**Proceedings of 10th IOE Graduate Conference**
***Peer Reviewed***
**ISSN: 2350-8914 (Online), 2350-8906 (Print)**
**Year: 2021 Month: October Volume: 10**

# Speech Emotion Recognition using Convolutional Recurrent Neural Network

Om Prakash Yadav [a], Laxmi Prasad Bastola [b], Jagdish Sharma [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pashchimanchal Campus, IOE, Tribhuvan University, Nepal*
**Corresponding Email**: [a] prakash12.opy@gmail.com, [b] laxmiprasad.bastola@wrc.edu.np, [c] jagdishpkr@gmail.com

**Abstract**
Deep learning for Speech Emotion Recognition (SER) has advantages over traditional approach using machine learning algorithm, having capability to detect emotional feature within speech signal. Convolutional Recurrent Neural Network (CRNN), consisting Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (BiLSTM) is proposed to learn emotional feature from log-mel scaled spectrograms of speech utterances. Local features are learned by convolution kernels of CNN and a layer of BiLSTM is selected to learn temporal dependencies from learned local features. Speech utterances are pre-processed to remove background noise and non-informative portions. Also data augmentation techniques are investigated and selected the best techniques to enrich number of data samples improving the recognition rate of the model. The proposed model is tested with two widely used datasets i.e. RAVDESS in North American English and Berlin EMO-DB in German language. It is observed that constructed model performance improve to 85.76% for RAVDESS and 91.59% for Berlin EMO-DB and the model is language independent.

**Keywords**
Bidirectional Long Short Term Memory, Convolutional Recurrent Neural Network, Log-Mel Spectrogram, Speech Emotion Recognition

## 1. Introduction

Speech Emotion Recognition or SER in short is generally a task to detect emotional states of speaker from voiced speech by extracting selected features for final classification. Emotions are reflection of intense mental activity and can be observed by many human activities such as facial expression, body gesture and voiced speech. Speech being fastest and generally used pathway of human communication, received large concentration of researchers in SER. Speech generally contains verbal information along with non-verbal cues or paralinguistic information [1]. Verbal information make listener to understand the meaning of speech and non-verbal cues contains the expressive messages such as speaker's emotion.

The factor that motivates for SER is recognition of emotion from voiced speech is an essential part of Human Computer Interaction (HCI). These automated system facilitate direct interaction between machine and human through spoken speech [2] and enables computer to understand emotional states expressed by human subjects. SER finds application in psychological studies for knowing human behaviour in the field of medicine, education, security, entertainment, etc.

There are several problems for SER as emotions are highly dependent on speaker's language, speaking style, cultural background, context and environmental condition. So, distinguishing paralinguistic features must be extracted to represent these characteristics of speakers. Many researchers studied different speech features, but no one can identify which one is the best one for emotion representation [3].

In order to extract features representing emotional content of speech, deep learning is found to be effective compared to traditional machine learning algorithm [4] as they have capability to detect complex pattern of data. The deep learning models can extract emotional features directly from time domain speech data or from handcrafted features such as spectrograms as shown in figure 1. The recognition rate is high using handcrafted features compared to raw speech data [5]. The method of extracting a set of handcrafted features, i.e. statistics of short-term

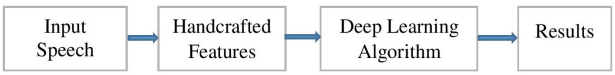


**Figure 1:** Deep learning approach for SER

feature vectors are computed in the temporal, spectral, or cepstral domains. Handcrafted feature computed in spectral or cepstral domain such as chromagram, MFCC, mel-sclaled spectrogram, etc. are mostly selected to train deep learning model [6]. While selecting 2D spectrogram either linear scale [7] or mel-scale[5], problem is modeled as image recognition task with DNN, making SER easier.

Deep learning algorithms require large amount of data for training the model. The available dataset used by authors are mostly acted voice utterances from different professional actors. The speech utterences available in datasets contains silences and background noise. These noise need to be removed using appropriate pre-processing steps as they impact extraction of emotion feature [8]. Also, there are only few number of data samples in datasets and size is increased by augmentation methods to enrich the number. Convolutional Recurrent Neural Network (CRNN) is proposed for extracting deep emotional features from perceptually relevant two-dimensional (2D) time-frequency representation of speech signal i.e. log-mel spectrograms for final classification.

## 2. Related Work

In literature, various machine learning and deep learning approaches are investigated with particular features and dataset. The final recognition rate is highly influenced by the choice of dataset and features selected as model input. Issa et. al. in [6] selected five different spectral features i.e. MFCC, mel-spectrogram, Chromagram, Tonnetz and Spectral contrast to train one dimensional CNN. They selected three widely used dataset i.e. RAVDESS, Berlin EMO-DB and IEMOCAP. They also used four different techniques of data augmentation with only Berlin EMO-DB to increase the number of training data. Authors in [9] used Head Fusion implementing multi-headed mechanism with Attention-based Convolutional Neural Network(ACNN) to improve the emotion recognition rate on IEMOCAP and RAVDESS data selecting MFCCs as model input. They carried out empirical experiments injecting noises to clean data to improve robustness of model.

Eldin et. al.in [10] proposed hybrid convolutional neural networks (CNN) and feed forward deep neural networks using MFCCs and bag-of-acoustic-words to train the proposed network. The combined output from the constructed hybrid networks is presented as input to a dense layer with softmax activation function that provide a probability values for different categories of emotions for SER. They have used both speech and songs samples of RAVDESS dataset. Author of [4] proposes a novel hybrid architecture that learns deep features combined with other acoustic features assisting for improved recognition. Deep features are extracted from linearly scaled spectrogram using pre-trained deep network architectures. Feature selection technique is implemented to reduce the dimensionality of extracted feature and Support Vector Machine (SVM) is utilized for final classification. Proposed model is investigated using IEMOCAP, Berlin EMO-DB and RAVDESS dataset with two data augmentation technique improving the performance [4].

Lakomkin et. al.[2] compares the performance of CNN and GRU (Gated Recurrent Unit) without using their combination and improves the performance with data augmentation while testing on iClub robot. Authors in [5, 7, 8] combined CNN and different variants of RNN to extract local features with temporal information. Zhao et. al. in [5] combined CNN and LSTM to extract emotional feature from log-mel spectrogram with temporal correlations. Similarly, 1D CNN is combined with Bidirectional GRU in [8] to tune the global weights and to recognize the correlated temporal information. Authors in [7] combined ResNet 101 architecture with Bidirectional LSTM to learn the temporal information for recognizing the final state of emotions. These authors carried out experiments with IEMOCAP, Berlin EMO-DB and RAVDESS datasets and did not use any data augmentation methods.

In this work, CNN is combined with BiLSTM to extract local feature with contextual correlation from handcrafted mel-scaled spectrogram of pre-processed speech utterances. The performance of model with handcrafted feature is better than that of using raw speech directly as depicted in [5]. Speech utterances are pre-processed to enhance speech quality. Several data augmentation techniques are investigated and selected four best methods to enrich the training data samples. The performance of proposed model improve after training with augmented data.

## 3. Methodology

A deep learning model called Convolutional Recurrent Neural Network (CRNN) is constructed in this article to extract deep features representing emotions. The flow diagram of methodology of this research work is as in figure 2 and all the steps involved are elaborated in following subsections.

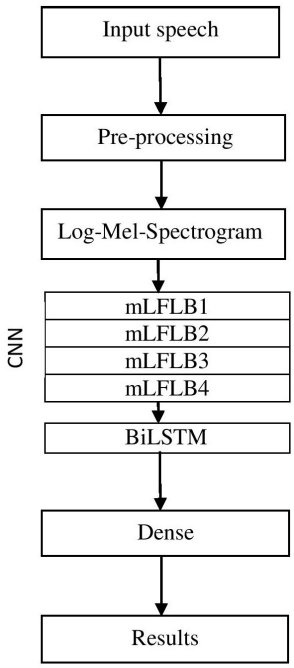


**Figure 2:** Methodology

### 3.1 CRNN Model

The CRNN model proposed is combination of Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (BiLSTM), a variant of Recurrent Neural Network (RNN). Each of them are explained in following subsections.

#### 3.1.1 CNN

The proposed network is inspired by the Local Feature Learning Blocks (LFLB) feature extraction unit of Zhao et. al. in [5], which is designed utilizing concepts of CNN to extract local emotional features. Each of these LFLBs consists of a layer of Convolution kernels followed by Batch Normalization (BN) layer, an activation function layer called Exponential Linear Unit (ELU) layer and a Pooling layer utilizing max-pooling. Layer with Convolution kernels and Pooling layer are crucial for feature learning.

The LFLB proposed by Zhao et. al. in [5] is modified here and named modified LFLB (mLFLB). Four mLFLB are used for local feature extraction. Each LFLB's convolution kernel is the same size (3x3). But, the size of convolution kernel is varied in each mLFLB. First mLFLB convolution kernel size is 9x9, second 7x7, third 5x5 and fourth 3x3. Smaller filter sizes collect as much local information as possible, while larger filter sizes represent more global, high-level, and representative information [10]. For this reason bigger convolution kernel size is used and the size is decreased in subsequent mLFLB to capture global to local feature effectively. Next, the BN layer normalizes the values generated by the convolution layer at each batch, improving deep network performance and stability. It maintains mean value near to zero and standard deviation nearly equivalent to one.

The second modification is, instead of ELU activation function ReLU is deployed. This is a result of initial experimentation where it was found that ReLU performed better than ELU. Next, values from ReLU layer is provided to Pooling layer. The learned features can be made more robust against noise and distortion by using a pooling layer. Max-pooling is mostly adopted non-linear function for Pooling. It separates the input into non-overlapping regions and outputs the maximum value of each of these sub-regions. No Dropout is used for the network, based on several reports discussing potential harm of generalization when combining two regularization methods i.e. when using both batch normalization and Dropout. The structure of modified LFLB is shown in figure 3.

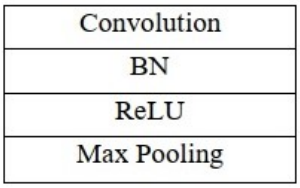


**Figure 3:** Modified Local Feature Learning Block

#### 3.1.2 BiLSTM

The proposed CRNN model consists of one layer of Bidirectional LSTM (BiLSTM) with 128 units that is constructed by combining LSTM. The bidirectional structure assists modeling the sequence from the previous sequence as well as from future sequence which may have an effect on the present state [7].

Hence, this bidirectional network could calculate to get and restore the hidden layer output from first-time to t-th time in the forward layer, then repeat the process in the backward layer to reverse computing from t-th time to first-time, and combine the output from forward and backward to get the final output as shown in figure 4.

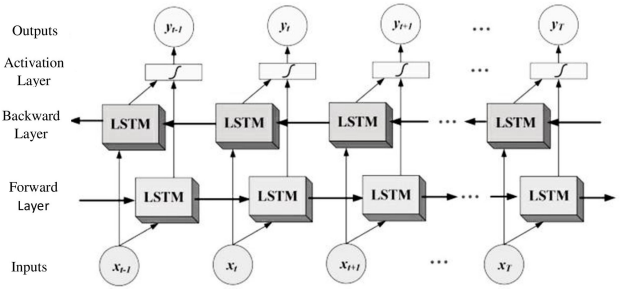


**Figure 4:** Bidirectional LSTM

The features output from last mLFLB block of CNN are presented as input to the BiLSTM layers. Then long-term temporal dependencies are learned from local features. So, the features output from the BiLSTM layer contains locally correlated global temporal information.

Dense layer with K neurons (K is equivalent to no. of emotion classes in dataset) having softmax activation function, implemented to make the prediction using those features for final classification [10]. It is also known as the fully connected layer, that is coupled to a loss function estimating the final classification error and is in charge of updating the network weights during back-propagation.The overall layer parameters determined from experiment of proposed CRNN model is presented in table 1.

**Table 1:** Layer parameter of CRNN (Conv = Convolution and Pool = Max-Pooling)

| Layers | Name | No. of Filters | Kernel Size | Stride |
|---|---|---|---|---|
| mLFLB1 | Conv<br>Pool | 64 | (9 x 9)<br>(2 x 2) | (1 x 1)<br>(2 x 2) |
| mLFLB2 | Conv<br>Pool | 64 | (7 x 7)<br>(4 x 4) | (1 x 1)<br>(4 x 4) |
| mLFLB3 | Conv<br>Pool | 128 | (5 x 5)<br>(4 x 4) | (1 x 1)<br>(4 x 4) |
| mLFLB4 | Conv<br>Pool | 128 | (3 x 3)<br>(4 x 4) | (1 x 1)<br>(4 x 4) |
| BiLSTM | BiLSTM | | 128 | |
| Dense | Softmax | | K | |

## 3.2 Datasets

Following two datasets of two different languages are selected to test language independence of proposed model.

### 3.2.1 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [11] dataset is recorded by 24 professional actors (12 females and 12 males), speaking two lexically-matched statements in North American English. Speech data have calm (192), happy(192), sad(192), angry(192), fearful(192), surprise(192), disgust(192) and neutral(96) number of emotional utterances. Only speech samples are used that contains 1440 files: 60 utterances per actor x 24 = 1440 speech audio utterances. It comprises eight classes, which makes categorization more difficult and classes are relatively balanced.

### 3.2.2 Berlin EMO-DB

The Berlin EMO-DB is a dataset developed by the University of Berlin recorded in German language [12]. It contains 535 voiced speech samples in seven emotions, that are: angry(127), bored(81), disgust(46), fear(69), happy(71), neutral(79) and sad(62). It is recorded from voice of ten professional actors including five males and five females, everyone speaking 49, 58, 43, 38, 55, 35, 61, 69, 56, 71 utterances respectively. It is used to make more detailed comparisons with earlier work.

## 3.3 Pre-processing

Speech signal is pre-processed to suppress noise and transform speech utterances to required length of 3s. Speech samples of both datasets are of variable duration having leading and trailing silence with noise as shown in figure 5.

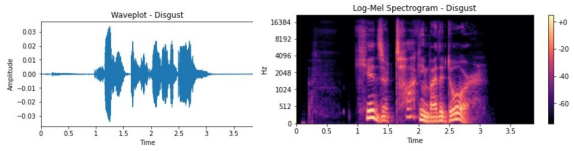


**Figure 5:** Waveplot and log-mel spectrogram before pre-processing

Steps involved in pre-processing is shown in figure 6. The speech utterances of RAVDESS is available at 48 KHz and that of Berlin EMO-DB is available at 16 KHz. First of all, the audio samples are resampled at

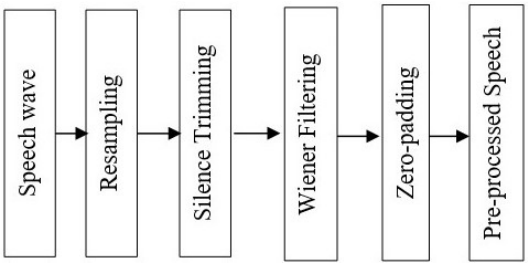


**Figure 6:** Steps in pre-processing

44.1 KHz as this sampling rate is used for CD quality audio. The leading and trailing silence with noise as shown in figure 5 are trimmed from each speech utterances in dataset using Librosa python library. After that Wiener filter is used to enhance the quality of each speech utterances removing background noise within the voiced audio segments. Wiener filtering is a linear estimating method to suppress noise in speech signal by estimating minimum mean-squared error between the original signal and its enhanced counterpart. Zero-padding is employed here to get speech utterance of fixed duration i.e. 3s. Zero value is padded at the end of each speech samples of length less than 3s. Now, all zero-padded samples are of equal in length and helps to generate fixed dimension log-mel spectrogram for model input. Waveplot and its corresponding log-mel spectrogram of a speech sample after pre-processing is shown in figure 7.

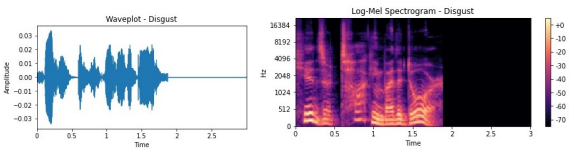


**Figure 7:** Waveplot and log-mel spectrogram after preprocessing

### 3.4 Log-Mel Spectrogram

A perceptually relevant 2D handcrafted feature called Log-mel spectrogram is extracted from pre-processed speech utterances for input to CRNN model. Computation of log-mel spectrogram from pre-processed speech waveform involves steps shown in figure 8. The pre-processed speech signal is windowed using hanning window of length 2048 and hop length is fixed to 512 samples. After that Short Time Fourier Transform (STFT) is performed to calculate the power spectrum of each frame. The mathematical equation (1) describes the discrete STFT, where y[n] and w[n] are discrete input signal and windowing function shifted by k samples. Here,

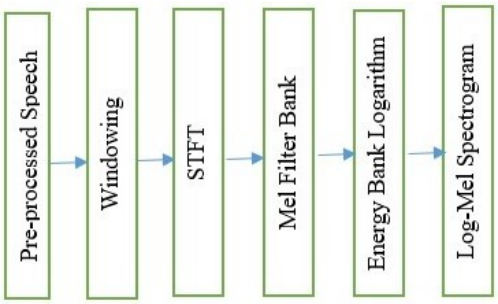


**Figure 8:** Computation steps of log-mel spectrogram

$\omega$ is discrete angular frequency.

$$Y(k,\omega) = \sum_{n=-\infty}^{\infty} y[n]w[n-k]e^{-j\omega n} \quad (1)$$

After computing spectrogram, the linear frequency scale ($f$ hertz) is converted into the Mel scale ($m$ mels), as we human do not perceive frequency on linear scale, using equation (2) by the help of 128 triangular band-pass mel filter bank. It reduce the dimension of spectrogram from (1025 x 259) to (128 x 259). Values of each mel bins are scaled to logarithmic scale (d decibels) using equation (3) as loudness of sound is perceived according to logarithmic scale.

$$m = 2595\log_{10}(1+\frac{f}{700}) \quad (2)$$

$$d = 10\log_{10}(\frac{m}{r}) \quad (3)$$

where $r$ is reference power for log-scaling and its value is set to one ($r$ = 1). Thus, the log-mel spectrogram is perceptually relevant representation of speech signal with 128 mel bins and 259 time frames as shown in figure 7. Hence, the input shape of each speech utterance is (128 x 259) for the CRNN network in the form of matrices and learns emotional features from this 2D representation. The spectrogram values are normalized using z-score normalization before presented to the model for learning deep features efficiently. Author in [13] showed that performance of deep learning models for SER improve after using different normalization techniques and best for z-score normalization. Hence, for a give feature value d, the mean, $\mu$ and standard deviation, $\sigma$ are calculated and the standardized feature value is computed as in equation (4).

$$\hat{d} = \frac{d-\mu}{\sigma} \quad (4)$$

It removes the mean and adjusts the data to unit variance in most cases.

## 4. Experiments and Results

### 4.1 Experiment Parameters

Following parameters are selected during training:

| | |
|---|---|
| Input shape: | 128 x 259 x 1 |
| Class labels: | One-hot encoding |
| Loss function: | Categorical-cross-entropy |
| Dataset split: | 80:20 |
| Optimizer: | Adam |
| Batch size: | 16 |
| Learning rate: | 0.0001 |
| Performance metric: | Accuracy |

All the training hyper-parameter are estimated from training data performing experiments and test data is only used for final prediction. Various techniques are implemented to reduce the chance of over-fitting during training of model such as regularization, batch normalization and early stopping. All experiments are done on Google Colab using Python programming language.

### 4.2 Experiment Results

#### 4.2.1 RAVDESS

The experimental result are displayed by the confusion matrix and recognition report. Also precision, recall and F1-score is calculated. For convenience class labels in confusion matrix are written in short form (Ang=Angry, Cal=Calm, Dis=Disgust, Fea=Fear, Hap=Happy, Neu=Neutral, Sur=Surprise, Bor=Bored, Sad=Sad) with actual labels along vertical axis and predicted labels along horizontal axis.The confusion matrix in figure 9 depicts the confusion among classes predicted by model on test data of RAVDESS dataset. It is normalized confusion matrix and principal diagonal shows the actual recognition rate of each class.

| | Ang | Cal | Dis | Fea | Hap | Neu | Sad | Sur |
|---|---|---|---|---|---|---|---|---|
| Ang | **94.59** | 0.0 | 2.7 | 0.0 | 0.0 | 0.0 | 2.7 | 0.0 |
| Cal | 0.0 | **91.67** | 0.0 | 0.0 | 0.0 | 2.78 | 5.56 | 0.0 |
| Dis | 4.88 | 0.0 | **75.61** | 7.32 | 2.44 | 0.0 | 9.76 | 0.0 |
| Fea | 2.44 | 2.44 | 4.88 | **80.49** | 2.44 | 0.0 | 7.32 | 0.0 |
| Hap | 5.0 | 2.5 | 0.0 | 15.0 | **67.5** | 0.0 | 7.5 | 2.5 |
| Neu | 0.0 | 16.67 | 0.0 | 0.0 | 0.0 | **70.83** | 12.5 | 0.0 |
| Sad | 0.0 | 5.88 | 5.88 | 14.71 | 2.94 | 0.0 | **70.59** | 0.0 |
| Sur | 0.0 | 0.0 | 0.0 | 0.0 | 2.86 | 0.0 | 11.43 | **85.71** |

**Figure 9:** Confusion matrix for RAVDESS

**Table 2:** Recognition report for RAVDESS

| Class(Support) | Precision | Recall | F1-score |
|---|---|---|---|
| Angry(37) | 87.50 | 94.59 | 0.9091 |
| Calm(36) | 80.49 | 91.67 | 0.8571 |
| Disgust(41) | 86.11 | 75.61 | 0.8052 |
| Fear(41) | 70.21 | 80.49 | 0.7500 |
| Happy(40) | 87.10 | 67.50 | 0.7606 |
| Neutral(24) | 94.44 | 70.83 | 0.8095 |
| Sad(34) | 54.55 | 70.59 | 0.6154 |
| Surprise(35) | 96.77 | 85.71 | 0.9091 |
| Accuracy | | | **79.86** |
| Macro Avg | 82.15 | 79.62 | 0.8564 |
| Weighted Avg | 81.72 | 79.86 | 0.8016 |

Table 2 visualizes support count along with precision, recall and F1-score for each class of RAVDESS dataset with macro (unweighted) and weighted averages. Test data has 288 utterances which is 20% of total utterances in dataset i.e. 1440. The recognition accuracy is 79.86%. The unweighted and weighted precision, recall and F1 score have nearly same value indicating samples in each emotion class are nearly balanced.

#### 4.2.2 Berlin EMO-DB

The confusion matrix for Berlin EMO-DB is shown in figure 10 showing confusion among emotion classes. From the confusion matrix it is observed that 30.77% samples of happy class are mis-classified as angry as both of these emotions are considered as strong emotion. All the samples of sad emotion are correctly recognized without any mis-classification. The

| | Ang | Bor | Dis | Fea | Hap | Neu | Sad |
|---|---|---|---|---|---|---|---|
| Ang | **85.29** | 0.0 | 0.0 | 8.82 | 5.88 | 0.0 | 0.0 |
| Bor | 0.0 | **80.0** | 0.0 | 10.0 | 0.0 | 10.0 | 0.0 |
| Dis | 0.0 | 0.0 | **80.0** | 0.0 | 20.0 | 0.0 | 0.0 |
| Fea | 0.0 | 0.0 | 0.0 | **91.67** | 8.33 | 0.0 | 0.0 |
| Hap | 30.77 | 0.0 | 0.0 | 7.69 | **61.54** | 0.0 | 0.0 |
| Neu | 0.0 | 11.76 | 0.0 | 5.88 | 0.0 | **82.35** | 0.0 |
| Sad | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **100** |

**Figure 10:** Confusion matrix for Berlin EMO-DB

average accuracy is 83.18% as shown in table 3. The test dataset has 107 utterances which is 20% of available dataset. Weighted and unweighted values of precision, racall and F1-score have slight difference as classes are not equally balanced.

**Table 3:** Recognition report for Berlin EMO-DB

| Class(Support) | Precision | Recall | F1-score |
|---|---|---|---|
| Angry(34) | 87.88 | 85.29 | 0.8657 |
| Bored(10) | 80.00 | 80.00 | 0.8000 |
| Disgust(10) | 100.0 | 80.00 | 0.8889 |
| Fear(12) | 64.71 | 91.67 | 0.7586 |
| Happy(13) | 61.54 | 61.54 | 0.6154 |
| Neutral(17) | 93.33 | 82.35 | 0.8750 |
| Sad(11) | 100.0 | 100.0 | 1.0000 |
| **Accuracy** | | | **83.18** |
| Macro Avg | 83.92 | 82.98 | 0.8291 |
| Weighted Avg | 84.59 | 83.18 | 0.8346 |

## 4.3 Data Augmentation

Augmentation processes on data are performed to increase the number of training samples as per the requirements of Deep Neural Networks [2] where the size of available original data is not enough. It generates additional training data samples by deforming the original data in the training dataset. The labels of the supplemented data cannot be changed, which is a condition for data augmentation [4]. For audio data augmentation, techniques mainly used are Background Noise Injection, Time Shifting, Pitch Shifting, Time Stretching, etc. Here, the experimental datasets are first randomly splitted into training and testing sets with the ratio of 80:20 and techniques of data augmentation are applied on training data only. Following four techniques are found to be more effective from initial experimentation. They are listed below:

- **Background Noise Injection:** Add random noise to speech samples with threshold value 0.1.
- **Time Stretch Up:** Stretch speech samples along time axis with the rate of 1.1.
- **Time Stretch Down:** Stretch speech samples along time axis with the rate of 0.8.
- **Pitch Sifting:** Shift the pitch of speech samples with the pitch factor of 1.5.

Sample log-mel spectrogram is shown in figure 11 after augmentation with its original one shown in figure 7. After data augmentation the training data size increases by four times the original training data size. The performance of model after it is discussed in following subsections.

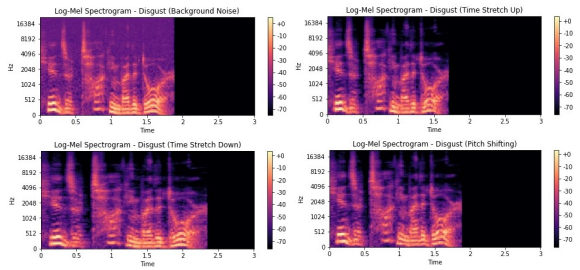


**Figure 11:** Sample log-mel spectrogram after data augmentation

## 4.4 Results and Discussion

### 4.4.1 RAVDESS

Confusion matrix in figure 12 shows the confusion among the classes on test data for RAVDESS dataset after augmentation. The recognition rate greatly increased from 79.86% to 85.76% showing the essence of data augmentation.

| | Ang | Cal | Dis | Fea | Hap | Neu | Sad | Sur |
|---|---|---|---|---|---|---|---|---|
| Ang | **97.3** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.7 | 0.0 |
| Cal | 0.0 | **88.89** | 0.0 | 0.0 | 0.0 | 2.78 | 8.33 | 0.0 |
| Dis | 2.44 | 0.0 | **95.12** | 0.0 | 0.0 | 0.0 | 2.44 | 0.0 |
| Fea | 0.0 | 2.44 | 0.0 | **70.73** | 4.88 | 0.0 | 19.51 | 2.44 |
| Hap | 5.0 | 0.0 | 0.0 | 0.0 | **82.5** | 0.0 | 7.5 | 5.0 |
| Neu | 0.0 | 12.5 | 0.0 | 0.0 | 4.17 | **75.0** | 8.33 | 0.0 |
| Sad | 0.0 | 2.94 | 5.88 | 0.0 | 2.94 | 2.94 | **85.29** | 0.0 |
| Sur | 0.0 | 0.0 | 0.0 | 0.0 | 11.43 | 0.0 | 0.0 | **88.57** |

**Figure 12:** Confusion matrix for RAVDESS after data augmentation

**Table 4:** Recognition report for RAVDESS after data augmentation

| Class(Support) | Precision | Recall | F1-score |
|---|---|---|---|
| Angry(37) | 92.31 | 97.30 | 0.9474 |
| Calm(36) | 86.49 | 88.89 | 0.8767 |
| Disgust(41) | 95.12 | 95.12 | 0.9512 |
| Fear(41) | 100.0 | 70.73 | 0.8286 |
| Happy(40) | 80.49 | 82.50 | 0.8148 |
| Neutral(24) | 90.00 | 75.00 | 0.8182 |
| Sad(34) | 61.70 | 85.29 | 0.7160 |
| Surprise(35) | 91.18 | 88.57 | 0.8986 |
| **Accuracy** | | | **85.76** |
| Macro avg | 87.16 | 85.43 | 0.8564 |
| Weighted avg | 87.49 | 85.76 | 0.8598 |

The detailed recognition report is shown in table 4 with weighted and unweighted averages for precision, recall and F1-score. The weighted average of recall has

highest value of 87.49% and weighted and unweighted values differs with negligible amount.

#### 4.4.2 Berlin EMO-DB

Confusion matrix in figure 13 shows the confusion among the classes on test dataset after training the model with augmented data along with original data. The average recognition rate increased from 83.18% to 91.59% after data augmentation. Weighted

| | Ang | Bor | Dis | Fea | Hap | Neu | Sad |
|---|---|---|---|---|---|---|---|
| Ang | **97.06** | 0.0 | 0.0 | 2.94 | 0.0 | 0.0 | 0.0 |
| Bor | 0.0 | **90.0** | 10.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Dis | 0.0 | 0.0 | **100.0** | 0.0 | 00.0 | 0.0 | 0.0 |
| Fea | 0.0 | 0.0 | 0.0 | **83.33** | 16.67 | 0.0 | 0.0 |
| Hap | 15.38 | 0.0 | 0.0 | 7.69 | **76.92** | 0.0 | 0.0 |
| Neu | 0.0 | 11.76 | 0.0 | 0.0 | 0.0 | **88.24** | 0.0 |
| Sad | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **100** |

**Figure 13:** Confusion matrix for Berlin EMO-DB after data augmentation

**Table 5:** Recognition report for Berlin EMO-DB after data augmentation

| Class(Support) | Precision | Recall | F1-score |
|---|---|---|---|
| Angry(34) | 94.29 | 97.06 | 0.9565 |
| Bored(10) | 81.82 | 90.00 | 0.8571 |
| Disgust(10) | 90.91 | 100.0 | 0.9524 |
| Fear(12) | 83.33 | 83.33 | 0.8333 |
| Happy(13) | 83.33 | 76.92 | 0.8000 |
| Neutral(17) | 100.0 | 88.24 | 0.9375 |
| Sad(11) | 100.0 | 100.0 | 1.0000 |
| **Accuracy** | | | **91.59** |
| Macro Avg | 90.53 | 90.79 | 0.9053 |
| Weighted Avg | 91.74 | 91.59 | 0.9155 |

average for precision reached up to 91.74% as shown in recognition report in table 5 for Berlin EMO-DB. The values of unweighted and weighted averages for precision, recall and F1-score are not far apart with recognition rate of 91.59%.

### 4.5 Performance Comparison

The performance of proposed model is compared with the latest methods in literature in table 6 in terms of achieved recognition accuracy and it is observed that our proposed CRNN model outperform all other methods. Comparison is made for both datasets with selected features as model input for training the model. For RAVDESS dataset, performance improves

**Table 6:** Performance comparison with other methods in literature

| Year/ Ref. | Features Input | RAVD-ESS | Berlin EMO-DB |
|---|---|---|---|
| 2020[6] | 5-spectral Features | 71.61 % | 86.1 % |
| 2020[7] | Spectrogram | 82.02 % | 91.14 % |
| 2020[4] | Spectrogram and Acoustic Features | 79.41% | 90.21% |
| 2021[10] | MFCC | 83 % | - |
| 2021[9] | MFCC | 77.8 % | - |
| 2021[8] | Enhanced Speech | 78.01 % | 91.14 % |
| **Our model** | Log-mel Spectrogram | **85.76 %** | **91.59 %** |

by around 3% as compared to other method in literature and performance of model with Berlin EMO-DB dataset is comparable to [4, 7, 8] with small improvement in accuracy. From the table 6 it is observed that log-mel spectrogram is an appropriate feature representation of emotional content as compared to other features such as MFCC, linear scale spectrogram, chromagram, etc. used by other researchers to train the model. Our CRNN model is able to extract deep emotional feature from it for final classification of emotion.

The performance of DNN models improves with enriching the number of training samples with the help of data augmentation as indicated by Bilal in [4]. He showed that recognition rate of best performing model improves from 77.26% to 79.41% for RAVDESS data and from 87.68% to 90.21% for Berlin EMO-DB using two data augmentation techniques. Isaa et. al. in [6] also used different augmentation techniques to increase the training data size of Berlin EMO-DB by four times to its original one. They achieved the recognition rate of 86.1% on test data after training one dimensional CNN model with augmented data. The performance of proposed model is in compliance with these studies reaching the recognition rate of 85.76% for RAVDESS dataset and 91.59% for Berlin EMO-DB dataset. The model is able to recognize the emotional contents of utterances in North American English language and German language with same hyper-parameter settings. Hence, achieved results indicate that the proposed model is language independent.

## 5. Conclusion and Future Work

SER is a difficult task due to lot many variability in speech data and it is also uncertain about the best feature representing emotional content of human speech. Our proposed method has tried to address this issue by selecting log-mel spectrogram, a perceptually relevant 2D representation of speech waveform, to train the CRNN model. The proposed model shows improved performance on North American English and German language datasets. Also the decision for pre-processing techniques and data augmentation techniques proved to be beneficial.

Though the proposed model is better at learning emotional content from log-mel spectrogram, the extraction process of log-mel spectrogram is computationally expensive and require lots of domain expertise and computation steps, making the process complicated. In future, we will extend this work by eliminating computationally expensive manual handcrafted feature extraction step. The model that can learn appropriate emotional content direct from raw audio waveform is to be investigated with other feature set increasing recognition rate. Several other data augmentation techniques can be used to further increase the training data size. Finally, the model can also be experimented and tested with other language speech data including Nepali by developing and using respective language datasets.

## Acknowledgments

Authors are thankful to IOE Pulchowk campus for providing this precious opportunity. The authors also express their gratitude for the support of the faculty and administrative members from IOE-Paschimanchal Campus who have helped directly and indirectly to make this research successful.

## References

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[2] Egor Lakomkin, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter. On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 854–860. IEEE, 2018.

[3] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[4] Mehmet Bılal Er. A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, 8:221640–221653, 2020.

[5] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[6] Dias Issa, M Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 2020.

[7] Muhammad Sajjad, Soonil Kwon, et al. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875, 2020.

[8] Mustaqeem and Soonil Kwon. 1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features. *CMC-COMPUTERS MATERIALS & CONTINUA*, 67(3):4039–4059, 2021.

[9] Mingke Xu, Fan Zhang, and Wei Zhang. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset. *IEEE Access*, 9:74539–74549, 2021.

[10] Mai Ezz-Eldin, Ashraf AM Khalaf, Hesham FA Hamed, and Aziza I Hussein. Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition. *IEEE Access*, 9:19999–20011, 2021.

[11] Steven R Livingstone and Frank 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*, 13(5):e0196391, 2018.

[12] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth european conference on speech communication and technology*, 2005.

[13] Tshephisho Joseph Sefara. The effects of normalisation methods on speech emotion recognition. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8. IEEE, 2019.