is used. The results of the SNN-SSD algorithm are related to artificial neural networks and long short term memory with different emotions to refine the classification for authorization.**

Keywords*—Brain-inspired decision-making spiking neural network (BDM-SNN); deep belief network; social ski-driver (SSD) optimization; emotion recognition

I. INTRODUCTION

Speech recognition is gaining a lot of attention, which deals with the recognition of speech and conversion into text by the computer. This origination of speech recognition can expand human-computer communication [1]. The speech recognition principle has been improved to speech emotion recognition (SER) which is proved to be a developing investigation area [2]. This, in turn, attempts to decide the emotion from the speech signals. The advancement in emotion recognition will convert everything to ease and hence making our lifestyle more comfortable through various researches [3]. Emotion recognition is actually very tricky in certain criteria's since reactions may be in accordance with the surroundings, principles, singular face response leads to vague discoveries; the emotion cannot be concluded just by using speech quantity and if there is a deficiency of speech databank in voluminous languages [4].

The focal investigation disputes in speech emotion recognition are an optimal feature set selection from the provided speech signals [5, 6]. The speech emotion recognition deals with a greater part of the past through various investigations of speech rhythm features and ethereal data [7, 8]. The speech emotion recognition uses some novel feature parameters which may include the Fourier parameters. Numerous validated acoustic parameters are found in order to hold emotional data, execution of various triumphs such that a lot of features that are executed dependably over various conditions [9]. In the same way, most of the analysts want to utilize a unification feature set that is made out of numerous sorts of features comprising of contemporary emotional data [10]. Utilizing the blending feature set may cause increment too high measurement and reduction of speech features, thus the learning procedure is being elaborated for most machine learning calculations and develops the probability of overfitting. Utilizing an assortment of modalities SER frameworks are being created by the analysts, a few examples of varying media signals are sound, pictures, video, and electroencephalogram (EEG). There are various reasons why the identification of emotions from human speech is fascinating. Sound nearness of emotions is one such reason in the acoustic channel of speech which is similar to semantic channel and this as the simple way to show off the emotion. In real-time, speech can be effectively acquired and prepared yet different modalities, for example, video or EEG are hard to get, which is a crucial factor remembering potential applications [11]. Emotions depend on language and culture, appropriate acoustic features and their correlation.

Traditional machine learning strategy and the deep learning strategy are the two classes of the SER techniques. In traditional machine learning techniques for automatic emotion recognition (AER), the strategy shadowed is feature determination, which is forthrightly branded with the precision of recognition [12]. The pitch frequency feature, the vitality related feature, the formant feature, the ghastly feature, etc. are integrated into the most eminent feature extraction strategy [13]. Artificial neural network (ANN), Bayesian network model, hidden Markov model (HMM), support vector machine (SVM) [14], Gaussian mixture model (GMM) [14],

*Corresponding Author

and multi-classifier fusion [15] are primed by utilizing the machine learning strategy and all these are done after separation of the features.

The above-used techniques are cooperative for recognizing explicit emotion; there is no authoritative strategy to reveal intertwined adoring states. Due to two prime reasons, the existing speech emotion recognition innovation is veracity. The only cause is that the recurrence of speech emotion data prompts the fall of the conclusive recognition rate and the long training time of sample data, and handling the information of numerous speech data obstructs the ongoing feedback of the planned framework. An additional cause is that the overall efficacy of the calculations hooked on the speaker-free highlights which can be connected to SER is comparatively compact and it additionally sways the practicability of speech feeling recognition innovation.

The essential bit of leeway of this strategy is to train a model in the lack of exceptionally huge information. While the inconvenience is that it is hard to pass judgment on the nature of the feature and some key features may be absent, which will diminish the precision of acknowledgment. Meanwhile, it is hard to guarantee the great outcomes that can be accomplished in an assortment of databases. Contrasted and the customary artificial intelligent (AI) technique, the deep learning can separate the abnormal state features, and it has been appeared to surpass human execution in visual assignments. Right now, deep learning has been connected to the SER by numerous researchers. Many optimization techniques have been used for increasing the performance of the system by optimizing the weights, hidden layers and other parameters of neural network. The results are very promising [16]-[23]

This paper recaps the allied work in Section 2, however, information concerning the speech recognition and deep neural network (DNN) is accessible in Section 3. The methodology used to conduct this review is précised in Section 4. In Section 5 the results are demonstrated, where Section 6 accomplishes this paper.

II. Related Work

Certain overviews lead the territory of speech emotion recognition. Badshah et al. [24] in 2017 presented a technique to identify emotions in the speech by means of convolutional neural network (CNN) along with rectangular kernels. Outcomes publicized that rectangular kernels and max pooling operations in rectangular neighborhoods are appropriate for SER via spectrograms. This technique effectively learns discriminative features from speech spectrograms. The propounded technique can be additionally improved if further labeled data can be collected and a much deeper CNN having rectangular kernels could be successfully trained. Zhao, J., et al. [25] in 2019 describes 1D and 2D CNN long short term memory (LSTM) networks for Speech emotion recognition. The results obtained determine that the designed networks accomplish the best performance on the task of distinguishing speech emotion. The investigational outcomes prove that the designed networks attain tremendous performance while recognizing speech emotion; particularly the 2D CNN LSTM network is better compared to the traditional methods, deep belief network (DBN) and CNN on the chosen databases. The proposed method provides less accuracy which can be enhanced in future works.

Gupta, S., et al. [26] in 2019 proposed a novel CNN architecture with a spatial pyramid pooling (SPP) layer that function on varying length feature representation of speech signals to accomplish emotion classification task. A constraint of the propounded kernel is that it necessitates a CNN model to attain varying size feature maps. These restrictions found in the proposed techniques are reduced in upcoming works. Xu, H., et al. [27] in 2019 proposed an attention mechanism with the ASR system to learn the alignment between the original speech and the recognized text, which is then used to fuse features from two modalities. The outcomes prove that the projected method is better than other methodologies concerning emotion identification ability. But the computational time is high and so it has taken under consideration in forthcoming works.

Hook, J., et al. [28] in 2019 describes a minor set of features presented a competitive performance in SER. The feature set used in this work performs well for both male and female speakers. The primary results appear to be hopeful and permit an additional investigation. In future works, supplementary testing on other ESDs to evaluate the quality of the recommended features must be accomplished moreover enhanced discrimination amid anger and happiness owing to the worldwide nature of the difficulty in the SER field for both machines and humans. Mohanty, M.N. et al. [29] in 2019 presented few models for the recognition which are depending upon the NN frameworks. The elementary organization as a multilayer perceptron (MLP), radial basis function network (RBFN), and probabilistic neural network (PNN) are utilized with various spectral features. The DNN model is confirmed with all these features. Consequently, the description of speech emotions in dissimilar levels and various domains is to be performed in the future work which has the ability to decide the combination type that can be enhanced than the current work.

III. Proposed Methodology

SER is an interdisciplinary research region which aims to naturally recognize the emotional condition of a person. It is critical to improving human-PC connection in numerous perspectives, for instance, a non-human communicator should be aware of the proper emotional condition of the voice to recognize dual implications of a similar term. Emotions are not that much easier to detect by machines that need a lot of intelligence and training. If controllable intelligence is developed, human-machine interaction can be made much easier than before. It is found that some emotions are recognized accurately, other emotions lead to an ambiguous condition in inferring and classifying the emotion. The efficient characterization of dissimilar emotions by the extraction of appropriate features comes under the design of an SER system which is a significant dispute. An accurate selection of features significantly distresses the performance of classification since the pattern recognition techniques are hardly independent of the problem domain.

The defect can be erased but we need an efficient deep learning technique that is necessary which is a need of deep learned automated speech recognizer with a combined accuracy rate for the effective results with improved recognition rate. In this work, a system using SNN is proposed that performs emotion recognition from standard raw speech database and report results with the help of the SAVEE database as shown in Fig. 1. The noteworthy stepping tool in accomplishing a decent presentation of the SER system is feature extraction. It is the process of extraction of certain types of emotion from the speech.

This paper utilizes a learning approach deep belief network (DBN) that normally assumes huge operations on data to extract out meaningful information based on the training condition of raw data. The classification or recognition stage is the decision making part of the recognition system. A DBN-SNN is used in this work for classifying and recognizing the speech emotion using the advantage of automatic differentiation. This technique can automatically and analytically afford the derivatives to the training algorithm, which will be optimized using an SSD optimization algorithm.

A. Formulation of a Dataset for Speech Recognition

Initially, the voice database D_s is utilized to extract relevant information from the database containing different emotions. *Ds* comprises both structured and unstructured data which comprises sonic-visual data which are recorded with English articulations in six distinct emotions namely anger, disgust, fear, happiness, sadness, surprise plus neutral. $d_1, d_2, d_3 \ldots d_n$ are the categorical labels of each utterance and can be displayed as follows.

$$D_s = d_1 + d_2 + d_3 + \ldots . d_n \quad (1)$$

The above equation (1) is selected because of the consideration of speech data. Experiments make use of speech alone as the input. The information from all speakers which is part of the training set is 70%, the validation set is 15% and a test set 15%. The audio data are having noises, and this data is fed into the preprocessing steps for the accurate extraction that is explained in the sections.

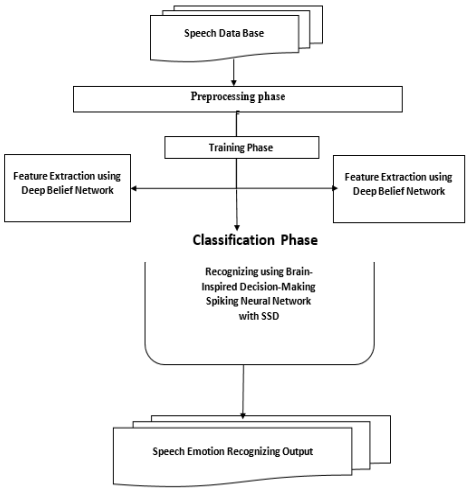


Fig. 1. Overall Schematic Diagram for Speech Emotion Recognition.

B. Data Augmentation

Over-fitting and aid generalization is diminished by means of a significant advance [21] termed as data augmentation. Since numerous data are necessary for DNNs, the preparation data was increased to orchestrate additional data. This analysis, this progression has fundamentally enhanced the generalization of the connected technique. Data in the preparation set was amplified by performing resampling the first sound at four diverse sampling frequencies, the first sampling recurrence. Few data augmentation procedures were investigated, for example, including Gaussian noise, yet were observed to be less effective than the detailed technique.

C. Preprocessing for Noise Removal

During the progress of the ASR system, pre-processing is the first phase in order to differentiate the voiced or unvoiced signal and to construct feature vectors in speech recognition. The speech signal, x (n), is being modified by preprocessing only then a noise-free input can be given for feature extraction analysis. Here x (n) must be tested to eradicate the background noise d (n).

$$x(n) = s(n) + d(n) \quad (2)$$

Where s (n) is the clean speech signal.

Different methods to diminish the noise are approved to work on a noisy speech signal. Spectral subtraction and adaptive noise cancellation are the two key methods under a noise reduction algorithm to develop a perfect speech recognition system.

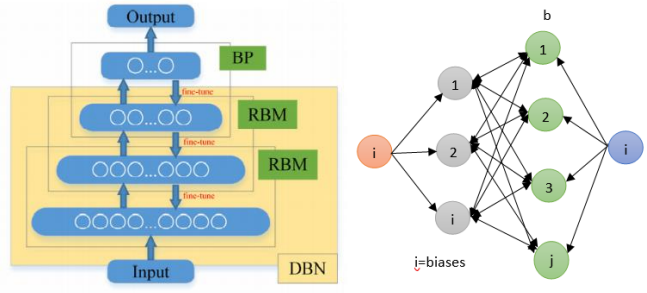
D. Feature Extraction using Deep belief Network

In this paper, DBN is utilized, utilizing speech acquired from the crude speech databases. Reenactment of the progressive way that the cerebrum forms the information is cultivated by a deep encoder or deep learning; it is the current propelled machine learning technique that connects with the deep structure to display the information circulation and inward structure. At first extraction of features from low level to abnormal state is executed by presenting deep learning technique which starts progressive structure. The next level is a grouping of the info. Incredible achievement is being envisioned quite a while back in PC vision and programmed speech acknowledgment through deep learning techniques. Indeed, even now the conventional machine learning technique is picked to be deep learning and just consideration on the precision with various low dimensional physically unmistakable features. The main sort of deep learning technique is shared by DBN, is acquainted with extracting meaningful emotional features from the data which is displayed in Fig. 6.

a) A brief introduction to DBN

By stacking various restricted Boltzmann machines (RBMs), DBN which is a generative model is being constructed, as illustrated in Fig. 2. Three components are needed to compose a common recognition model: collecting detecting signals, removing features and building connections; all these components need many manual efforts. DBN gives a foundation to make sense of the model straight from what we see to what we need to know. A sort of hierarchical feature

representation is resolved as the layer-by-layer structure. The brainwork-consuming feature-extracting segment is replaceable in light of the fact that the network training procedure is self-versatile. The layer by layer training procedure is directed via massive unlabeled samples.



(a) (b)

Fig. 2. Architecture of (a) RBM and (b) DBN.

Learning a likelihood conveyance over the preparation set is set up by RBM which is a generative stochastic counterfeit neural system dependent on statistical mechanics. Two layers of two-fold stochastic units are perceived as an obvious layer and a concealed layer. As to as an undirected graphical model, every single noticeable unit is related to every shrouded unit and nonattendance associations inside each layer. The visible unit biases G, hidden units biases I and connection weight J are the parameters included. The hypothetical inference of RBM starts from the meaning of framework vitality for a particular framework state, which describes a likelihood circulation over the joint state of the noticeable units and the shrouded units, formulated as:

$$J(b,m) = x(n) - \sum_{j=1}^{sB} g_j b_j - \sum_{i=1}^{sM}\sum_{j=1}^{sB} m_i C_{ij} b_j - \sum_{i=1}^{sM} h_i m_i \tag{3}$$

The visible unit i and the hidden unit j has the binary states as b_i and m_i. The number of visible units and hidden units is represented as sB and sM respectively. The joint distribution of these units is formulated as:

$$J(b,m) = \frac{1}{A} e^{-J(b,m)} \tag{4}$$

Where $A = \sum_{b,m} e^{-J(b,m)}$ is called the partition function. Among the associated two margin distributions are

$$U(b) = \frac{\sum_m e^{-J(b,m)}}{A} = U(m) = \frac{\sum_b e^{-J(b,m)}}{A}$$
$$U(b \mid m) = \frac{U(b,m)}{U(m)} = \frac{e^{-J(b,m)}}{\sum_{\tilde{m}} e^{-J(b,\tilde{m})}} \tag{5}$$

$$= \frac{\exp\left(\sum_{j=1}^{sB} g_j b_j + \sum_{i=1}^{sM}\sum_{j=1}^{sB} m_i C_{ij} b_j + \sum_{i=1}^{sM} h_i m_i\right)}{\sum_{\tilde{m}} \exp\sum_{j=1}^{sB} g_j b_j + \sum_{i=1}^{sM}\sum_{j=1}^{sB} \tilde{m}_i C_{ij} b_j + \sum_{i=1}^{sM} h_i \tilde{m}_i} \tag{6}$$

Similar derivation can be done to U(b|m)

$$U(b \mid m) = \prod_{j=1}^{sM} \frac{\exp\left(g_i b_j + \sum_{j=1}^{sB} m_i C_{ij} b_j\right)}{\sum_b b_j + \exp\left(g_i \tilde{b}_j + \sum_{j=1}^{sB} m_i C_{ij} \tilde{b}_j\right)} \tag{7}$$

$$U(b \mid m) = \prod_{j=1}^{sM} \frac{\exp\left(g_i b_j + \sum_{j=1}^{sB} m_i C_{ij} b_j\right)}{\sum_b b_j + \exp\left(g_i \tilde{b}_j + \sum_{j=1}^{sB} m_i C_{ij} \tilde{b}_j\right)}$$

Because of the nonattendance certain connections, the units in a single layer are restrictively free while other layers are provided, so we can obtain.

$$U(m_i = 1 \mid b) = \frac{1}{1 + \exp(-\sum_{j=1}^{sM} C_{ij} \tilde{b}_j + \sum_{i=1}^{sB} m_i C_{ij} - h_i)} \tag{8}$$

$$U(b_i = 1 \mid m) = \frac{1}{1 + \exp(-g_i \sum_{i=1}^{sB} m_i C_{ij}} \tag{9}$$

A maximum likelihood estimation is the best technique to train the RBM. The log-likelihood of the model for a single training sample is given below.

$$Q(\theta) = \log U(b \mid m) = \log \sum_m e^{-J(b,m)} - \log \sum_{b,m} e^{-J(b,m)} \tag{10}$$

Where Θ = {C, g, h} is the parameters to be evaluated. The gradient can be given as:

$$\frac{\partial(Q)}{\partial(\theta)} = \frac{\partial}{\partial(\theta)}\left(\sum_m e^{-J(b,m)} - \log \sum_{b,m} e^{-J(b,m)}\right) \tag{11}$$

$$= \frac{\sum_m e^{-J(b,m)}}{\sum_{b,m} e^{-J(b,m)}}\left(-\frac{\partial J(b,m)}{\partial(\theta)}\right) - \sum_{b,m} \frac{e^{-J(b,m)}}{\sum_{b,m} e^{-J(b,m)}}\left(-\frac{\partial J(b,m)}{\partial \theta}\right)$$

$$= \sum_m U(b \mid m)\left(-\frac{\partial J(b,m)}{\partial(\theta)}\right) - \sum_m U(b \mid m)\left(-\frac{\partial J(b,m)}{\partial \theta}\right)$$

Dual symbols are introduced to simplify the equation which is equated below:

$$\langle\theta\rangle_{data} = \sum_m U(b \mid m)\left(-\frac{\partial J(b,m)}{\partial(\theta)}\right) \quad \langle\theta\rangle_{model} = \sum_m U(b \mid m)\left(-\frac{\partial J(b,m)}{\partial \theta}\right) \tag{12}$$

The partial derivative of energy function to model parameters is briefed as follow:

$$-\frac{\partial J(b,m)}{\partial(\theta)} = b_i m_j \quad -\frac{\partial J(b,m)}{\partial \theta} = b_j \quad \frac{\partial J(b,m)}{\partial} = m_i \tag{13}$$

$\langle\theta\rangle_{data}$ can be calculated easily while the $\langle\theta\rangle_{model}$ needs to traverse all the probable value mixtures of the hidden units and visible units, this is called an NP-hard problem. The Gibbs inspecting begins with a training test, and on the other hand tests the shrouded units and unmistakable units utilizing condition (8) and (9) by k ventures, as delineated underneath:

$$b^{(0)} = t, \quad m^{(0)} \sim U(m \mid b^{(0)})$$

$$b^{(1)} \sim U(b \mid m^{(0)}), = t, \quad m^{(0)} \sim U(m \mid b^{(1)})$$

$$b^{(k)} \sim U(b \mid m^{(k-1)}), = t, \quad m^{(k)} \sim U(m \mid b^{(k)}) \tag{14}$$

When k->∞ the accurate model distribution can be gained and $\langle\theta\rangle_{model}$ can be calculated. In practice, Pro. Hinton brought up that the CD learning with k =1 can give satisfactory outcomes to appropriately gauge the model gradient. Therefore the subsequent term can be assessed utilizing $\langle\theta\rangle_{model}$ Gibbs sampling as

$$\langle\theta\rangle_{model} = \sum_{b,m} U(b,m)(-\frac{\partial E(b,m)}{\partial\theta})$$

$$\langle\theta\rangle_{model} = \sum_{b} U(b) \sum_{m} U(b \mid m)(-\frac{\partial E(b,m)}{\partial\theta}) \tag{15}$$

$$\langle\theta\rangle_{model} = \frac{1}{l} \sum_{b} \sum_{m} U(b \mid m)(-\frac{\partial E(b^{(1)}, m^{(1)})}{\partial\theta})$$

In handy application, the training data is partitioned into mini-batches to upgrade the registering effectiveness. What's more, a typical system is set l equivalent to the size of the mini-group. Partition of "mini-batch" from the total gradient by the information size mini-clump to abstain from varying the learning rate when the size of mini-group changes. Along these lines, with the stochastic gradient descent algorithm, the refreshing standards of the parameters can be formulated as:

$$\theta = \theta + \varepsilon\Delta\theta = \theta + \varepsilon(<\theta>)_{data} - <\theta>_{model} \tag{16}$$

Where ε is considered as the learning rate. The gradient for a size 1 mini-batch is equated as:

$$\Delta C_{ij} = \frac{\sum_{x=1}^{l} (m_{(x)}^{(0)} b_{(x)}^{(0)} - m_{(x),i}^{k} b_{(x),j}^{(k)}}{l} \tag{17}$$

$$\Delta g_j = \frac{\sum_{x=1}^{l} (b_{(x),j}^{(0)} - b_{(x),j}^{(k)})}{l}$$

$$\Delta h_i = \frac{\sum_{x=1}^{l} m_{(x),j}^{(0)} - m_{(x),j}^{(k)})}{l}$$

Where the $(.)_{(x),i}^{k}$ notation signifies the parameter of the x[th] training sample's i[th] element, Furthermore, k infers the example got after k-step Gibbs sampling. The entire construction is arranged avariciously layers one after the other using unlabeled preparing data on a course of action RBM units, and RBMs are all around arranged, their parameters are then spread out to the DBN network, and BP algorithm is accomplished to align the entire system using a much little game plan of named data. We accept that the degrees of deliberation associated with mapping a matrix of pixel esteems to an emotion class make it especially appropriate for examining profound structures. When these highlights have been learned in solo preparing, we expect that the model will effectively get familiar with the specific characterization task. Also, we anticipate that more profound designs will outflank shallow systems since our assignment was picked so advance progressive portrayals. We have less refined expectations concerning the impact of the size of concealed layers, yet for the most part, anticipate that driving data pressure by lessening the number of units in the shrouded layer will bring about better abstractions.

E. Brain-Inspired Decision-Making Spiking Neural Network

Deep learning uses engineering with numerous layers of trainable parameters and has exhibited remarkable execution in AI and AI applications. DNNs are prepared to start to finish by utilizing advancement calculations typically dependent on BP. The multi-layer neural design in the primate's cerebrum has enlivened specialists to focus on the profundity of non-straight neural layers as opposed to utilizing shallow networks with numerous neurons. Additionally, hypothetical and test results show preferable execution of deep rather over wide structures. DNN extricate complicated highlights by means of consecutive neuron layers prepared by nonlinear, differentiable activation functions to give a suitable stage to the BP calculation.

For most portrayal issues, a softmax module is utilized as the yield layer of a significant framework. One-hot encoding is utilized during the preparation vector. In a one-hot encoding, every vector segment is matched to the potential classes. This vector is parallel with precisely one section set to 1 that identifies with the ideal objective class. The softmax module for the yield layer provides assurances that the estimations of all of the yield units fall inside the range (0, 1) and total to 1. This gives a great deal of on a very basic level inconsequential and far-reaching probability regards. The softmax recipe now and again called the normalized exponential,

$$D_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{18}$$

Where, x_i, is the net input to a particular output unit, j indexes the set of output units, and D_i is the value of output unit i, which falls in the range (0, 1).

The input undergoes pre-processing through the input layer. The data is then sent to a progression of hidden layers, the quantity of which can fluctuate. As the data proliferates through hidden layers, includes that are increasingly mind-boggling are separated and learned.

a) Social Ski-Driver (SSD) Optimization Algorithm

The conduct of SSD which is a novel optimization algorithm was motivated by various evolutionary optimization algorithms. Its name compliments to the way that its stochastic exploration in some way or another looks like the ways that ski-drivers take downhill. SSD has numerous parameters; a short depiction of these parameters is given beneath.

1) The places of the agents (Xi € Ř^n) are utilized to figure the target work in the same area, where n is the search space element.

2) The best position previously P_i: The fitness function helps in determining the fitness value for all operators.

Contrasting the fitness value for every specialist is performed and it puts away both the present position and the best position. This is like the PSO algorithm.

3) Mean global solution M_i: In our algorithm, as in the GWO, the agents are directed toward the global point which signifies the mean of the finest three solutions given by eq (19).

$$R_i^t = \frac{x_\alpha + x_\beta + x_\gamma}{3} \tag{19}$$

Where X_α, X_β, and X_γ are the three solutions that are considered to be the finest.

The velocity of the agents (V_i): The agents' positions are updated by adding the velocity Vi as follows

$$X_i^{t+1} = X_i^t + B_i^t \tag{20}$$

Where

$$B_i^{t+1} = \begin{Bmatrix} h\sin(w_1)(u_i^t - x_i^t) + \sin(w_1)(R_i^t - x_i^t) & if\ w_2 \le 0.5 \\ h\cos(w_1)(u_i^t - x_i^t) + \cos(w_1)(R_i^t - x_i^t) & if\ w_2 \rhd 0.5 \end{Bmatrix} \tag{21}$$

Where B_i is the velocity of X_i, w_1 and w_2 are consistently created random numbers in the range of [0, 1], u_i is the superlative clarification of the ith agent, R_i is the mean global solution for the entire population, and h is a parameter makes the exploration constancy to the exploitation which is calculated as $h^{t+1} = \alpha h^t$, where the current iteration is denoted as t and $0<\alpha<1$ is used to decrease h value. Hence, h $\sin(w_1) <$ 0and h $\cos(w_1) > 0$, where t -> t_{max} and t_{max} is the highest times the iteration occurs. Equation (21) indicates, the moving directions for the agents are reversed as in GWO or PSO, and the reason behind this is the sine and cosine functions. Fig. 4 pictures an example with double agents moving in the SSD algorithm. Thus an offered calculation provides a superior guided investigation capacity and creates the hunt bearings to be differentiated,

Searching for near-optimal solutions in space is the main objective of the SSD. The dimension of that space is being determined by the number of parameters that must be optimized which is displayed in Fig. 3. In SSD, the agent's positions (Xi) are arbitrarily set, where the number of agents is fixed by the user. By the accumulation of the velocity to the previous position, the new position is being updated as in equation (20). The agents' velocities are also haphazardly initialized, and it is adapted based on equation (21).In this equation, the agent's adjusted velocity relies upon the partition amid the present positions, Xti, other than the past best position Pi, (2) the detachment between the present position, Xti, moreover the mean worldwide arrangement Mi. Hence, the agents in SSD advance in the direction of the mean of the finest three arrangements which makes the SSD calculation additional social compared to PSO. Furthermore, the SSD administrators are moved not a reasonable way, which in turn provides the SSD computation enhanced investigation precision. Fig. 4 clearly describes the flow of our proposed framework.

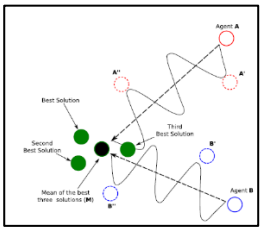


Fig. 3. Diagrammatic Representation of an Instance of how Two Agents (A and B) using the SSD Algorithm.

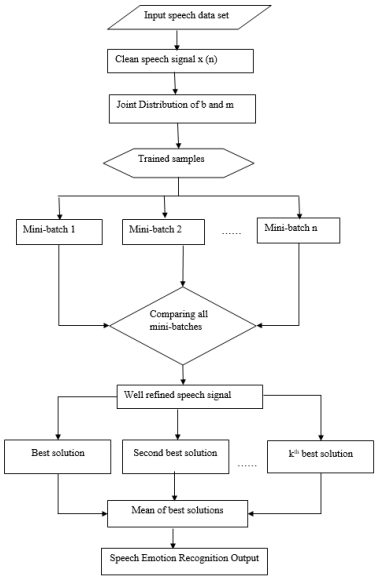


Fig. 4. Flow of our Proposed Framework.

IV. Simulation Results

System configuration: Operating System: Windows 8, Processor: Intel Core i3, RAM: 4 GB and Platform used is MATLAB.

A. Dataset Description

SAVEE [20] is an audio-visual database that includes 480 English articulations from four male entertainers in seven unique feelings, which are anger, disgust, fear, happiness, sadness, surprise plus neutral. Utterances were categorically labeled. The information collected from almost all the speakers as audio was randomly part in training (70%), validation (15%) and test (15%) sets. The mentioned sets are fundamentally unrelated.

B. Performance Evaluation

The performance of the proposed method was employed with some criteria such as accuracy, precision, recall F-measure. The formulation for precision, recall, accuracy, F-measure, is given below.

a) Precision

The precision is equated as follows

$$\Pr ecision(p) = \frac{Sum of\ relevant\ data\ \ detected}{Total\ sum of\ data\ \ detected} \tag{22}$$

b) Recall

The recall is equated as follows.

$$\text{Recall}(r) = \frac{\text{Sum of accurate data detected}}{\text{Total sum of relevant data in the database}} \quad (23)$$

c) Accuracy

The data accuracy is equated by an equation which is given below.

$$\text{Accuracy} = \frac{p+r}{2} * 100 \quad (24)$$

C. Simulation Results

Exploration of various deep learning architectures and their influences on the classification is by making use of SAVEE database. Solidly, the influences of the quantity of hidden layers neurons for a permanent quantity of layers and the number of layers for a permanent quantity of neurons per hidden layer will be taken under consideration for the study. In this way, a thorough report on measurements for the picked design for the SAVEE databases is seen. Fig. 5 portrays the input SAVEE voice dataset which is converted to be a speech signal that is the mixture of different emotions like anger, fear, happiness, sad and neutral, etc. Fig. 6 shows the feature extracted using DBN.

D. Comparative Analysis

a) Proposed different parameter value with a different emotion

Table I describes the performance of the SAVEE database on the proposed system with different emotions to attain the required values and its bar graph is displayed in Fig. 7. The overall precision value is 96.94%, recall value attained is 97.42% and accuracy values is 98.21%. The different emotion values of proposed parameters are depicted in Fig. 8 which describes precision, recall, f-score and accuracy for different emotions.

b) Performance evaluation with the proposed and existing technique

In Table II, the CNNs and LSTM network-based SER results are compared with our proposed methodology that is depicted in Fig. 9 which describes the different emotion with different attributes. The training and testing accuracy graph is displayed in Fig. 10.

Fig. 9 describes the parameter values of different emotions compared with different algorithms, like CNN [30], LSTM [30] and proposed technique. When compared to all other methodology, the recommended technique gives the better result. The overall average of CNN precision value is 87.49% and LSTM value is 80.78%, average recall value of CNN is 85.94% and LSTM is 83.83% and average F-score values for CNN and LSTM is 85.77% and 79.85, respectively. The proposed methodology has average precision value of 95.64%, average recall value is 96.06% and average F-score value 97.59% which is far better compared to CNN and LSTM techniques. Training accuracy values of proposed technique 96.32% and testing accuracy value is 94.25% compared to various existing works like support vector machine and DNN along with extreme learning machine (ELM) technique plotted in Fig. 10.

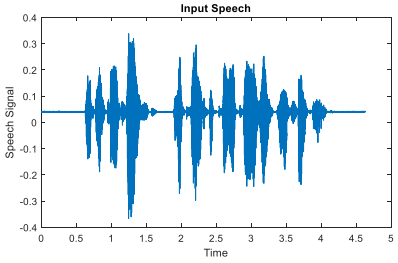


Fig. 5. Input Voice Database.

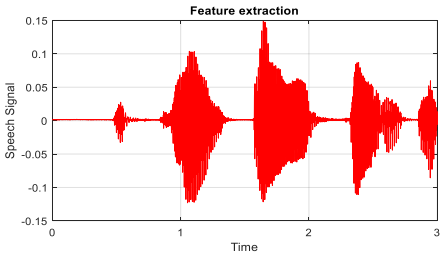


Fig. 6. Feature Extraction using Deep belief Network.

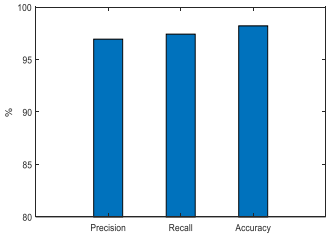


Fig. 7. Overall Proposed Performance Values.

TABLE. I. PERFORMANCE METRICS ON THE SAVEE DATABASE WITH A DIFFERENT EMOTION

EMOTION	PRECISION	RECALL	F-SCORE	ACCURACY
Anger	95.36	96.32	97.36	97.25
Disgust	97.25	95.32	96.87	96.32
Fear	95.36	97.25	97.98	96.21
Happiness	94.25	95.32	97.25	96.21
Neutral	96.32	95.32	98.78	96.32
Sadness	95.31	95.32	97.32	97.32
Surprice	95.32	96.32	97.25	98.32
All	98.32	94.56	97.22	98.57

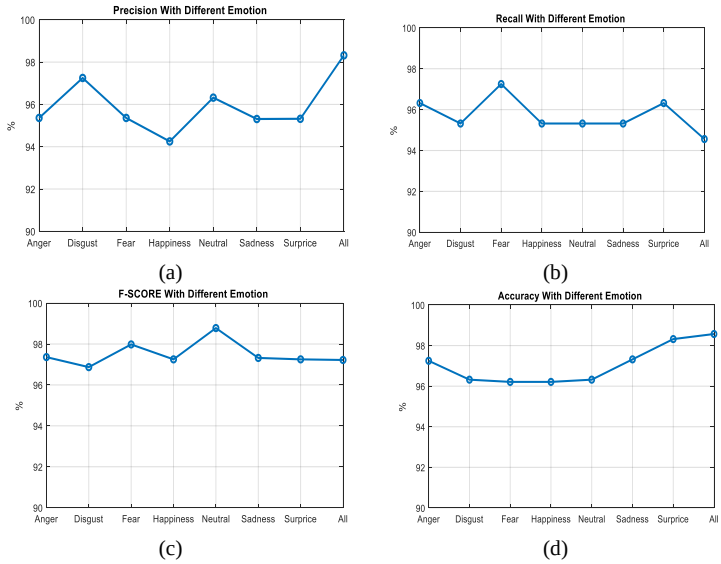


Fig. 8. Proposed Values with different Parameter (a) Precision (b) Recall (c) F-Score and (d) Accuracy.

TABLE. II. PERFORMANCE ANALYSIS OF THE PROPOSED AND EXISTING TECHNIQUE

	CNN [30]			**LSTM[30]**			**PROPOSED**		
EMOTION	P	R	F	P	R	F	P	R	F
Anger	87.16	92.58	89.48	84.82	89.04	86.46	95.36	96.32	97.36
Disgust	87.76	90.30	88.48	80.18	82.60	80.56	97.25	95.32	96.87
Fear	87.56	80.22	82.70	84.90	94.08	79.76	95.36	97.25	97.98
Happiness	80.88	68.86	73.26	73.90	66.84	68.78	94.25	95.32	97.25
Neutral	91.58	85.32	87.16	75.28	76.36	74.58	96.32	95.86	98.78
Sadness	90.04	98.40	93.56	85.62	94.08	88.96	95.31	96.32	97.32

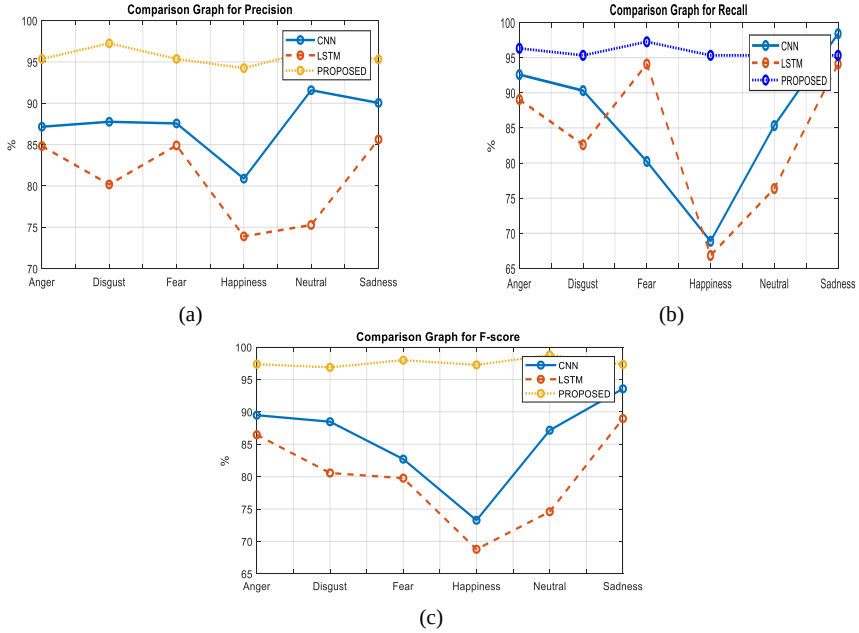


Fig. 9. Comparison Values of Existing and Proposed Methodology (a) Precision (b) Recall and (c) F-Score.

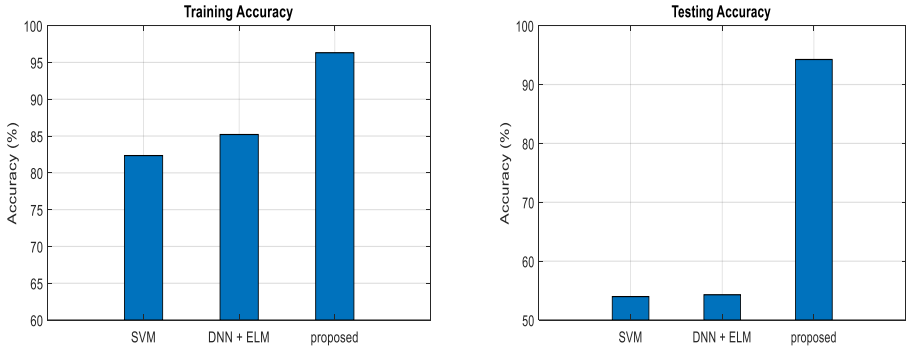


Fig. 10. Training and Testing Accuracy Graph.

V. CONCLUSION

The field of profound learning is adequately new to at present be quickly growing, frequently from development in new formalizations of profound learning issues driven by reasonable applications. Be that as it may, perceiving emotions from discourse is as yet a difficult issue. In this paper, we proposed the profound learning system brain-inspired decision-making spiking neural network with a social ski-driver (SSD) optimization-based network without utilizing any conventional hand-made highlights to characterize emotional discourse. For SER, we examined the recognition result by contrasting and the essential CNNs and LSTM based emotion recognition results. In all the existing methods adequate upswing in the computational costs was noticed which is being eradicated by introducing the deep learning-based feature extraction technique. Thus the results obtained by implementing the SNN-SSD algorithm related to artificial neural networks and long short term memory provides a refined classification of different emotions. At this point when contrasted with existing work our proposed work achieved better outcomes with a precision of about 96.94%, recall value of 97.42% and accuracy of about 98.21%. In the future the sound/video-based multimodal emotion recognition task can be implemented.

REFERENCES

[1] Morganti, F., & Riva, G. Ambient Intelligence for Rehabilitation. In G. Riva, F. Vatalaro, F. Davide, & M. Alcañiz (Eds.), Ambient Intelligence: The evolution of technology, communication and cognition towards the future of human-computer interaction (p. 281–292), 2005.

[2] Petrushin, “Emotion in speech: Recognition and application to call centers,” In Proceedings of ANN in engineering, vol. 710, pp. 22, 1999.

[3] Picard, Rosalind & Vyzas, Elias & Healey, Jennifer. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[4] Parkinson, M. H. Reviews : The Dual Voice. Free Indirect Speech and Its Functioning in the Nineteenth-Century European Novel. By Roy Pascal. Manchester: Manchester University Press, and Totowa, New Jersey: Rowman and Littlefield, 1977. Journal of European Studies, 9(35), 210–211,1979.

[5] M. El Ayadi, & F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” in Journal Pattern Recognition vol. 44, no. 3, pp. 572-587, 2011.

[6] J. Edwards, & P.E. Pattison, “Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review,” CPR, vol. 22, no. 6, pp. 789-832, 2002.

[7] Y.L. Lin, “Speech emotion recognition based on HMM and SVM,” ICMLC, vol. 8, pp. 4898-4901, 2005.

[8] A.V. Haridas, & V.G. Sivakumar, “A critical review and analysis on techniques of speech recognition: The road ahead,” KBIES, vol. 22, no. 1, pp. 39-57, 2018.

[9] K. Wang, & L. Li, “Speech emotion recognition using Fourier parameters,” vol. 6, no. 1, pp. 69-75, 2015.

[10] B. Schuller, & D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” Speech Communication , pp. 1062-1087, 2011.

[11] S. Pouyanfar, S.C. Chen, & S.S. Iyengar, “A survey on deep learning: Algorithms, techniques, and applications,” CSUR, vol. 51, no. 5, pp. 92, 2019.

[12] A. Hassan, “On automatic emotion classification using acoustic features”, Doctoral dissertation, University of Southampton, 2012.

[13] L. Li, & H. Sahli, “Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition,” HACACII, pp. 312-317, 2013.

[14] M. Sheikhan, & D. Gharavian, “Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method,” NCA, vol. 23, no. 1, pp. 215-227, 2013.

[15] X. Cheng, & Q. Duan, “Speech emotion recognition using gaussian mixture model,” In Proceedings of the 2012 International Conference on CASM, 2012.

[16] S. Shukla, M. Jain & Dubey R,K, “Increasing the Performance of Speech Recognition System by Using Different Optimization Techniques to Redesign Artificial Neural Network,” Journal of Theoretical and Applied Information Technology, vol. 97, no. 8, pp. 2404-2415, 2019.

[17] Shukla, S. & Jain, M. International Journal of Speech Technology, 22: 959. https://doi.org/10.1007/s10772-019-09639-0 ,2019.

[18] M. Gupta, & B. Kumar, “Novel class of stable wideband recursive digital integrators and differentiators,” IET Signal Processing, vol.4, iss.5, pp.560–566, 2010.

[19] M. Jain, M. Gupta, and N. Jain, “Linear Phase Second-Order Recursive Digital Integrators and Differentiators,” in Radioengineering, vol. 21, no. 2, 2012.

[20] M. Gupta, & B. Kumar, “Wideband digital integrator and differentiator,” IETE Journal of Research, pp. 166-170, 2012.

[21] M. Jain, & N.K. Jain, “The Design of the IIR Differentiator integrator and its Application in Edge Detection,” Journal of Information Processing Systems, vol. 10, iss. 2, pp. 223 - 239, 2014.

[22] M. Jain, M. Gupta, and N.K. Jain, “Design of half sample delay recursive digital integrators using trapezoidal integration rule,” International Journal of Signal & Imaging Systems Engineering, vol. 9, iss. 2, pp. 126 - 134, 2016.

[23] M. Jain, & N.K. Gupta, “Analysis and design of digital IIR integrators and differentiators using minimax and pole, zero, and constant optimization methods,” ISRN Electronics, vol. 2013, pp. 1 - 14, 2013.

[24] A.M. Badshah, & S.W. Baik, “Deep features-based speech emotion recognition for smart affective services,” MTA, vol. 78, no. 5, pp. 5571-5589, 2017.

[25] J. Zhao, X. Mao, & L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," BSPC, vol. 47, pp. 312-323, 2019.

[26] S. Gupta, K. De, D.A. Dinesh, & V. Thenkanidiyoor, "Emotion Recognition from Varying Length Patterns of Speech using CNN-based Segment-Level Pyramid Match Kernel based SVMs", NCC, pp. 1-6, 2019.

[27] Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., Li, X. "Learning Alignment for Multimodal Emotion Recognition" from Speech. Proc. Interspeech pp 3569-3573, DOI: 10.21437/Interspeech.2019-3247,2019.

[28] J. Hook, F. Noroozi, O. Toygar, & G. Anbarjafari, "Automatic speech-based emotion recognition using paralinguistics features," Bulletin of the polish academy of sciences technical sciences, vol. 67, no. 3, 2019.

[29] M.N. Mohanty, & H.K. Palo, "Segment based emotion recognition using combined reduced features," IJST, vol. 22, no. 4, pp. 865-884, 2019.

[30] W. Lim, D. Jang, & T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," APSIPA, pp. 1-4, 2016.