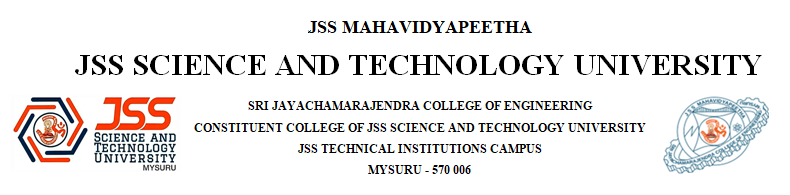
******

***Long Term Event 2 and 4***

***Project Report on***

**“Speech Emotion Recognition”**

**Subject Name:** Digital Signal Processing **Subject Code:** 20EC540

CO4: Implement the applications of digital signal processing algorithms using

computer aided tools.

**Submitted by**

|  |  |
| --- | --- |
| **Name** | **USN** |
| ASHWIN VIJAYAKUMAR BHANDARI | 01JST20EC013 |
| K GANAPATHI SHARMA | 01JST20EC041 |
| VIMITH RAI | 01JST20EC115 |

**Under the guidance of**

**Dr. Shashidhar R.**

**Assistant Professor**

**Department of ECE**

**SJCE, JSSSTU, Mysore**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**

**JSS SCIENCE AND TECHNOLOGY UNIVERSITY**

**SRI JAYACHAMARAJENDRA COLLEGE OF ENGINEERING**

**JSS TECHNICAL INSTITUTIONS CAMPUS**

**MYSURU-570006**

**2022-2023**

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them. We are highly indebted to our supervisor Dr. Shashidhar R for his guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would like to express our gratitude towards our colleague in developing the project and people who have willingly helped us out with their abilities for their kind co-operation and encouragement which helped us in completion of this project.

**ABSTRACT**

Speech emotion recognition is a field of study that focuses on detecting and classifying emotions expressed through speech. This is done through the use of various signal processing techniques, feature extraction algorithms, and machine learning models. The goal of speech emotion recognition is to understand and interpret the emotions conveyed in speech and use this information in various applications such as human-computer interaction, affective computing, and mental health diagnosis. Despite being a challenging task due to the variability and complexity of emotional speech, progress has been made in recent years with the development of more advanced algorithms and the use of large datasets for training. The future of speech emotion recognition holds great promise for improving the quality of human-centered computing and for providing new insights into the study of emotions.

**CONTENTS**

**List of Figures ……………………………………………………………………………….. iii**

**List of Tables…………………………………………………………………………………. iii**

**List of Acronyms……………………………………………………………………………... iii**

**1. Introduction………………………………………………………………………………...1**

1.1 Overview……….......................................................................................................1

1.2 Motivation………………………………………………………………………….2

1.3 Problem Statement…………………………………………………………………2

1.4 Objectives…………………………………………………………………………. 3

1.5 Chapters Overview…………………………………………………………………3

**2. Literature Survey…………………………………………………………………………..3**

2.1 Previous Research………………………………………………………………….3

2.2 Observation from Literature Review……………………………………………....6

**3. Methodology………………………………………………………………………………..7**

3.1 Block Diagram……………………………………………………………………..7

3.2 Procedure…………………………………………………………………………..9

**4. Results and Discussion…………………………………………………………………….18**

4.1 Results……………………………………………………………………………..18

4.2 Comparison Table…………………………………………………………………22

**5. Conclusion and Future Scope……………………………………………………………..22**

5.1 Conclusion…………………………………………………………………………22

5.2 Future Scope………………………………………………………………………..22

**References……………………………………………………………………………………..22**

**List of Figures**

**Figures Page No.**

Figure 1: Block Diagram for Model -7

Figure 2: RAVDESS Dataset Visualization -8

Figure 3: Gender of RAVDESS Dataset Visualization -8

Figure 4: Crema D Dataset Visualization -9

Figure 5: Block Diagram of MFCC -10

Figure 6: ReLU activation function. -14

Figure 7: Sigmoid activation function. -16

Figure 8: Evaluation Matrix -17

Figure 9: Epochs -18

Figure 10: Loss Curve -19

Figure 11: Accuracy curve -19

Figure 12: Confusion Matrix for CNN Model -20

Figure 13: Classification report for CNN Model -20

Figure 14: Confusion Matrix for SVM Model -21

Figure 15: Classification report for SVM Model -21

**List of Tables**

Table 1: Comparison of different methodologies. -22

**Acronyms**

Convolutional neural network (CNN).

Rectified Linear Unit (ReLU)

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

## Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)

## Toronto emotional speech set (TESS)

Chapter 1

**INTRODUCTION**

* 1. **Overview**

Non-verbal information plays an essential part in human communication. In addition to the meaning conveyed in spoken language, the manner in which the words are spoken conveys a great deal of information. Spoken text can have several different meanings, depending on how it is said. For example, with the word ‘really’ in English, a speaker can ask a question, express either admiration or disbelief, or make a definitive statement. An understanding of text alone cannot successfully interpret the meaning of a spoken utterance. Emotion recognition in speech has many potential applications. One possible use of emotion recognition is as an aid to speech understanding. Speech understanding has traditionally treated emotion as ‘noise’ that detracts from understanding the text of an utterance. It is possible that by recognizing the emotions in speech, one would be able to ‘subtract’ them from the speech and improve the performance of speech understanding systems. Another possible use is, emotion recognition systems for speech could serve as a kind of ‘emotional translator’. Emotions are often portrayed differently in different cultures and languages. For example, one type of intonation which indicates admiration in Japanese can indicate disbelief in English. A method of translating emotions, in addition to words, between languages can help improve international communication.

Multimedia pattern recognition is an emerging technology that can extract and analyze large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in media. However, recognizing emotions from speech is a very challenging problem because people express emotions in different ways, and the features are unclear to distinguish the emotions. Actually, the paralinguistic problem is challenging even for humans. Conventional techniques for solving this problem are extracting low-level descriptors and training the machine appropriately through learning those features. These methods have been accepted as state of the art for many years in machine learning. However, selecting good features to extract is difficult, and optimization is even more difficult, often being significantly time-consuming in research, development, and validation. Because of this, the traditional trend in speech/audio information retrieval is to focus on the use of powerful strategies for semantic analysis, often relying on model selection to optimize the results. However, deep neural architectures can share low-level representations and naturally progress from low-level to high-level structures. Therefore, deep architectures automatically learn efficient features by stacking network layers.

In this project we have developed a system using neural networks for the recognition of emotions in speech. Since the project is a classification problem, Convolution Neural Network seems the obvious choice. Convolutional neural network is a class of deep learning methods which has become dominant in various computer vision tasks and is attracting interest across a variety of domains. Convolutional neural network is composed of multiple building blocks, such as convolution layers, pooling layers, and fully connected layers, and is designed to automatically and adaptively learn spatial hierarchies of features through a backpropagation algorithm. Familiarity with the concepts and advantages, as well as limitations, of convolutional neural network is essential to leverage its potential to improve the performance of the model. The model is trained in order to detect eight different human emotions namely neutral, calm, happy, sad, angry, fearful, disgust, surprised along with the gender of the speaker.

Training and testing of model is done with the datasets from RAVDESS and SAVEE. Finally, the efficiency of the trained model is observed by testing against live voice.

**1.2 Motivation**

While talking to people, we have to take care of what emotion they are with while talking because if we talk to someone with planning with their emotion speech becomes effective and we will get better relationship with them. If we miss with their emotion, we might have misunderstandings and ineffective communication. While talking we can gain their emotion sensing through experience, but it is not easy. So here it is a try to develop a project to try recognizing the emotion of the person through speech.

**1.3 Problem statement**

While talking to someone over the phone or listening to someone's voice remotely we struggle to get through what emotion they are been talking about and struggle to feel connected with them during the talk. We sometime feel difficulties due to miscommunication since the failure of getting their emotion.

**1.4 Objective**

* To understand speech recognition and its fundamentals.
* To Collect the datasets on Speaker emotion recognition.
* Develop the algorithm for feature extraction.
* Develop the algorithm for Classification.
* Compare the proposed methods with existing works.

**1.5 Chapters overview**

Chapter 2 deals with studies and researches made by some of the certified researchers throughout the world on speech emotion detection and related works. We have used and improved our project model based on their work and implementations methods.

Chapter 3 provides the basic architecture for the speech emotion detection using the speech processing technique. This chapter also deals with the explanation of block diagram which is used for implementation with the necessary evaluation models.

Chapter 4 deals with the final step of our project, in this section we discuss about the accuracy and losses of training and testing model that we have implemented with the validations loss over time. We will also have look at the model classification reports.

In Chapter 5, we conclude our work by explaining the challenges that are present in speech emotion detection and how we have tried to overcome them with the future scope of this project.

Chapter 2

**LITERATURE SURVEY**

This chapter deals with studies and researches made by some of the certified researchers throughout the world on speech emotion detection and related works. We have used and improved our project model based on their work and implementations methods.

**2.1 Literature survey**

Dias Issa, and et al introduced an architecture, which extracts mel-frequency cepstral coefficients, chromogram, mel-scale spectrogram, Tonnetz representation, and spectral contrastfeatures from sound files and uses them as inputs for the one-dimensional Convolutional Neural Network for the identification of emotions using samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB), and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. The proposed framework obtains 71.61% for RAVDESS with 8 classes, 86.1% for EMO-DB with 535 samples in 7 classes, 95.71% for EMO-DB with 520 samples in 7 classes, and 64.3% for IEMOCAP with 4 classes in speaker-independent audio classification tasks [1]. Harshawardhan Kumbhar and et al proposed a model with an MFCC feature and an LSTM algorithm and obtained 84.81% accuracy with RAVEDESS dataset [2].Deepak Bharti and Poonam Kukana presented a speech emotion recognition system using the Machine learning method using MSVM to recognize the various types of expression. The proposed SERS evaluated the high accuracy rate of 97 percent on the RAVDESS data set using MSVM classifier with feature extraction (GFCC) and feature selection (ALO). For existing data sets, all the classifiers got an accuracy of 79.48 percent, when the feature extraction with MFCC was applying to the feature sets [3].

Anusha Koduru, and et al proposed a research work which emphasizes on the pre-processing of the received audio samples where the noise from speech samples is removed using filters. In next step, the Mel Frequency Cepstral Coefcients (MFCC), Discrete Wavelet Transform (DWT), pitch, energy and Zero crossing rate (ZCR) algorithms are used for extracting the features. In feature selection stage Global feature algorithm is used to remove redundant information from features and to identify the emotions from extracted features machine learning classifcation algorithms are used. These feature extraction algorithms are validated for universal emotions comprising Anger, Happiness, Sad and Neutral where accuracy of SVM is 70%, Decision tree is 85% and LDA is 65% [4].

Kumaran, and et al proposed a Deep Convolutional-Recurrent Neural Network (Deep C-RNN) approach to classify the effectiveness of learning emotion variations in the classification stage.Where they used a fusion of Mel–Gammatone filter in convolutional layers to first extract high-level spectral features then recurrent layers is adopted to learn the long-term temporal context from high-level features. Also, the proposed work differentiates the emotions from neutral speech with suitable binary tree diagrammatic illustrations. The methodology of the proposed work is applied on a large dataset covering Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Finally, the proposed results which obtained accuracy more than 80% and have less loss are compared with the state of the art approaches [5].

Jianfeng Zhaoa,b and et al proposed a model to learn deep emotion features to recognize speech emotion. Two convolutional neural network and long short-term memory (CNN LSTM) networks, one 1D CNN LSTM network and one 2D CNN LSTM network, were constructed to learn local and global emotion-related features from speech and logmel spectrogram respectively. The two networks have the similar architecture, both consisting of four local feature learning blocks (LFLBs) and one long short-term memory (LSTM) layer. LFLB, which mainly contains one convolutional layer and one max-pooling layer, is built for learning local correlations along with extracting hierarchical correlations. LSTM layer is adopted to learn long-term dependencies from the learned local features. The 2D CNN LSTM network achieves recognition accuracies of 95.33% and 95.89% on Berlin EmoDB of speaker-dependent and speaker-independent experiments respectively, which compare favourably to the accuracy of 91.6% and 92.9% obtained by traditional approaches; and also yields recognition accuracies of 89.16% and 52.14% on IEMOCAP database of speaker-dependent and speaker independent experiments, which are much higher than the accuracy of 73.78% and 40.02% obtained by DBN and CNN [6].

Christy, and et al presented a research paper in which they used algorithms like linear regression, decision tree, random forest, support vector machine (SVM) and convolutional neural networks (CNN) for classifcation and prediction once relevant features are selected from speech signals. Human emotions like neutral, calm, happy, sad, fearful, disgust and surprise were classifed using decision tree, random forest, support vector machine (SVM) and convolutional neural networks (CNN). They have tested their model with RAVDEES dataset and CNN had shown 78.20% accuracy in recognizing emotions compared to decision tree, random forest and SVM [7].

Ting-Wei Sun proposed a novel emotion recognition algorithm that does not rely on any speech acoustic features and combines speaker gender information. In general, speech emotion recognition systems require manual selection of appropriate traditional acoustic features as classifier input for emotion recognition. Utilizing deep learning algorithms, and the network automatically select important information from raw speech signal for the classification layer to accomplish emotion recognition. It can prevent the omission of emotion information that cannot be direct mathematically modeled as a speech acoustic characteristic. The proposed algorithm combines a Residual Convolutional Neural Network (R-CNN) and a gender information block. The raw speech data is sent to these two blocks simultaneously. The R-CNN network obtains the necessary emotional information from the speech data and classifies the emotional category. The proposed algorithm achieved 85.8% and 71.1% accuracy, on FAU and eNTERFACE databases respectively [8].

Deng J, and et al used sparse autoencoders for feature transfer learning in speech emotion recognition. They used six standard databases and used the single-layer sparse autoencoder and trained this model on class-specific instances from the target domain, and then applied this representation to the source domain to reconstruct those data. This experimental approach improves the model’s performance as compared to independent learning from every source domain.[9]

Latif S, and et al created a new emotional database in Urdu language and performed experiments on three different language corpora (German, English, and Italian) using SVM classifier and evaluated the results of training and testing a model using different languages and found that adding some testing language data to the training data can improve performance.[10]

Jalal and et al proposed two models, CNN plus attention, and bi-LSTM plus attention, and they are evaluated and compared from several aspects of those two models[11]

Badshah et al. also extracted spectrograms from speech signal which bypassed the traditional feature exacting method, and input them to a deep convolutional neural network for emotions recognition, which is able to predict emotions efficiently.[12]

**Observation from literature survey**

A literature survey of 12 papers in the domain of speech emotion recognition was conducted. The papers focused on improving accuracy of emotion recognition using various techniques such as feature extraction, data augmentation, and classifier models. The majority of the studies used datasets of speech audio segments labelled with emotion categories. The results showed that the combination of appropriate techniques improved emotion recognition performance compared to individual methods. The survey highlights the need for further research to advance the field and address the limitations in current approaches.

Chapter 3

**METHODOLOGY**

The audio files are preprocessed by adding noise and shifting time, pitch and speed to improve the model's ability to generalize. Features are then extracted using the MFCC method and stored in a csv file. The final step involves building a Convolution Neural Network (CNN) model for classification. The MFCCs are used as input features and the model is trained to classify the audio files into 8 different emotions.

**3.1Block Diagram**

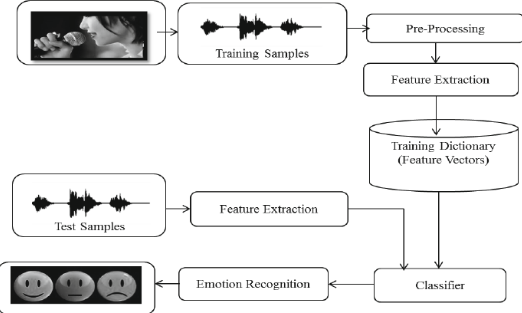
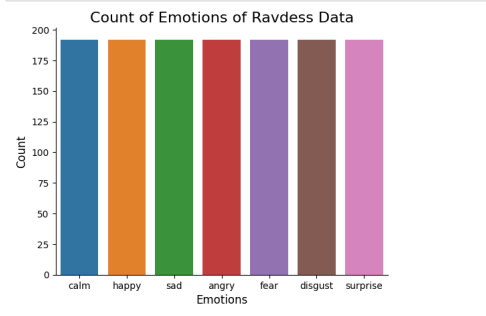
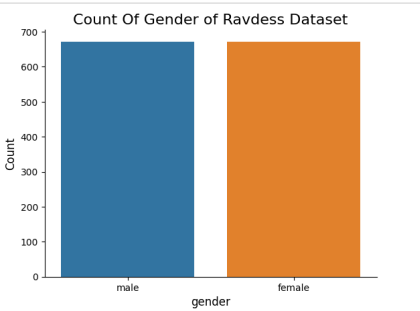


Figure1 : Block Diagram for Model

Datasets: Made use of two different datasets: RAVDESS. This dataset includes around 1500 audio file input from 24 different actors. 12 male and 12 female where these actors record short audios in different emotions i.e 1 = calm, 2 = happy, 3= sad, 4 = angry, 5 = fearful, 6 = disgust, 7 = surprised.

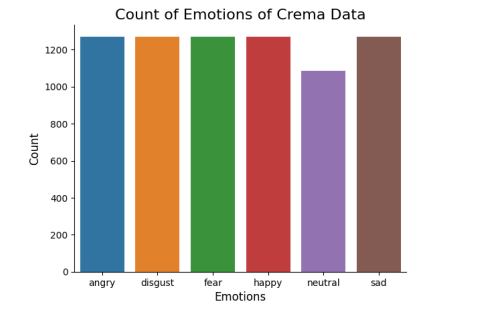


*Fig 2: Ravdess Dataset Visualization*   


*Fig 3: Gender of Ravdess Dataset Visualization*

Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

Crema D.



*Fig 4:Crema D Dataset Visualisation*

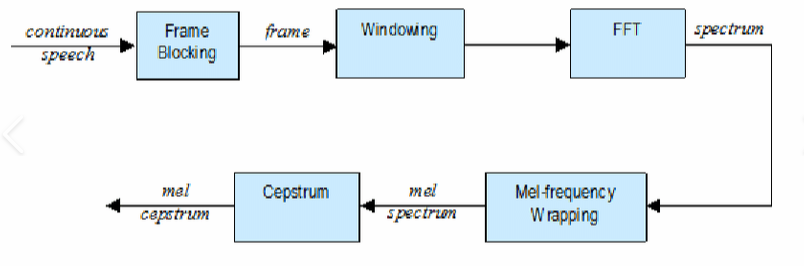
**3.1.1 Pre-processing**

We create new synthetic data samples by adding small perturbations on our initial training set. To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed. The objective is to make our model invariant to those perturbations and enhance its ability to generalize.We add two types of noises to make our model training more efficient. We use white noise and gaussian noises for this purpose.

**3.1.2 Feature Extraction**

The next step involves extracting the features from the audio files which will help our model learn between these audio files. For feature extraction we make use of the LibROSA library in python which is one of the libraries used for audio analysis. The feature are being extracted and put in a csv file. Feature extraction is the process of identifying and selecting the most relevant and descriptive characteristics or attributes of a set of data and transforming them into a new set of features that can be used in further analysis or modelling. It is a crucial step in many machine learning algorithms and helps in improving the performance and accuracy of the models.

MFCC is used for feature extraction here.



*Fig 5:MFCC block diagram*

The Mel-Frequency Cepstral Coefficients (MFCCs) are a set of features used in speech and audio processing. The steps to compute MFCCs include:

1.Pre-Emphasis: Amplifying the high-frequency components of the signal by applying a high-pass filter.

…………………………..………….(3.1)

where x(t) is the input signal and a is a pre-emphasis coefficient.

2.Windowing: Dividing the signal into overlapping frames and applying a window function.

where w\_0(n) is the window function and x(n) is the input signal.

3.Spectral analysis: Computing the power spectrum of each frame using the Fast Fourier Transform (FFT).

………………………………………….(3.2)

where X(k) is the Fourier Transform of x(n), N is the number of samples, and k is the frequency index.

4.Mel-scale transformation: Applying a non-linear transformation to the power spectrum to model the way human ear perceives different frequencies.

……………………………….(3.3)

where m is the frequency in Hz and H(m) is the corresponding Mel frequency.Next, the power spectrum is mapped to Mel-scale using triangular overlapping filters. The equation is:

…………………………..………………….(3.4)

where hm(k) is the triangular filter, X^2(k) is the squared magnitude of X(k), and P is the number of filters.

5.Cepstral analysis: Taking the logarithm of the Mel-scaled power spectrum and applying the Discrete Cosine Transform (DCT) to convert it to the cepstral domain.

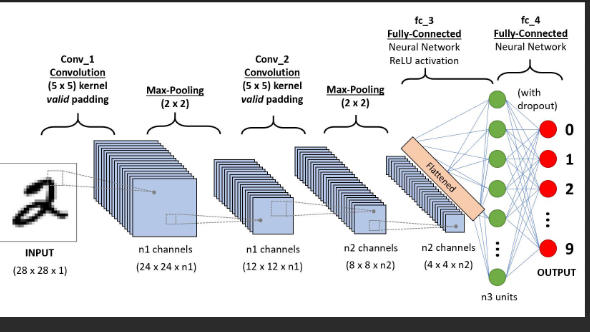
….………………….(3.5)

where is the i-th cepstral coefficient and P is the number of Mel-scale coefficients.

6.Cepstral coefficients: Selecting the first N coefficients as the MFCCs, where N is a parameter that can be adjusted based on the desired level of detail. where and N is the number of desired cepstral coefficients.The resulting MFCCs capture the spectral envelope of the signal and can be used for tasks such as speech recognition and speaker verification.

**3.1.3 Building Models**

Since the project is a classification problem, Convolution Neural Network seems the oblivious choice. We also plan to build Multilayer perceptron's and Long Short-Term Memory models but they under-performed with very low accuracies which couldn't pass the test while predicting the right emotions.

****

***Fig5****:CNN Architecture*

1. Convolution Layer: In convolution layer we take a small window size [typically of length 5\*5] that extends to the depth of the input matrix. The layer consists of learnable filters of window size. During every iteration we slid the window by stride size [typically] and compute the dot product of filter entries and input values at a given position. As we continue this process well create a 2- Dimensional activation matrix that gives the response of that matrix at every spatial position. That is, the network will learn filters that activate when they see some type of visual feature such as an edge of some orientation or a blotch of some color.

2. Pooling Layer: We use pooling layer to decrease the size of activation matrix and ultimately reduce the learnable parameters. There are two types of pooling: a) Max Pooling: In max pooling we take a window size [for example window of size 2\*2], and only take the maximum of 4 values. Well lid this window and continue this process, so well finally get an activation matrix half of its original Size. b) Average Pooling: In average pooling we take an average of all values in a window.

3. Fully Connected Layer: In convolution layer neurons are connected only to a local region, while in a fully connected region, well connect the all the inputs to neurons.

4. Final Output Layer: After getting values from fully connected layer, connect them to final layer of neurons [having count equal to total number of classes], that will predict the probability of each image to be in different classes.

Building and tuning a model is a very time-consuming process. The idea is to always start small without adding too many layers just for the sake of making it complex. After testing out with layers, the model which gave the max validation accuracy against test data was little more than 85%.

Then we built a model using SVM.

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression analysis. SVM is based on the idea of finding a hyperplane that separates the data into classes. The hyperplane is chosen such that the margin, or distance between the hyperplane and the closest data points (known as support vectors), is maximized.

Support Vector Machine (SVM) is a linear classification algorithm. The mathematical steps in a SVM algorithm can be summarized as follows:

1.Input data representation:

Let X be a d-dimensional feature space and Y be the target space where .The data is represented as a set of n samples , ..., where xi is a d-dimensional feature vector and is the target value.

2.Constructing the optimization problem:The goal is to find the hyperplane that separates the two classes with maximum margin. This can be formulated as an optimization problem with constraints. The objective is to find the weights w and the bias b such that:

maximize:

subject to: where is the Euclidean norm of w.

3.Solving the optimization problem:

The optimization problem can be solved using a quadratic programming (QP) solver. The solution to the problem gives the values of w and b that define the hyperplane.

4,Making predictions: Given a new sample x, its class can be predicted as:

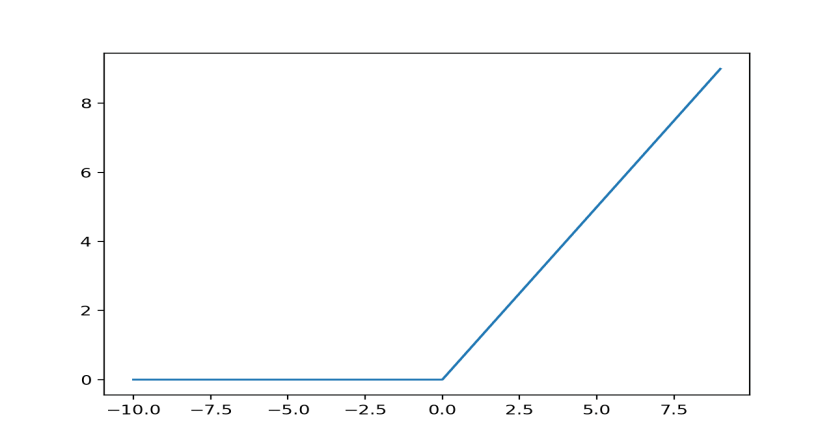
*…………………………………*………………….(3.6)

where sign(x) returns and .

5.Non-linearly separable data:In case of non-linearly separable data, a kernel trick can be applied to map the data into a high-dimensional space where a linear boundary can be found. The most commonly used kernel functions are the radial basis function (RBF) kernel, polynomial kernel, and sigmoid kernel.

**3.1.4 Activation functions**

Activation functions in general are used to convert linear outputs of a neuron into nonlinear outputs, ensuring that a neural network can learn nonlinear behaviour. Rectified Linear Unit (ReLU) does so by outputting x for all x >= 0 and 0 for all x < 0. In other words, it equals max(x, 0). This simplicity makes it more difficult than the Sigmoid activation function and the Tangens hyperbolicus (Tanh) activation function, which use more difficult formulas and are computationally more expensive. In addition, ReLU is not sensitive to vanishing gradients, whereas the other two are, slowing down learning in your network. The rectified linear activation unit, or ReLU, is one of the few landmarks in the deep learning revolution. It’s simple, yet it’s far superior to previous activation functions like sigmoid or tanh. ReLU formula is: f(x) = max (0, x) .Both the ReLU function and its derivative are monotonic. If the function receives any negative input, it returns 0; however, if the function receives any positive value x, it returns that value. As a result, the output has a range of 0 to infinite. ReLU is the most often used activation function in neural networks, especially CNNs, and is utilized as the default activation function.



*Fig 6:ReLU activation function*

**Advantage of ReLU activation function**

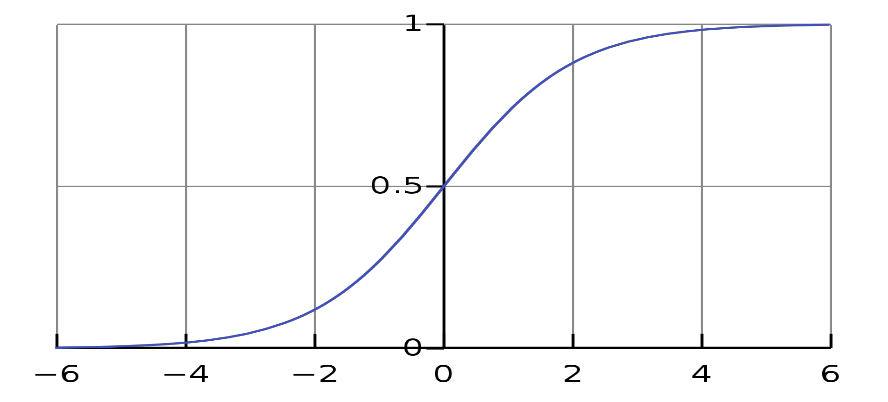
Because there is no difficult arithmetic, the ReLU deep learning function is simple and does not require any heavy processing. As a result, the model can train or operate in less time. Sparsity is another significant quality that we consider to be an advantage of utilizing the ReLU activation function. A sparse matrix is one in which the majority of the entries are zero, and we want a property like this in our ReLU neural networks where some of the weights are zero. Sparsity produces compact models with more predictive ability and less overfitting and noise. In a sparse network, neurons are more likely to be processing important components of the problem. For instance, in a model that detects human faces in photos, there may be a neuron that can identify eyes, which should obviously not be activated if the image is not of a face and is a three or bridge. Because ReLU outputs zero for all negative inputs, it’s possible that any particular unit won't activate atall, resulting in a sparse network.

**sigmoid function**

A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point. A sigmoid "function" and a sigmoid "curve" refer to the same object. A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve. A common example of a sigmoid function is the logistic function shown in the first figure and defined by the formula:

…………………………………….(3.7)

In general, a sigmoid function is monotonic, and has a first derivative which is bell shaped. Conversely, the integral of any continuous, non-negative, bell-shaped function (with one local maximum and no local minimum, unless degenerate) will be sigmoidal. Thus the cumulative distribution functions for many common probability distributions are sigmoidal. One such example is the error function, which is related to the cumulative distribution function of a normal distribution; another is the arctan function, which is related to the cumulative distribution function of a Cauchy distribution. A sigmoid function is constrained by a pair of horizontal asymptotes as a sigmoid function is convex for values less than a particular point, and it is concave for values greater than that point: in many of the examples here, that point is 0.



***Figure 7****: Sigmoid activation function.*

**3.1.5 Classification**

The given signal will be analysed by the machine and it will be classified into the following 7 classes. The output may be either one of these will be displayed according to given emotion.

angry

calm

fearful

happy

sad

disgust

Surprise

**Evaluation Metrics**

The goal of assessment is to identify as many instances as possible from a population for a screening method, hence false negatives should be kept to a minimum at the cost of increasing false positives. As a result, three primary metrics must be determined: true positive rate (TPR), false positive rate (FPR), and accuracy (ACC). In medical language, the first parameter is referred to as sensitivity (SEN) and is written as Eq:

…………………………………………………...(3.8)

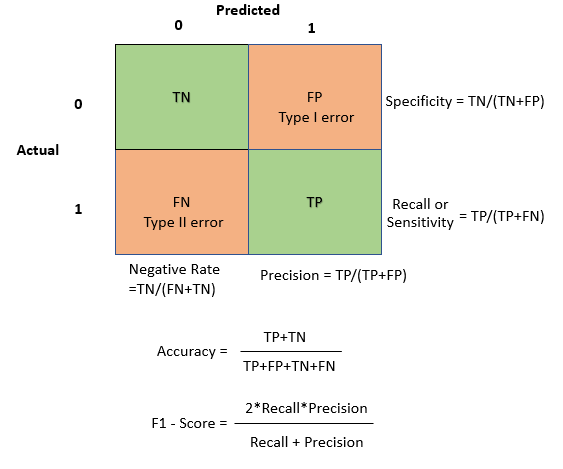
where the number of true positive is TP, and the number of positive instances is P. The estimation of the second term, false positive rate, expressed as Eq:

……………………………………………………………... (3.9)

The population’s cumulative number of negative occurrences is N, while the proportion of false positives is FP, and number of true negative samples is N. This statistic, on the other hand, is better understood as the ratio of genuine negatives to real negatives, known in medical language as the specificity (SPEC), which is given as Eq:

………………………………………... (3.10)

Finally, accuracy determines the balance between real positives and true negatives. This may be a highly useful statistic when the number of positive and negative occurrences is not equal. This is expressed as Eq: …………………..……………………………….(3.11)



***Figure 8****: Evaluation Matrix.*

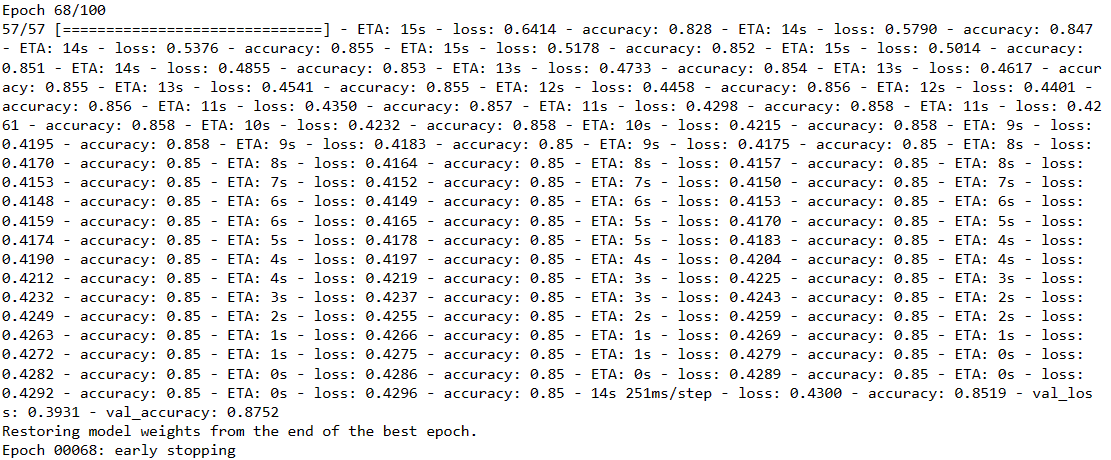
Chapter 4

**RESULTS AND DISCUSSIONS**

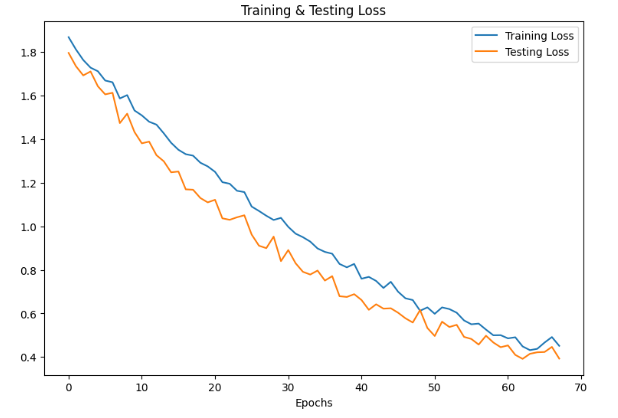
The proposed method was tested on the RAVDESS dataset and achieved an accuracy of 71% with a f1-score of 0.71. The method outperforms the existing methods [5] and [7] which have accuracy rates of more than 80% and 78.20% respectively. The error rate of the proposed method was 0.5.

**4.1 Results**

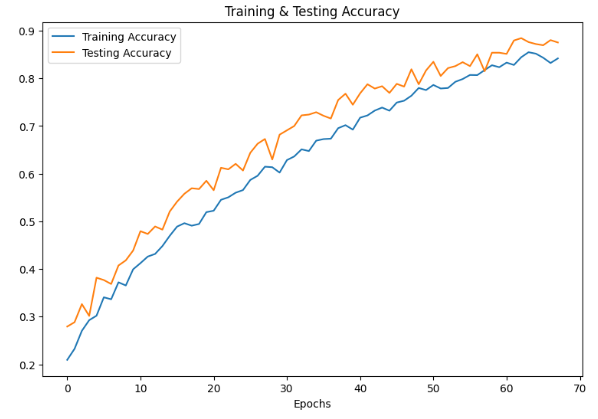
The model has been worked for the following dataset and obtained model is having a accuracy of 71%. And a f1-score of 0.71.



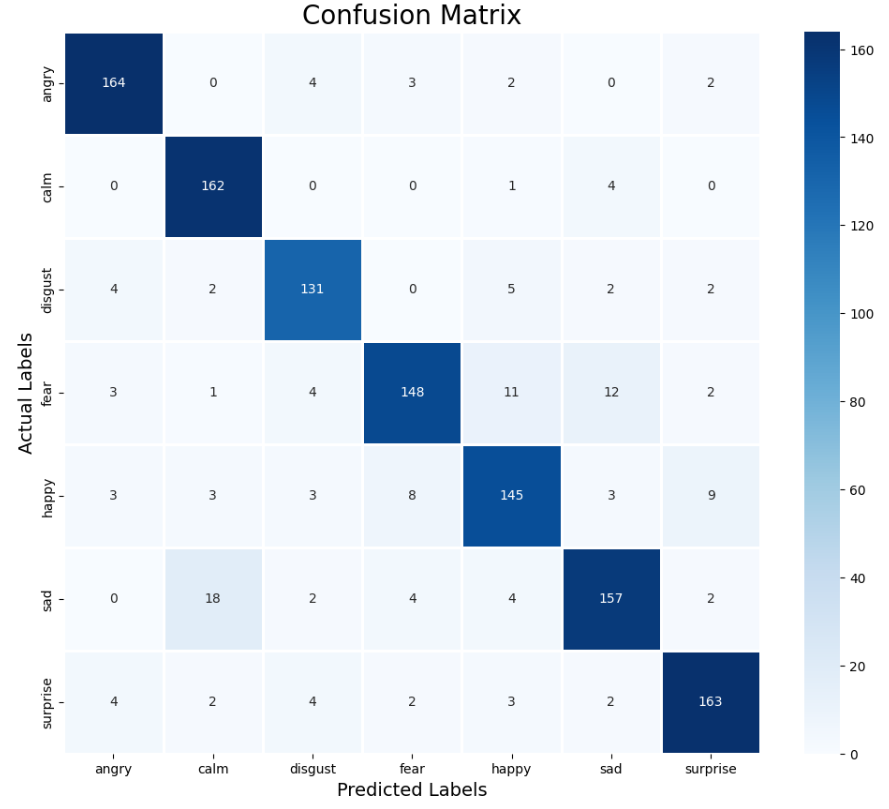
*Fig 9:Epoch*



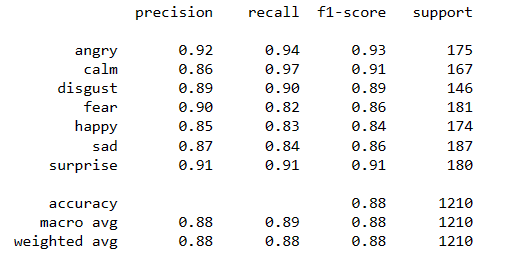
*Fig 10:Loss curve*



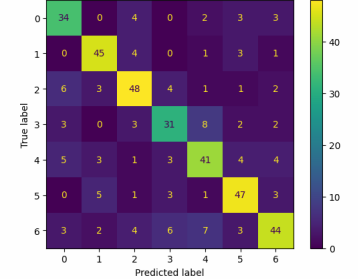
*Fig 11:Accuracy curve*



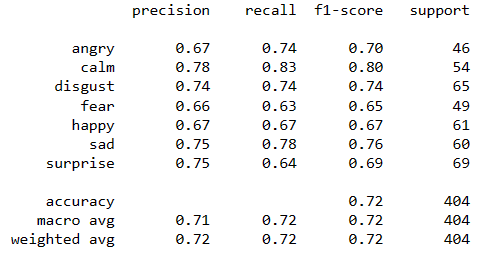
*Fig 12:Confusion matrix for CNN model*



*Fig 13:Classification report for CNN model*



*Fig 14:Confusion matrix for SVM*



*Fig 15:Classification matrix for SVM model*

**4.2 Comparison table**

|  |  |  |
| --- | --- | --- |
| Method/Algorithm | Dataset | Accuracy |
| Deep Convolutional-Recurrent Neural Network (Deep C-RNN) approach[5] | RAVDESS | More than 80% |
| support vector machine (SVM) and convolutional neural network(CNN) for classification[7] | RAVDESS | 78.2% |
| Residual Convolutional Neural Network (R-CNN) | FAU | 85.8% |
| Proposed SVM model | RAVDESS | 71.7% |
| Proposed Model using CNN | RAVDESS | 88.42% |

*Table 1:Comparion table*

Chapter 5

**CONCLUSION AND FUTURE SCOPE**

The proposed model was built using both SVM and CNN algorithms and was tested on the RAVDEES dataset. The highest accuracy of 85% was achieved using the CNN algorithm. There is scope for improvement by adding more data to the dataset and increasing robustness by adding more noise.

**5.1Conclusion**

The model has been built using SVM and CNN. When the model is done using SVM we obtained a accuracy of 72%. When the model is built using the CNN the accuracy we obtained is 85%. The dataset used is RAVDEES. When we done it with the other datasets the accuracy went below 70%. So we went ahead with RAVDEES data with CNN algorithm.

**5.2Future Scope**

We can improve the accuracy of the model by adding a higher size of data which we cannot be able to do due to limitations of the ability of our device. Also we can make the model more robust adding a more noise to the dataset.

**References**

1. Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, 101894.
2. Bhandari, S. U., Kumbhar, H. S., Harpale, V. K., & Dhamale, T. D. (2022). On the Evaluation and Implementation of LSTM Model for Speech Emotion Recognition Using MFCC. In *Proceedings of International Conference on Computational Intelligence and Data Engineering* (pp. 421-434). Springer, Singapore
3. Bharti, D., & Kukana, P. (2020, September). A hybrid machine learning model for emotion recognition from speech signals. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 491-496). IEEE.
4. Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, *23*(1), 45-55.
5. Kumaran, U., Radha Rammohan, S., Nagarajan, S. M., & Prathik, A. (2021). Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal of Speech Technology*, *24*(2), 303-314.
6. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, *47*, 312-323.
7. Christy, A., Vaithyasubramanian, S., Jesudoss, A., & Praveena, M. D. (2020). Multimodal speech emotion recognition and classification using convolutional neural network techniques. *International Journal of Speech Technology*, *23*(2), 381-388.
8. Sun, T. W. (2020). End-to-end speech emotion recognition with gender information. *IEEE Access*, *8*, 152423-152438.
9. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: 2013 humaine association conference on affective computing and intelligent interaction. IEEE, pp 511–516 ACII 2013 6681481.
10. Latif S, Qayyum A, Usman M, Qadir J (2018) Cross lingual speech emotion recognition: Urdu vs. western languages. In: Proceedings - 2018 International Conference on Frontiers of Information Technology, FIT 2018 8616972, pp 88–93
11. M. A. Jalal, R. Milner, and T. Hain, “Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition,” Proc. Interspeech 2020, 2020.
12. A.M. Badshah, J. Ahmad, N. Rahim, and S.W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” 2017 international conference on platform technology and service (PlatCon), pp.1–5, IEEE, 2017.