International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *18 August 2023* ***Revised:*** *08 October 2023* ***Accepted:*** *20 October 2023*

Integration of MFCC Extraction and LSTM Algorithm on PYNQ-Z2 for Enhanced Audio Analysis

Sheetal U. Bhandari[1], Deepti Khurge[1], Rajani PK[1], Varsha Bendre[1], Ashwini S. Shinde[1]
[1]Department of Electronics and Telecommunication Engineering, Pimpri Chinchwad College of Engineering, Pune, India.
sheetal.bhandari@pccoepune.org, dipti.khurge@pccoepune.org, rajani.pk@pccoepune.org, varsha.bendre@pccoepune.org, ashwinik09@gmail.com

Abstract— The need for Speech Emotion Recognition (SER) is growing since researchers have found it difficult to interpret human emotions from speech data. SER is very interesting yet very challenging task of human-computer interaction (HCI). The SER application can be benefitted depending on the type of feature extraction technique and model used for classification. Deep Learning has made a great impact in the field of audio, image, video, EEG and ECG classification. The speech signal characteristics and classification model affect how well the SER application performs. The paper briefs about deploying Deep Learning Algorithm on FPGA based board i.e., PYNQ-Z2. MFCC feature extraction technique and LSTM model used for classification of human emotion is implemented on the board. Emotion can be predicted using led buttons on the board.

Keywords- Speech; Emotion; Feature Extraction; Deep Learning; MFCC; LSTM.

I. INTRODUCTION

The task of recognizing the emotional components in speech is known as speech emotion recognition (SER). Studies into automating emotion recognition in HCI are still ongoing, but it is possible to do so using natural language processing (NLP). This is something that humans are good at doing naturally, but is still being researched in this area. Making machines act and appear more like human by giving them emotions is considered to be challenging task. Robots that are able to understand emotions would be able to provide appropriate responses in human machine interaction. In some situations, machines that can imitate human emotions and have conversations that seem very natural can take the place of people. The ability for machines to recognize the emotions expressed in speech is essential. With just this feature, it is possible to have a fully meaningful conversation based on mutual HCI trust and understanding.

The aim is to create a network that can take raw data as input and produce a class label as output. The network handles everything by itself. The network parameters, such as weights and bias values, which operate as features effectively classifying the data into the appropriate categories, are optimized during the training stage. Applications that require for human-machine interaction, including online movies and computer tutorials, can benefit from the usage of SER. The system's reaction to the user in these applications is based on the emotion it has identified. In order to ensure safety, it is also useful for in-car board systems to be aware of the driver's mental condition while driving. Therapists can also utilize it as a diagnostic tool.

The system's accuracy rate is dependent on the features and classification model used for SER application. Pitch, energy, Zero Crossing Rate (ZCR), Discrete Wavelet Transform (DWT), Mel Frequency Cepstral Coefficients (MFCC) are some of the features that can be considered for SER application. Learning Algorithms like CNN, RNN, SVM, LSTM can be considered for classification. Table I shows a literature survey for SER application with different classification model. Out of these MFCC is considered for feature extraction because it can extract more information required for prediction and LSTM for classification of human emotion as it can store information about previous state.

Table I. LITERATURE SURVEY

Reference No.	Feature Extraction	Algorithm Used	Accuracy (%)
[2]	MFCC	LSTM	90
[5]	MFCC	CNN	83
[8]	MFCC	LSTM-RNN	80
[9]	STFT	CNN	86
[9]	STFT	LSTM	78.3
[10]	MFCC	CNN	75

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 18 August 2023 **Revised:** 08 October 2023 **Accepted:** 20 October 2023*

II. METHODOLOGY

The 32-bit integer must first be converted to 16-bits for the voice processing phase. After that, split 16-bit words into 8-bit words with 1 bit sample in each. The sample rate is reduced from 3MHz to 32kHz. This is done to reduce the buffer size error. Because it provides superior frequency resolution in the low frequency range than any other feature extraction approaches, MFCC is taken into consideration. As a result, it can be used for all signal types and is unaffected by noise. Here MFCC coefficients are calculated and trained using Deep learning model. LSTM and CNN algorithm are for classification. Comparison of both algorithm is done. Here for both algorithm three models are trained and tested. Here LSTM algorithm has better performance, thus it is considered to implement on PYNQ-Z2 board. Board consists of 4 LEDs; the predicted emotion is displayed on the board using led as shown in Fig.1.

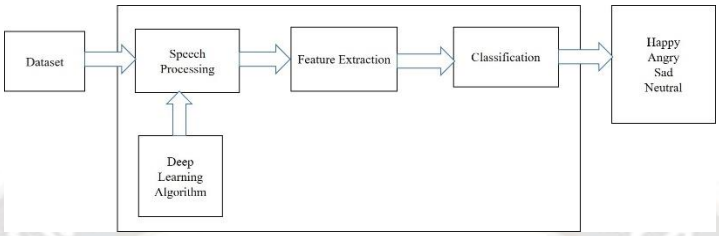


Figure 1.SER Block Diagram

A. Dataset

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database contains recordings of 24 actors portraying 7 different emotions. Happy sentiments are calm feelings; sad or angry feelings are furious or frightening feelings. Emotions like surprise and contempt are also frequent. From each emotional category, 30 to 40 examples of male and female speech are selected. Every sample is produced with a specific identification name and is either of strong or normal intensity. Name of the file: 03-09-06-01-02-01-12.wav

B. MFCC

The MFCC calculation process involves several steps to transform the raw audio signal into a set of representative coefficients.

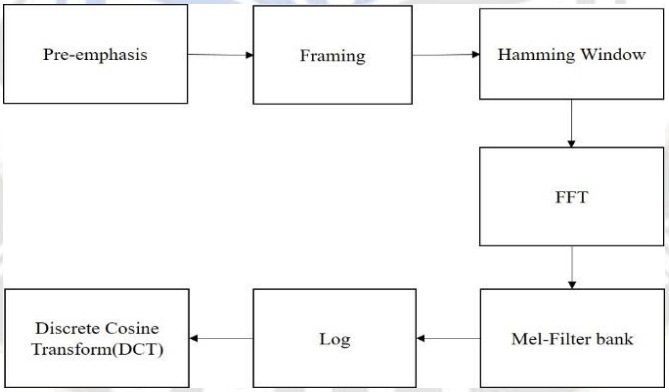


Figure 2. MFCC Calculation Process

- Pre-emphasis: At low frequency region the speech signal has more energy as compared to that of high frequency region. Pre-emphasis step is done to remove white noise and also to improve energy at high frequency region.

$$Y(n)=X(n) -0.95\ X(n-1) \qquad (1)$$

- Framing: Framing is done to divide signal into short frames so that signal can be analyzed properly. To provide a near-steady state signal, the signal is split into 20ms long, 50 percent overlapping segments.
- Windowing: This technique is used to minimize spectral distortion and frame discontinuities at the beginning and completion of the data. Every frame of the signal is subjected to windowing, because it eliminates side lobes and offers a precise frequency spectrum, the hamming window approach is used. The Hamming Window Equation is presented as:

$$W(n)=W0 = (n - \frac{N-1}{2}) \qquad 0 \leq n \leq N-1 \quad (2)$$

N= In each frame number of samples
Y(n) = Output Signal, X(n) = Input Signal, W(n) = Hamming Window

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: 18 August 2023 Revised: 08 October 2023 Accepted: 20 October 2023

- Fast Fourier Transform [FFT]: The windowing signal is transformed into the frequency domain and the signal spectrum using the Fast Fourier Transform (FFT). The FFT of the signal is shown in equation. The Fourier Transform of y(t), h(t), and x are shown here as Y(w), H(w), and X(w) (t).

$$Y(w)=FFT[h(t)*x(t)]=H(w).X(w) \qquad (3)$$

- Mel filter bank and frequency wrapping: The frequency scale of Mel is spaced linearly for 1000Hz frequency and it is spaced logarithmically above 1000Hz. Pitch or frequency of the signal is specified by a Mel unit.
  Triangular filters that have been overlapped comprise the Mel filter bank. The center frequencies of 2 adjacent filters are used to calculate each filter's cutoff frequency. The filters' center frequencies are linearly spaced and have fixed bandwidths on the Mel scale. The Mel function converts the actual frequency (Hz) of the voice signal to the Mel frequency.

$$F = [2595 * log10(1 + \frac{f}{700}) \qquad (4)$$

- Discrete Cosine Transform [DCT]: DCT is utilized to ascertain the time domain of the signal by analyzing the log Mel spectrum. The output of the DCT is the Mel Frequency Cepstral Coefficients. The collection of coefficients is known as an acoustic vector. An audio vector sequence is created from each input utterance.

$$C(n) = \sum_{k=1}^{k} (\text{logSk}) \cos\left(\text{n} * (\text{k} - 0.5) * \frac{\pi}{k}\right) \quad (5)$$

$$\text{n} = 1, 2\ldots., \text{k}$$

Whereas S_k = 1, 2, …..,k are the outputs of last step.

As shown in Fig 2. The resulting set of MFCCs, along with optional delta and delta-delta coefficients, forms a feature vector that represents the spectral characteristics of the audio signal within each frame.
Table II shows MFCC parameters required for the SER application

Table II. MFCC Parameters
Parameters Value

MFCC Parameters	Value
Pre-emphasis filter	0,9
Window	Hamming Window
Analysis window length	20ms
Number of Cepstrum	13 coefficients

C. LONG SHORT-TERM MEMORY (LSTM)

Long-term memory (LTSM) is a type of neural network that is able to remember the order of events. The output from the prior step is used as the input for the current phase of the RNN. LSTM solved the problem of the RNN's long-term reliance, in which the RNN can predict words based on current input but cannot predict words stored in long-term memory.

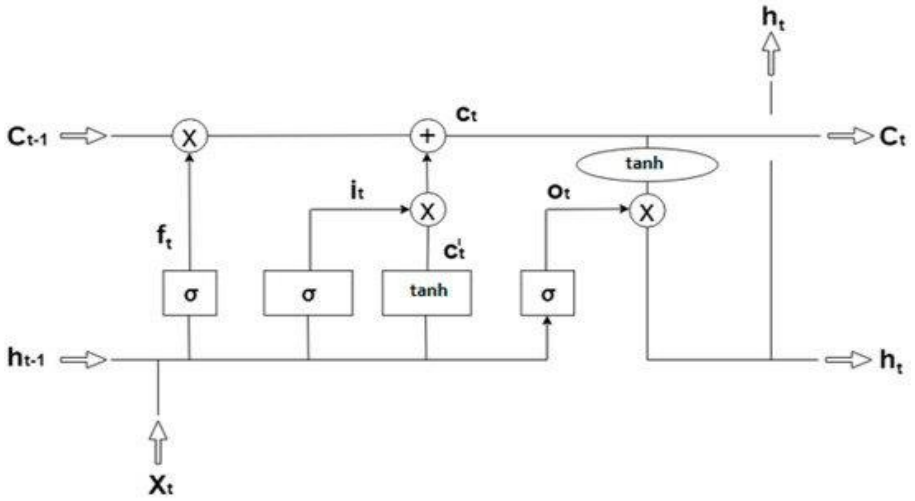


Figure 3. LSTM Network

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 18 August 2023 **Revised:** 08 October 2023 **Accepted:** 20 October 2023*

Higher gap lengths do not result in efficient performance from RNN. By default, the LSTM may retain data for a long time. The application is used to process, predict, and categorize time-series data.
The inputs provided to the hidden layer make up the weight matrix U in the LSTM structure shown in Fig. 3. Xt, Ct-1, ht-1 is the inputs given to the network and Ct and ht are the outputs. Table III shows LSTM parameters used in the application.

Table III. LSTM Parameters

Layer	Parameters
LSTM	26624
dropout	0
dense	2080
Dense_1	264
Dense_2	36

III. SER IMPLEMENTATION

Figure 3 shows the block diagram of proposed system. From the speech, this system predicts four emotions: happy, angry, sad, and neutral. The system needs an Ethernet connector, an SD
card with an installed operating system for PYNQ, and a USB cable for power. Using a virtual Jupyter notebook, speech processing, feature extraction, and emotion prediction is implemented on board.

- SER Implementation Results:

Table IV. LSTM Parameters

Model No	Model 1	Model 2	Model 3
MFCC Length	39	28	12
Hidden layer	2	2	2
LSTM layer output size	32	32	32
1st hidden layer	16	16	16
2nd hidden layer	8	8	8

Tables IV shows three machine learning models for the LSTM algorithms that have been evaluated for speech emotion recognition systems using various performance parameter combinations.

A. Performance Parameters

One of the key elements in creating a powerful DL model is evaluating the performance of Deep Learning. Different metrics—also referred to as performance metrics or evaluation metrics—are used to assess the effectiveness or quality of the model. These performance indicators make it easier to comprehend how well the model has done given the available data. Precision.

The proportion of true positives to all expected positives is known as precision.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \qquad (6)$$

Where TP = True Positive FP = False Positive

If a model's precision score is close to 1, it can accurately distinguish between correct and erroneous labeling and didn't miss any true positives. A classifier with a poor precision score (!0.5) produces a lot of false positives, which may be the result of an unbalanced class.

- Recall

A Recall is essentially the ratio of true positives to all the positives in ground truth.

$$Recall = \frac{\text{TN}}{(\text{TN}+\text{FP})} \qquad (7)$$

Where TP = True Positive FN = False Negative

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 18 August 2023 **Revised:** 08 October 2023 **Accepted:** 20 October 2023*

Recall toward 1 indicates that the model didn't overlook any genuine positives and can distinguish between cancer patients who have been mislabeled and those who have been accurately identified. A classifier with a poor recall score (!0.5) makes a lot of erroneous negative predictions, which may be the result of an unbalanced class or improperly configured model hyperparameters.

- F1-Score

The harmonic mean of precision and recall, also known as the traditional F-measure or balanced F-score, is a metric that combines precision and recall.

$$F1\ score = \frac{2\times TP}{2\times TP+FN+FP} \quad (8)$$

A high F1 score denotes both high recall and great precision. It has a great balance between recall and precision and produces favorable outcomes. Poor F1 performance indicates performance at a threshold. Low recall indicates that the entire test set was improperly provided. A low degree of precision indicates that some of the examples that were considered affirmative cases are incorrect.

Accuracy One of the simplest Classification metrics to use is accuracy, which is calculated as the proportion of accurate predictions to all other predictions.

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

Table V. CNN Model Parameters

Model	**I**	**II**	**III**
MFCC Length	28	39	12
MaxPooling 2D	2	2	2
Dense	64	32	64
Dropout Rate	0.2	0.2	0.2
Activation Function	'relu'	'relu'	'relu'

Table VI. LSTM Model Parameters

Model	**I**	**II**	**III**
MFCC Length	39	28	12
Hidden Layer	2	2	2
LSTM Layer	32	32	32
1st Layer	16	16	16
2nd Layer	8	8	8

Tables V and VI show three deep learning models for the LSTM and CNN algorithms evaluated for speech emotion recognition systems using various performance parameter combinations.80 percent of data was used for training and 20 percent for testing. Python IDLE was used to compile and train and test the dataset.

Table VII. Performance Parameters of CNN Model

Model Number	**CNN**	**I**	**II**	**III**
Precision	Neutral	0.61	0.45	0.62
	Angry	0.77	0.81	0.69
	Happy	0.49	0.55	0.47
	Sad	0.49	0.44	0.4
Recall	Neutral	0.39	0.46	0.29
	Angry	0.79	0.6	0.57
	Happy	0.51	0.34	0.49
	Sad	0.57	0.73	0.63
F1 score	Neutral	0.48	0.46	0.39
	Angry	0.48	0.68	0.62
	Happy	0.5	0.42	0.48
	Sad	0.52	0.55	0.49
Accuracy	Neutral	0.8	0.74	0.79
	Angry	0.85	0.82	0.78
	Happy	0.73	0.75	0.72
	Sad	0.77	0.73	0.7

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *18 August 2023* ***Revised:*** *08 October 2023* ***Accepted:*** *20 October 2023*

Average Accuracy	0.78	0.76	0.74

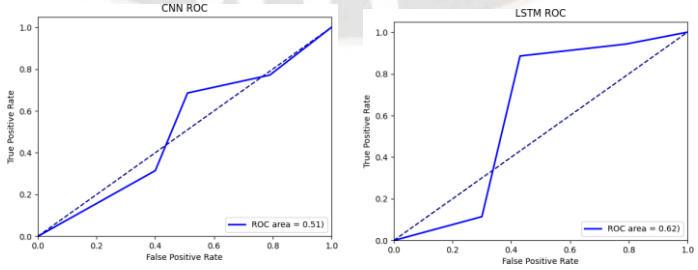
Table VII and VIII shows the comparison for Performance parameters for LSTM and CNN. For CNN algorithm I model gave an average accuracy of about 0.78 which is highest among the two models i.e for model II and III is 0.76 and 0.74.Even though the accuracy of model I is highest but its recall value and f1 score is less. For LSTM algorithm II model is considered better and accuracy obtained is 0.85. For standalone prediction of human emotion, accuracy for happy emotion obtained in model I is 0.9 which is highest among all predicted emotion.

Table VIII. Performance Parameters of LSTM Model

Model Number	**LSTM**	**I**	**II**	**III**
Precision	Neutral	0.67	0.88	0.78
	Angry	0.82	0.84	0.79
	Happy	0.82	0.75	0.68
	Sad	0.57	0.55	0.56
Recall	Neutral	0.36	0.5	0.64
	Angry	0.79	0.76	0.71
	Happy	0.8	0.77	0.77
	Sad	0.87	0.8	0.63
F1 score	Neutral	0.47	0.64	0.71
	Angry	0.8	0.8	0.75
	Happy	0.81	0.76	0.72
	Sad	0.68	0.65	0.59
Accuracy	Neutral	0.82	0.87	0.88
	Angry	0.88	0.88	0.85
	Happy	0.9	0.87	0.84
	Sad	0.82	0.8	0.8
Average Accuracy		**0.85**	**0.85**	**0.84**

- Model Performance Parameters

LSTM II model gave better results as compared to other, so as for CNN, I model gave better results. Model were trained for 100 epochs and results were obtained. As seen from result accu- racy for trained data is more as compared to test data. In fig 4. Accuracy for LSTM model is stabilized after 40 epochs. Model loss for test data for both algorithms increases with increase in epochs. For LSTM, the observed average loss is 0.5, but for CNN, it is more like 0.8. The receiver operating characteristic curve (ROC curve) is a graph that displays how well a classification model performs across all categorization levels. Two parameters are plotted on this curve: % True Positive. For CNN ROC is 0.51 and so for LSTM it is 0.62. To achieve better performance results, need to reduce model loss and false positive rate.



CNN I LSTM II

Figure 4: Observed Model ROC (Receiver operating characteristic curve)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *18 August 2023* ***Revised:*** *08 October 2023* ***Accepted:*** *20 October 2023*

IV. LSTM IMPLEMENTATION ON PYNQ Z2 BOARD

To Implement SER application on board, LSTM is considered because it has long term dependencies which is required for emotion prediction. One of the key benefits of utilizing LSTM is that huge quantities of data can be trained and tested without expanding the network size. While performing SER application for CNN algorithm, training of data takes lot of time as compared to LSTM and other disadvantage of using CNN is its performance result.

```
[5]: import time
     import numpy as np

     start = time.time()
     af_uint8 = np.unpackbits(pAudio.buffer.astype(np.int16)
                              .byteswap(True).view(np.uint8))
     end =  time.time()

     print("Time to convert {:,d} PDM samples: {:0.2f} seconds"
           .format(np.size(pAudio.buffer)*16, end-start))
     print("Size of audio data: {:,d} Bytes"
           .format(af_uint8.nbytes))

     Time to convert 4,608,000 PDM samples: 0.04 seconds
     Size of audio data: 4,608,000 Bytes

[6]: import time
     from scipy import signal

     start = time.time()
     af_dec = signal.decimate(af_uint8,8,zero_phase=True)
     af_dec = signal.decimate(af_dec,6,zero_phase=True)
     af_dec = signal.decimate(af_dec,2,zero_phase=True)
     af_dec = (af_dec[10:-10]-af_dec[10:-10].mean())
```

Figure 5: Preprocessing Signal

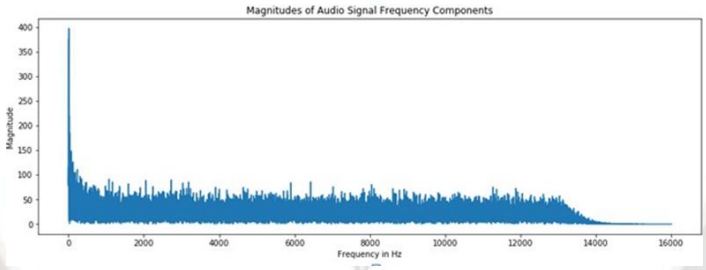
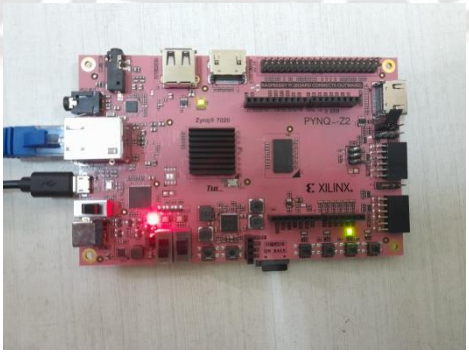


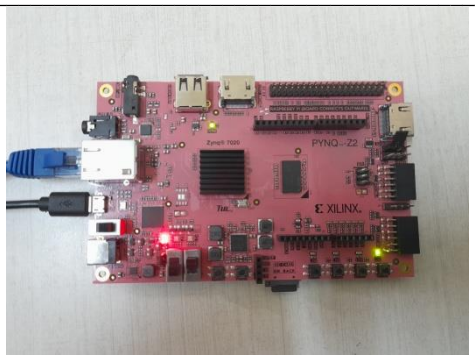
Figure 6Magnitude of Signal

RAVDESS dataset is applied as input to the board,.80 percent of data was used to train and 20 percent for testing. First preprocessing of signal was done i.e 32 bit was divided into 16 bit and then to 8 bit. Then sample rate was converted to from 3MHz to 32 KHz. This is done because of buffer size. After that LSTM weights are applied to each sample. Tan h function is used as activation function for the model.

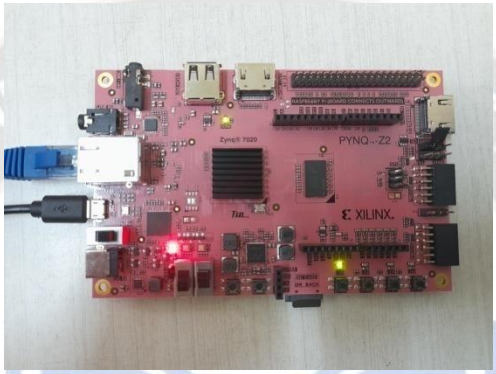


a. Happy Emotion Predicted

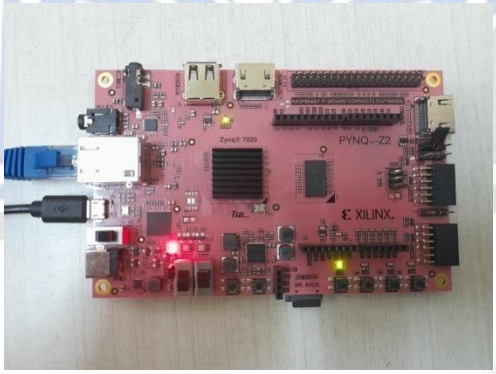
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *18 August 2023* ***Revised:*** *08 October 2023* ***Accepted:*** *20 October 2023*



b. Sad Emotion Predicted



c. Angry Emotion Predicted



d. Neutral Emotion Predicted

Figure 7 Emotion Predicted

V. CONCLUSION

Different types of speech features are considered for speech emotion recognition. For this application MFCC is considered as it can extract more information required for emotion prediction.39MFCC coefficients are extracted. To implement SER application on PYNQ-z2 RAVDESS dataset was considered. MFCC feature extraction and LSTM algorithm is implemented on board. Speech preprocessing is done first and

prediction of emotion is done. Output on board is shown with the help of led. led (0) indicates sad emotion is predicted, led (1) indicates angry emotion, led (2) indicated neutral emotion, led (3) indicates happy emotion is predicted. Further Improvement can be done by obtaining accuracy result using PYNQ-z2 board. Still lot of research is needed to achieve better performance results on board.

REFERENCES

[1] Luca S, M.Santambrogia, D.Sciuto, ”On How to Efficiently Implement Deep Learning Algorithm On PYNQ platform.”, Published in 2018.

[2] H S Kumbhar,S.U Bhandari , ” On the Evaluation and Implementation of LSTM Model for Speech Emotion Recognition using MFCC ”,Published in February 2022.

[3] Anusha Koduru, Hima Bindu Valiveti1, Anil Kumar Budati1, "Feature extraction algorithms to improve the speech emotion recognition rate " International Journal of Speech Technology March 2020,
[4] Zhang Wanli, Li Guoxin, "The Research of Feature Extraction Based on MFCC for Speaker Recognition ", 2013 3rd International Conference on Computer Science and Network Technology
[5] Alif Bin Abdul Qayyum ,AsifulArefeen,"Convolutional Neural Network (CNN) Based Speech-Emotion Recognition " ,IEEE International Conference on Signal Processing ,2019.
[6] Shreya Narang , Ms. Divya Gupta," Speech Feature Extraction Techniques: A Review," in International Journal of Computer Science and Mobile Computing, Vol.4 Issue.3, March-2015.
[7] Moataz El Ayadi "Survey on speech emotion recognition: Features, classification schemes, and databases," in Pattern Recognition 44 (2011) 572–587
[8] Devi C Akalya "Affective Model Based Speech Emotion Recognition Using Deep Learning Techniques" in Indian Journal of Computer Science October 2020.
[9] Wootael Lim, Daeyoung Jang" Speech Emotion Recognition Using Convolutional and Re-current Networks"in Asia -Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) in 2016.
[10] Apoorv Kumar, Kshitij Kumar's," Speech Emotion Recognition Using CNN" in International Journal of Psychosocial Rehabilitation 2020.
[11] Srinivas Parthasarathy and Ivan Tashev, "Convolutional Neural Network Techniques for Speech Emotion Recognition," in 16thInternational Workshop on Acoustic Signal Enhancement (lWAENC) in 2018.
[12] Haytham M. Fayek , Margaret Lech , Lawrence Cavedonb, "Evaluating deep learning architectures for Speech Emotion Recognition" Neural Networks,Elsevier,Volume 92, August 2017, Pages 60-68.
[13] Anuradha D Thakare, Sheetal Umesh Bhandari, " Artificial Intelligence Applications and Reconfigurable Architectures", John Wiley & Sons, , Pages 1-224, 2023.