Turkish Journal of Computer and Mathematics Education Vol.12 No. 5 (2021), 1378-1388

Research Article

An Enhanced CNN-2D for Audio-Visual Emotion Recognition (AVER) Using ADAM Optimizer

D.N.V.S.L.S. Indira[a], Ch. Suresh Babu[b], Ch. Venkateswara Rao[c], Lakshmi Hari Prasanna[d]

[a]Associate Professor, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, AP, India. E-mail: indiragamini@gmail.com
[b]Associate Professor, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, AP, India. E-mail: sureshdani2004@gmail.com
[c]Assistant Professor, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, AP, India. E-mail: venkatgecit@gmail.com
[d]UG Student, Department of Computer Science and Engineering, Gudlavalleru Engineering College, Gudlavalleru, AP, India. E-mail: lakshmi.hariprasanna@gmail.com

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: The importance of integrating visual components into the speech recognition process for improving robustness has been identified by recent developments in audio visual emotion recognition (AVER). Visual characteristics have a strong potential to boost the accuracy of current techniques for speech recognition and have become increasingly important when modelling speech recognizers. CNN is very good to work with images. An audio file can be converted into image file like a spectrogram with good frequency to extract hidden knowledge. This paper provides a method for emotional expression recognition using Spectrograms and CNN-2D. Spectrograms formed from the signals of speech it's a CNN-2D input. The proposed model, which consists of three layers of CNN and those are convolution layers, pooling layers and fully connected layers extract discriminatory characteristics from the representations of spectrograms and for the seven feelings, performance estimates. This article compares the output with the existing SER using audio files and CNN. The accuracy is improved by 6.5% when CNN-2D is used.

Keywords: Human Emotion Recognition, Audio-Visual Emotion Recognition (AVER), Spectrograms, SER, CNN-2D, ADAM Optimizer.

1. Introduction

Our eyes are the most appropriate place to look. Hearing for our ears. If eyes were hypothetically wiser and quicker than ears, wouldn't it be more useful for our eyes to send sound signals for processing?[1][2]. Human Emotion Recognition is classified into various methods. Among them three are playing very crucial role in the recent days. 1. Facial Emotion Recognition 2. Speech Emotion Recognition and the 3. Audio Visual Emotion Recognition. Now a day's Artificial Intelligence and Neural Networks[3][8] can be implemented using different software like Python, Jupitor, Anakonda etc. These softwares are well suited for Image Analytics in a smooth manner. Here in this paper we implemented CNN[4][7][11] and CNN-2D on RAVDESS dataset to find out the human emotion. When we implemented CNN-2D we converted the whole dataset into image files from wav files. It is a part in Audio-Visual Emotion Recognition.

Basically, convolution neural networks are used for image data classification. The arrays consisting of pixel values are given as input to the convolution neural networks. The operations performed by 1, 2 or 3 dimensional convolution neural networks are the same. The difference is convolution direction rather than the input or filter dimension. In convolution 2D networks, the kernel moves in 2 directions. Conv-1D[5] is mainly used for time series or signal that is audio whereas conv-2D is used for image data analysis.

In this, we give the input features mfcc, mel, tonnetz, contrast in the array form to CNN-1D[7] sequential model. The structure of the model is 4 convolutional layers(with the activation function relu) in which the first two of them have a max_pooling layer with them. Then a flatten layer and the dense layer are stacked. The output layer activation function is softmax. The optimizer is rmsprop, learning rate is 0.00005 and the loss is sparse_categorical_crossentropy. Batch size is 20 and the epochs is 500.

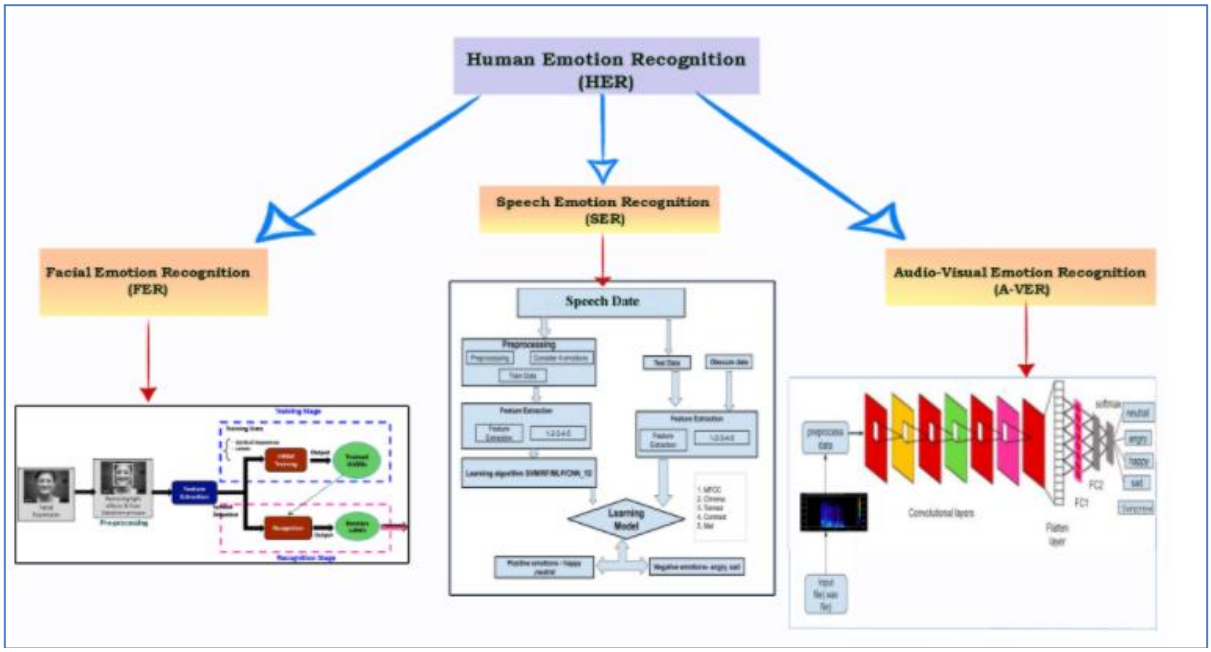

Figure 1. Categories of Human Emotion Recognition

2. Existing Survey

Table 1. Various Methods used in SER from 2010-2020

Reference No.	Year	Methods Used
[1]	2010	Regularized non-negative matrix factorization (NMF) problem with a regularization
[2]	2012	Modeling source separation using MRFs(Markov random fields), MRF inference
[3]	2012	Row mean vector of the spectrograms, Euclidean distance and Manhattan distance,
[4]	2014	Deep CNN
[5]	2014	Hilbert-Huang transform (HHT) and Teager Energy Operator (TEO)
[6]	2015	Single-Pass Spectrogram Inversion (SPSI) algorithm, the magnitude spectra using quadratic interpolation, Griffin-Lim algorithm.
[7]	2017	Spectrograms and Deep Convolutional Neural Network (D-CNN)
[8]	2017	Data Synchronization Management and Automatic Evaluation
[9]	2017	Feature extraction, Image edge detection
[10]	2017	Enhanced kernel isometric mapping (EKIsomap)
[11]	2018	SVM, CNN
[12]	2018	discrete category and Dimensional struc- ture theories discrete category and Dimensional struc- ture theories discrete category and Dimensional struc- ture theories discrete category and Dimensional struc- ture theories Discrete category and dimensional structure theory
[13]	2018	Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), K-nearest neighbor (KNN)
[14]	2019	Firefly Algorithm
[15]	2019	Data normalization and data augmentation techniques
[16]	2020	Attribute cryptosystem and blockchain technology
[17]	2020	Novel fast convolution algorithms
[18]	2020	Batch Normalization, Max Pooling, ReLU Activation Function
[19]	2020	GMM, HMM models
[20]	2020	Support Vector Machine

Here we are expressing literature survey in different model like a table format. We studied various papers and placed the list of algorithms from 2010 to 2020. But we observed the same type of algorithms before decade. We identified that there is a need of work on audio- visual mode. Image analytics also places a vital role in these days.

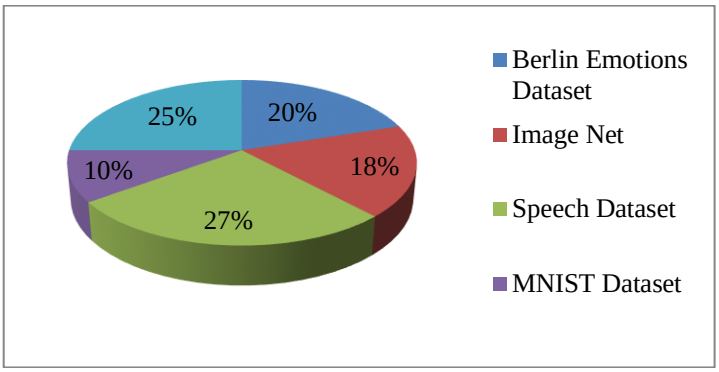

Figure 2. Various Datasets used in Different Articles

3. Proposed Work

3.1. Dataset

Here in this paper we used RAVDESS dataset[18] for training and testing our work which consists of '.wav' files. The RAVDESS audio dataset is converted into spectrogram dataset by using spek tool. The images are in '.png' form and the file name is explained as follows:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

3.2. Methodology

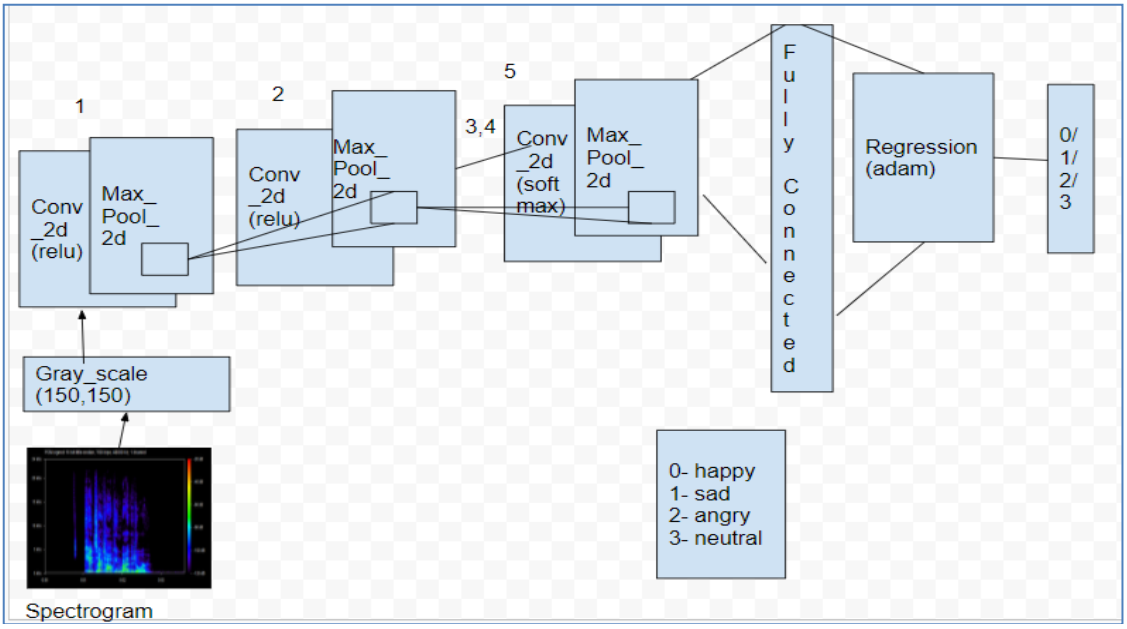

Figure 3. Flow Diagram of CNN-2D Proposed Methodology

Convolution Layer — The convolution layer (CONV)[19][20] uses filters that perform convolution operations with respect to its dimensions while scanning the input picture. Its hyperparameters include the size of the filter that can be 2x2, 3x3, 4x4, 5x5 (but not limited to them alone) and the size of the stride (S). The resulting output (O) is called the feature map or activation map and uses the input layers and filters to compute all the features.

Pooling Layer — For down sampling of the characteristics, the pooling layer (POOL) is used and is usually added after a convolution layer. Max and average pooling are the two types of pooling operations, where the maximum and average value of functions is taken, respectively. Max pooling was used.

Fully Connected Layers — On a flattened input, where each input is connected to all the neurons, the completely connected layer (FC) operates. These are normally used to link the hidden layers to the output layer at the end of the network, which helps to maximize class scores.

Each model has the same structure, i.e., it has a stack of five convolution and pooling layers (as a pair) without any dropout, then a completely connected layer with a drop of 0.8, followed by a regression layer that is fully connected.

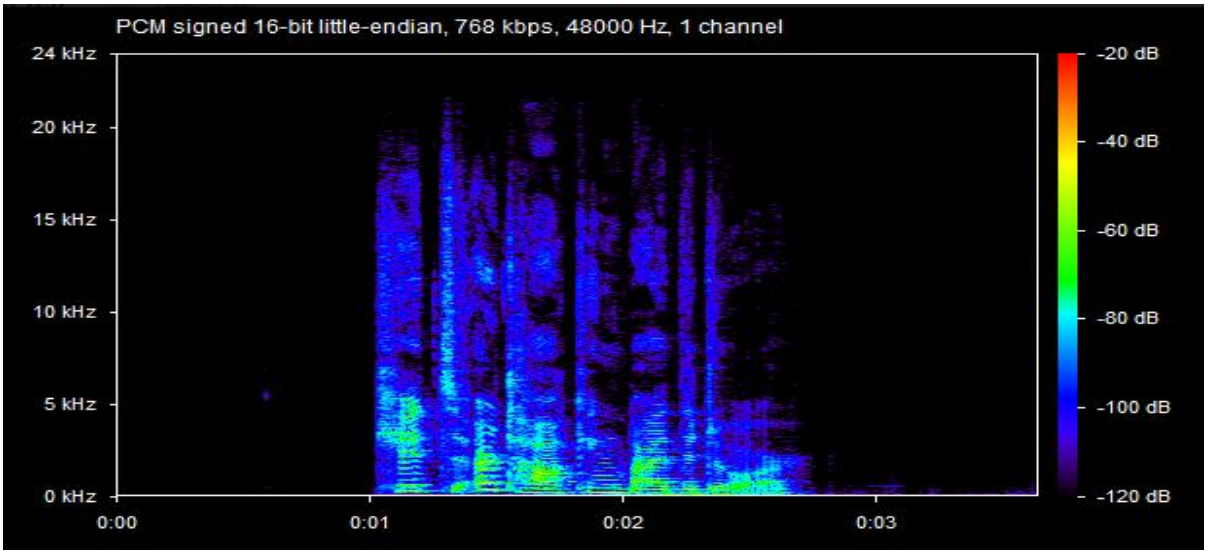

Figure 4. In the above Spectrogram File, the Dialogue is “Kids are Talking by the Door”

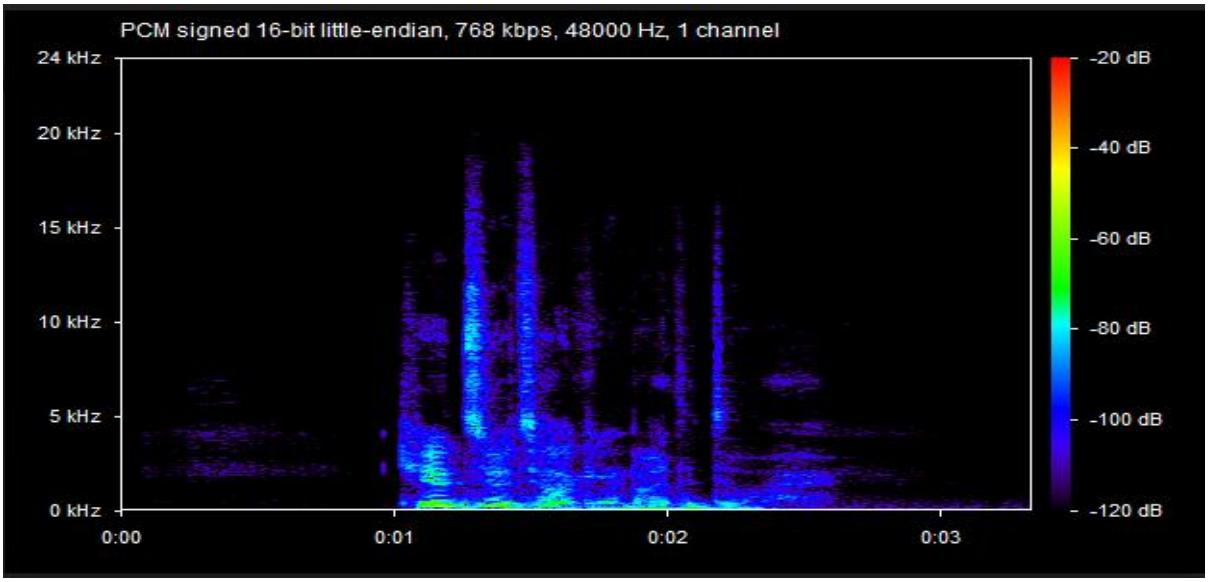

Figure 5. In the above Spectrogram File, the Dialogue is "Dogs are Sitting by the Door"

3.3. CNN-2D

In the CNN-2D network, we give the image dataset values as input. The RAVDESS dataset[18] is converted into spectrogram dataset i.e., all the audio files are converted into spectrogram images[6][9] manually by using the spek tool. The images are converted to gray scale with 150 size. The model has a stack of five convolution and pooling layers(as pair) without any dropout then fully connected layer with drop out of 0.8 followed by fully connected layer with regression. Relu activation function is used for all the layers of CNN and the last layer has softmax activation function. The regression part has adam optimizer, loss is categorical_cross_entropy, learning rate is 0.005 and the number of epochs are 50.

3.4. Optimizers

Here in this paper three optimizers are used to enhance the output. Those are ADAGARD, RMSPROP and ADAM. Among all these ADAM optimizer gave best result when the learning rate is 0.005.

3.4.1. ADAGARD - Adaptive Gradient Algorithm

$$v_t^w = v_{t-1}^w + (\nabla w_t)^2$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t^w + \epsilon}} * \nabla w_t$$

$$v_t^b = v_{t-1}^b + (\nabla b_t)^2$$
$$b_{t+1} = b_t - \frac{\eta}{\sqrt{v_t^b + \epsilon}} * \nabla b_t$$

Formula 1. Rule for AdaGrad

3.4.2. RMS Prop - Root Mean Square Propagation

$$v_t^w = \beta * v_{t-1}^w + (1-\beta)(\nabla w_t)^2$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t^w + \epsilon}} * \nabla w_t$$

$$v_t^b = \beta * v_{t-1}^b + (1-\beta)(\nabla b_t)^2$$
$$b_{t+1} = b_t - \frac{\eta}{\sqrt{v_t^b + \epsilon}} * \nabla b_t$$

Formula 2. Rule for RMSProp

3.4.3. Adam - Adaptive Moment Estimation

$$m_t = \beta_1 * m_{t-1} + (1-\beta_1) * \nabla w_t$$
$$v_t = \beta_2 * v_{t-1} + (1-\beta_2) * (\nabla w_t)^2$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} * \hat{m}_t$$

Formula 3. Rule for Adam

3.5. Proposed Algorithm

1. BEGIN
2. Load spectrograms Dataset.
3. List emotions that have to be considered i.e.emotions(angry, sad, neutral, happy).
4. Function label_image(emotion):
    if emotion == 'happy': return [1, 0,0,0]
    elif emotion == 'sad': return [0, 1,0,0]
    elif emotion=='angry' : return [0,0, 1,0]
    elif emotion=='neutral' : return [0,0,0,1]
5.. for img in file_path:
    i.convert into grayscale
    ii. if emotion in emotions:
    label = label_img(emotion)
    training_data.append([np.array(img), np.array(label)])
6. CNN model cm
i. Training_data is given as input
ii. (N x N) * (F x F) = (N-F+1)x(N-F+1) here F is filter and N x N is image size
iii. Pooling layer
W2 = (W1−F)/S+1
H2 = (H1−F)/S+1
D2 = D1
Where W2, H2 and D2 are the width, height and depth of output.
iv. Fully connected layer
Softmax activation
v. output layer
7.result= cm.predict(imag1)[0]:
    i.if(result=='happy') :
        Print "Your emotion comes under positive category... your emotion is happy"
    elif result=='sad' :
        Print "Your emotion comes under negative category...your emotion is sad"
    elif result=='angry' :
        Print "Your emotion comes under negative category... your emotion is angry"
    else :
        Print "Your emotion comes under positive category.... your emotion is neutral"

    ii. Print "Emotion categories : Positive & Negative"
    iii. Print "Positive emotions : happy, neutral"
    iv. Print "Negative emotions : angry, sad"
8.END

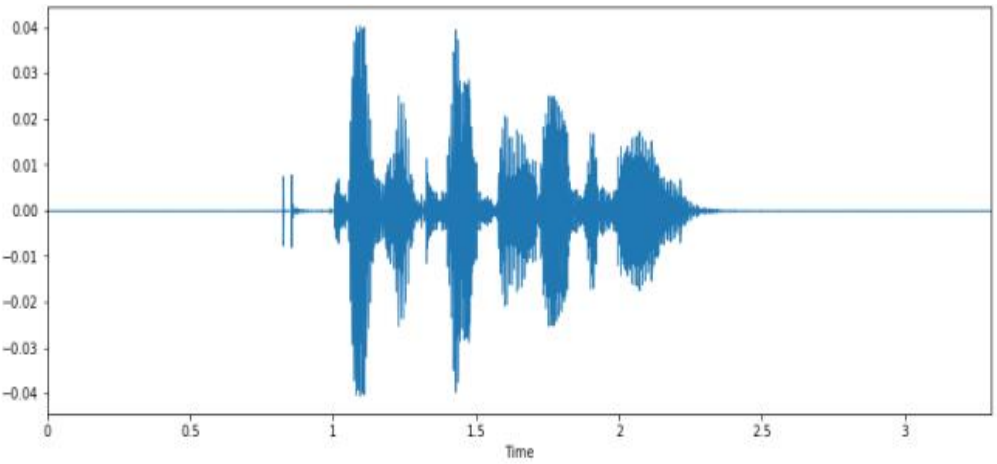

Figure 6. Sample Audio Signal from the Input Dataset

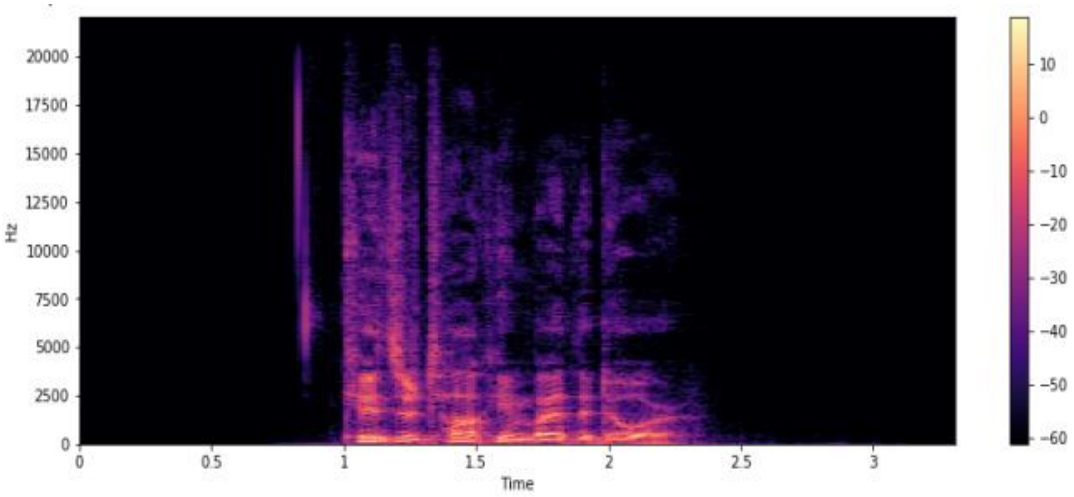

Figure 7. Corresponding Spectrogram for the above Audio File.

4. Results and Discussion

The convolution neural network model is used for the generated image dataset.

i) Image_Size =150, Gray Scale Image, Learning Rate= 0.001

a) Relu activation function [18] is used for all the layers of CNN and the last layer has softmax activation function.
The regression part has ADAGARD optimizer, loss is categorical_cross_entropy,
The accuracy obtained for this model is **70 %.**

The classification report is:

Table 2. a

Precision	Recall	f1-Score	Support
0.61	0.83	0.7	54
0.66	0.57	0.65	61
0.82	0.75	0.78	60
0.65	0.63	0.64	35

Accuracy 0.70 210
Macro Avg 0.70 0.70 0.69 210
Weighted Avg 0.71 0.70 0.70 210

b) Relu activation function, last layer has softmax activation function.
The regression part has RMSPROP optimizer, loss is categorical_cross_entropy,
The accuracy obtained for this model is 74 %.

The classification report is:

Table 2. b

Precision	Recall	f1-score	Support
0.71	0.72	0.72	54
0.79	0.67	0.73	61
0.78	0.93	0.85	60
0.68	0.6	0.64	35

c) Relu activation function,last layer has softmax activation function.
The regression part has ADAM optimizer, loss is categorical_cross_entropy,
The accuracy obtained for this model is **76 %.**

The classification report is:

Table 2. c

Precision	Recall	f1-Score	Support
0.74	0.74	0.74	54
0.75	0.80	0.78	61
0.92	0.78	0.85	60
0.60	0.69	0.64	35

ii) Image_Size =150, Gray Scale Image, Learning Rate= 0.005

d) Relu activation function is used for all the layers of cnn and the last layer has softmax activation function. The regression part has ADAGRAD optimizer, loss is categorical_cross_entropy, The accuracy obtained for this model is **79 %.**

The classification report is

Table 2. d

Precision	Recall	f1-Score	Support
0.77	0.85	0.81	54
0.80	0.84	0.82	61
0.87	0.80	0.83	60
0.71	0.63	0.67	35

e) Relu activation function, last layer has softmax activation function. The regression part has RMSPROP optimizer, loss is categorical_cross_entropy, The accuracy obtained for this model is **82 %.**

The classification report is:

Table 2. e

Precision	Recall	f1-Score	Support
0.84	0.78	0.81	54
0.85	0.85	0.85	61
0.84	0.87	0.85	60
0.7	0.74	0.72	35

f) Relu activation function, last layer has softmax activation function. The regression part has ADAM optimizer, loss is categorical_cross_entropy, The accuracy obtained for this model is **89 %.**

The classification report is:

Table 2. f

Precision	Recall	f1-Score	Support
0.89	0.83	0.86	54
0.90	0.90	0.90	61
0.89	0.93	0.90	60
0.75	0.79	0.77	35

Table 2. a, b, c, d, e, f Describes the Outputs from RMSPROP, ADAGARD and ADAM Optimizer Where Learning Rate is 0.001 and 0.005

Table 3. Distinguishes from Existing Algorithms vs Proposed CNN-2D

Models	Parameters	Dataset	No. of Training & Testing Samples	Report			
					Precision	recall	F1-score
Multi Layer Perceptron	alpha : 0.01, batch_size:256, h_l:300 epsilon:1e-0.8	RAVDESS	985 & 325	accuracy	-	-	0.79
				Macro avg	0.73	0.79	0.80
				weighted avg	0.80	0.79	0.78
CNN-1D	b_s:20,epoch:500, Optimizer: RMSPROP, 5e-6,d_o=.1		985 & 325	accuracy	-	-	0.83
				macro avg		0.82	0.82
				weighted avg	0.83	0.83	0.83
CNN-2D	Image-size : 150 Learning rate :0.005 Optimizer : ADAM		985 & 325	accuracy	-	-	0.89
				macro avg		0.88	0.86
				weighted avg	0	0.89	0.89

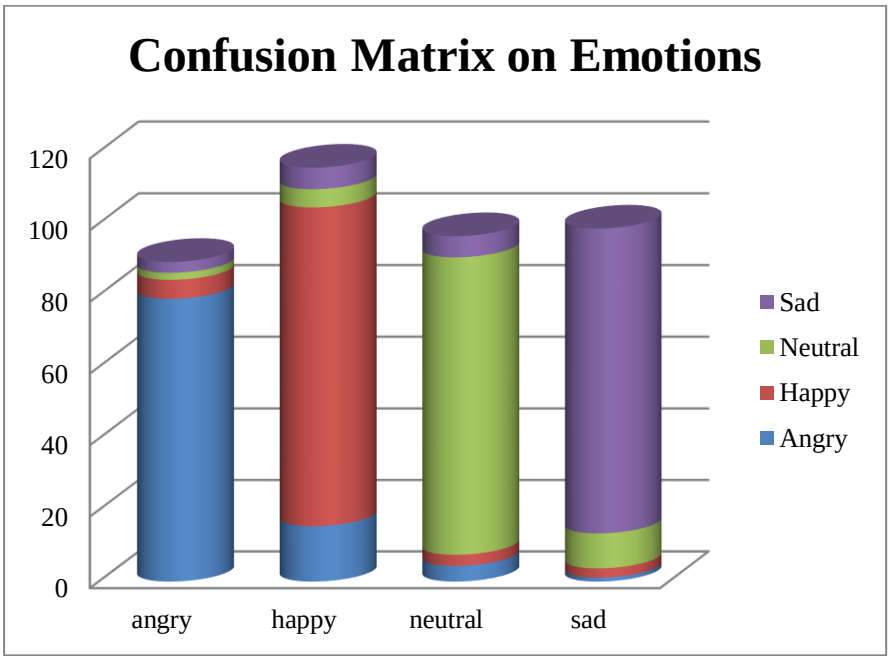

Figure 8. CNN-2D Confusion Matrix

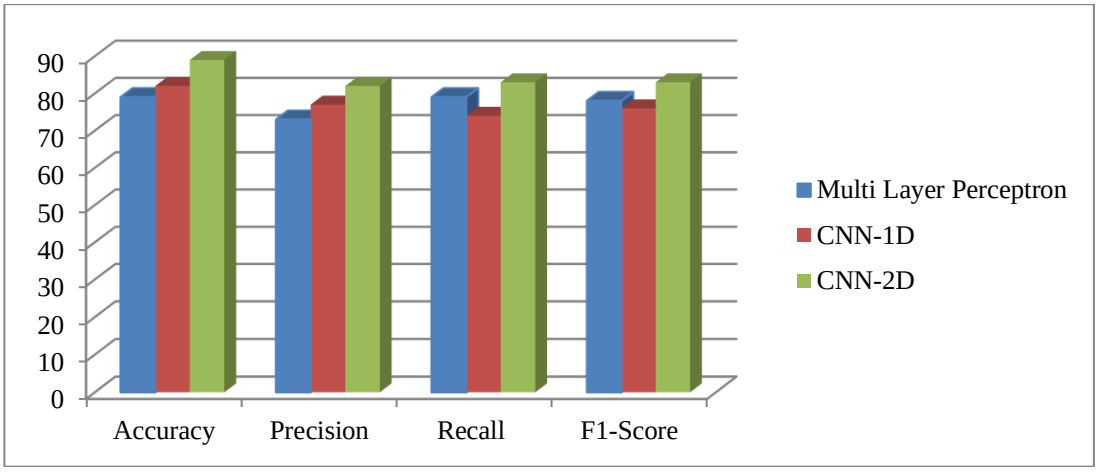

Figure 9. Comparison of Proposed CNN-2D with CNN-1D and MLP for Detecting Emotions

5. Conclusion and Future Work

The optimizers adam, rmsprop and adagrad are used in the whole procedure of designing the model. Relu activation function is used for all the layers and softmax function is opted for the last layer. The loss is categorical cross entropy. By considering all the optimizers and tuning the learning rate as 0.001 and 0.005, we observed that Adam optimizer with 0.005 learning rate model outperforms all the remaining models by achieving 89% accuracy. The next model with more accuracy is model with learning rate 0.005 and rmsprop optimizer i.e., 82%. Remaining models with 0.001 learning rate gave 70% and 74% and 76% for adagrad, rmsprop and adam respectively.

This paper's contribution is three-fold. First, we suggested a novel speech amplification model based on CNN-2D, which transforms audio files into spectrograms. The spectrograms are then used with more accurate phase information to synthesize enhanced speech waveforms. Second, we derive an optimizer that has taken into account various metrics in the objective function. We will investigate the integration of video and voice input data as a spectrogram in the future. We also have to explore this application in various fields.

References

1. Wilson, K.W., & Raj, B. (2010). Spectrogram dimensionality reduction with independence constraints. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, 1938-1941.
2. Kim, M., Smaragdis, P., Ko, G.G., & Rutenbar, R.A. (2012). Stereophonic spectrogram segmentation using markov random fields. *In IEEE International Workshop on Machine Learning for Signal Processing*, 1-6.
3. Kekre, H.B., Kulkarni, V., Gaikar, P., & Gupta, N. (2012). Speaker identification using spectrograms of varying frame sizes. *International Journal of Computer Applications*, *50*(20).
4. Kang, L., Kumar, J., Ye, P., Li, Y., & Doermann, D. (2014). Convolutional neural networks for document image classification. *In 22nd International Conference on Pattern Recognition*, 3168-3172.
5. Lalitha, S., Patnaik, S., Arvind, T.H., Madhusudhan, V., & Tripathi, S. (2014). Emotion Recognition through Speech Signal for Human-Computer Interaction. *In Fifth International Symposium on Electronic System Design*, 217-218.
6. Beauregard, G.T., Harish, M., & Wyse, L. (2015). Single pass spectrogram inversion. *In IEEE international conference on digital signal processing (DSP)*, 427-431.
7. Badshah, A.M., Ahmad, J., Rahim, N., & Baik, S.W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. *In international conference on platform technology and service (PlatCon)*, 1-5.
8. Karol, M., & Michal, K. (2017). Medical data management. *In IEEE 14th International Scientific Conference on Informatics.*
9. Albawi, S., Mohammed, T.A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *In International Conference on Engineering and Technology (ICET)*, 1-6.
10. Zhang, S., Zhao, X., & Lei, B. (2013). Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. *International Journal of Advanced Robotic Systems*, *10*(2), 114.

11. Sultana, F., Sufian, A., & Dutta, P. (2018). Advancements in image classification using convolutional neural network. *In Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 122-129.
12. Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93-120.
13. Gulnaz Nazir Peerzade, Ratnadeep R. Deshmukh, S.D Waghmare. (2018). Speech Emotion Recognition, *International Journal of Computer Science and Engineering*, *6*(3).
14. Strumberger, I., Tuba, E., Bacanin, N., Zivkovic, M., Beko, M., & Tuba, M. (2019). Designing convolutional neural network architecture by the firefly algorithm. *In International Young Engineers Forum (YEF-ECE)*, 59-65.
15. Tariq, Z., Shah, S.K., & Lee, Y. (2019). Speech emotion detection using iot based deep learning for health care. *In IEEE International Conference on Big Data (Big Data)*, 4191-4196.
16. Yang, X., Li, T., Pei, X., Wen, L., & Wang, C. (2020). Medical data sharing scheme based on attribute cryptosystem and blockchain technology. *IEEE Access*, *8*, 45468-45476.
17. Cheng, C., & Parhi, K.K. (2020). Fast 2D convolution algorithms for convolutional neural networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *67*(5), 1678-1691.
18. Darshan K.A, Dr. B.N. Veerappa. (2020). Speech Emotion Recognition, *International Research Journal of Engineering and Technology* (IRJET), *7*(9).
19. Dipankar Dutta, Ridip dev Choudhury, Swapnanil Gogoi. (2020). Speech Databases, Features Extraction Techniques and Classifiers with Special Reference to Automatic Speech Emotion Recognition, *International Journal of Scientific & Technology Research*, *9*(2).
20. Nagaraju Naik M. Dr, Merugu Suresh Dr. (2020). Speech Based Emotion Recognition Through Cepstral Features and Support Vector Machine Algorithm, *International Journal of Advanced Science and Technology*, *29*(6).