Convolution neural network-based Speech Emotion Recognition

[1]Shalini Singhal, [2]Meenakshi Nawal, [3]Vipin Jain, [4]Anurish Gangrade

[1]Assistant Professor, [2]Associate Professor, [3]Associate Professor, [4]Web Developer
[1,3]Department. of IT, [2]Department. of CSE, [4]Web Developer
[1,2,3]Swami Keshvanand Institute of Technology, Management & Gramothan Rajasthan Department of IT, Jaipur-302017, India
[4]Green Arrow Consultancy, Cardiff, Wales,United Kingdom

***Abstract*- Automatic speech emotion recognition has grown in popularity because it allows for natural human-computer connection. One way to recognize emotion is voice. Speech, however, also includes silence that cannot be related to emotion. The elimination of silence and/or ignoring silence while paying greater attention to the segment of speech is two ways to improve performance. This Paper propose a combination of silence elimination and a care model in this paper to enhance the performance of speech emotion. An improved CNN model is presented here which consists of combination of convolution 1d layers and generalized to form a 9 layer architecture of CNN (convolutional neural network), model accuracy has been checked with respect to emotion classes such as considering 5 emotions considered as angry, calm, fearful, happy, sad for male as well as female, likewise included use of classes such as positive, negative, neutral to achieve optimum accuracy. The results show that silence cancellation and attention model combinations are better than just the noise cancellation model or just the attention model. In the realm of human-computer interaction, speech emotion recognition is a critical and difficult job. Various models and feature sets for training the system have been proposed in previous work. Using input signals of various lengths, a novel speech-emotion detection system based on Convolutional Neural Networks (CNN) is presented in this research. With the use of a powerful GPU, a model is created and fed with unprocessed speech from a specified dataset for training, classification, and testing purposes. Finally, it achieves a convincing accuracy of 89.00%, which is far higher than any other comparable job on this dataset. This work will have an impact on the creation of social and conversational robots that can convey all the subtleties of human emotion.In terms of accuracy of the model the results are comparatively improved as compared to previous models using same dataset. This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

***Index Terms*- Convolution neural network, MFCC, speech emotion recognition**

I. INTRODUCTION

The synthesis of audio-visual emotional emotions has drawn the interest of human-computer interaction researchers for many years. Despite the significant advancements in human-computer interaction, the majority of the current user interfaces are unidirectional and hostile in that they prevent the computer from comprehending the user's emotional state [1-3].Designing interfaces that can express and pick up on emotions has been a challenge for many academics in recent years.Due to this, a new branch of study called Speech Emotion Recognition has emerged, with the goal of determining a speaker's emotional state from their speech.

Due to the large number of social media users, low cost, and quick bandwidth of the Internet, the SER have encountered various difficulties and limitations in this research period.

Researchers have worked to propose new techniques to extract the most salient elements from speech signals and trained models to precisely identify the speaker's emotion during speech in order to close the semantic gap in this field.The technology is improving every day, giving researchers fresh, adaptable platforms on which to test out novel AI-based methodologies.

The improvement of human-computer interaction (HCI), such as emotion identification, depends heavily on the growth of skills, technology, and employment of AI and deep learning techniques.

It may be used in contact centers to gauge customer satisfaction, in human-computer interaction to gauge human emotion, in emergency call centers to gauge a user's emotional state to determine the best course of action, and in virtual reality, among many other real-time uses.

The results of this study will be useful to participants in the market who are interested in creating end-to-end models for SER or enhancing the performance of current models. The creation of a model for video tagging or metadata production might be one application case.
The outcomes can be utilized as either a standalone model or as a component of a complete model to decide how to mix the audio components in order to utilize all of the data present in the video.
Improving crucial systems, such as emergency call centers, identifying callers' emotions when they contact 911, and seeking efficient assistance, is a notable good use. Automated customer support systems, in which the caller's emotions may influence the system's response and whether a human should be dispatched, are among the intriguing uses of SER. This may significantly lessen friction and annoyance, which will increase customer satisfaction.
Additionally, SER can provide and improve feedback from educational software agents, enabling the system to recognize and respond to student input.

II. Related Work

Numerous research have been done on extracting emotions from aural information. In order to recognize emotions from speech, it is necessary to first extract the features from a corpus of chosen or implemented emotional speech. Then, the retrieved characteristics are used to classify the emotions.
The successful extraction of the features has a significant impact on how well the classification of emotions performs.
The original RECOLA dataset sound information is used in a start to finish manner by Ali Bakhshi et al. to present a flexible emotion identification system based on Conv-BiGRU layer for speech motion recognition.In order to more fully grasp the effect that these factors have on the efficiency of network prediction, the experiment looks at various combinations of time length and batch size.[2]. Jerry Joy et al. uses MLP classifier to classify emotions according to a given fluctuation signal, so the choice of learning rate is adaptive, the dataset used was RAVDESS with an accuracy rate of 70.28% was achieved These five parameters, which are as follows, will captivate the characteristics to be retrieved from the given audio input, Chroma, Tonnetz, Mel Spectrograph Frequency, Contrast, and MFCC[3].
Darshan K.A et al uses CNN(convolution neural network) approach on dataset such as RAVDESS & SAVEE and achieved overall accuracy of 77% from it.[4]. MinSeop Lee et al uses PPG signal in a standardized10s PPG signal and 10 s NN interval to extract features19 statistical traits with a high Pearson correlation value were chosen, and 10 of them were statistical features. The NN interval and the properties of the normalized PPG signal are also extracted using the CNN model.
SujayAngadi et al uses hybrid neural network SER structure using CNN & BLSTM Methods with a 9.83% improvement in SER. [5].

III. Methods

The two processes of feature extraction and classification make up the standard building blocks of a pattern recognition workflow, which includes voice emotion identification tasks.
Using the MFCC parameters, we suggest an emotion recognition system in this study.
Below Fig.1, the suggested architecture is displayed

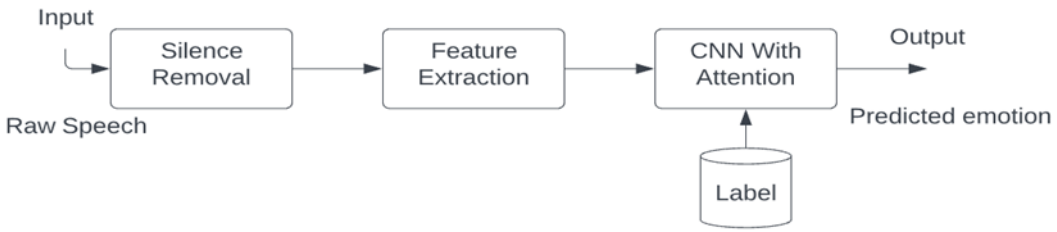

Figure 1 System with silence removal & classifier

Elimination of Silence

When using audio knowledge, the first thing you should do is view the audio file as a vector or matrix. After viewing each sound file in the dataset as a vector, we perform silence removal on each file using the minimal number of samples and the threshold. Every spoken vocalization in the speech dataset is examined using those 2 parameters. Filtered speech is the result of this silence reduction approach.
Formula 1 contains the whole formula for silence removal.

"Algorithm 1 Algorithm for silence removal

Require: speech dataset
Ensure: filtered speech

1: minimum threshold= threshold
2: minimum samples = n i min
3: n i = 0
4: for speech in speech dataset do
5: for i in speech do
6: if abs(amplitude[i]) < threshold then
7: n i = n i + 1
8: end if
9: if n i = n i min then
10: remove n i samples
11: end if
12: end for
13: end for “

Extraction of Features

We adhere to the feature extraction steps.Each speech-to-auditory communication is first divided into window frames and impacted by overlapping ones.The processes for feature extraction were carried out over each frame at regular intervals during each auditory transmission.At first, we used 256 as the conv 1d filter and an input shape of 259 x train shape parameter with a kernel size of 8.Consequently, each auditory communication's feature vector has the following dimensions: (None,259,256) CNN networks are provided with these alternatives.
In this work MFCC (Mel-frequency Cepstral Coefficient) has been used in the feature extraction

Mel-frequency Cepstral Coefficient
The logarithm of the Fast Fourier Transform (FFT) module of the signal is filtered using the Mel scale, and then the inverse Fast Fourier Transform (IFFT) is used to obtain the MFCC coefficients [15].

The process to find MFCCs typically entails the stages below. These actions are shown in the image (Fig.2.)Start by applying the Fast Fourier Transform on the input signal. The next step is to convert the spectrum's power, which was derived in the step before, to Mel scale.Next, calculate the speech signal's logarithms of power at each of the Mel frequencies.Next, use Discrete Cosine Transform on the Mel Log Powers bank.Finally; we translate the log Mel spectrum back to time. The Mel frequency cestrum coefficients are what we get as a consequence (MFCC).

To validate the models 9-fold cross-validation method is used. In other words, the data were divided up into nine folds.The first fold served as a test set, while the other folds were utilized to train our models. The next fold is then utilized to test our models, and the remaining are used for training, etc.

Architecture

A convolutional neural network with nine convolutional layers and one totally linked layer with 1024 hidden neurons made up the basic configuration of the deep neural network used in the current investigation. To estimate the probability distribution of the categories A 5-way or a 7-way SoftMax unit is used. A maxpooling or average-pooling layer came after each layer of convolutional network layer. Convolutional and totally linked layers used rectified Linear Units (ReLU) as activation functions to add the nonlinearity .The kernel size of the pooling layers was set to pool size of eight when the first range of kernels was established.

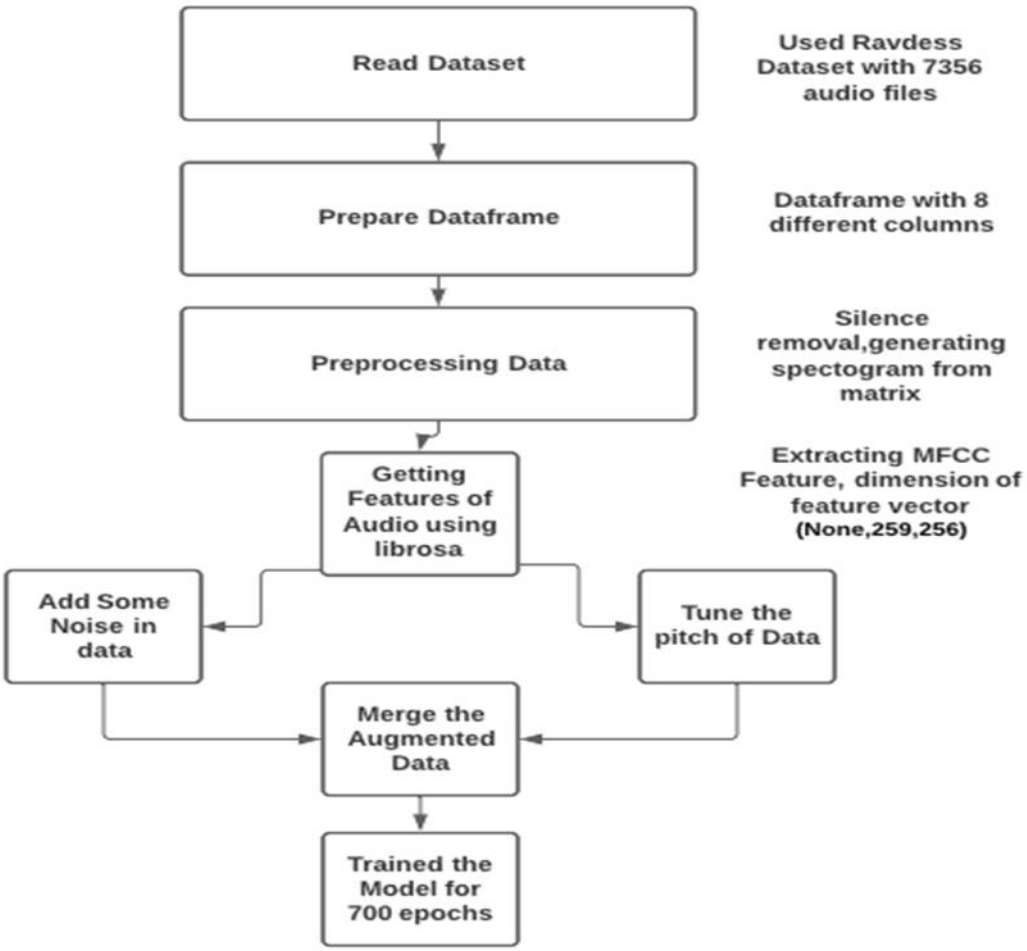

Figure 2. Architecture Workflow

.

IV Experiment & Results

In this work we employed RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) emotional dataset which contains Eight emotional states—neutral, calm, happy, sad, angry, afraid, disgusted, and surprised—are covered for spoken data, while six emotions are covered for song recording (neutral, calm, happy, sad, angry, and fearful).There are just 5 emotions employed in this (angry,calm,fearful,happy,sad)Total files in the dataset number 7356.There are twelve male and twelve female artists on the tape.In our study, we often just use speech recordings.In the RAVDESS dataset, twenty-four artists are given the task of singing and speaking two sentences—"Kids are talking by the door" and "Dogs are sitting by the door"—while evoking a variety of distinct emotions.

The model is trained in the first experiment for 700 epochs 30% of the dataset is utilized for testing, while 70% is used for training.The model has been taught to distinguish between the two types of male voice, namely male positive and male negative.This model's accuracy is 93.75%, F1 score is 92%, and loss value is around 0.0148, which is a very excellent result. For the three classes of data—positive, negative, and neutral—accuracy, F1 score is 89.86%, loss value is 0.02 and F1 score is 89.86%.The emotion distribution and confusion matrix are displayed in the figures below.

The second attempt used a CNN architecture that was lighter. The model has 700 training epochs 20% of the dataset is utilized for testing, while 80% of the dataset is used for training.The model is taught to distinguish between the 10 categories of male and female voice.The accuracy of this model is around 89%, while the average F1 score is about 78%.

For the train valid loss graph on three classes of emotions, validation was performed on 528 samples, while training on 2112 samples and validating on 480 samples for two classes of emotions, respectively. With an 8:2 splitting ratio, training and validation sets employ 1–20 actors.

Actors 21–24 are segregated for testing purposes, Neutral, Disgust, and Surprised are removed in 10 class recognition from the dataset in order to reduce the complexity of the model used to understand male emotions. Next, two actors were selected for testing and subjected to an 8:2 stratified shuffle split.The model is trained using a batch size of 16 and a parameter range of 700 epoch

The model is trained using training data, and after that it is tweaked using metrics (accuracy, loss, etc.) obtained from the validation set.

Truth labels have been defined Based on the number of classes to classify the speech labels are defined. Some of the classes are as follows:
Class: positive and negative & neutral
Positive: Calm, Happy.
Negative: Fearful, Sad, Angry.

Emotions are categorized and divided on the basis of these above classes emotions such as Happy,calmetc are referenced to positive class likewise for other classes and these classes are further used in calculating model performance and accuracy , confusion matrix from these classes helps the model in training and testing complex pattern and recognizing emotion from large datasets and certain intervals.

As compared to other approaches in this work CNN has been used by applying 9 layer design in such a way that it acts as an cnn-lstmlayer using only the MFCC features the efficiency of accuracy and improvements compared to previous approaches where MFCC HNR ZCR were used the accuracy achieved by this current approach is effective in predicting emotions from dataset in profound manner.

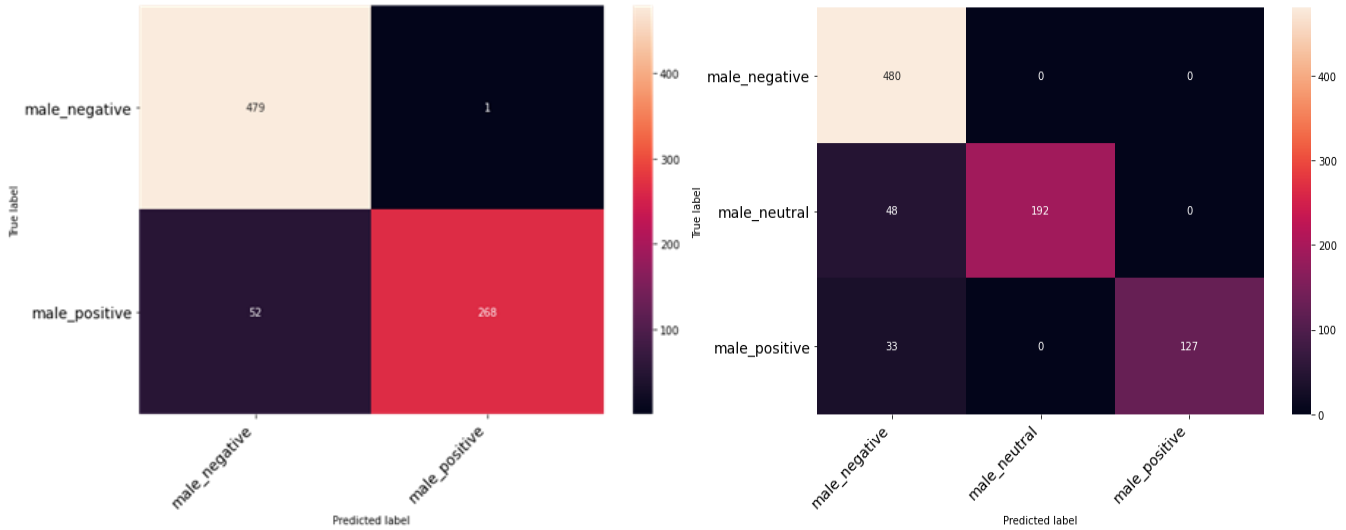

Figure 3 Confusion matrix for positive, negative, neutral class

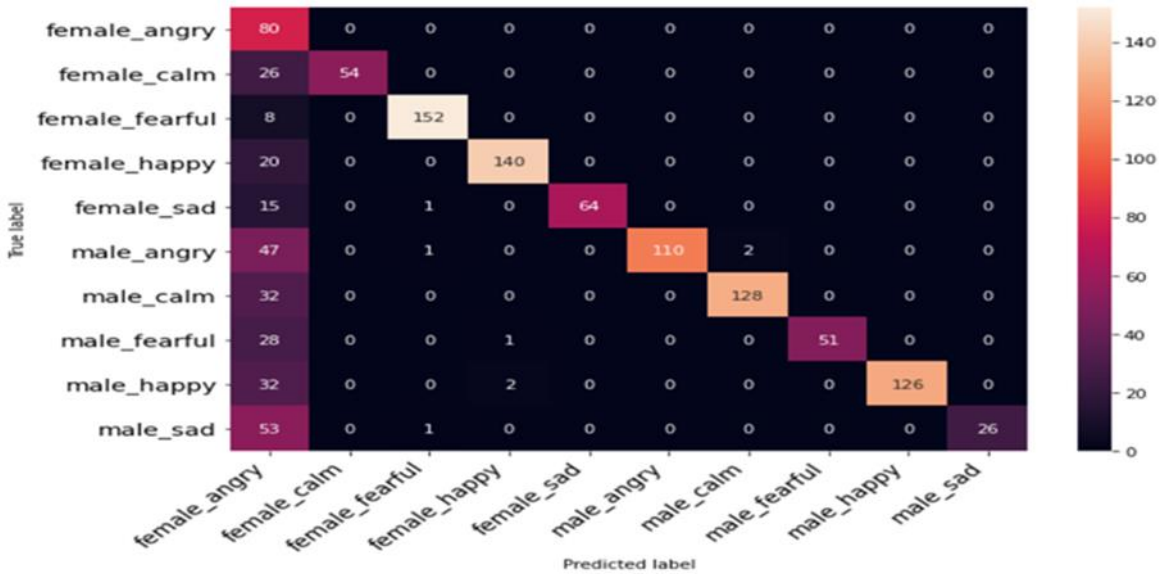

Figure 4 Confusion matrix describing 10 class matrix with support values

Classes	*precision*	*Recall*	*f1_score*	*support*
male_ negative	0.86	1.00	0.92	480
male _neutral	1.00	0.80	0.89	240
male_ positive	1.00	0.79	0.89	460

From the confusion matrix of 10 class these results show that our model can distinguish various emotions accurately for example Angry, Happy, Fearful, Sad Here female angry has less prediction accuracy compared to other emotion.

For 2 class confusion matrix positive class contains calm, happy Negative class contains fearful, angry, and sadhere positive class precision has best prediction.
For 3 Class confusion matrix positive class contains Happy, Neutral class contains calm, and negative class contains sad, angry, fearful, neutral and positive class precision here has best prediction.

Table 3. Comparison between proposed approach and other methodology

SNo.	*Author*	*DatasetUsed*	*Methodology*	*Gaps|Drawbacks*
[1]	Jerryjoy,Aparnakannan,Shreyaram,S.rama[3]	RAVDESS	Neural network with multilayer perceptron	On the dataset accuracy is low even on using cnn with decisiontree 70.28%
[2]	DarshanK.A, Dr.B.N. Veerappa[4]	RAVDESS & SAVEE	CNN(convolutional neural network)	Achieved Accuracy 77%
[3]	LingjieShen,Wei Wang[16]	IEMOCAP	CNN(convolution neural network)	With phonological representation CNN provides 60.22% of UAR on categorical emotions on utterance level
[4]	Hendrik purwins,BoLi,TuomasVirtanen,JanSchluter,Shuo-YiinChang,TaraSainath[21]	Private Series of Chinese emotional speech dataset	Decision Tree & Random forest	Best 82.54% & worst 16%
[5]	Our proposed system	Ravdess dataset	Convolutional neural network & MFCC Features	Achieved accuracy of 89% for all classes ,93.75% for positive, negative class & 93.37% for positive ,negative,neutral.

V LIVE EMOTION PREDICTION

The CNN (convolutional neural network) model is trained on the dataset RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) and it can recognize up to 5 emotions from the dataset as per the training provided the model is generalized to predict speech emotion from live spoken feeded voice. Currently, in order to effectively predict emotion with gender instance from the data, the recorded voice must be in the mentioned accent since the dataset used for training is for north American English.
The base CNN module is given the stored wav file once a separate audio recorder module has been established to record the live voice according to the model's duration standards. As demonstrated by the model's performance with the confusion matrix, the model can accurately anticipate from the taught emotion. The output is presented as an array with the gender and voice-retrieved emotion displayed.
By training on several other datasets prediction accuracy can be increased for various other vocal languages.

VI. CONCLUSION

A typical emotion recognition system design involves the use of high-dimensional features on a curated data set. The disadvantage of this method is that the data set is limited and it is difficult to analyze in the high-dimensional feature space. The proposed system is based on a 9-layer CNN design, and the proposed work accuracy is 89%. The accuracy

of the traditional system is 77%, so the proposed work is improved by approx 12%.This suggested approach is based on a 9-layer CNN architecture, with an accuracy rate of 89.00% for 10 classes of emotions, 93.75% for two classes (positive, negative), and 93.37% for three classes (positive, negative, neutral)

REFERENCES:

1. Janu, Neha, Sunita Gupta, Meenakshi Nawal, and Pooja Choudhary. "Query-based image retrieval using SVM." In International Conference on Emerging Technologies in Computer Engineering, pp. 529-539. Cham: Springer International Publishing, 2022.
2. Nawal, Meenakshi, et al. "11 A Critical Analysis of Graph Topologies for Natural Language." Graph Learning and Network Science for Natural Language Processing (2022): 189.
3. Awasthi, Charu, Meenakshi Nawal, and Prashant Kumar Mishra. "Security concerns of fog computing in field of healthcare using blockchain: A review." 2021 International Conference on Communication information and Computing Technology (ICCICT). IEEE, 2021.
4. Apoorv Singh, Kshitij Kumar Srivastava, HariniMurugan "Speech Emotion Recognition Using Convolutional Neural Network AliBakhshi and AaronS.W.Wong and StephanChalup "End-To-End Speech Emotion Recognition Basedon Time and Frequency Information Using Deep Neural Networks" ECAI 2020.
5. Jerry Joy, AparnaKannan, Shreya Ram, S. Rama "Speech Emotion Recognition using Neural Network and MLP Classifier''IJESC, April 2020.
6. Darshan K.A, Dr. B.N. Veerappa "SPEECH EMOTION RECOGNITION" IRJET 2020.
7. MinSeop Lee, Yun Kyu Lee, Myo-TaegLim , and Tae-Koo Kang, "Emotion Recognition Using Convolutional Neural Network with Selected Statistical Photoplethysmogram Features" Adpl2020.
8. Gangrade, Anurish, and Shalini Singhal. "A Research of Speech Emotion Recognition Based on CNN Network." SKIT Research Journal 26.07.2022.