WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Analyzer

Varshini P[1], Soundarya R[1], Assistant Prof.Merin Meelet[2], Professor Dr. Anala M R[3], Assistant Prof.Smitha G.R.[2]

[1]UG Student, Information Science Engineering, RV College of Engineering, India
[2] Assistant Prof, Information Science Engineering, RV College of Engineering, India
[3] Professor, Information Science Engineering, RV College of Engineering, India

ABSTRACT

Perceiving the feeling from speech has become one the of the active research topics in speech processing and in applications dependent on human-PC collaboration. This paper gives the implementation of the model using Convolutional Neural Networks (CNN). The architecture was adapted using image preprocessing CNN. The hypothetical foundation that lays the establishment of the grouping of feelings dependent on voice boundaries is momentarily introduced and distinguishability of emotional features in speech were first studied and later followed by emotion classification on a custom dataset was made.This paper contributes towards the adaptation of deep learning model for processing the audio files,applying data augmentation techniques and training the files using CNN model which predicts the gender and emotion of the speaker. This paper also has the comparison between the performance of a plain CNN model and the performance of a CNN model with greater dataset and augmentation methods applied.

Keywords - Speech emotion recognition,Convolutional Neural Networks,Speech Processing, Mel-Frequency Cepstral Coefficients

I.INTRODUCTION

Speech is one of the most characteristic approaches to communicate our thoughts. We depend so much on it because we perceive its significance when turning to other correspondence structures like messages and instant messages where we regularly use emotions to communicate the feelings related with the messages. As feelings assume a crucial part in correspondence, the recognition and investigation of the equivalent is of fundamental significance in the present advanced world. Feeling identification is a difficult assignment, since feelings are abstract. As feelings play an imperative part in communication, the discovery and examination of the same is of significance in today's computerized world of communication. Feeling discovery could be challenging, since feelings are subjective.In Order to communicate viably with people,the systems need to get the feelings in the speech. Hence, there is a need to create machines that can recognize the paralinguistic data like feeling to have effective clear communication like people[1]. One imperative information in paralinguistic data is Feeling. A parcel of machine learning calculations have been developed and tried in order to classify these emotions carried present in the speech. The objective of this work is to characterize a SER framework as an assortment of strategies that interact to recognize feelings inserted in them and also to predict the gender of the speaker. Such a framework can discover use in a wide assortment of regions like intelligent voice based-collaborator or guest specialist discussion investigation. In this examination we endeavor to identify basic feelings in recorded speech by investigating the acoustic highlights of the sound information of accounts[2].
Convolution Neural Systems are used to identify the emotions in speech. The model uses RAVDEES[16] , SAVEE[17] and TESS[18] dataset which contains speech samples.

II.RELATED WORKS

Over the last years, an investigation has been made to recognize emotions by using speech statistics. Here well-known papers are studied to understand the previous work and how they operate.
In [3] the author has implemented a DL model of CNN to determine the emotion of speech signal.The architecture which was designed consisted of the adaptation which includes image processing CNN, which was programmed in Python using Keras library and TensorFlow was used as backend. The main drawback of this paper was that it was necessary to have a system which had more reliability and real time emotion recognition could be developed using the same architecture.
A curious strategy was proposed in [4], the creators depict the effect of visual methodology in development to discourse and substance for moving forward the exactness of the feeling discovery framework. In this regard, a neural arrangement was associated to get covered up representations of the

Corresponding Author E mail :sonur15400@gmail.com

modalities,tests were performed on the standard IEMOCAP dataset utilizing all three modalities (sound, substance, and video). The achievement showed up a noteworthy advancement of 3.65% in terms of weighted exactness compared to the standard framework. The biggest hole of this show was that it got confused between the energized and upbeat lesson since there exists a report of cover in recognizing these two classes indeed within the human appraisals.It presents a model in which it could recognize emotion in a discourse sample. The paper points to developing a machine to interpret paralinguistic data, like emotion. Convolution Neural Networks were utilized to predict the feelings in speech tests. To train the network RAVDEES and SAVEE dataset are used which contain speech tests that were created by the 24 on-screen characters and 4 actresses respectively. The demonstrated model showed the different emotions according to the different inputs and showed accuracy of 77%[6].

Another method proposed in [5] includes a multimodal speech emotion recognition and ambiguity resolution model which was a multi-class classification problem were execution of two categories of models were compared.Within the to begin with approach, the extricated highlights are utilized to get ready six conventional machine learning classifiers, while the moment approach is based on profound learning wherein a standard feed-forward neural organize and an LSTM-based classifier were arranged over the same highlights. He made a conclusion expressing that the lighter machine learning based models which were arranged over some hand-crafted highlights were able to realize execution when compared to the current significant learning based state of-the-art strategy for emotion recognition.
In the paper [9] there's an interaction-aware consideration organization (IAAN) which consolidates relevant data within the learned vocal representation through a novel consideration component. The proposed strategy accomplishes 66.3% precision (7.9% over standard strategies) in four lesson feeling acknowledgment and is additionally the current state-of-art acknowledgment rate obtained on the benchmark database.

In [10] paper the creator made acknowledgment utilizing sound and content and outlined a novel significant double repetitive encoder show which utilizes content information and sound signals at the same time to get distant better;a much better;a higher;a stronger;an improved" a distant better understanding of discourse data.In differentiate to these approaches, the issue has been handled by presenting an thought instrument to combine the data. In this respect, a neural organize was associated to get covered up representations of the modalities. The proposed demonstrate beats past state-of-the-art strategies in allotting information to one of four feeling categories (i.e., perturbed, cheerful, pitiful and impartial) when the illustration is connected to the IEMOCAP dataset, as reflected by exactnesses extending from 68.8% to 71.8%. The drawback of this model was that text and alignment information which were required for the Cross Attention Network(CAN) did not work properly and plans for research by integrating the CAN with the automatic speech recognition system which yields the text and alignment information in a given speech signal could have been more suitable.

The strategy proposed in [2] included Neural Organize as a classifier to classify the particular enthusiastic states such as cheerful, pitiful, shock etc from enthusiastic discourse databases. For the execution of classification they utilized the discourse which incorporate Mel Recurrence cepstrum coefficient (MFCC). Further increment in accuracy may be brought by including filters before the feature extraction process and further enhancements could be brought to the incorporation of speaker and gender based emotion recognition.

III.METHODOLOGY

The followed methodology can be discussed in the following points-

A. Dataset

a. Ryerson Audio-Visual Database of Passionate Speech and Tune (RAVDESS)

The Ryerson Audio-Visual Database of Enthusiastic Speech and Songs contains 24 proficient performing artists (12 females, 12 male), articulating two lexically-comparable sentences in a fair-minded North American supplement. Discourse tests consolidate sentiments like quiet, cheerful, dismal, irate, shocking, shock, and repugnance articulations, Each feeling is made at two phases of enthusiastic force (ordinary, strong), with an extra impartial feeling.

b. Study Audio-Visual Communicated Feeling (SAVEE)

The sound organizer of this dataset comprises speech samples recorded by four male speakers. For each emotion class there are 15 sentences. The beginning letter of the record name represents an emotion class. The letters 'su', 'sa', 'n', 'h', 'f', 'd' and 'a' represent 'surprise' 'sadness' 'neutral', 'happiness', 'fear', 'disgust', and 'anger' feeling classes individually.

c. Toronto emotional speech set (TESS)

A set of 200 target words were talked within the carrier express "Say the word _' by two performing artists and recordings were made of the set depicting each of seven feelings (outrage, appall, fear, joy, wonderful shock, pity, and impartial). There are 2800 files in total.

B. Data Preprocessing

The initial step includes arranging the sound records. The feeling in a sound example can be dictated by the exceptional identifier of the record name at

the third position, which addresses the kind of feeling. The dataset comprises seven distinct feelings. 1. Calm 2. Happy 3. Sad 4. Angry 5. Fearful 6. Disgust 7. Neutral.

a. Defining Labels

In light of the quantity of classes to arrange the discourse names are characterized. A portion of the classes are as per the fig 1.

```
female_neutral     676
female_sad         584
female_angry       584
female_fear        584
female_happy       584
female_surprise    496
female_disgust     496
male_neutral       408
male_happy         252
male_angry         252
male_fear          252
male_sad           252
male_surprise      156
male_disgust       156
```

Fig 1 - Examples of the labels

b. Data Augmentation

The datasets utilized have two fundamental disadvantages: class imbalance and little size. To adapt to both obstacles several data augmentation methods were followed.

a. Adding White Noise

Each time a preparing test is uncovered to demonstrate, arbitrary clamor is included to the input factors making them diverse each time it is uncovered to the show.

b. Random Shifting

Moving time is exceptionally basic. It simply moves sound to left/directly with an irregular second. On the off chance that moving sound to the left (quick forward) with x seconds, first x seconds will stamp as 0 (for example quietness). On the off chance that moving sound to right (back forward) with x seconds, last x seconds will stamp as 0 (for example quiet).

c. Stretching

The preparation of changing the speed/duration of sound without influencing the pitch of sound.

d. Time Shifting

Here, the wave by sample_rate/10 is calculated. This will move the wave to the correct by a given calculation along the time axis.

e. Pitch Shifting

It could be a handle of changing the pitch of sound without influencing its speed.

Fig. 2 shows the waveplot of an audio before data augmentation is applied and Fig. 3 shows the waveplot of the same audio after data augmentation is applied.

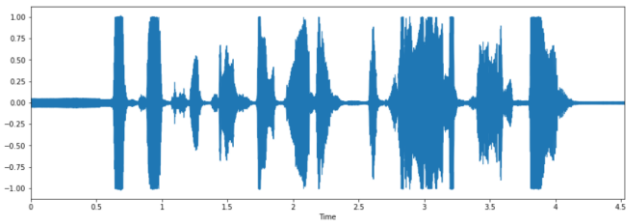


Fig 2 - Waveplot before data augmentation

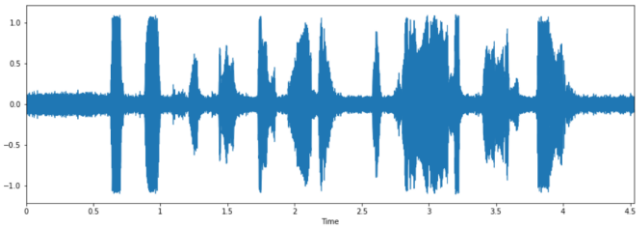


Fig 3 - Waveplot after data augmentation

Fig. 4 shows the flowchart of speech emotion analyzer model.

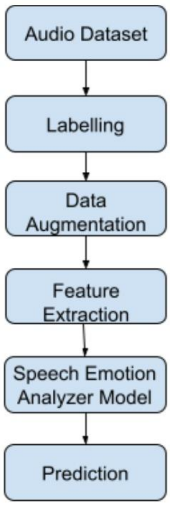


Fig 4 - Flowchart of speech emotion analyzer

C. Feature Extraction

A sound is composed of a wide range of features like frequency, pitch, tone and feature extraction is the main advance to investigate it. The wav files are first changed over into an array containing the samples of amplitude and the sample rate. This is then used to identify the acoustic features. In this work, we have utilized the Mel Frequency Cepstral coefficient (MFCC).

MFCC represents elements of human speech. MFCC depicts the logarithmic insight of loudness and pitch of the human auditory system. The Mel-scale is used to fit the frequency perceived through human ears with the real frequency. The MFCC is calculated through splitting the audio into multiple frames. Then Fourier transform and power spectrum are calculated for every frame and associated with the Mel-Scale. The discrete cosine transform (DCT) is calculated at the Mel log energies and the coefficients are estimated.

D. Model Architecture

In this paper, we aim to show the impact of expanding the data available and also show the impact data augmentation methods had on the performance of the model.

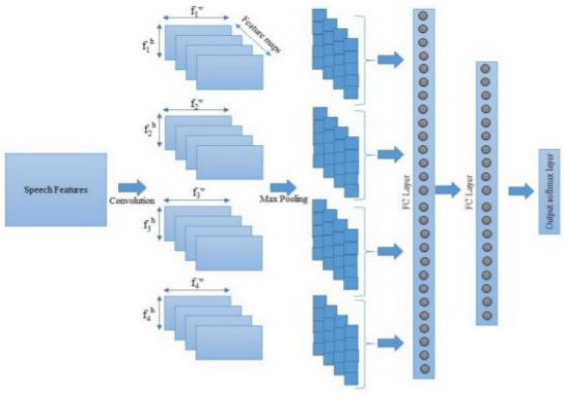


Fig 5 - MFCC based CNN architecture[19]

The proposed CNN model comprises 4 convolutional layers and each convolutional layer is taken after by batch normalization, activation, dropout and max pooling layer. To avoid neural systems from overfitting, dropout layers are utilized. For activation function Rectified Linear Units (ReLU) is used for the initial layers and softmax is used for the final layer. The CNN show rundown is given in Fig. 5. "Adam" is the optimizer utilized with a learning rate of 0.001 amid training.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 30, 216, 1)]	0
conv2d_4 (Conv2D)	(None, 30, 216, 32)	1312
batch_normalization_5 (Batch	(None, 30, 216, 32)	128
activation_5 (Activation)	(None, 30, 216, 32)	0
max_pooling2d_4 (MaxPooling2	(None, 15, 108, 32)	0
dropout_6 (Dropout)	(None, 15, 108, 32)	0
conv2d_5 (Conv2D)	(None, 15, 108, 32)	40992
batch_normalization_6 (Batch	(None, 15, 108, 32)	128
activation_6 (Activation)	(None, 15, 108, 32)	0
max_pooling2d_5 (MaxPooling2	(None, 7, 54, 32)	0
dropout_7 (Dropout)	(None, 7, 54, 32)	0
conv2d_6 (Conv2D)	(None, 7, 54, 32)	40992
batch_normalization_7 (Batch	(None, 7, 54, 32)	128
activation_7 (Activation)	(None, 7, 54, 32)	0
max_pooling2d_6 (MaxPooling2	(None, 3, 27, 32)	0
dropout_8 (Dropout)	(None, 3, 27, 32)	0
conv2d_7 (Conv2D)	(None, 3, 27, 32)	40992
batch_normalization_8 (Batch	(None, 3, 27, 32)	128
activation_8 (Activation)	(None, 3, 27, 32)	0
max_pooling2d_7 (MaxPooling2	(None, 1, 13, 32)	0
dropout_9 (Dropout)	(None, 1, 13, 32)	0
flatten_1 (Flatten)	(None, 416)	0
dense_2 (Dense)	(None, 64)	26688
dropout_10 (Dropout)	(None, 64)	0
batch_normalization_9 (Batch	(None, 64)	256
activation_9 (Activation)	(None, 64)	0
dropout_11 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 14)	910

Total params: 152,654
Trainable params: 152,270
Non-trainable params: 384

Fig 6 - Summary of the CNN model

Fig 6 shows the summary of the CNN model built.

IV RESULTS AND DISCUSSIONS

The aim of this work is to show the impact of expanding the data available and also show the impact data augmentation methods had on the performance of the model.

In experiment 1, only RAVDESS and SAVEE datasets were used and data augmentation was not performed on these audio samples. The same CNN model was used and the model was trained for 50 epochs and it had a training accuracy of 86.67% and a validation accuracy of 56.67% which is quite less. Fig 7 shows the model accuracy and loss graph and fig 8 shows the confusion matrix of experiment 1.

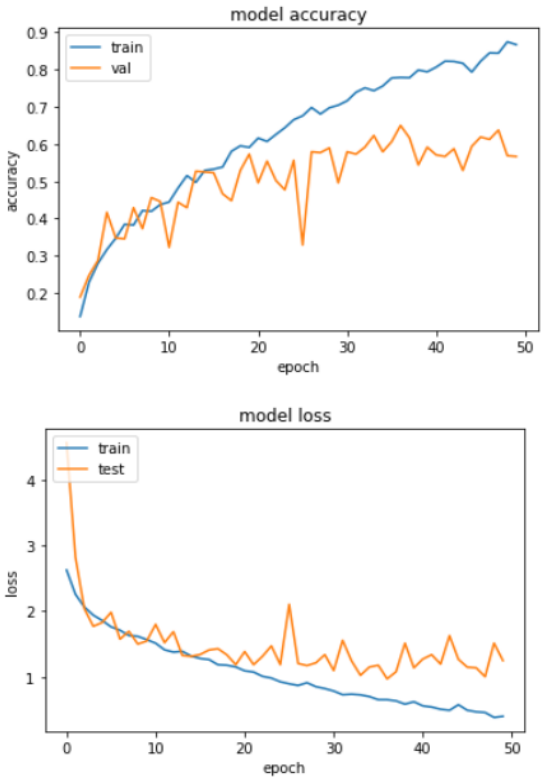


Fig 7 - Model accuracy and loss graph of experiment 1

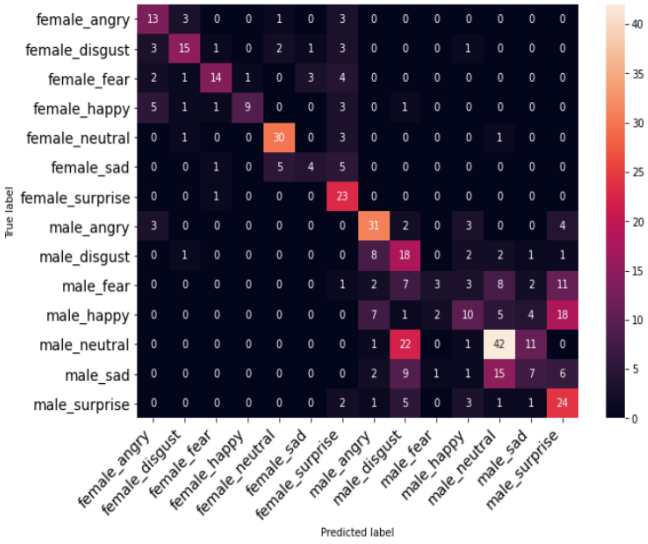


Fig 8 - Confusion Matrix of experiment 1

In experiment 2, along with RAVDESS and SAVEE, an additional dataset TESS was also used. This gave an additional 2800 audio files to train and along with this, data augmentation methods like adding white noise, random shifting, stretching,time shifting and pitch shifting were also applied. The same CNN model was used and the model was trained for 50 epochs. It was observed that the training accuracy was nearly 100 and the validation accuracy was 87.43% which is a significant improvement when compared to experiment 1. Fig 9 shows the model accuracy and loss graph and fig 10 shows the confusion matrix for experiment 2.

For training, 70% of the dataset is utilized and for testing 30% of dataset is utilized.

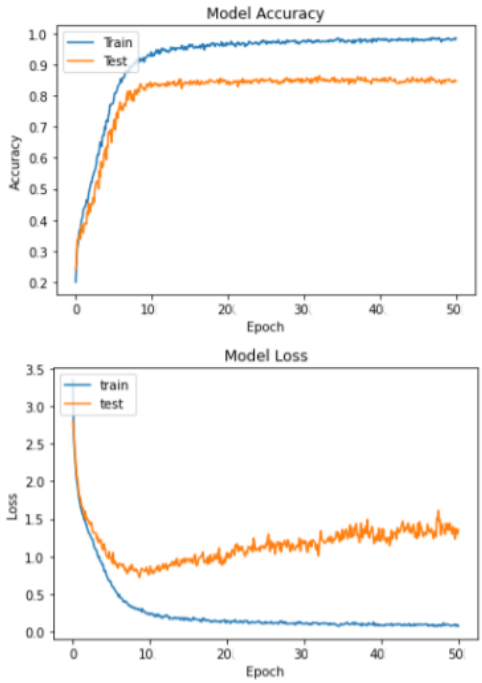


Fig 9 - Model accuracy and loss graph for experiment 2

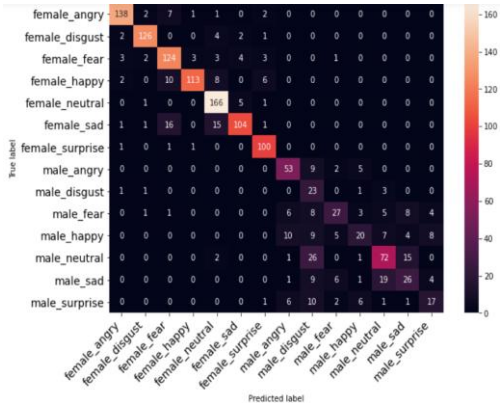


Fig 10 - Confusion Matrix for experiment 2

Table II shows the Actual values vs Predicted values of experiment 2 for a few audio files.

TABLE II: Actual Values vs Predicted Values

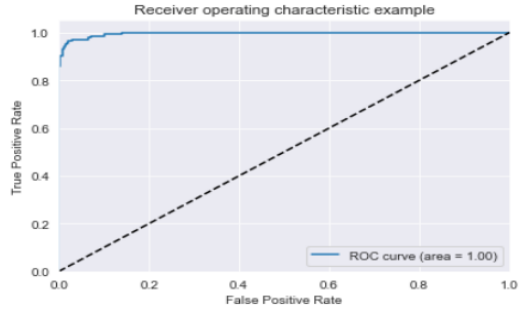
Actual Values	Predicted Values
male_fearful	male_happy
male_fearful	male_fearful
male_fearful	male_fearful
male_sad	male_sad
male_fearful	male_fearful
male_happy	male_happy
female_angry	female_angry
female_angry	female_fearful
male_angry	male_angry

Table III shows the performance of the experiment 2 model when different optimizers were used.

TABLE III: Performance of the model for different optimizers

Optimizer/Learning Rate	Loss	Accuracy
Adam	48.16	87.43
RMSProp	72.21	74.64

Fig 11. Shows the ROC curve of experiment 2 for male fear.



male fear

Fig 11: ROC curve for male fear

A. Real Time Analysis

The user records the sound utilizing a pyaudio worked in the library and the chronicle takes for 4 secs. Later the sound record gets downloaded. Then, at this point choosing the recorded speech takes place to test and to discover the emotion in it. The model concentrates the MFCC highlights from the example and predicts the feeling in the class according to the pre-characterized feeling class as shown in the figure 12 and 13.

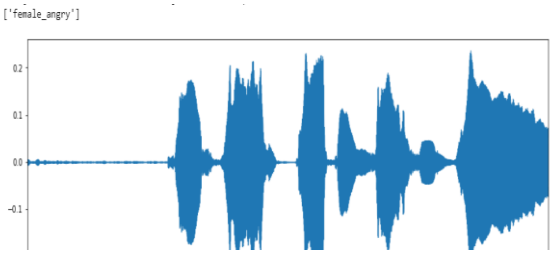


Fig 12 - Real time analysis of the prediction

['male_happy']

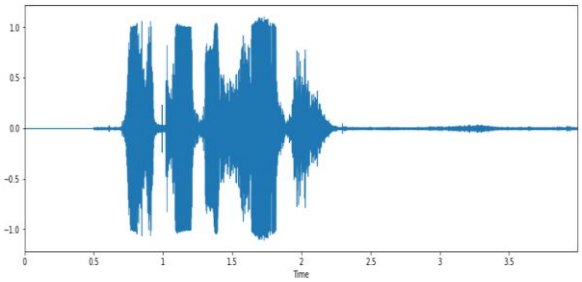


Fig 13 - Real time analysis of the prediction

V CONCLUSION

Two experiments were performed on the same CNN model by varying the amount of training data used and augmentation methods applied. It is clearly observed that the experiment 2 was the superior one among the two experiments as a high validation accuracy 87.43% was achieved and the overfitting of the model also seems to be quite low. The performance of the CNN model in experiment 2 was found to be much better. From this observation, it can be concluded that by adding more data and applying many such augmentation methods, it is possible to increase the accuracy of such speech emotion recognition models significantly and this can also be taken up as a future scope for this work.

REFERENCE

[1] Panagiotis Tzirakis, AnhNguyen, Stefanos Zafeiriou, Bjorn W. Schuller "Speech emotion recognition using semantic information" arXiv:2103.02993 (4 March 2021).

[2] Abdul Ajij Ansari, Ayush "Speech emotion recognition using CNN" International Journal of Psychosocial Rehabilitation (2020).

[3] Darshan K.A, Dr. B.N. Veerappa "Speech emotion recognition" International Research Journal of Engineering and Technology (IRJET)Sep 2020.

[4] Panagiotis Tzirakis, AnhNguyen, Stefanos Zafeiriou, Bjorn W. Schuller "Speech emotion recognition using semantic information" arXiv:2103.02993 (4 March 2021)..

[5] Seunghyun Yoon , Subhadeep Dey , Hwanhee Lee and Kyomin Jung, "Attentive Modality Hopping Mechanism For Speech Emotion Recognition" IEEE Signal Processing Society Resource Center (2020).

[6] Gaurav Sahu "Multimodal Speech Emotion Recognition And Ambiguity Resolution" arXiv:1904.06022 12 Apr 2019.

[7] Darshan K.A, Dr. B.N. Veerappa "Speech Emotion Recognition" International Research Journal of Engineering and Technology (IRJET)Sep 2020.

[8] Alif Bin Abdul Qayyum, Asiful Arefeen*, Celia Shahnaz."Convolutional Neural Network (CNN) Based Speech-Emotion Recognition"

[9] Panagiotis Tzirakis, AnhNguyen, Stefanos Zafeiriou, Bjorn W. Schuller "Speech emotion recognition using semantic information" arXiv:2103.02993 (4 March 2021).

[10] Seunghyun Yoon , Subhadeep Dey , Hwanhee Lee and Kyomin Jung, "Attentive Modality Hopping Mechanism For Speech Emotion Recognition" IEEE Signal Processing Society Resource Center (2020).

[11] Gaurav Sahu, "Multimodal Speech Emotion Recognition And Ambiguity Resolution" arXiv:1904.06022 12 Apr 2019.

[12] Agarap, Abien Fred, "Deep learning using rectified linear units (relu)" March 2015

[13] Sethu, Vidya Saharan and Epps, Julien and Ambikairajah, Eliathamby,"Speech Based Emotion Recognition," pp. 197-228, September 2015.

[14] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "iVectors for Continuous Emotion Recognition," Training, vol. 45, p. 50.

[15] S. Ioffe, "Probabilistic linear discriminant analysis," in Computer Vision–ECCV 2006. Springer, 2006, pp. 531–542.NCE

[16] Sam Roweis, "Em algorithms for pca and spca," Advances in neural information processing systems, pp. 626–632, 1998.

[17] Steven R Livigstone, RAVDESS emotional speech audio, Kaggle (2017).

[18] Tarun Sunkaraneni, SAVEE database, Kaggle (2018).

[19] Eu Jin Lok, Toronto emotional speech set, Kaggle (2018).

[20] Ian Jolliffe, Principal component analysis, Wiley Online Library, 2002.