IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 7, 2020, accepted April 21, 2020, date of publication April 27, 2020, date of current version May 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990405

Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM

MUSTAQEEM[1], MUHAMMAD SAJJAD[2], AND SOONIL KWON[1]
[1]Interaction Technology Laboratory, Department of Software, Sejong University, Seoul 05006, South Korea
[2]Digital Image Processing Laboratory, Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan

Corresponding author: Soonil Kwon (skwon@sejong.edu)

This work was supported in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea Government through the Ministry of Science and ICT (MSIT) (Voice emotion recognition and indexing for affective multimedia service) under Grant 2017-0-00189, and in part by the 2020 Faculty Research Fund of Sejong University.

ABSTRACT Emotional state recognition of a speaker is a difficult task for machine learning algorithms which plays an important role in the field of speech emotion recognition (SER). SER plays a significant role in many real-time applications such as human behavior assessment, human-robot interaction, virtual reality, and emergency centers to analyze the emotional state of speakers. Previous research in this field is mostly focused on handcrafted features and traditional convolutional neural network (CNN) models used to extract high-level features from speech spectrograms to increase the recognition accuracy and overall model cost complexity. In contrast, we introduce a novel framework for SER using a key sequence segment selection based on redial based function network (RBFN) similarity measurement in clusters. The selected sequence is converted into a spectrogram by applying the STFT algorithm and passed into the CNN model to extract the discriminative and salient features from the speech spectrogram. Furthermore, we normalize the CNN features to ensure precise recognition performance and feed them to the deep bi-directional long short-term memory (BiLSTM) to learn the temporal information for recognizing the final state of emotion. In the proposed technique, we process the key segments instead of the whole utterance to reduce the computational complexity of the overall model and normalize the CNN features before their actual processing, so that it can easily recognize the Spatio-temporal information. The proposed system is evaluated over different standard dataset including IEMOCAP, EMO-DB, and RAVDESS to improve the recognition accuracy and reduce the processing time of the model, respectively. The robustness and effectiveness of the suggested SER model is proved from the experimentations when compared to state-of-the-art SER methods with an achieve up to 72.25%, 85.57%, and 77.02% accuracy over IEMOCAP, EMO-DB, and RAVDESS dataset, respectively.

INDEX TERMS Speech emotion recognition, deep bidirectional long shot term memory, key segment sequence selection, normalization of CNN features, radial-based function network (RBFN).

I. INTRODUCTION OF SER

Automatic recognition and identification of emotions from speech signals in speech emotion recognition (SER) using machine learning is a challenging task [1]. SER is a quick and usual method of communication and exchanging information among humans and computers and has many real world applications in the domain of Human-computer interaction (HCI). Currently, researchers are facing a major challenge in feature extraction i.e., how to select a robust method to extract salient and discriminative features from speech signals to represent the emotional state of a speaker from their acoustic contents. In the past decade, many researchers have investigated low-level handcrafted features for SER such as energy, zero-crossing, pitch, linear predictor coefficient, Mel-frequency MFCC, and nonlinear features such as tiger energy operator. Nowadays, mostly researchers utilize deep learning techniques for SER using Mel-scale filter bank speech spectrogram as an input feature. A spectrogram is a 2-D representation of speech signals which is widely used in convolutional neural networks (CNNs) for extracting the salient and discriminative features in SER [2] and other signal processing applications [3], [4]. Mostly 2-D CNNs are specially designed for visual recognition tasks [5]–[7]

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

and researchers are inspired by their performance to explore 2-D CNNs in the field of SER. Spectrograms are suitable representations of speech signals for CNNs model to extract high-level salient information to recognize emotions in speech signals. Similarly, some researchers have developed fully convolutional networks (FCNs) with the help of CNN's to handle fix input of variable size. The FCNs achieved a good performance output in time series classification tasks based on fix input variable size [8]. The lack of FCNs is not able to learn temporal information regarding this issue, the LSTM-RNNs is suitable for learning special and temporal features among sequences [9]. In the field of SER in this era CNN-LSTM and LSTM-RNNs are widely used for extracting hidden temporal information [10]. Some researchers are working to improve the recognition performance of SER to select some salient segments from speech signals and to learn temporal features using the CNN-LSTM model [11]. Badshah *et al.* [12] proposed a method for SER using the CNN features for smart effective devices to recognize the emotional state of the person in health care centers.

SER is an active area of research, recently researchers are utilizing deep learning techniques to develop a variety of methods to recognize the emotional state of speakers. Typically, researchers utilize the CNNs model, to learn high-level salient and discriminative features and feed them to the LSTM network to learn hidden temporal features to recognize emotions among sequences. The usage of CNNs and artificial intelligence increases recognition accuracy, but computation cost also increased with the usage of huge networks weight. The present traditional CNNs and LSTM architectures have not shown the substantial enhancement for increasing the level of accuracy and reducing the cost complexity of the existence SER systems. In this research, we proposed a novel deep learning-based approach for SER using RBF based K-mean clustering with a deep BiLSTM network. In the proposed method, we select the emotional segments from whole audio, utilizing RBF based similarity measurement technique to select one segment from each cluster. The selected sequence of segments is converted into spectrograms using the STFT algorithm. Furthermore, we extract the high-level discriminative features from selected segments utilizing the ''FC-1000'' layer of the Resnet101 [13] model. After that, we use the mean and standard deviation strategy to normalize the features and feeds to deep BiLSTM network for extracting temporal information and recognize the final state. The Softmax classifier is used for producing the probability among speech emotions. The main contributions of the proposed technique are documented below:

1. We proposed an efficient and novel framework for SER that is able to learn spatial and temporal information from speech spectrogram by leveraging CNN with deep bidirectional LSTM. Our model is capable to learn features and automatically model the temporal dependencies. To the best of our knowledge, the CNN model used in our research is a novel one in SER domain, therefore, we aim to contribute to the SER literature by using ResNet101 features in an effective manner, integrated with sequential learning mechanism.
2. We proposed a new strategy for SER by using sequence selections and extraction via non-linear RBFN based method to find similarity level in clustering. We select one key segment from the whole cluster which is near to the centroid of the cluster and represents the rest of the segments. Furthermore, we process these key segments to ensure the accurate recognition of emotion and reduce the processing time, as proved from the experimentations.
3. We endorsed, that the presented technique is a recent success of a deep learning approach based on key segments sequence selection and normalization of CNN features based on mean and standard deviation that can easily improve the existing state-of-the-art methods. To the best of our knowledge, this is a new deep learning approach for SER based on RBFN with CNN and deep BiLSTM. Thus, the key contribution of our framework lies in the usage of normalization technique to enhance the usage of features.
4. We tested the proposed SER model over different benchmark datasets and evaluated from different perspectives with baseline methods, the results are encouraging and are suitable for monitoring to recognize the real time emotions of the speakers. The achieved accuracy for IEMOCAP, EMO-DB, and RAVDESS dataset is 72.25%, 85.57%, and 77.02%, respectively.

The rest of the manuscript is distributed into the following folds: literature about the existing techniques of SER is documented in Section II, the detail explanation of the suggested framework of SER is elaborated in Section III, the experimental result of the mentioned technique are given in Section IV, and the detail discussion of the experimentations is mentioned in Section V, in the last Section VI, including on conclusion and future work of the proposed SER.

II. LITERATURE REVIEW OF SER

Digital signal processing is an emerging field of research in this era. Recently, many researchers have developed a various approaches in this area for SER from over the past decade. Typically, the SER task is divided into two main sections: features selection and classification. The discriminative features selection and classification method that correctly recognizes the emotional state of the speaker in this domain is a challenging task [14]. With the increase in data and cost computation deep learning approach is rapidly used for SER [15] and many researchers are used deep learning approaches for robust features representation in various fields [11]. Due to their enormous achievement in recognition of visual tasks, Huang *et al.* [4] presented a CNN based approach for SER and similarly, [16] used CNN to learn high-level discriminative features from spectrograms of speech signals and recognize the emotional state of speakers. Some researchers are used the Gaussian mixture model to classify the emotional state of speaker with robust features [17].

Nowadays, mostly researchers have worked with 2-D CNNs to extract high-level discriminative features from speech signals. Hence, extracting spectrograms, plotting speech signals with respect to time and feeding to CNNs to learn hidden information has become a new trend of research in this era for SER [2], [18]. Similarly, we can utilize the transfer learning strategies for SER using speech spectrograms passing through pre-trained CNNs models like VGG [5] or Alex-Net [19]. Spectrogram is a suitable representation for CNNs model to extract high-level discriminative features from speech signals to recognize the emotional state of the speaker in the SER system [20]. Similarly, LSTM-RNNs are mostly used to learn hidden temporal information in speech signals which is cyclically employed in the SER system [21], [22]. Nowadays, deep learning approaches play a crucial role to increasing the research interest in SER. Recently in [23] presented an end to end LSTM-DNN based model for SER with the combination of LSTM layers and fully connected layers to directly extract representation from raw data rather than obtaining hand-crafted features.

The joint approach of CNN-LSTM is presented in [24] to extract the deep salient high-level features from raw speech data using CNN and passed to the LSTM network for capturing the sequential information similar to [25]. Ma *et al.* [26] presented a neural network structure to take the variable-length speech for SER. In this method, CNN was used to represent the features of speech spectrograms and RNNs handled the variable-length speech segments. Zhang *et al.* [27] presented a technique for SER by utilizing the pre-trained Alex-Net model for features representation and traditional support vector machine (SVM) for emotions classification. Similarly, Liu *et al.* [28] used the CNN-LSTM model for spontaneous SER using the RECOLA [29] natural emotion dataset.

In the field of SER, many methods utilize CNN models with different types of input to extract salient features from speech signals to boost the recognition accuracy [30]. Similarly, some researchers used the pre-trained model to extract the high-level features from speech spectrograms and trained a separate classifier [31] for recognition, which boosts the cost computation of the system. In this paper, we developed a novel SER technique to process some useful segments from the whole audio file which are selected through K-mean clustering algorithm using RBF based similarity measurement. The selected segments of speech are converted into spectrograms and extract high-level discriminative features utilizing the CNN model called Resnet101. Furthermore, we normalized these features using mean value and standard deviation then feed them to deep BiLSTM network to learn hidden temporal information from speech segments to recognize the final emotional state of speakers. The proposed system reduces the execution time due to process selected segments rather than all segments and increases the level of accuracy due to used salient and normalized features with deep BiLSTM network. According to the best of our knowledge, the proposed architectures are novel and efficient than all other methods which are described in the literature.

III. PROPOSED TECHNIQUE OF SER

In this section, the proposed methodology of the SER framework and its main components are discussed in detail including the emotion recognition in speech. The suggested framework consists of the main three blocks. The first block consists of two parts; in the first part, we divide the audio file into multiple segments with respect to time and find the difference between consecutive segments. The obtained difference is to pass from a threshold to ensure the similarity and find out the value of "K" for clustering utilizing the shot boundary detection method [32]. Primarily start K = 1, and estimate the pairwise difference if the consecutive segment difference within threshold when the difference exceeds from threshold the "K" value automatically increases by one unit. Due to this process, we select the value of "K" dynamically for clustering to make groups accordingly. Furthermore, we select one segment from each group or sequence as a key segment that is near to the center of the cluster. We utilized the RBF, strategy for similarity measurement inside the clustering algorithm which is explained in section III (B) with detail. In the second part, the selected sequence of key segments is converted into spectrograms, plotting the frequency with respect to time using STFT. In the second main block, we work with features learning to extract the salient and discriminative features from speech spectrograms with transfer learning strategy utilizing the "FC-1000" layer of pre-trained Resnet101 [13]. The detailed specification of each unit and layers of the proposed Resnet model is mention in **Table 1**. The learned features are normalized with the help of mean and standard deviation for better performance. In the last block, we feed the extracted normalized CNN features to the suggested deep bi-directional LSTM to learn temporal cues and recognize the sequential information in a

TABLE 1. The overall specification of Resnet 101 is illustrated with set of convolutions layers, output size, and number of units which consist of kernel size, stride, and number of channels.

Layer	Output Size	No. of Units
Conv1	112×112	[7×7, stride 2, channel 64] × 1 [3×3 max pooling, stride 2] ×1
Conv2_x	56×56	[1×1, channel 64] × 3 [3×3, channel 64] × 3 [1×1, channel 256] × 3
Conv3_x	28×28	[1×1, channel 128] × 4 [3×3, channel 128] × 4 [1×1, channel 512] × 4
Conv4_x	14×14	[1×1, channel 256] × 23 [3×3, channel 256] × 23 [1×1, channel 1024] × 23
Conv5_x	7×7	[1×1, channel 512] × 3 [3×3, channel 512] × 3 [1×1, channel 2048] × 3
Output	1×1	Average pooling Fc-1000, & softmax

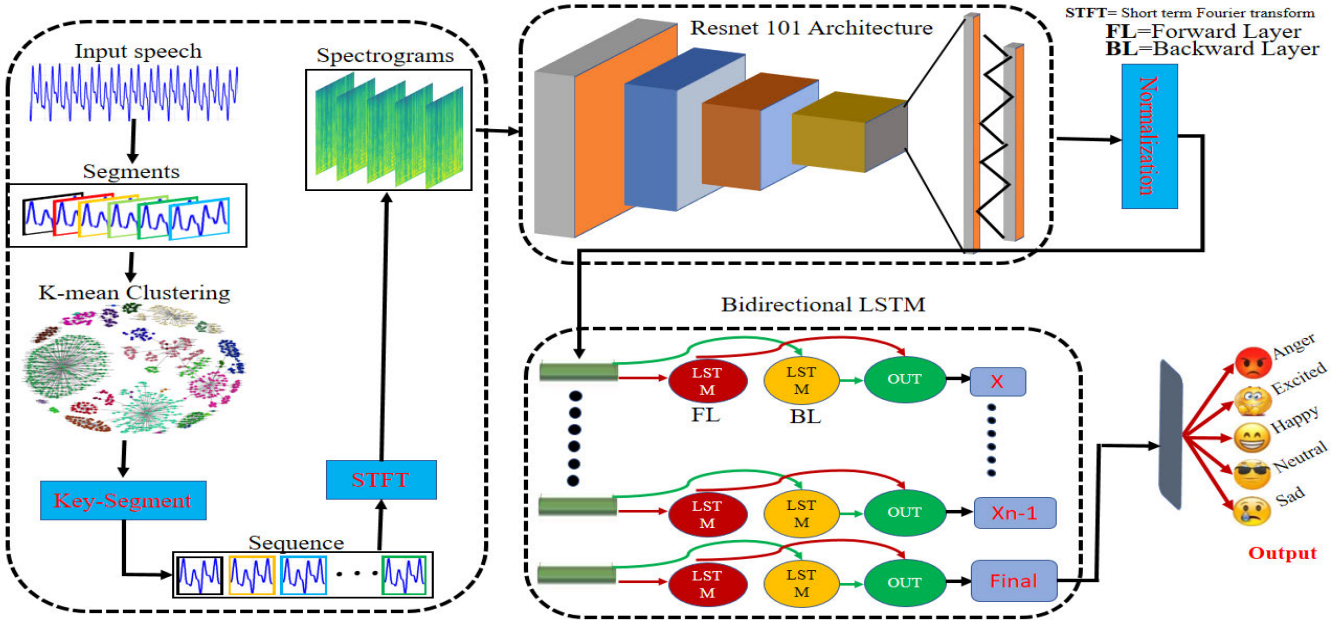

FIGURE 1. The proposed architecture for speech emotion recognition using dynamic clustering based Key segments selection with deep bi-directional LSTM and normalized CNN features.

sequence and analyze the final emotional state of the speaker in speech signals. The proposed framework diagrammatical representation is shown in **Figure 1**. The detailed description of each block of the framework are discussed in the subsequent sections.

A. PRE-PROCESSING AND SEQUENCE SELECTION

In this section, we split the audio file into multiple chunks (frames) concerning a suitable time and convert the whole utterance into segments. The selection of suitable time for the audio segment is a challenging problem in this era. Many researchers have worked, how to select a suitable time for each speech segment which has found some reasonable solution, that a segment of a speech signal is longer than 260ms that have more information to recognize the emotions in his/her speech [33], [34]. In this paper, we have done different observations on multiple frame durations to optimally select 500ms window size to convert single utterance into several segments. Single label is assigned to all segment of one utterance and give to K-mean clustering [35] algorithm to group the similar segment with each other. The K-mean clustering algorithm is most widely used for grouping the big data [36]. The Euclidean distance matrix [37], [38] is conventionally used in K-mean clustering technique for computing difference within elements. But in this work, we used the Radial Basis Functions (RBF) [37], [38] replaced by Euclidean distance matrix in K-mean for computing the difference between two frames. Because the RBF approach has been used for a non-linear method just like human brain's to compute the difference and recognize the patterns. The other important part is the selection of "K" value for partitioning the data into "K" groups. K-mean algorithm uses the random initialization technique to select the value of "K", but in this approach, we select the "K" value for each file dynamically by using the shot boundary detection method to estimate the similarity [32]. The pairwise difference is computed in the consecutive frame and if the difference is greater than the selected threshold value then increment the "K" value by one unit. After the total segments have been clustered using K-mean algorithm and one segment is selected from each cluster as key segment which near to the centroid of the cluster based on the RBF distance method, which is explained in the upcoming section. The selected key-segments are converted into spectrograms based on STFT algorithm for 2-D representation.

B. SIMILARITY MEASURING BASED ON RBF

In this section, we documented the detailed description of the non-linear similarity measure within audio signal segments. We also discuss the RBF based similarity approach for audio signal processing. The RBF uses the non-linear approach to compute the similarity between segments based on nonlinearity [39]. The visual perception section of the human brains also works on the non-linear processing system to differentiate and recognize the patterns. Hence, we use this approach in our proposed framework for finding the similarity measurement within audio segments.

We explore the RBF to simulate the non-linear human perception model to capture and compute the similarity between audio segments. Our model is also working as a non-linear model based on RBFN [40]. We use a mapping function to find the degree of similarity between audio segments. The concept of regularization is applied to estimate the mapping function of basic RBF. 1-D Gaussian shaped model [41] that meets an important requirement of the regularization method which smoothens the mapping function for the similar inputs

consistent to similar outputs which is given by:

$$\Phi(x) = \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) \tag{1}$$

The center and width of the function are denoted by the parameters "z" and "σ", and the transformation of Gaussian is performed by Φ(x) that finding the distance and the degree of similarity between input "x" and center "z". the different RBFs are generated from an RBFN which is an exceptional ability for non-linear approximation [42] function f(x) which is given below that obtain by RBF:

$$f(\underline{x}) = \sum_{i=1}^{N} w_i \, \Phi(x, z_i) \tag{2}$$

The expanded form of mapping function in it is given as:

$$f(\underline{x}) = \sum_{i=1}^{N} w_i \exp\left(-\frac{1}{2\sigma_i^2}\sum_{j=1}^{P}(x_j - z_{ij})^2\right) \tag{3}$$

$\Phi(\underline{x}, z_i)$ represents the width and "σ_i" presents the center of the function respectively and the mapping function f(x) is defined by the sum of "N" Gaussians. To reduce the computational cost of the network we utilize the 1-D Gaussian RBF for every segment of the speech signal.

$$f(\underline{x}) = \sum_{i=1}^{P} \Phi_i(x_i, z_i) \tag{4}$$

$$f(\underline{x}) = \sum_{i=1}^{P} \exp\left(-\frac{(x_i - z_i)^2}{2\sigma_i^2}\right) \tag{5}$$

In the above equations, $\underline{x} = [x_1, ..., x_P]^T$ is a particular part of speech signal in utterance and $\underline{z} = [z_1, ..., z_P]^T$ is the center points of the RBF and the width of the particular speech segment of RBF is denoted by $\sigma_i (i = 1, \ldots, P)$. We utilize the equation. 5 to calculate the similarity among two signal segments and characterize it by an adjustable width of each RBF to making "P" basis functions $\{\Phi_1(\sigma_1), \Phi_2(\sigma_2) \ldots\ldots \Phi_P(\sigma_P)\}$. The parameters tuning, non-linear weighting and sample variance estimation of the relevance set is obtained by:

$$\sigma_i = \exp(\alpha, S_i) \tag{6}$$

$$S_i = \sqrt{\left(\frac{1}{Q-1}\sum_{j=1}^{Q}(x_{ji} - \overline{x}_i)^2\right)} \tag{7}$$

If the specific segment of the speech signal is more relevant, then the expected value of the standard deviation will be small among the speech segments. If the standard deviation value is high it means the speech segments are irrelevant, so the change in distance is more sensitive for those segments which have a small parameter "σ".

C. CNN FEATURE EXTRACTION AND RECURRENT NEURAL NETWORK

In this section, we discuss the feature extraction and RNN process in detail for sequential, audio data for recognizing the emotions of a speaker from his/her speech. CNN is the most powerful source in this era for representation and recognition of hidden information in data. In contrast, we converted the speech signals into multiple segments, each individual segment is represented by CNN features, followed by deep BiLSTM for finding the sequential information. The speech signals have many redundant information, which are computationally expensive and defect the overall model efficiency. Considering this constraint, we proposed a novel technique for selecting a most dominant sequence from utterance based on K-mean and RBF, the detail explanation is mentioned in the above sections. The selected sequence each segment is converted into spectrograms, plot the frequencies with respect to time for 2-D representation using STFT algorithm. The sequence of spectrograms [43] is fed to the pre-trained parameters of CNN, Resnet101 [44] model to extract high-level discriminative features by transfer learning strategy utilizing the last "FC-1000" layer. The features of each segment are considered as one RNN step with respect to time interval. RNNs is the most dominant source for analyzing hidden information in both spatial and temporal sequential data [45]. We process all key segments of every utterance and the final state of RNN is counted for each utterance as a final recognition of emotion. RNNs can easily learn the sequential data but forget the earlier sequence in terms of long sequences. This is a vanishing gradient problem in RNNs which is solved by LSTM [46]. It is a special type of RNNs having input, output and forget gates to learned long sequences that explain in the following equations.

$$i_t = \sigma\left((x_t + s_{t-1})\, w^i + b_i\right) \tag{8}$$

$$f_t = \sigma((x_t + s_{t-1})w^f + b_f) \tag{9}$$

$$o_t = \sigma((x_t + s_{t-1})w^o + b_o) \tag{10}$$

$$g = tanh((x_t + s_{t-1})w^g + b_g) \tag{11}$$

$$c_t = c_{t-1}.f_t + g.i_t \tag{12}$$

$$s_t = tanh\,(c_t)\,.o_t \tag{13}$$

$$final\ state = soft\ max\ (v_{st}) \tag{14}$$

x_t Represents the input at time "t" and f_t represent the forget gate in the LSTM, which needs to clear information form cell and keeps the records of the previous one. "o_t" represents the output gate which responsible for keeping info about imminent step, and "g" represents the recurrent unit having "tanh" activation function to computed from the present input segment and previous segment s_{t-1}. The memory cell "c_t" show the hidden state of RNN which is calculated in every step through the "tanh" activation function. The final state of the RNN step feeds to the Softmax classifier for taking the final decision of the RNN network. Training a huge amount of data with large and complex sequences is not correctly recognized by a simple LSTM network. Hence,

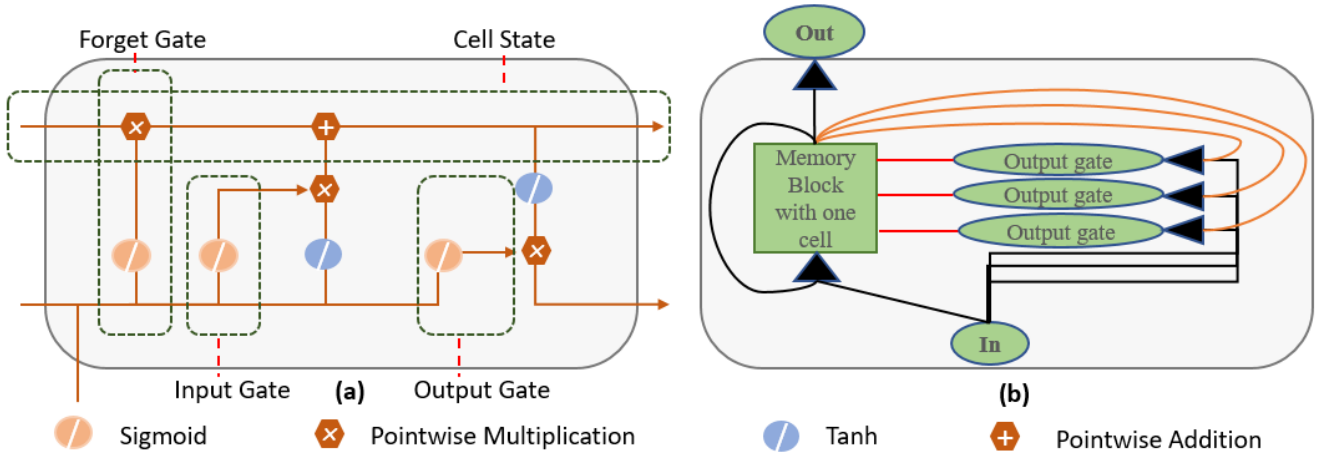

FIGURE 2. Left (a) represents the internal structure of LSTM and right (b) show the LSTM network with memory blocks (one block is shown).

in this paper, we proposed a multi-layer deep BiLSTM to learn and recognized long term sequences in audio data for recognizing emotions. The internal structure and memory blocks information is illustrated in **Figure 2**.

D. BI-DIRECTIONAL LSTM

In BiLSTM, the output at time "t" is dependent on both, previous and next segments of the sequence not only dependent in a single segment [47]. Bidirectional RNNs including two stacked of RNNs, one goes to forward, and another goes to the backward direction and calculates the joint output of both RNNs built on their hidden state. In this paper, we utilize the multi-layer concept of LSTMs network, in our method we used the two-layer network for both backward and forward pass. The overall concept of the suggested multi-layer bidirectional LSTM is shown in **Figure 3.** The external architecture is shown in the given figure which represents the training phase of the bidirectional RNN and combined both forward and backward pass hidden state in the output layer. After the output layer, the cost and validation are computed and adjust the weights and biases through back propagation. The network is validated on 20 % data, which is separated from training data and compute the error rate in the validation data using cross-entropy. Adam optimization [48] is used for minimization of cost with a 0.001 learning rate. In the deep BiLSTM network, the forward and backward pass consists of cells, which make deep our network to compute the output from the previous and next sequence with respect to time because the network performed in both directions.

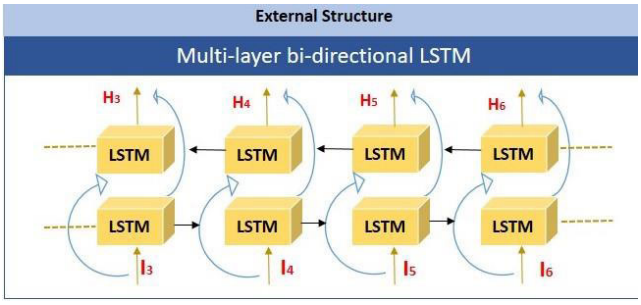

FIGURE 3. External architecture of the suggested deep bidirectional LSTM network.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we evaluated the effectiveness of the proposed system for SER and compared it with other baseline methods on publicly available benchmark speech emotions dataset. In this paper, we utilize the three public speech emotions datasets, the IEMOCAP [49] interactive emotional dyadic motion capture dataset, Emo-DB [50] berlin emotional dataset, and RAVDESS [51] Ryerson audio-visual dataset of emotional speech and songs. The detailed description of the datasets is explained in the upcoming sub sections.

A. IEMOCAP DATASET

The IEMOCAP [49] is a well-known dataset which is commonly used for recognition of emotional speeches, which has two types of dialogs, scripted and improvised. The dataset consists of 10 experienced actors to records 12 hours of audiovisual data including audio, videos, motion of faces, speech and text transcriptions. The IEMOCAP dataset has five sessions and each session consists of 2 actors (one male and one female) to record the emotional script with 3 to 15 second long with a 16 kHz sampling rate. Each session has different categories of emotions like; anger, sad, happy, neutral, surprise, disgust, frustrated, excited and fearful which is annotated by three expert persons. Individually labeled the data, we select those utterances that two experts are agreed upon them. In this paper, we evaluated our system on four emotions anger, sad, happy and neutral for comparison which is mostly used in literature. The detailed description of emotions distribution is mentioned in the given table 2.

Table 2 shows the distribution of four emotions of all five sessions of the IEMOCAP dataset for evaluating the model. We utilize the 5-fold cross-validations technique to train the

TABLE 2. The detailed description of emotions distribution in different classes and each class data participation in percentage of the IEMOCAP dataset.

Class	Total utterances	Participation in (%)
Anger	1103	19.94
Sad	1084	19.60
Happy	1636	29.58
Neutral	1708	30.88

speaker-independent model, the four sessions are used for training and one session is used for testing the system in each fold.

B. EMO-DB BERLIN EMOTION DATASET

The Berlin emotion database Emo-DB [50] contains 535 utterances recorded by ten actors: 5 male and 5 female. Each actor read the pre-selected sentences with different emotions like anger, fear, boredom, disgust, happy, neutral and sadness. In the Emo-DB approximately 2 to 3 seconds utterances having a 16 kHz sampling rate. The detailed descriptions of emotions are mention in **Table 3**.

TABLE 3. The detailed description of emotions distribution in different classes and each class data participation in percentage of the EMO-DB dataset.

Class	Total utterances	Participation in (%)
Anger	127	23.74
Sadness	62	11.59
Happy	71	13.27
Neutral	79	14.77
Disgust	46	8.60
Fear	69	12.90
Boredom	81	15.14

Table 3 represents the description of all emotions of the Emo-DB dataset which is a small dataset having limited emotions. We utilize the 5-fold cross-validation technique for training the speaker-independent model to recognize the emotions in daily conversations. We used the sentences of 8 speakers for training the system and the other 2 speakers are used for testing the system.

C. RAVDESS DATASET

The RVDESS (Ryerson audiovisual database of emotional speech and songs) [51] is an acted dataset, which is recorded in English language, which broadly utilize for expressive music and dialog reactions. The dataset contains (8) emotions having 24 professional actors, 12 in each category, male and female. The emotions like sad, calm, happy, angry, surprise, neural, fearful, and disgust recorded by different male and female. The total 1440 audio files are recorded with 48000 Hz sampling rate. We performed experiments using 5-folds cross-validation technique to split the dataset for training and testing parts. The explanation is remark in **Table 4.**

TABLE 4. The detailed description of emotions distribution in different classes and each class data participation in percentage of the RAVDESS dataset.

Class	Total utterances	Participation in (%)
Anger	192	13.33
Calm	192	13.33
Sadness	192	13.33
Happy	192	13.33
Neutral	96	6.667
Disgust	192	13.33
Fear	192	13.33
Surprise	192	13.33

D. EXPERIMENTAL EVALUATION

In this section, we evaluate our system for speaker-dependent and independent emotions recognition. We separated each utterance into multiple segments "fs" with respect to time "t" with 25% overlapping to select the sequence ($s = fs_1, fs_2, fs_3, \ldots\ldots, fs_n$) from each utterance. The RBF based similarity method was used in K-mean clustering to select one segment as a key-segment from each cluster, which is near to centroid of the cluster that represents the whole cluster. The detail description is mentioned in Section III. After selecting key-segments, we extracted the high-level discriminative features utilizing the "FC-1000" layer of the Resnet101 model and normalize the extracted features with global mean and standard deviation for boosting the accuracy of the overall model. The normalized features feed to deep BiLSTM network step by step to learn the hidden patterns and recognize the emotion in the given sequence. The final state of the proposed deep BiLSTM network was followed by the Softmax classifier to produce the probability for emotions. The recommended system was implemented in MATLAB 2019b utilizing the neural network toolbox for features extraction, model training, and evaluations. The data are divided into training and testing folds with an 80:20 % ratio and generated spectrograms of every segment. The suggested model was trained and evaluated on a single NVIDIA GeForce GTX 1070, 8 GB on-board memory GPU system. The detailed description of speaker-dependent and independent experiments is in upcoming sub sections.

E. MODEL OPTIMIZATION

In the training stage, we tuned the model with different parameters to make it sufficient and optimal for SER. We performed different experiments with multiple batch sizes, learning rate, number of LSTM and BiLSTM layer to choose the optimal solution. We selected the Adam optimizer for model optimization and the best bias correction for better effect. We also did experiments with normalized features and un-normalized features to check the model efficiency. We selected the batch size, 512 and learning rate, 0.001 for this model which is empirically proved from extensive experiments over three different speech emotional datasets. We performed two types of experiments, with normalized features

and without normalized features and obtained the results of both to select the features for model training. The detail description of diverse parameters and the corresponding result of the proposed model is shown for normalized and un-normalized features in the below tables for every dataset. Each table represents individual dataset result with different batch size and learning values. We select the best learning rate and batch size for whole model before these extensive experiments for all datasets

In the **Tables, 5-7** represents the results of the proposed model using normalized and un-normalized features. The features normalization improves the overall recognition accuracy for IEMOCAP is (0.4%), for EMO-DB is (0.23%) and for RAVDESS is (0.19%) respectively from un-normalized features. Hence, the normalized features recognition accuracy is better and the processing time for model testing and training is lower than other baseline models.

TABLE 5. Model performance with different parameters, learning rate, batch size, and normalized value for emotion recognitions of speaker-dependent (SD) and speaker-independent (SI) on IEMPOCAP dataset.

Model	Batch-Size	Learning rate	SD (%)	SI (%)
Proposed Model + Un-normalized features	256	0.01	73.73	68.31
		0.001	73.96	68.32
		0.0001	73.65	68.11
	512	0.01	76.63	70.21
		0.001	**76.92**	**70.31**
		0.0001	76.55	70.01
	1024	0.01	75.53	69.11
		0.001	75.76	69.22
		0.0001	75.45	69.06
Proposed Model + Normalized features	256	0.01	78.78	71.22
		0.001	78.96	71.36
		0.0001	78.55	71.12
	512	0.01	80.43	71.08
		0.001	**81.01**	**72.25**
		0.0001	80.25	72.03
	1024	0.01	79.63	71.41
		0.001	79.66	71.52
		0.0001	79.05	71.31

TABLE 6. Model performance with different parameters, learning rate, batch size, and normalized value for emotion recognitions of speaker-dependent (SD) and speaker-independent (SI) on EMO-DB dataset.

Model	Batch-Size	Learning rate	SD (%)	SI (%)
Proposed Model + Un-normalized features	256	0.01	86.71	83.33
		0.001	87.92	84.48
		0.0001	86.55	84.23
	512	0.01	89.63	84.52
		0.001	**90.72**	**85.01**
		0.0001	89.55	84.01
	1024	0.01	87.53	84.01
		0.001	88.66	84.08
		0.0001	88.45	84.06
Proposed Model + Normalized features	256	0.01	89.77	84.22
		0.001	89.95	85.06
		0.0001	89.54	85.12
	512	0.01	90.33	85.08
		0.001	**91.14**	**85.57**
		0.0001	90.25	85.01
	1024	0.01	88.73	83.71
		0.001	88.76	84.42
		0.0001	87.45	82.81

TABLE 7. Model performance with different parameters, learning rate, batch size, and normalized value for emotion recognitions of speaker-dependent (SD) and speaker-independent (SI) on RAVDESS dataset.

Model	Batch-Size	Learning rate	SD (%)	SI (%)
Proposed Model + Un-normalized features	256	0.01	79.83	74.81
		0.001	80.02	74.92
		0.0001	79.65	74.61
	512	0.01	80.63	75.51
		0.001	**81.42**	**76.21**
		0.0001	81.45	75.45
	1024	0.01	80.73	75.31
		0.001	80.96	75.72
		0.0001	79.45	75.16
Proposed Model + Normalized features	256	0.01	81.38	76.22
		0.001	81.76	76.36
		0.0001	81.25	76.12
	512	0.01	81.83	76.68
		0.001	**82.01**	**77.02**
		0.0001	81.65	76.53
	1024	0.01	80.83	75.48
		0.001	81.86	76.35
		0.0001	80.45	74.91

Similarly, we compare our model processing time with other baseline methods using the diverse parameter for proving the model effectiveness and feasibility. We set the batch size to be 512 and select the 0.001 learning rate with Adam optimizer and analyze the processing time for IEMOCAP, EMO-DB and RAVDESS dataset utilizing the normalized features. The details are mention in **Table 8.**

TABLE 8. The processing time evaluation and comparison of the proposed model with other baseline models using SER corpuses.

Models	IEMOCAP	RAVDESS	EMO-DB
ACRNN [52]	13487 sec	-	6811 sec
ADRNN [53]	13887 sec	-	7187 sec
Prop model	**10452 sec**	**6250 sec**	**5396 sec**

Table 8 illustrated the processing time of the model which indicates that the proposed model takes less time in training and testing due to the efficient strategy of the model. In the proposed model we didn't take all segments of each utterance, but we just select one segment form each cluster as a key segment that represent the whole cluster and train a model on that selected cluster. So, that's the reason for less processing time, our model processes the selected segment not all segments of utterance and extract the CNN feature which feeds to deep BiLSTM network for classification.

F. SPEAKER INDEPENDENT PERFORMANCE OF THE PROPOSED MODEL

We performed experiments on spontaneous emotional data of the IEMOCAP, EMODB dataset and also evaluated the effectiveness of the model on RAVDESS corpus. The IEMOCAP

and EMODB corpus have 10 speakers and the RAVDESS dataset has 24 speakers. We follow 5th-fold cross validation technique to split the data with an 80:20 % ratio according to the number of speakers, the 80% data are used for model training and the remaining data are used for test the model. We evaluated the proposed system over these datasets and check the prediction performance on testing data. The overall model performance are presented in term of class level precision, recall, and F1 score for each emotion. Similarly, we find out the weighted accuracy, the ratio between correctly classified emotion and total emotion in consistent class. The un-weighted accuracy, mean the ratio with in correct predicted emotion and total emotion in whole dataset. The detail description and quantitative or numerical results of each dataset is given in **Tables 9-11**.

TABLE 9. The performance of the proposed model for speaker independent emotion recognition using IEMOCAP dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.85	0.83	0.84
Happiness	0.53	0.58	0.55
Neutral	0.88	0.70	0.78
Sadness	0.66	0.78	0.72
Weighted	0.74	0.74	0.74
Un-weighted	0.73	0.72	0.72

TABLE 10. The performance of the proposed model for speaker independent emotion recognition using EMO-DB dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.82	0.91	0.86
Boredom	0.99	0.90	0.84
Disgust	0.93	0.87	0.90
Fear	0.88	0.92	0.90
Happiness	0.88	0.66	0.75
Neutral	0.64	0.85	0.73
Sadness	0.77	0.88	0.82
Weighted	0.86	0.86	0.86
Un-weighted	0.86	0.84	0.85

TABLE 11. The performance of the proposed model for speaker independent emotion recognition using RAVDESS dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.81	0.95	0.87
Clam	0.57	0.95	0.71
Disgust	0.99	0.86	0.92
Fearful	0.88	0.91	0.90
Happiness	0.98	0.43	0.60
Neutral	0.99	0.50	0.67
Sadness	0.91	0.61	0.73
Surprised	0.76	0.95	0.85
Weighted	0.80	0.86	0.81
Un-weighted	0.86	0.77	0.77

Measuring the proposed system by weighted and un-weighted accuracy and show the precision and recall values of each category in confusion matrix. The confusion matrix show the actual predicted emotions and model confusion result of each class. The classification performance of

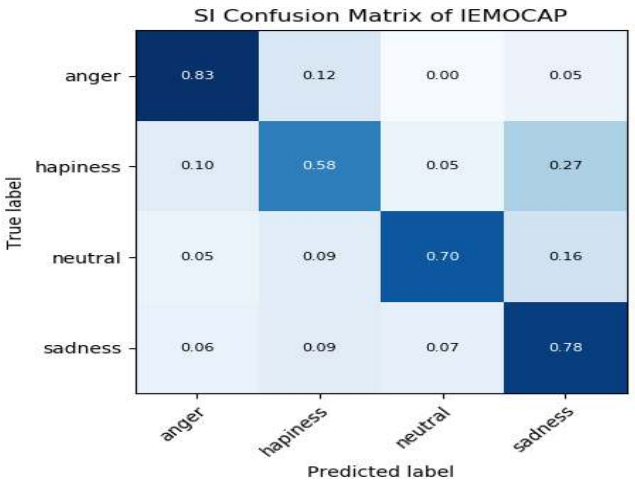

FIGURE 4. Confusion matrix of speaker independent emotions prediction on IEMOCAP corpus with (72.25%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

the suggested system for speaker-independent evaluation was conducted in the given **Figure 4.** Which shows the recognition performance of the proposed model on the IEMOCAP challenging dataset for speaker-independent SER. In this experiment, we obtained 83% accuracy for anger emotion and 78% for sad, 70% for neutral and 58% for happy emotion respectively. The recognition rate of happy and neutral emotions is low in this experiment, but we obtained better results from state of the art. The results of the EMO-DB dataset are shown in **Figure 5**.

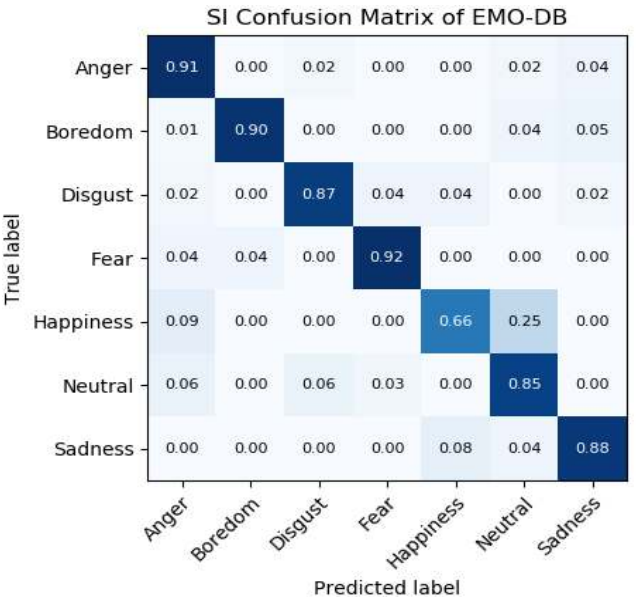

FIGURE 5. Confusion matrix of speaker independent emotions prediction on EMO-DB corpus with (85.57%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

In the above figure, the overall emotion recognition performance is increased as compared with other baseline methods,

but the recognition rate of happy emotion is increased but still lower. Hence, the happy emotion mostly confused with other emotions in classification. The anger, fear, and boredom have high, greater than 90% accuracy and disgust, neutral and sad have greater than 80% accuracy respectively. Our proposed system overall achieved high recognition (85.75%) score for the EMO-DB dataset. The RAVDESS dataset confusion matrix is shown in **Figure 6**. We evaluated the effectiveness of our proposed system on the RAVDESS dataset, which is mostly used for emotional songs and speech.

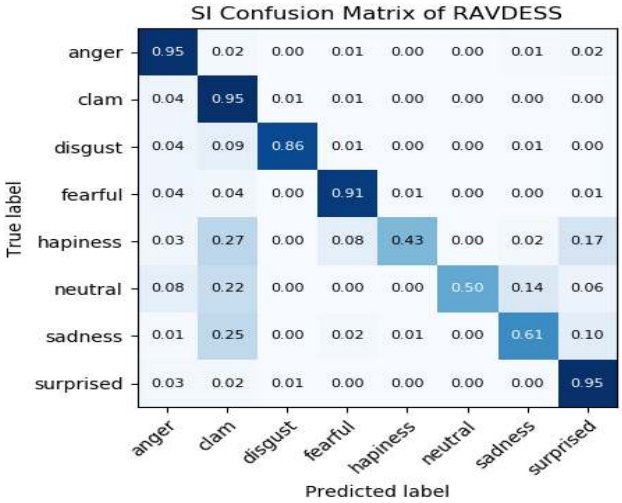

FIGURE 6. Confusion matrix of speaker independent emotions prediction on RAVDESS corpus with (77.02%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

The performance of the suggested model is better than other baseline techniques. The system recognized anger, clam fear, and surprise with high priority and happy, neutral, and sad emotions were recognized with lower priority. The system mostly confused in happy, neutral, and sad emotions and recognized these emotions as a calm due to the minimum diversity with each other. The recognition rate of calm is high and the system confused with other emotions and recognize it as a calm. The overall accuracy of the system for speaker-independent emotion recognition is better than other baseline methods on IEMOCAP, EMO-DB, and RAVDESS corpuses.

G. SPEAKER DEPENDENT PERFORMANCE OF THE PROPOSED MODEL

In this type of experiment, we don't split the dataset individually like speaker independent. In the speaker-dependent system, we combine all speeches (dataset) in a single file and make a whole set and trained them respectively. We divide the whole set into an 80:20 % split ratio for model training and testing. We shuffle the data and randomly select 80% data for model training and 20% data is used for validation and testing. Similarly, we used the most normalized features for model training to reduce the overfitting and achieve the goal, to get the most reliable result of SER. Furthermore, we investigated the speaker dependent model for all datasets and also mention the qualitative result and statistic in term of precision, recall, F1 score, weighted, and un-weighted accuracy. The detail numerical results of the each dataset is given in **Table 12-14**.

TABLE 12. The performance of the proposed model for speaker dependent emotion recognition using IEMOCAP dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.91	0.92	0.92
Happiness	0.68	0.64	0.66
Neutral	0.98	0.79	0.88
Sadness	0.73	0.89	0.81
Weighted	0.83	0.84	0.83
Un-weighted	0.83	0.81	0.81

TABLE 13. The performance of the proposed model for speaker dependent emotion recognition using EMO-DB dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.90	0.96	0.93
Boredom	0.99	0.91	0.95
Disgust	0.99	0.89	0.94
Fear	0.88	0.96	0.92
Happiness	0.94	0.75	0.83
Neutral	0.75	0.94	0.84
Sadness	0.86	0.97	0.91
Weighted	0.92	0.93	0.92
Un-weighted	0.92	0.91	0.91

TABLE 14. The performance of the proposed model for speaker dependent emotion recognition using RAVDESS dataset.

Emotion	Precision	Recall	F1 Score
Anger	0.83	0.99	0.90
Clam	0.63	0.95	0.76
Disgust	0.99	0.90	0.94
Fearful	0.90	0.93	0.91
Happiness	0.99	0.50	0.66
Neutral	0.99	0.67	0.80
Sadness	0.94	0.66	0.77
Surprised	0.89	0.96	0.92
Weighted	0.85	0.89	0.85
Un-weighted	0.90	0.82	0.82

We selected the best model which give best results in SER with a high preference for generalization. The classification result of speaker-dependent model in term of confusion matrix is illustrated in **Figure 7, 8**, and **9**.

Figure 7 presents the class level accuracy of the proposed model in a confusion matrix which indicated the original emotion label and predicted emotion label. In this experiment, the model highly recognized the anger and sad emotion with 92% and 89% respectively. The happy emotion recognition rate was relatively low from other emotions but better than the speaker-independent model. The happy and neutral emotions were mostly confused with sadness in both speaker-dependent and independent experiments. The speaker-dependent confusion matrix of EMO-DB dataset is shown in **Figure 8.**

The speaker dependent experiments of the proposed model showed outperform results on the EMO-DB dataset and

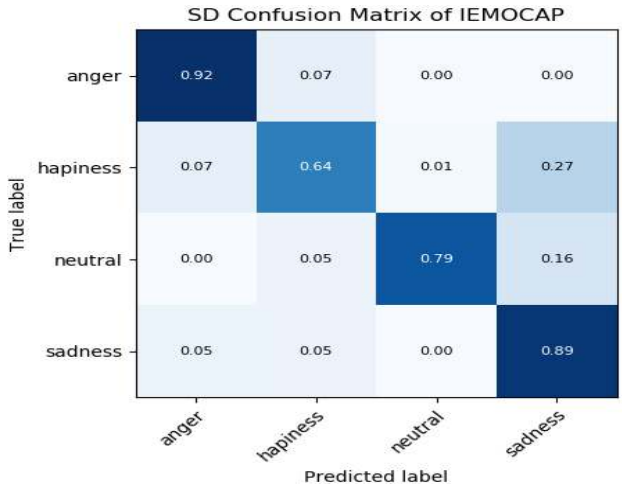

FIGURE 7. Confusion matrix of speaker dependent emotions prediction on IEMOCAP corpus with (81.02%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

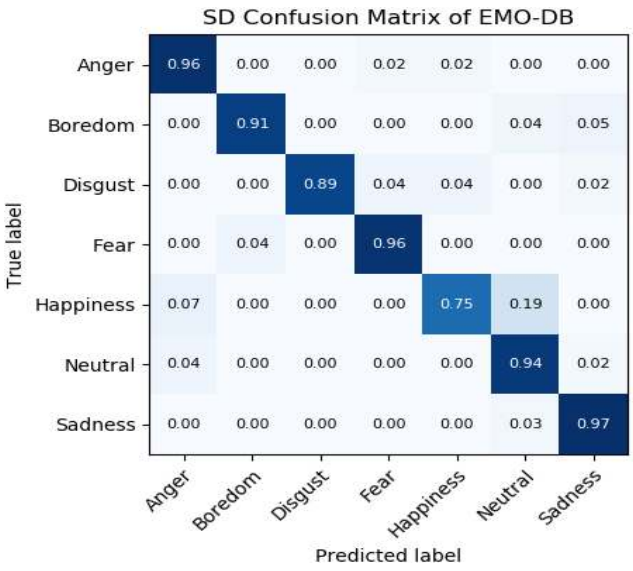

FIGURE 8. Confusion matrix of speaker dependent emotions prediction on EMO-DB corpus with (91.14%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

recognized the emotions with 91.14 % average recall. In this experiment the system recognized anger, fear and sadness emotion with high rank and disgust, neutral, boredom had more than 85% recognition rate and the happy emotion is recognized with a 75% ratio respectively. The system was confused among happy and neutral emotion and mostly happy emotions were recognized as neutral similarly, like a speaker independent. The overall performance of the proposed system is better, affective and significant than other baseline techniques. The speaker-dependent performance of the suggested system for RAVDESS is illustrated in **Figure 9**.

We evaluated our model on the REVDESS dataset to show the performance and generalization of the model for SER.

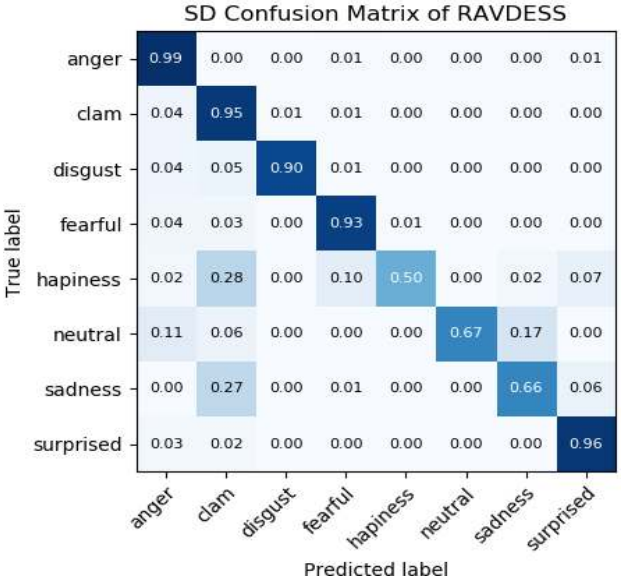

FIGURE 9. Confusion matrix of speaker dependent emotions prediction on RAVDESS corpus with (82.02%) un-weighted accuracy and confusion among actual and predicted emotions are showed in the corresponding row.

We obtained the record results of the model on multiple benchmark datasets which outperform output respectively. The emotion recognition rate of the proposed model was 95% for anger, 93% for fear, 96% for surprised, 95% for calm and 90 % for disgust respectively. The happy emotion rate was relatively low but better than previous work. The proposed system misrecognized the happy emotion as compared to other classes. According to our opinion, the features of happy emotion are easily confused with others and as a result the suggested model misrecognized them. Another reason for misrecognized the happy emotion is the limitation of data, the size of the datasets is less than other pattern recognition datasets like images, video, and text. Hence, in SER, to increase the accuracy of happy emotion is a very significant improvement in this field. Many researchers have worked to develop new techniques to extract discriminative features and efficient way of classification to enhance the accuracy of this field, SER.

V. DISCUSSION

In the proposed framework, the efficient sequence selection using K-mean clustering based on RBF similarity and normalized discriminative features with deep BiLSTM are major contributions for SER utilizing speech signals. We performed, speaker-dependent and speaker independent experiments over three benchmarked datasets for recognizing the emotional state of his/her speech. We developed a new technique for SER, to select a sequence from utterance using RBF based K-mean clustering technique. We selected one segment from each cluster which is near to centroid as a key segment that represents the corresponding clusters and converted all key segments into spectrograms applying

STFT for 2-D representations. Furthermore, we extracted the high level discriminative (CNN) features from spectrograms utilizing the "FC-1000" layer of the Resnet101 CNN model. We normalized the extracted features using average mean and standard deviation algorithm and passed to deep BiLSTM for classification. We used this novel approach for SER to improve the classification accuracy and reduce the processing time as compared to other traditional CNN_ELM [54] and DNN_KLM networks [33]. We obtained better results on three benchmarks, IEMOCAP, EMO-DB and RAVDESS datasets using this novel approach of SER for speaker-dependent and speaker-independent experiments. The comparison of the proposed approach with the baseline methods are shown in the below Table. **Table 15-17** represents the comparative analysis of the proposed system with other baseline SER methods on IEMOCAP, EMO-DB and RAVDESS datasets respectively. The proposed system boosts the overall accuracy up to (6.14%), (2.14%) and (7.01%) in speaker-dependent and (3.07%), (1.57%) and (2.41%) in speaker-independent experiments on IEMOCAP, EMO-DB, and RAVDESS datasets to recognize the emotional state of speaker, respectively.

TABLE 15. Speaker-dependent and speaker-independent comparison of the proposed technique with baseline methods based on spectrograms using IEMOCAP dataset.

Baseline method	Speaker-dep (UA %)	Speaker-indep (UA %)
L Guo et al. [33]	-	57.01
W.Zheng at al. [55]	-	40.02
K. Han et al. [56]	-	51.24
H Meng et al. [53]	74.96	69.32
Z Zhao et al. [57]	-	66.50
D Luo et al. [58]	-	63.98
Tripathi et al. [59]	-	61.60
M Chen et al. [52]	-	64.74
Proposed model	**81.01**	**72.25**

TABLE 16. Speaker-dependent and speaker-independent comparison of the proposed technique with baseline methods based on spectrograms using EMO-DB dataset.

Baseline method	Speaker-dep (UA %)	Speaker-indep UA %)
L Guo et al. [33]	87.85	84.49
H Meng et al. [53]	90.37	84.99
M Chen et al. [52]	-	82.82
Badshah et al[12]	89.46	80.79
P jaing et al. [60]	86.44	84.53
Proposed model	**91.14**	**85.57**

TABLE 17. Speaker-dependent and speaker-independent comparison of the proposed technique with baseline methods based on spectrograms using RAVDESS dataset.

Baseline method	Speaker-dep (UA %)	Speaker-indep (UA %)
Y Zeng et al.[61]	-	64.48
M. Jalal et al. [62]	-	69.4
A Bhavan et al. [63]	-	75.69
A Zamil et al. [64]	-	67.14
Proposed model	**82.01**	**77.02**

We reduce the processing time of the suggested model due to process one segment from each cluster and the usage of normalized features. In the state-of-the-art methods, researchers have used traditional and un-normalized features process for classification. The proposed system evaluated on three standard datasets which outperformed and demonstrated significant results that proved the robustness and effectiveness of the system. The performance of the proposed system has evaluated over different pre-trained CNN models as a features extractor. The comparative analysis of multiple CNN models is given in **Figure 10** and **11.**

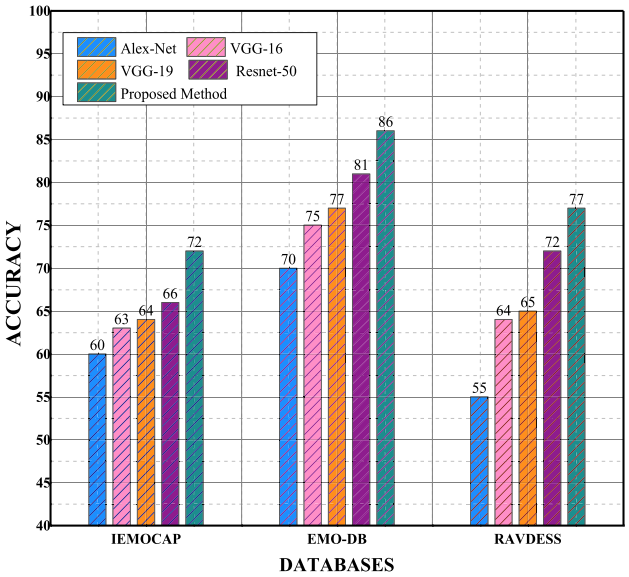

FIGURE 10. Speaker-independent comparative analysis of the proposed model with other pre-trained CNNs models on IEMOCAP, EMO-DB, and RAVDESS datasets.

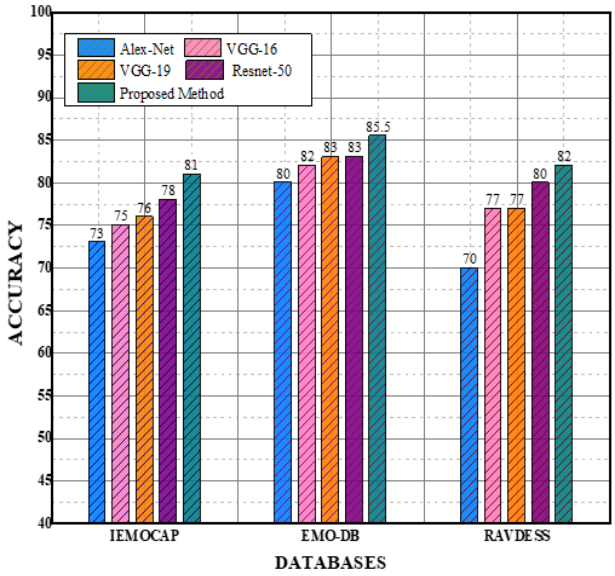

FIGURE 11. Speaker dependent dependent comparative analysis of the proposed model with other pre-trained CNNs models on IEMOCAP, EMO-DB, and RAVDESS datasets.

We utilized different pre-trained CNN models as features extractor in the proposed technique and evaluated over three benchmarks: IEMOCAP, EMO-DB, and RAVDESS datasets for speaker-dependent and speaker-independent experiments. The recognition results of the proposed model are illustrated in **Figures 10** and **11** with recognition accuracy over suggested speech emotion datasets. The recognition accuracy of the proposed system outperforms other CNN models that clearly indicated the robustness and significance of the model for SER using spectrograms of speech signals [3].

VI. CONCLUSION AND FUTURE WORK

The existing CNNs system of SER has too many challenges such as improvement in accuracy and reduce the computational complexity of the whole model. Due to these limitations, we planned a novel approach for SER to improve the recognition accuracy and reduce the overall model cost computation and processing time. In contrast, we suggested a new technique to select a more efficient sequence from speech using RBF based K-mean clustering algorithm and convert it into spectrograms by applying STFT algorithm. Hence, we extracted the discriminative and salient features from spectrograms of speech signal by utilizing the ''FC-1000'' layers of the CNN model, called Resnet and normalize it by applying mean and standard deviation to remove the variation. After normalization, we feed these discriminative features to deep BiLSTM to learn the hidden information and recognize the final state of sequence and classify the emotional state of speakers. We evaluated the proposed system on three standard IEMOCAP, EMO-DB, and RAVDESS datasets to check the robustness of the system. We improve the recognition accuracy for IEMOCAP dataset as 72.25%, obtain 85.57% for EMO-DB dataset and for RAVDESS dataset, we achieved 77.02%. We reduce the processing time of our system, which process the selected segments for emotion recognition rather than all segments that yielding a computational friendly system. The experimental results of the proposed system proved the robustness and significance for SER to correctly recognize the emotional state of the speaker using spectrograms of speech signals.

The proposed architecture can be further used in future for other applications and can explore speech emotion recognition using DBN, GRU and spike networks to get better accuracy with less computational complexity. The proposed model can be an aspiration for speaker recognition and speaker identification that is used in many real-world problems.

REFERENCES

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, ''EERA-ASR: An energy-efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing,'' *IEEE Access*, vol. 6, pp. 52227–52237, 2018.

[2] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, ''An image-based deep spectrum feature representation for the recognition of emotional speech,'' in *Proc. 25th ACM Multimedia Conf. (MM)*, 2017, pp. 478–484.

[3] Mustaqeem and S. Kwon, ''A CNN-assisted enhanced audio signal processing for speech emotion recognition,'' *Sensors*, vol. 20, no. 1, p. 183, 2020.

[4] J. Huang, B. Chen, B. Yao, and W. He, ''ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network,'' *IEEE Access*, vol. 7, pp. 92871–92880, 2019.

[5] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[6] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, ''Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 77–86, Jan. 2020.

[7] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, ''Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies,'' *Appl. Sci.*, vol. 9, no. 22, p. 4963, 2019.

[8] F. Karim, S. Majumdar, and H. Darabi, ''Insights into LSTM fully convolutional networks for time series classification,'' *IEEE Access*, vol. 7, pp. 67718–67725, 2019.

[9] A. Zhang, W. Zhu, and J. Li, ''Spiking echo state convolutional neural network for robust time series classification,'' *IEEE Access*, vol. 7, pp. 4927–4935, 2019.

[10] P. Tzirakis, J. Zhang, and B. W. Schuller, ''End-to-End speech emotion recognition using deep neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5089–5093.

[11] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, ''Speech emotion recognition using deep learning techniques: A review,'' *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[12] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, ''Deep features-based speech emotion recognition for smart affective services,'' *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] S. Jiang, ''Memento: An emotion-driven lifelogging system with wearables,'' *ACM Trans. Sensor Netw.*, vol. 15, no. 1, p. 8, 2019.

[15] H. Wang, Q. Zhang, J. Wu, S. Pan, and Y. Chen, ''Time series feature learning with labeled and unlabeled data,'' *Pattern Recognit.*, vol. 89, pp. 55–66, May 2019.

[16] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, ''Sound classification using convolutional neural network and tensor deep stacking network,'' *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[17] M. Navyasri, R. RajeswarRao, A. DaveeduRaju, and M. Ramakrishnamurthy, ''Robust features for emotion recognition from speech by using Gaussian mixture model classification,'' in *Proc. Int. Conf. Inf. Commun. Technol. Intell. Syst.* Cham, Switzerland: Springer, 2017, pp. 437–444.

[18] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. Schuller, ''Attention-based convolutional neural networks for acoustic scene classification,'' in *Proc. DCASE*, 2018, pp. 1–5.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''Imagenet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, ''Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition,'' *IEEE Access*, vol. 7, pp. 93847–93857, 2019.

[21] M. Zeng and N. Xiao, ''Effective combination of DenseNet and BiLSTM for keyword spotting,'' *IEEE Access*, vol. 7, pp. 10767–10775, 2019.

[22] Y. Xie, R. Liang, H. Tao, Y. Zhu, and L. Zhao, ''Convolutional bidirectional long short-term memory for deception detection with acoustic features,'' *IEEE Access*, vol. 6, pp. 76527–76534, 2018.

[23] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, ''Convolutional, long short-term memory, fully connected deep neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.

[24] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, ''End-to-End multimodal emotion recognition using deep neural networks,'' *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[25] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, ''Depaudionet: An efficient deep model for audio based depression classification,'' in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 35–42.

[26] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech*, Sep. 2018, pp. 3683–3687.
[27] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[28] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, Jan. 2018.
[29] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.
[30] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[31] P. Liu, K.-K.-R. Choo, L. Wang, and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification," *Soft Comput.*, vol. 21, no. 23, pp. 7053–7065, Dec. 2017.
[32] L. Wu, S. Zhang, M. Jian, Z. Lu, and D. Wang, "Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks," *IEEE Access*, vol. 7, pp. 77268–77276, 2019.
[33] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.
[34] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3682–3686.
[35] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.
[36] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering approach based on mini batch kmeans for intrusion detection system over big data," *IEEE Access*, vol. 6, pp. 11897–11906, 2018.
[37] Z. Yu, W. Chen, X. Guo, X. Chen, and C. Sun, "Analog network-coded modulation with maximum Euclidean distance: Mapping criterion and constellation design," *IEEE Access*, vol. 5, pp. 18271–18286, 2017.
[38] S. S. Chouhan, A. Kaul, U. P. Singh, and S. Jain, "Bacterial foraging optimization based radial basis function neural network (BRBFNN) for identification and classification of plant leaf diseases: An automatic approach towards plant pathology," *IEEE Access*, vol. 6, pp. 8852–8863, 2018.
[39] A. M. Sheri, M. A. Rafique, M. T. Hassan, K. N. Junejo, and M. Jeon, "Boosting discrimination information based document clustering using consensus and classification," *IEEE Access*, vol. 7, pp. 78954–78962, 2019.
[40] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowl.-Based Syst.*, vol. 117, pp. 56–69, Feb. 2017.
[41] P. K. Mishra, S. K. Nath, M. K. Sen, and G. E. Fasshauer, "Hybrid Gaussian-cubic radial basis functions for scattered data interpolation," *Comput. Geosci.*, vol. 22, no. 5, pp. 1203–1218, Oct. 2018.
[42] O. Fresnedo, P. Suarez-Casal, and L. Castedo, "Transmission of analog information over the multiple access relay channel using zero-delay nonlinear mappings," *IEEE Access*, vol. 7, pp. 48405–48416, 2019.
[43] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 360–371, Jan. 2006.
[44] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.
[45] K.-I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural Netw.*, vol. 6, no. 6, pp. 801–806, Jan. 1993.
[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[47] A. Ogawa and T. Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Commun.*, vol. 89, pp. 70–83, May 2017.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[49] 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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
[50] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.
[51] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American english," *PLoS ONE*, vol. 13, no. 5, 2018, Art. no. e0196391.
[52] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[53] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[54] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5150–5154.
[55] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 827–831.
[56] K. Han and D. I. Yu Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.
[57] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.
[58] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1–5.
[59] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," 2019, *arXiv:1906.05681*. [Online]. Available: http://arxiv.org/abs/1906.05681
[60] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.
[61] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019.
[62] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 1–5.
[63] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.
[64] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST)*, Jan. 2019, pp. 281–285.

MUSTAQEEM received the B.S. degree in computer science from the Institute of Business and Management Sciences (IBMS), Peshawar, Pakistan, and the M.S. degree from the Islamia College, Peshawar, Pakistan, with research in video analysis (content-based video retrieval). He is currently pursuing the Ph.D. degree in digital content from the College of Electronics and Information Engineering, Sejong University, Seoul, South Korea. He is also a Researcher with the Interaction Technology Laboratory (IT Lab), Sejong University. His research interests include digital signals and speech processing, emotion recognition image, and video processing.

MUHAMMAD SAJJAD received the master's degree from the Department of Computer Science, College of Signals, National University of Sciences and Technology, Rawalpindi, Pakistan, and the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea. He is currently an Assistant Professor with the Department of Computer Science, Islamia College Peshawar, Pakistan. He is also the Head of the Digital Image Processing Laboratory, Islamia College Peshawar, where students are involved in research projects under his supervision, such as social data analysis, medical image analysis, multimodal data mining and summarization, image/video prioritization and ranking, fog computing, the Internet of Things, virtual reality, and image/video retrieval. His primary research interests include computer vision, image understanding, pattern recognition, and robot vision and multimedia applications, with current emphasis on raspberry-pi and deep learning-based bioinformatics, video scene understanding, activity analysis, fog computing, the Internet of Things, and real-time tracking.

SOONIL KWON received the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, USA, in 2000 and 2005, respectively. He is currently a Professor at the Department of Software, College of Software Convergence, Sejong University, Seoul, South Korea, where he is also the Head of the Interaction Technology Laboratory, where students are involved in research projects under his supervision, such as social data analysis, audio analysis, multimodal data analysis and speech emotion recognition, speech synthesis, speaker recognition, and speaker diarization. His research interests include speech recognition, human–computer interaction, affective computing, and speech and audio processing. He served as a professional reviewer for several well-reputed journals, such as the *IEEE Communication Magazine*, *Sensors*, *Information Fusion*, *Information Sciences*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, MBEC, MTAP, SIVP, and JVCI.