Received: January 1, 2019

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hybrid Deep Network Scheme for Emotion Recognition in Speech

Sujay Angadi[1]* **Venkata Siva Reddy[1]**

[1]*Reva University, Bangalore, Karnataka, India*
* Corresponding author's Email: Sujayangadi90@gmail.com

Abstract: Speech Emotion Recognition (SER) is an active research area with wide range of applications like medical, entertainment, monitoring in the field of Human Computer Interface. Generally, speech signals include high dimensionality of the feature that degrades the performance of SER. This research paper concentrated on SER using hybrid network that is composed of Convolutional Neural Network and Bidirectional Long Short Term Memory Networks (CNN-BLSTM). The major objective of this study is to recognize more relevant temporal features rather than traditional feature learning. The proposed CNN-BLSTM techniques increases the precision, recall and accuracy of emotion recognition in speech signal significantly. IEMOCAP dataset has been used in experimental analysis of the proposed approach to classify the different emotions of human; those are happy, angry, sad, and neutral. The performance of the CNN-BLSTM method has been measured using parameters like precision, recall and accuracy. Compare to the existing Support Vector Machine (SVM), the proposed CNN-BLSTM achieved approximately 9.83% of true positives proportion enhancement in SER.

Keywords: Bidirectional long short-term memory, Convolutional neural network, Deep learning, Speech emotion recognition.

1. Introduction

Now a day, SER is the challenging task in the research field because speech is the one of the significant element of human interaction. The emotion recognition technique uses different kinds of inputs like body language, facial expressions, speech, etc. [1]. Human speech is the natural mode of communication, hence emotion recognition majorly focused on the human voice [2]. The speech signals include both linguistic and paralinguistic information in emotion. Although, the emotional state does not alter the linguistic content, it's an important factor in human communication and improve the voice based emotion recognition [3]. The Speech Emotion Recognition system includes different emotional attributes like sad, angry, joy, fear, happiness, etc. [4, 5]. The SER is used in the different applications, for example, doctors may hear the patient voice for diagnosing the illness, analyze the telephone conversation of criminals in the crime investigation department [6].

Many researchers concentrated on voice, audio, speech, etc. for emotion recognition in human interactions. There are many methods for SER, those are Neural Network [7], wavelet Packet Transform [8], Support Vector Machine [9], etc. The existing SER technique have several issues such as a number of unknown features increase the computational complexity [10], noisy data, language-dependent [11], presence of irrelevant data, curse of dimensionality [12]. This research paper presented an efficient hybrid SER technique, which is combination of deep learning convolutional Neural Network (CNN) with Bidirectional Long Short Term Memory (CNN-BLSTM). The proposed CNN-BLSTM technique significantly reshape the dimensions and select the most relevant features. After selecting the significant features, it classifies the emotions like sad, happiness, angry or neutral. The significant contribution of the proposed CNN-BLSTM method is addressed below.

- The BLSTM networks are combined with CNN, combined network is helps to extract

the both spectral and temporal features for improve the SER performance.

- The proposed CNN with BLSTM method significantly reduce the feature dimensions in first and second layers.
- In SER process, proposed method identifies the four emotion classes such as happy, angry, sad and neutral.
- In training process, the proposed CNN-BLSTM method helps to improve the SER accuracy and gradually decreases the loss with respect to different number of iterations.

This paper is composed as follows. Section 2 presents a survey of several recent papers on deep learning based SER strategies. Section 3 explains an effective SER method CNN with BLSTM. Section 4 shows comparative experimental result for proposed and existing SER using reputed dataset. The conclusion is made in section 5.

2. Literature review

In the recent past, researchers have proposed many works for speech emotion recognition. Traditional methods mainly involved in feature based, i.e. choosing optimal features for classification problem. Some researchers also proposed dimensionality reduction on those features. Few works proposed deep learning based convolution neural network and recurrent neural networks. A brief evaluation of some essential contributions to the existing literatures presented in this section.

L. Sun, J. Chen, K. Xie, and T. Gu, [13] presented an efficient feature fusion technique named as Deep Convolutional Neural Network (DCNN) for improving the speech emotion recognition. As other deeper convolutional layers provided the abstract information and not detailed information of human emotions hence not suitable for SER. Moreover, shallow features consist of only global information and not considers the deep layer's information. In order to overcome the addressed issues proposed DCNN method combined the deep features and shallow features. The DCNN classified the emotion attributes effectively but number of iterations were high.

S. Lalitha, S. Tripathi, and D. Gupta, [14] developed the Speech Emotion Recognition (SER) using Deep Neural Network (DNN). This DNN algorithm classified the emotions significantly along with that extracted the relevant perceptual features. This DNN strategy helped to design the simple SER system that handled the dynamic emotion features in speech. The performance of SER decreased because the dataset was imbalanced.

K. Mannepalli, P.N. Sastry, and M. Suman, [15] presented Adaptive Fractional Deep Belief Network (AFDBN) strategy for SER. At first, extracted the spectral features from the input signal and those features were forwarded to the AFDBN for classification. The input speech signal was forwarded to the feature extraction phase to extract the relevant features finally; score values were forwarded to the AFDBN. The AFDBN method helped to find the optimal weights because it improved efficiency of emotion recognition. In DBN, network weights were updated iteratively with the help of fractional theory. After that, updated weights as well as bias terms were used during the testing phase. The number of features increased then network layers also increased so, computational complexity was raised.

C.K. Yogesh, M. Hariharan, R. Ngadiran, A.H. Adom, S. Yaacob, and K. Polat, [16] developed hybrid optimization technique that is combinations of Biogeography Based Optimization (BBO), Particle Swarm Optimization (PSO) (BBO-PSO) to improve the SER performance. The objective of this optimization technique was to minimize the subject related features and maximize the recognition rate effectively. Hence, the multi-cluster feature selection process was adopted. This optimization technique decreased the dimension of feature space and detected most relevant features from the search space. The recognition rate was varied for different cross-linguistic databases.

Y. Huang, K. Tian, A. Wu, and G. Zhang, [17] presented an efficient feature fusion technique Weighted Wavelet Packet Cepstral Coefficients (W-WPCC) for SER. The W-WPCC technique merge the sub-band features like band energies and spectral centroids with the help of the weighting strategy to produce the audio features. The extracted features were forwarded to the DNN. The DNN able to handle the non-linear features of training data and forecast the probability distribution over classification labels. As, in the network-training model, initialization of the weight parameters was fall into local optimal hence training period increased.

Many of the recent researches carried out on traditional feature learning, spectral and temporal learning to improve the SER. In order to rectify the above discussed problems, combinations of CNN and BLSTM technique proposed in this paper.

3. Proposed methodology

The most of the traditional SER method used to analyse the speech signal in the audio by representing two-dimensional data. Time-frequency analysis helps

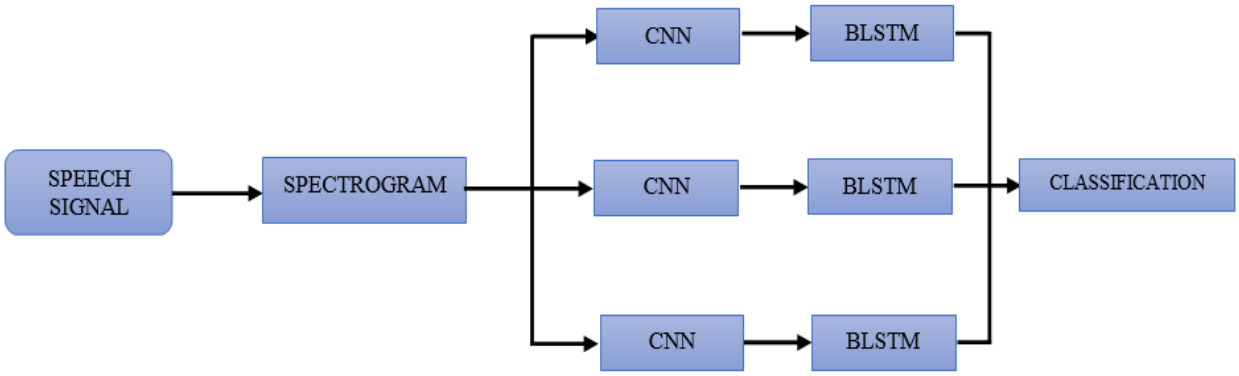

Figure.1 Proposed architecture

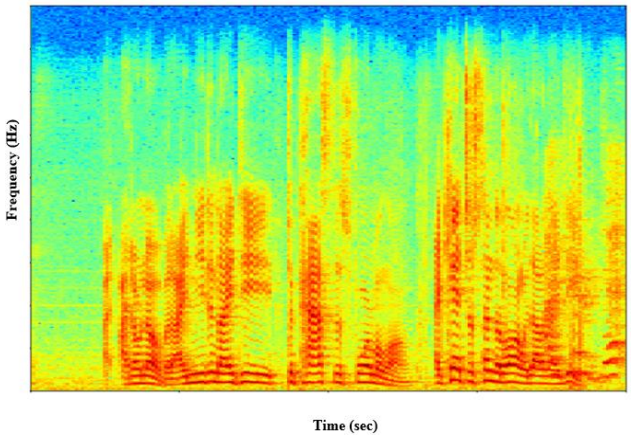

Figure.2 Generated spectrogram

to represent the spectrogram of speech signal. In this research work, the layers between CNN and BLSTM helps to handle the reshaping dimensions. The proposed CNN-BLSTM method helps to reduce the feature dimensionality in speech signal. The proposed system architecture is shown in the Fig. 1.

3.1 Spectrogram speech signal

The input signals converted into Spectrogram using Short Term Fourier Transform (STFT). These generated spectrograms fed as input to the proposed hybrid network based on Convolutional Neural Network and bidirectional Long Short Term Memory Networks (BLSTM).

Convolutional neural network alone can handle spectral feature extraction. As the BLSTM networks combined with CNN, it is able to extract both spectral and temporal features. The sample of generated spectrogram is shown in the Fig. 2.

3.2 Convolutional neural network and bi-directional long short term memory

The Convolutional Neural Network is the deep learning architecture and is widely used in many computer vision applications like image classification, scene labelling, face recognition and action recognition [18]. The image is divided into many tiles of fixed structure and is fed into the small neural network. Each extracted feature represented in the form of feature map, which formed by the set of arrays. The proposed block diagram of CNN-BLSTM is shown in the Fig. 3.

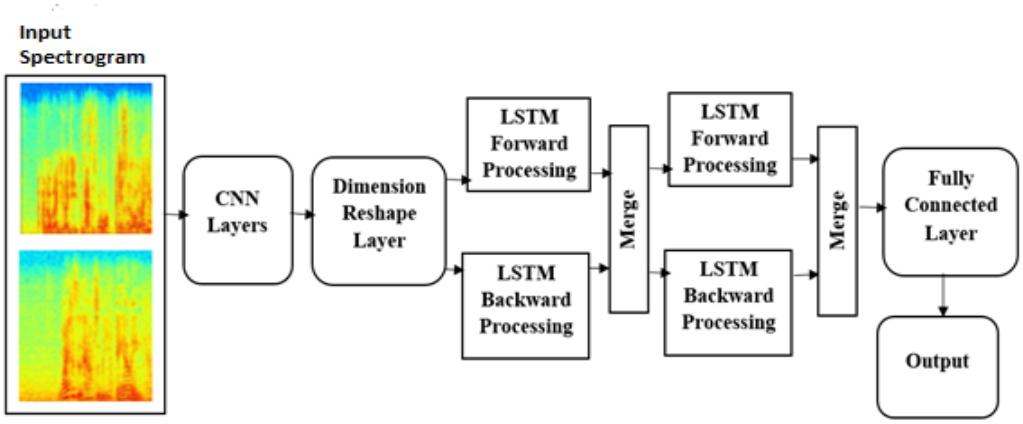

Figure.3 Proposed CNN-BLSTM setup

The convolutional layer is the primary element in the network; it uses the convolution kernel to extract the local information about the image texture, color, shape and other features. Moreover, convolutional operation improves the input features and reduce the noise interference. The mapping operation in convolution process is expressed in Eq. (1).

$$x_j^l = f_c\left(\sum_{i\in M_j} x_i^{l-1} \times k_{i,j}^l + \theta_j^l\right) \quad (1)$$

Whereas, x_j^l is indicated as l is the convolutional layer and j is the mapping set, x_i^{l-1} is the ith feature set indicating in the $(l-1)$ convolutional layer. $k_{i,j}^l$ is indicated as convolutional kernel between the i^{th} feature set and j^{th} mapping set in the layer l. The variable θ_j^l is the bias and f_c is the activation function. The next step is the pooling process; it reduces the possibility of overfitting during the training process. Pooling process is mathematically shown in the Eq. (2).

$$x_j^l = f_p\{\beta_j^l down(x_i^{l-1}) + \theta_j^l\} \quad (2)$$

Whereas, $down(.)$ represents the down sampling method from layer (l-1) to layer l^{th}. Commonly CNN includes maximum pooling and average pooling. β_j^l and θ_j^l indicates the multiplicative bias and additive bias respectively. In pooling layer, $f_p(.)$ is the activation function.

The matrix features of final pooling layer are sequentially taken out and arranged in a vector to form a rasterization layer and it's corresponded to the fully connected layer. The output of any node j can be expressed in Eq. (3).

$$h_j = f_h\left(\sum_{i=0}^{n-1} w_{i,j} x_i - \theta_j\right) \quad (3)$$

Whereas, $w_{i,j}$ indicates the connection weight of the input vector x_i. The node j, θ_j is the node threshold, $f_h(.)$ is the activation function. These fully connected layers are meant for classification process to produce class labels and the score. Also, it has been observed that number of filters in each convolutional layer has great impact on training speed as well as training effect. If there are a smaller number of samples and a greater number of filters, then this will lead to drastic decrease in the performance of the system.

LSTM and BLSTM are considered effective in learning sequential data as they utilize the data in the form of time steps. The normal LSTM is limited to the learning representation from the previous time steps that is considered as its major issue. The BLSTM method helps to understand the context and eliminate the ambiguity from the samples. To learn both forward and backward sequence data, BLSTM used two hidden layers in the same output layer. BLSTM proved better than unidirectional LSTM and used in several applications: phoneme classification and in speech recognition.

In the next stage, data from BLSTM fed to fully connected layer for the final classification. This fully connected layer contains soft max function. This soft max function provides the probability of each class label. Discrete probability function is indicated as p_d and given M classes in Eq. (4).

$$p_d = \sum_{m=1}^{M} p_m \quad (4)$$

Whereas, x is indicated as activation factor in previous layer in network and θ denotes the weight in the soft max function and o is the input to soft max function in Eq. (5).

$$o = \sum_i^{n-1} \theta_i x_i \quad (5)$$

And probability in Eq. (6),

$$p_m = \frac{\exp(o_m)}{\sum_{k=0}^{n-1} \exp(o_m)} \quad (6)$$

Thus, predicted class would be y,

$$y = \arg max\ p_i$$

$$i \in 1, \dots, N \quad (7)$$

Layer by layer filter setup composed of CNN, reshape layer and BLSTM in the proposed system is shown Table 1.

The CNN-BLSTM ensure the compatibility of the two algorithms by reducing the feature dimensions in speech signal. In later part, output of this layer fed to BLSTM that is composed of two

Table 1. Layer setup

Layer Number	Layers	Filters
1	Convolution	16
2	Max-Pooling	-
3	Convolution	32
4	Max-Pooling	-
5	Convolution	32
6	Reshape	-
7	BLSTM	64
8	BLSTM	64

layers. The 64 and 128 filters are composed in first and second layer in BLSTM respectively. The first layer between CNN and BLSTM handles the reshaping dimensions. This will ensure the compatibility of the two algorithms by decreasing the feature dimensions. The output of the first layer is given as input to the next layer BLSTM. The proposed CNN-BLSTM method gradually decreases the precision and recall of testing sets with the help of neural network strategy. Also, improves the SER performance with respect to different emotions.

4. Experimental result and discussion

For experimental simulation, the hardware configurations matter in the speed of training and testing of speech signals. The proposed work was implemented in GPU system, which configured with the GeForce GTX 1060, frame buffer of 6 GB and 16GB RAM. In order to estimate the efficiency of proposed algorithm, the performance of proposed method was compared with SVM algorithm on the reputed database: IEMOCAP dataset. The performance of the proposed method was compared in terms of precision, recall and accuracy.

4.1 Evaluation metrics

The relationship between the input and output variables of a system understand by employing the suitable performance metrics like precision, recall and accuracy. The general formula for calculating the precision, recall and accuracy of SER rate are given in the Eq. (8), Eq. (9) and Eq. (10).

$$Precision = \frac{TP}{(TP+FP)} \times 100 \quad (8)$$

$$Recall = \frac{TP}{(TP+FN)} \times 100 \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (10)$$

Where, TP is represented as true positive, FP is denoted as false positive, TN is represented as true negative and FN is stated as false negative.

4.2 Dataset description

The SER based Interactive Emotional Dyadic Motion Capture (IEMOCAP) that designed and developed at SAIL lab in university of south California [19]. This dataset contains 12 hours of data that derived from the five mixed gender pairs of actors. The actors were given script to perform. The whole dataset was divided into five sessions and the two actors per session. Again, these sessions were divided into spoken utterances. These utterances were labelled as angry, sad, neutral, frustrated, excited, fearful, surprised, disgusted, etc. Three annotators did the labelling. The majority vote was taken as the final sentiment for each utterance. The dataset provides segmented audio files in .wav format that was directly used for the processing. The dataset can be obtained from the SAIL USC website. In order to maintain consistency with the previous research, this work considered only four frequently found emotions in the dataset- Happy, anger, sad and neutral. The excitement class and happiness class are very similar so, it includes same features. The IEMOCAP dataset includes overall 5531 samples. Table 2 shows the selected emotions from the database. Detailed information about the dataset.

Table 2. Selected emotion attributes

Emotion Attributes	Training Samples
Happy	1636
Angry	1103
Sad	1084
Neutral	1708
Total	5531

Table 3. Training and testing split

Class	Train	Test	Total
Happy	1145	491	1636
Angry	772	331	1103
Sad	759	325	1084
Neutral	1196	512	1708
Total	3872	1659	5531

4.3 Performance of training and testing of different emotions

In this section, various emotion attributes of training and testing samples are tabulated in the Table 3. The class happy includes 1145 training samples and 491 testing samples, the class angry includes the 772 of training samples and 331 of testing samples. Similarly, sad attribute includes 759 and 325 training and testing samples respectively. Neutral emotion includes 1196 and 512 training and testing samples. The overall training and testing include the 3872 and 1659 samples respectively. Experimental setup for training and testing done in a ratio of 70-30 respectively. The train/test split is shown in Table 3.

All the parameters such as kernels, shift and number of epochs were fine-tuned to achieve better precision because of imbalanced data samples. Detailed results are tabulated in the confusion matrix is shown in Table 4.

Table 4. Confusion matrix of proposed and existing SER technique

Hybrid CNN-BLSTM method					**SVM [20]**			
Emotion Classes	Happy (%)	Angry (%)	Sad (%)	Neutral (%)	Happy (%)	Angry (%)	Sad (%)	Neutral (%)
Happy	79.83	6.52	9.17	4.48	56.80	9.75	7.54	25.88
Angry	6.37	80.64	7.0	5.99	12.65	59.83	8.21	19.29
Sad	10.31	8.31	72.76	8.62	14.03	7.75	60.75	17.45
Neutral	4.37	6.86	5.82	82.95	12.18	9.62	10.27	67.91

Table 5. Performance analysis of different classes of emotions

Different Methods	**Emotion Classes**	**Performance Parameters (%)**		
		Precision	**Recall**	**Accuracy**
SVM [20]	Happy	65.21	56.81	74.67
	Angry	61.53	59.83	75.80
	Sad	63.08	60.75	72.20
	Neutral	58.22	67.91	70.82
	Average	62.01	61.32	73.37
Hybrid CNN-BLSTM	Happy	79.83	79.13	81.6
	Angry	86.12	86.40	87.70
	Sad	72.76	76.79	79.77
	Neutral	82.95	81.29	82.4
	Average	80.41	80.90	83.2

4.4 Performance analysis of different emotion recognition techniques

In this section, performance of existing SER technique and proposed CNN-BLSTM approach is analyzed. SER analyzed and tabulated in Table 5. The performance of SER technique measured using efficient evaluation metrics such as precision, recall and accuracy with respect to different emotion classes such as happy, sad, angry and neutral. The proposed CNN-BLSTM method shows better results than the existing SVM method. In order to remain consistent with the existing research work we have considered same emotion classes for the analysis. According to the Table.5, Result analysis shows that emotion classes happy and neutral has yield better results compared to angry and sad classes considering recall and precision parameters. This is because of less training samples in angry and sad classes. However, in the IEMOCAP dataset, the speech of the two actors' overlaps with each other. This would have suppressed the performance to some extent.

According to the Table.5, Result analysis shows that emotion classes happy and neutral has yield better results compared to angry and sad classes considering precision and recall parameters. This is because of less training samples in angry and sad classes. However, in the IEMOCAP dataset, the speech of the two actors' overlaps with each other. This would have suppressed the performance to some extent.

The graphical representation of precision and recall is shown in the Fig. 4 and Fig. 5. According to Fig. 4, the x-axis represents the different emotion attributes and y-axis indicates the precision value (%). The SVM method showed the approximately 62.01% and 61.32% of precision and recall. The proposed CNN-BLSTM method achieved 80.41% and 80.90% of precision and recall respectively. The hybrid CNN-BLSTM approach significantly reduced the dimensionality of the features and selected the most relevant features. Hence, it showed better results compared to the other methods. Similarly, performance of recall is shown in the Fig. 5.

The graphical representation of accuracy performance is shown in the Fig. 6. The traditional SVM method showed the approximately 73.37% of accuracy for SER. In SVM method, more number of features are used and size of the samples are small hence, the SER performance was gradually decreased. The proposed CNN-BLTM method achieved approximately 83.2% of accuracy in SER. The proposed method used the more relevant features and significantly decreases the feature dimensions.

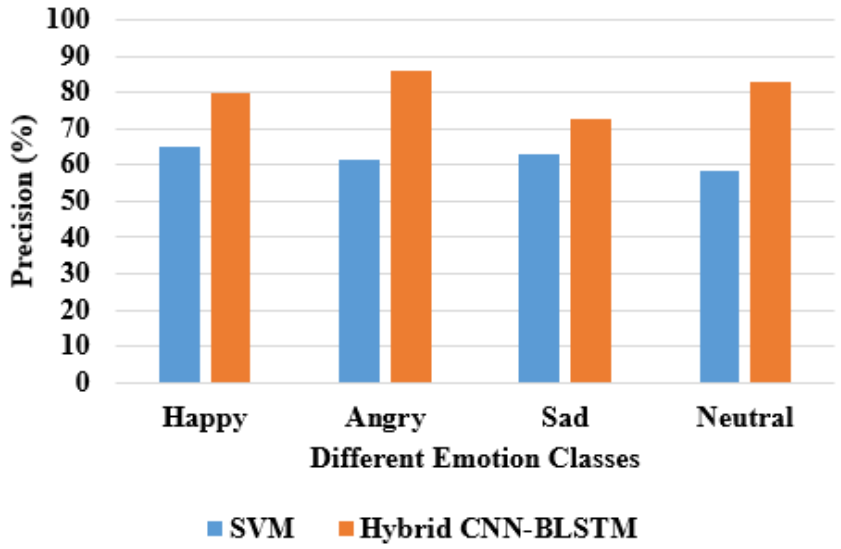

Figure.4 Performance of precision

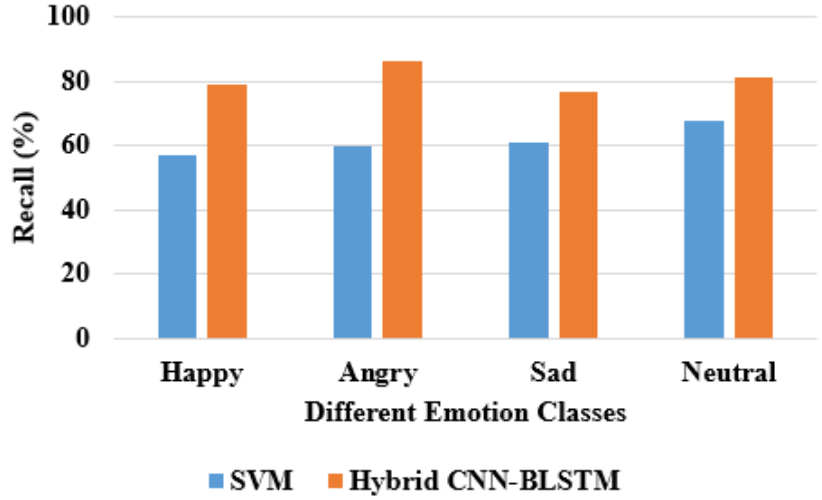

Figure.5 Performance of recall

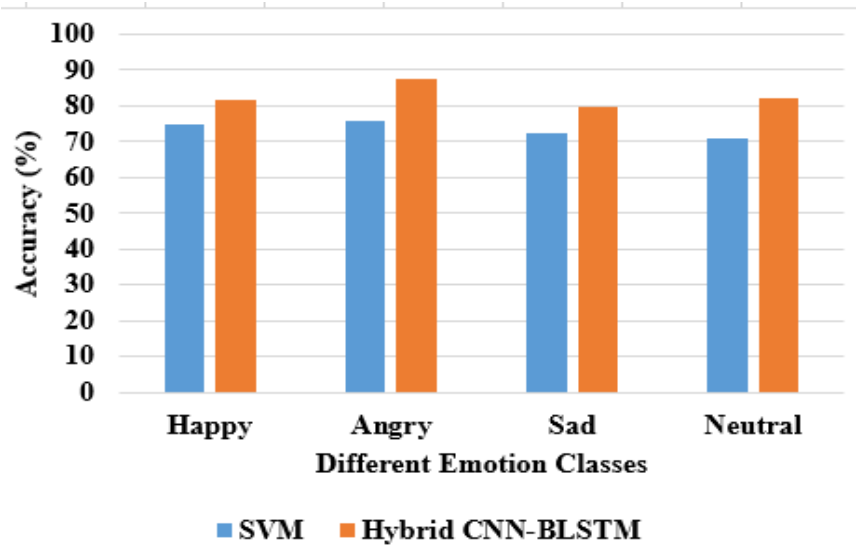

Figure.6 Performance of accuracy

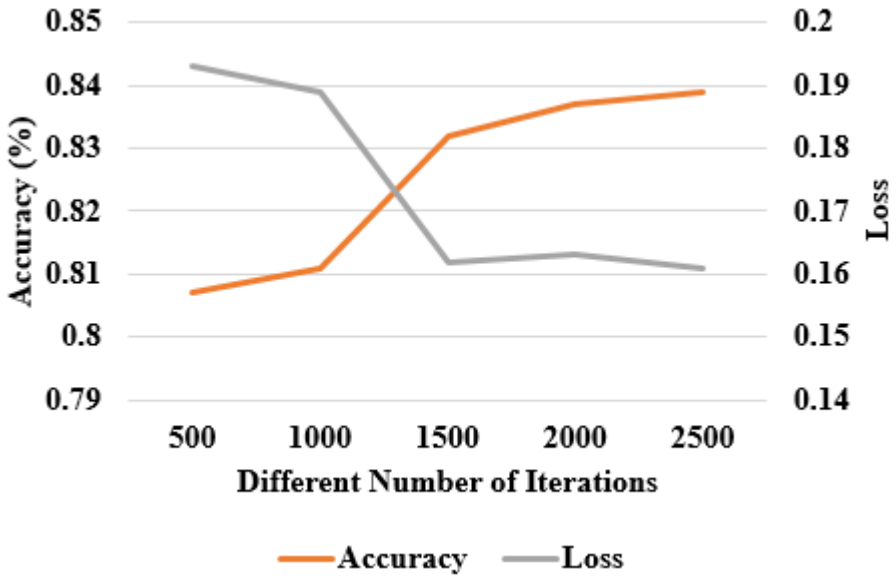

Figure.7 Performance of training process

In traditional AFDBN method [15] is employed in emotion recognition in speech signal data. The number of hidden neuron is varied from 200 to 600. These parameters are fixed by analysing their performance with various values and selected the best parameter value for the comparison. If the hidden neurons are increased, then computational complexity also gradually increased. The CNN-LSTM method approximately use the 1000 neurons and proposed CNN-BLSTM method use the approximately 400 neurons. The proposed method decreases number of the hidden layers without violating the recognition accuracy.

4.5 Analysis of emotion recognition efficiency based on CNN-BLSTM method

In this section, the training process of CNN-BLSTM method performed with respect to different number of iterations. In existing DCNN method [13] uses the shallow features and deep features for recognize the emotions in speech. But, the number iterations were maximum to obtain the accurate SER rate. Moreover, the SER accuracy is rate is maximum 64.78% with respect to 2500 iterations and training process takes approximately 45 min. But, proposed CNN-BLSTM method shows the better results and tabulated in Table 6.

Table 6. Performance of accuracy and loss

Number of Iterations	**500**	**1000**	**1500**	**2000**	**2500**
Accuracy	0.807	0.811	0.832	0.837	0.839
Loss	0.193	0.189	0.162	0.163	0.161

According to the training process of CNN-BLSTM method, 0.807% of accuracy is achieved and 0.193 of loss in terms of 500 iterations. In 1000 iterations, 0.811% of accuracy and 0.189 of loss. Similarly, 0.832%, 0.837% of accuracy and 0.162, 0.163 of loss with respect to 1500 and 2000 iterations respectively. In 2500 iterations, CNN-BLSTM achieved approximately 0.839% of accuracy and 0.161 of loss. The emotion recognition accuracy is approximately constant (0.83%) nearly 1500 iterations. Also, if the number of iterations are increased then loss is gradually decreased. The graphical representation of the accuracy and loss performance is shown in the Fig. 7. The training process graph is clearly shows the accuracy is gradually increased and loss is slowly decreased with respect to different number of iterations.

5. Conclusion

Machine learning field is rapidly expanding with the deep learning capabilities. SER is one of the challenged application in deep learning. In this research paper, an efficient SER framework based hybrid neural networks composed of CNN with BLSTM method is proposed. The effectiveness of the proposed system relies on the learning approach, which handles both spectral and temporal features. In real time, each person reveals his or her emotions with different degree and in certain manner. The hybrid CNN-BLSTM method achieved approximately 9.83% of improvements in SER. The recognition system identifies the differences between acted and spontaneous speech. In future, this research work can be extended using three-dimensional CNN

based SER technique that can be implemented on multiple languages of datasets.

References

[1] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol.1, pp.3, 2017.

[2] V. V. R. Vegesna, K. Gurugubelli, and A. K. Vuppala, "Application of Emotion Recognition and Modification for Emotional Telugu Speech Recognition", *Mobile Networks and Applications*, pp.1-9. 2018.

[3] M. Sheikhan, M. Bejani, and D. Gharavian, "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method", *Neural Computing and Applications*, Vol.23, No.1, pp.215-227, 2013.

[4] Z. Farhoudi, S. Setayeshi, and A. Rabiee, "Using learning automata in brain emotional learning for speech emotion recognition", *International Journal of Speech Technology*, Vol.20, No.3, pp.553-562. 2017.

[5] M. Sheikhan, D. Gharavian, and F. Ashoftedel, "Using DTW neural–based MFCC warping to improve emotional speech recognition", *Neural Computing and Applications*, Vol.21, No.7, pp.1765-1773. 2012.

[6] S. R. Krothapalli and S. G. Koolagudi, "Characterization and recognition of emotions from speech using excitation source information", *International Journal of Speech Technology*, Vol.16, No.2, pp.181-201, 2013.

[7] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification", *International Journal of Speech Technology*, Vol.15, No.3, pp.335-349, 2012.

[8] V. N. Degaonkar and S. D. Apte, "Emotion modeling from speech signal based on wavelet packet transform", *International Journal of Speech Technology*, Vol.16, No.1, pp.1-5, 2013.

[9] Y. Chen and J. Xie, "Emotional speech recognition based on SVM with GMM supervector", *Journal of Electronics (China)*, Vol.29, No.3-4, pp.339-344, 2012.

[10] C. Huang, B. Song, and L. Zhao, "Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering", *International Journal of Speech Technology*, Vol.19, No.4, pp.805-816, 2016.

[11] J. S. Park and J. H. Kim, "Emotional information processing based on feature vector enhancement and selection for human–computer interaction via speech", *Telecommunication Systems*, Vol.60, No.2, pp.201-213, 2015.

[12] A. Milton and S. T. Selvi, "Four-stage feature selection to recognize emotion from speech signals", *International Journal of Speech Technology*, Vol.18, No.4, pp.505-520, 2015.

[13] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition", *International Journal of Speech Technology*, pp.1-10, 2018.

[14] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks", *International Journal of Speech Technology*, pp.1-14, 2018.

[15] K. Mannepalli, P. N. Sastry, and M. Suman, "A novel adaptive fractional deep belief networks for speaker emotion recognition", *Alexandria Engineering Journal*, Vol.56, No.4, pp.485-497, 2016.

[16] C. K. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat, "Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech", *Applied Soft Computing*, Vol.56, pp.217-232, 2017.

[17] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition", *Journal of Ambient Intelligence and Humanized Computing*, pp.1-12, 2017.

[18] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features", *Signal length, and acted speech*. arXiv preprint arXiv:1706.00612, 2017.

[19] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", *Language Resources and Evaluation*, Vol.42, No.4, pp.335, 2008.

[20] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis", *Neurocomputing*, Vol.261, pp.217-2, 2017.