Detection of Emotion from Sound or Speech

A project plan and Literature review

Name: Divyadurga Alluri

Student id: N1037733

Supervisor Name: Sanei Saeid

Date of submission: 30/05/2022

Declaration of authorship

I am Divyadurga Alluri (N1037733), hereby declare that: [30-05-2022]

In submitting this work, I confirm that I am aware of, and am abiding by, the University's expectations for proof reading.

Abstract

The field of emotional speech has been expanding significantly for many years. Machines cannot sense or display emotions, and humans are despised for doing so. The system's ability to identify emotions is becoming more automated as a result of human-computer interaction. The need for human involvement was lessened. The author of this study examines the speech signals associated with various emotions, including joyful, sad, angry, and neutral. They have included speech analysis and several categorization techniques. The procedure for extracting and creating the different characteristics from the signals is applied to the sound speech signal. For the most effective finding technique, the author has employed data pre-processing, data cleaning, data visualization, and creating the models. The most common techniques utilized in SER detection are deep learning techniques. In order to detect emotional speech signals for this publication, the researcher used deep learning techniques including Convolutional Neural Networks (CNN), GRU (Gated Recurrent Unit), and Long-Term Short Memory (LSTM). The author has recognized the many emotions that a person experiences. They also observed emotions and extracted emotional features from speech. In the LSTM model, the author received the highest possible accuracy score of 87%. The optimal answer for this strategy was found to be the detection of emotional sound recognition and identification.

Table of Content

List of Figures

1 Part l-Project Plan

Detection of Emotion from Sound or Speech

1.1 Introduction

The emotion in a speech is the vital information that gets conveyed from the speaker to the listener. According to Micallef et al. (2022), the speaker conveys the intended emotions through acoustic cues and the perception of the cues heard by the receiver perceives the information. It is a well-known fact that of all living beings, human beings are the only beings with the ability to make different sounds using vocal cords and express different emotions. This project aims to use artificial intelligence algorithm-based approaches to detect different

and pilots.

Sound is identified mainly by the human-machine communication process. As a result, it can be assumed that the audio files will be the input to a DL model in this project and that packages linked to dealing with his audio files, as well as making and operations, will be an important duty in this study.

1.2 Human Emotions and Background

Human emotions are general and every human being's experiences show the actions of the face according to psychologists. However, the view of human emotions in the terms of psychology can be defined as something which can define the actions of the face and the characteristics of a human. As per the psychologists' terms and references it shows that there are 7 types of emotions which include beings that are disgusting, fearful surprising, happy, and even angry (Cosmides and Tobby, 2000). However, these emotions are a product of some of the emotions like feeling Pride excited and others.

Normally the basic emotions are happy, sad, fearful, and disgusting which are combined two words forming some mixed type of emotions that can lead to other emotions like love and excitement. Hence, from these evaluations, it can be said that the basic types of emotions which can be concentrated towards the later stages of the study are being happy, sad, fearful,

and disgusting. However, considering the later stages of the study it is very important to understand the faces of the humans while these emotions are detected such that it can be helpful

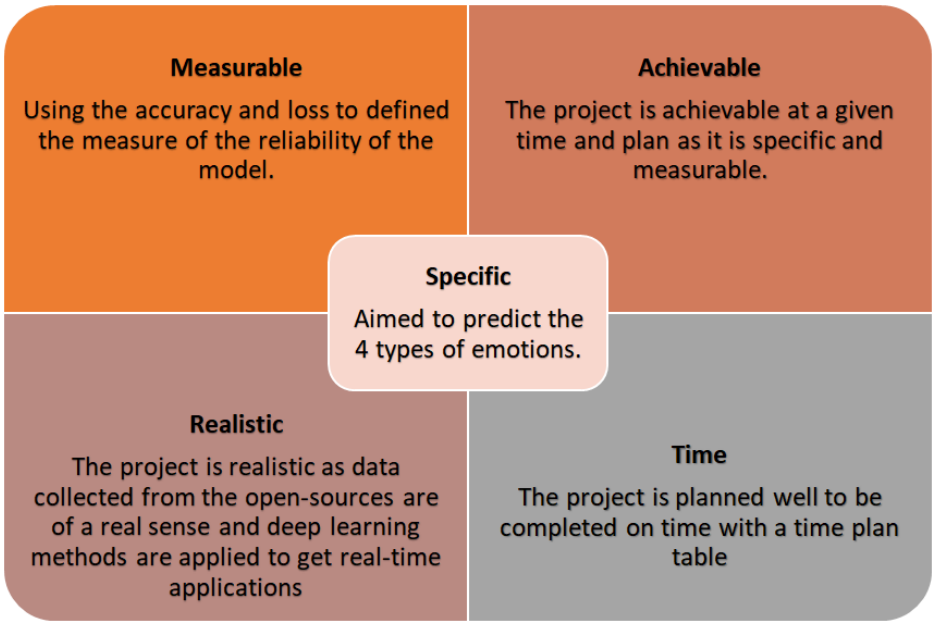
problem of data scarcity in the source data and the destination data as well (Jin et al.2022). However, deep neural networks and convolutional neural networks are used for image processing and to detect the emotion behind the speech.

1.3 Aim and Objectives

Aim: To detect emotions from sounds or speech using artificial intelligence algorithm-based approaches.

Objectives:

To understand the current systems adopted to understand human emotions via vocal


Measurable
Using the accuracy and loss to defined the measure of the reliability of the model.
Achievable
The project is achievable at a given time and plan as it is specific and measurable.
Specific
Aimed to predict the 4 types of emotions.
Realistic
The project is realistic as data collected from the open-sources are of a real sense and deep learning methods are applied to get real-time applications
Time
The project is planned well to be completed on time with a time plan table

with some historical data, will not be able to detect those emotions. which brings back the fact of being less ethical and less professional in which one can improve eyesight by considering highly reliable DL models and more and more data.

Socially this kind of application will bring many conclusions because if we talk about speech and detecting emotion from the speech in the context of someone crying in happiness becomes

very different and can give a wrong social message to people (Cowie, 2015). Hence, this emotion detection system has its own disadvantages but however when the DL models are incorporated with more and more examples of data this kind of misunderstanding socially professionally as well as physically can decrease.

1.8 Time Plan

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Introduction												
Literature Review												
Methodology												
Experiment Implementation												
Results												
Analysis and Discussions												
Conclusions												
Final Submission												

2 Part ll – Literature Survey

Speech processing techniques for identifying the emotions of a person using deep neural networks

2.1 Overview

In this section the author discussing about the review of literature, in the section 2.2 the author is describing about different methods used and identified for the emotions from the sounds. In the section 2.3 the researcher has discussed various DL methods are used in the identification of emotion from the speech from the various authors. In the 2.4 section, researcher has described about the voicing preprocessing from the sounds, and the process of implementation to detect the speech. In the section 2.5 the author is describing about the different author methods, results, and the identification of gaps for the various research. In section 2.6, the author has described the Existing Approach of these project as the comparing from the different author's research. In 2.7 section, the author discuses that these overall section of the literature from the various emotions of a person.

2.2 Speech Processing Techniques for Identifying the Emotions from Sound or Speech

Although there are many different kinds of speech processing algorithms, they still need to be improved in order to be accurate with low-quality or noisy signals (Xiao et al., 2022). The design of a computer system that aids in the recognition of the voice and words involved in

2	Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features (Abdel-Hamid, 2020)	Different ML techniques used in that case such as SVM and KNN.	The highest accuracy obtained by the author is approx. 98 percent.	Here the author created multiple models for multiple emotions but he didn't create a single model for this problem statement.
3	Speech emotion recognition considering nonverbal vocalization in affective conversations. (Hsu *et al., 2021)*	Different DL techniques used along with ResNets and LSTM model.	The highest accuracy obtained 61.92 percent accuracy using these techniques.	The accuracy is very less for this model.
4	Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network. (Puri *et al.,* 2022)	Here the author used CNN, DNN and LSTM models.	The author obtained almost 98 percent accuracy through these models.	Not applicable.
5	Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. (Alsabhan, 2023)	Here the author used 1D and 2D CNN models along with LSTM model.	The author obtained highest accuracy of 96 percent in different dataset.	Not applicable.

2.6 Existing Research

In these contexts, the author has described the two methods such as one is the traditional method, and another one is the modern approach. The traditional method is the various author was described about past researches about the emotion from the speech, and the modern methods are the present and future approaches of the researches.

2.6.1 Traditional Approach

To determine whether speech has been digitalized, the three main elements of signal pre-processing, future extraction, and classification are contrasted in the conventional recognition system. Segmentation is done via acoustic pre-processing, which starts with the unit of the signal. It is determined how much speech emotion is processed overall for each classification (Atmaja et al., 2019). The input data comprises of stage-specific speech recognition, feature extraction and selection based on their properties, the availability of data from the various speech emotions, measurement and computation of the sounds in SER, and emotion recognition from the speech.

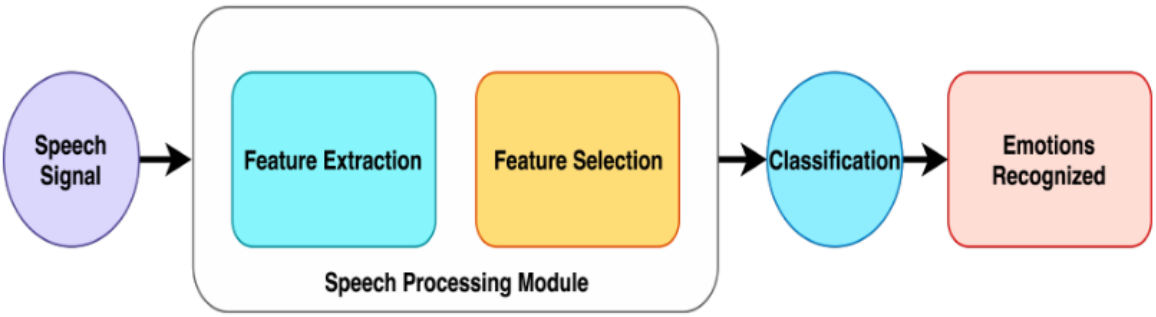


Figure 2.1 Traditional approach of the emotion from speech recognition

Source: Khalil *et al*., (2019)

2.6.2 Modern Approach

The detection of emotions from voice recognition can be done using a variety of modern techniques, including machine learning (ML), deep learning (DL), artificial intelligence (AI), and neural networks. Saravanan et al., (2019), The recognition system to distinguish between the different emotions of the four classes happy, sad, angry, and neutral to get an accuracy to be increased more to adapt from their feature extraction of the ANN approach, neural networks to detect and distinguish between the various emotions of the sounds.

2.7 Summary

Here the author gives a brief description of different techniques that are available to solve this problem of emotion detection from speech. After that describes along with different research proof related to this problem statement and that are the techniques they are used also contained in this section. Apart from that here the author also shows the gaps applied by the different researchers and some of the traditional and modern approaches to solve this problem. And in the next chapter, the author going to discuss different methods that are used to successfully complet8ing this work study.

3 Methodology

3.1 overview

In the part of the methodology, section 3.1 is overviewing all the sections in detail explanation of the methods. 3.2 section is about the data explanation in this topic the four data types are explained in a detailed way. In the 3.3 section, the method Librosa was explained as to how it was used and worked in speech recognition. In the 3.4 section, the library of TensorFlow how was used, and how it is working with the ML algorithm, in the 3.5 section, what is Keras, and how is it used in emotion recognition, in the 3.6 section, Long Short-Term Memory (LSTM) network and its applications and what is the logic behind it was explained in a detailed way. In the 3.7 section, the activation function was used on how it works, In the 3.8 section, the loss function categorical cross entropy how was used in ML and artificial intelligence, In the 3.9 section the Optimizer was used for the DL method and how it converting the data and how it was used in the emotion recognition was explained, and in the 3.10 section, the overall summary of the methodology was discussed in this part for the purpose of next step to get the results.

3.2 Data Explanation

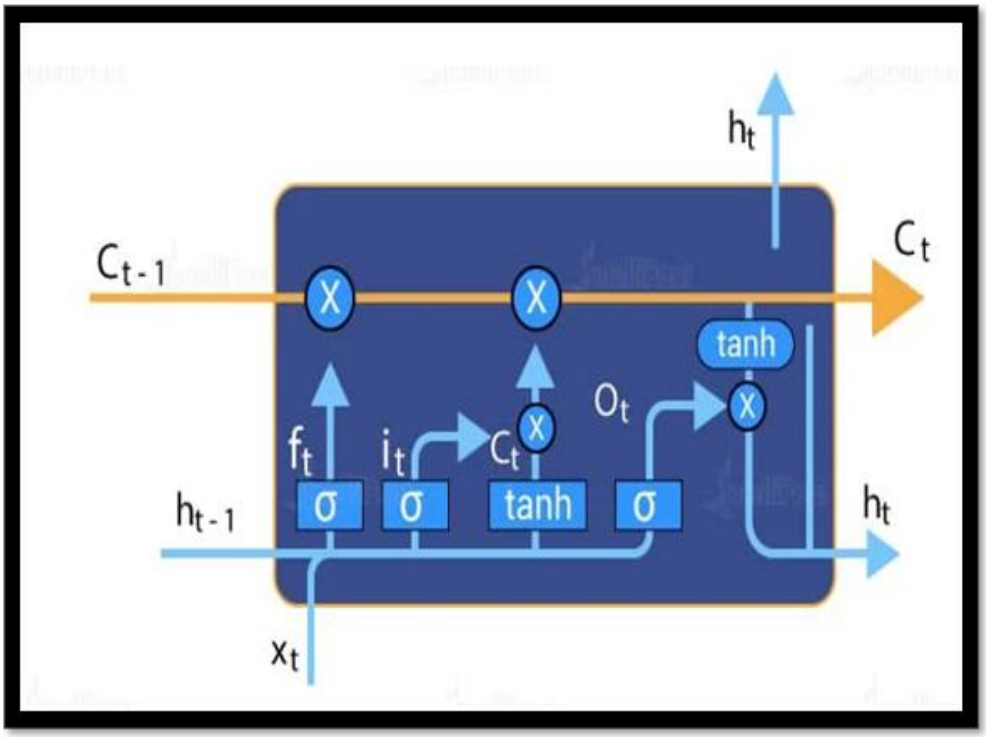
The emotion or recognition of a speech has four different types of databases such as like Ravdess, crema, Tess data, and crema-D data, which are given below:

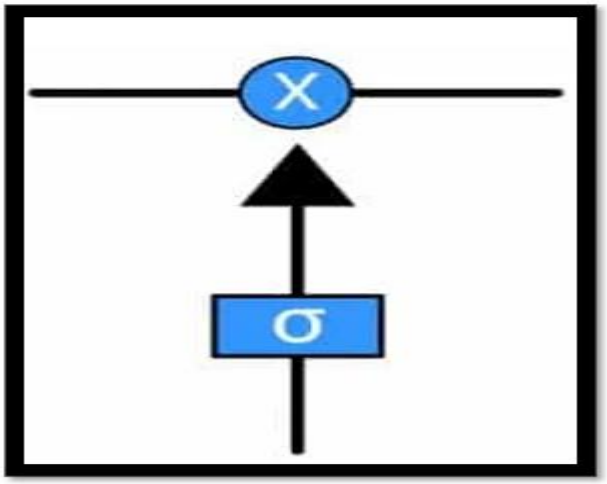
3.2.1 RAVDESS Data

The RAVDESS data is known for the validation of multimodal databases of the recognition of speeches, and sons (Livingstone and Russo, 2018). This data is a consisting for the purpose of actors, lexically it is vocalizing, and the neural north American accent is matching the statements. Ravdess is a method for the emotional speech audio for the dataset. The Ravdess is also known as Ryerson audio-visual database o emotional speech and song.

3.2.2 Surrey Data

Surrey Audio-visual expressed emotion is also known as SAVEE. The surrey audio-visual dataset explores or identifies an emotion for the interested one to have male data and a high level of qualitative audio. The male has a speaker for the imbalance of the representation of emotional data, the data will be advisable for the compliment while comparing to the other datasets with the more female speakers containing an emotional expression of the data.


h_t
C_{t-1}
C_t
X
X
tanh
O_t
X
f_t
i_t
C_t
X
h_{t-1}
σ
σ
tanh
σ
h_t
x_t


X
σ

can only be of limited assistance as Deep Neural Networks grow larger because more advanced techniques are needed to achieve outstanding results.

3.11 Summary

The overall summary of a methodology discussed all the models and methods connected with the ML and DL method, it also discusses the advantages and disadvantages, and applications of the functional methods to get maximum outcomes from it. The author discussed the overall

methodology section, for the better-obtained outcomes for the project discussed to get the best suitable method and technique was used to get an optimal solution in the results and analysis part, the project was explained about the analysis and overall best method used in the Results part for the solutions.

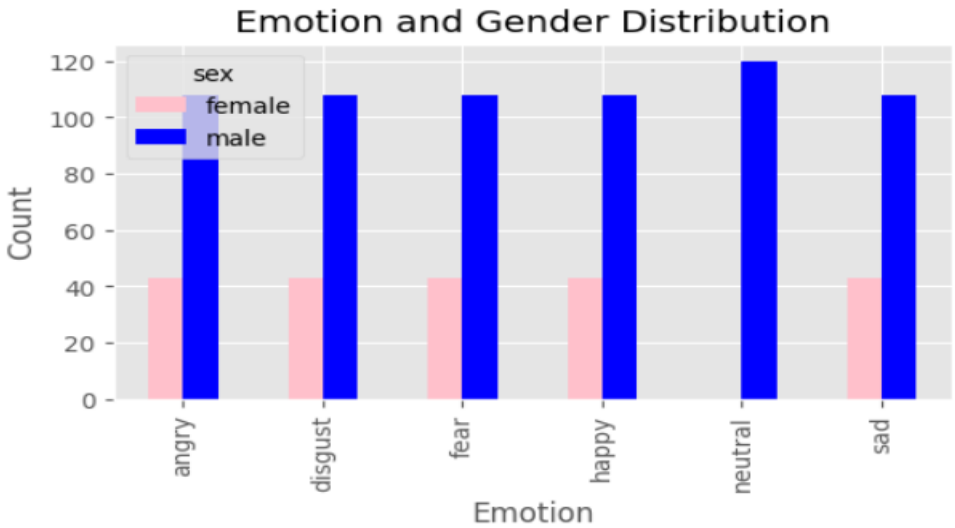
4 Result and Analysis

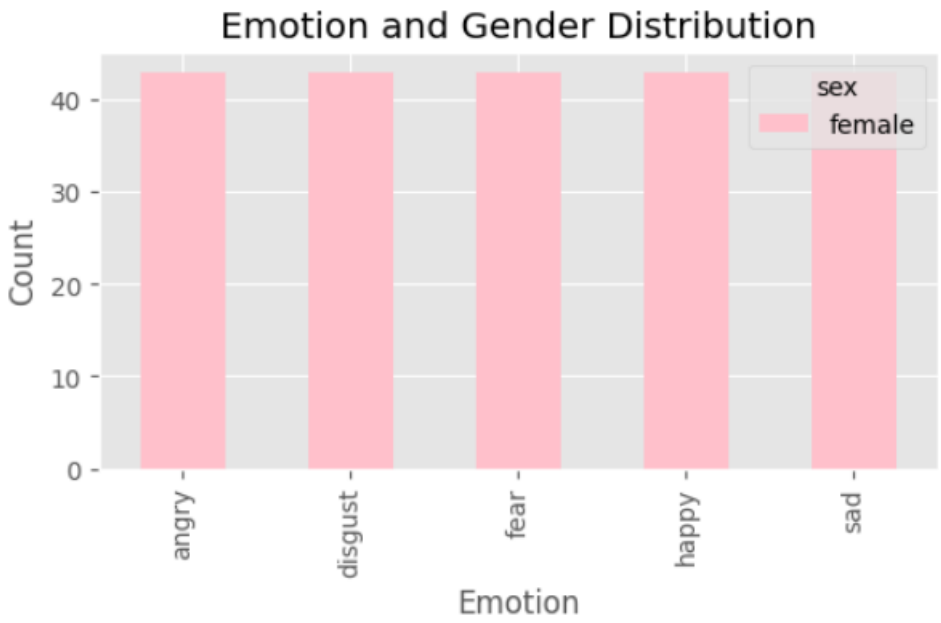
4.1 Overview

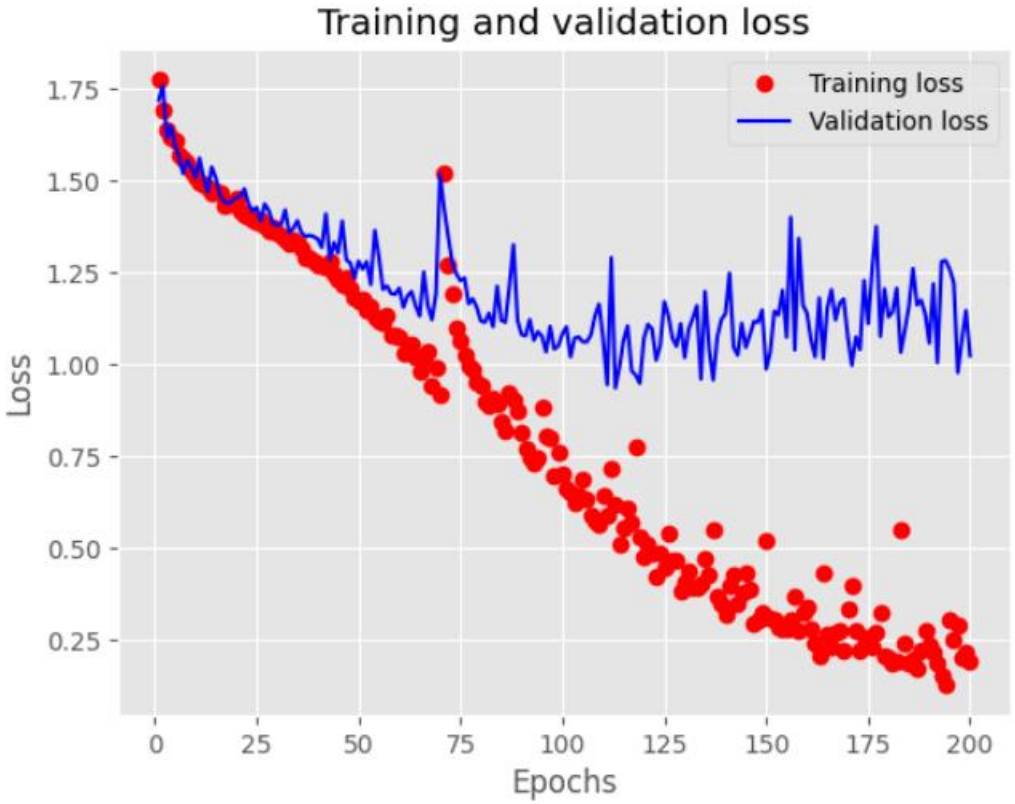
This chapter covered details information about the result and analysis of Voice Recognition. At the start of this chapter, the author provides some gist about the entire chapter. The author all the information about the dataset in section 4.2. Here, the author shows where they collect this dataset and what information is available inside the data. also, the author provides the source link of this selected data. After completing the data description, the author gives all the implementation steps in section 4.3 to execute the entire program. Because everyone knows that implementation steps are different for every program. So, this section helps to understand what techniques are used to run this project. Libraries / Packaged always played a vital role in every data science project and this project also run with the help of the library. So, before starting the code, the author added all the required libraries and all the library-related information is available in section 4.4. Another important thing is the dataset, with the help of the dataset model properly learn and after that, it is capable to test the model. All the data loading processes are available in section 4.5. Data pre-processing is one of the important things because sometimes more useless information is added and the data is not pure. So, if this type of data is directly used for creating a model, then the model quality is destroyed. So pure the data, the author used many techniques in this project and this type of processing related information is available in section 4.6. The author describes how split the data and for what reason they split the data in section 4.7. After completing all the processing steps, the author describes which techniques were used to build the model in section 4.8. In section 4.9, the author finalizes the model and shows how many percentages were found. After completing all this thing, the author also gives a small summary for concluding this chapter in section 4.10.

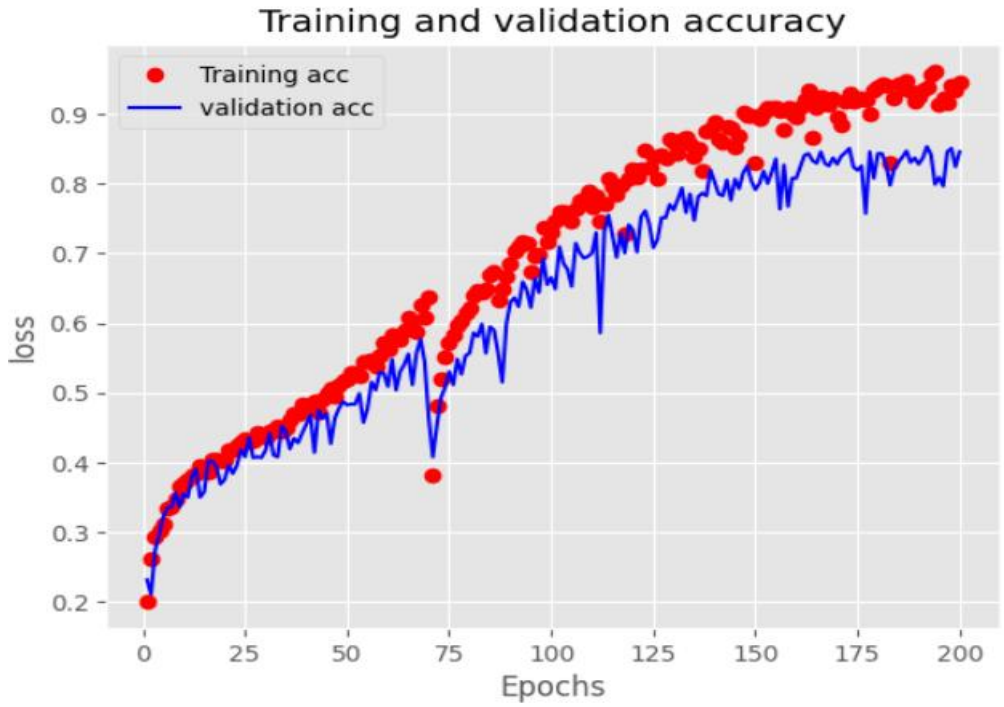
4.2 Data Description

In this work-


Emotion and Gender Distribution
sex
female
male
Count
120
100
80
60
40
20
0
angry
disgust
fear
happy
neutral
sad
Emotion


Emotion and Gender Distribution
sex
female
Count
0
10
20
30
40
angry
disgust
fear
happy
sad
Emotion


Training and validation loss
Training loss
Validation loss
Loss
Epochs
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
25
50
75
100
125
150
175
200




```
9/9 [==============================] - 0s 16ms/step - loss: 0.3867 - Accuracy: 0.9688
[0.38671982288360596, 0.96875]
```

by the models. The performance result of these models is given in below:

```
25/25 [==============================] - 1s 9ms/step - loss: 1.6957 - Accuracy: 0.3077
[1.695735216140747, 0.3076923191547394]
```

Figure 4.6 Model performance without validation data

It can be seen in Fig. 4.6 that the accuracy is very less in this model.

```
25/25 [==============================] - 0s 14ms/step - loss: 2.0210 - Accuracy: 0.1438
[2.021017551422119, 0.1437578797340393]
```

Figure 4.7 Model performance with validation data

It can be observed that there is worst performance by the model compared to without validation data and it can be seen in Fig. 4.7.

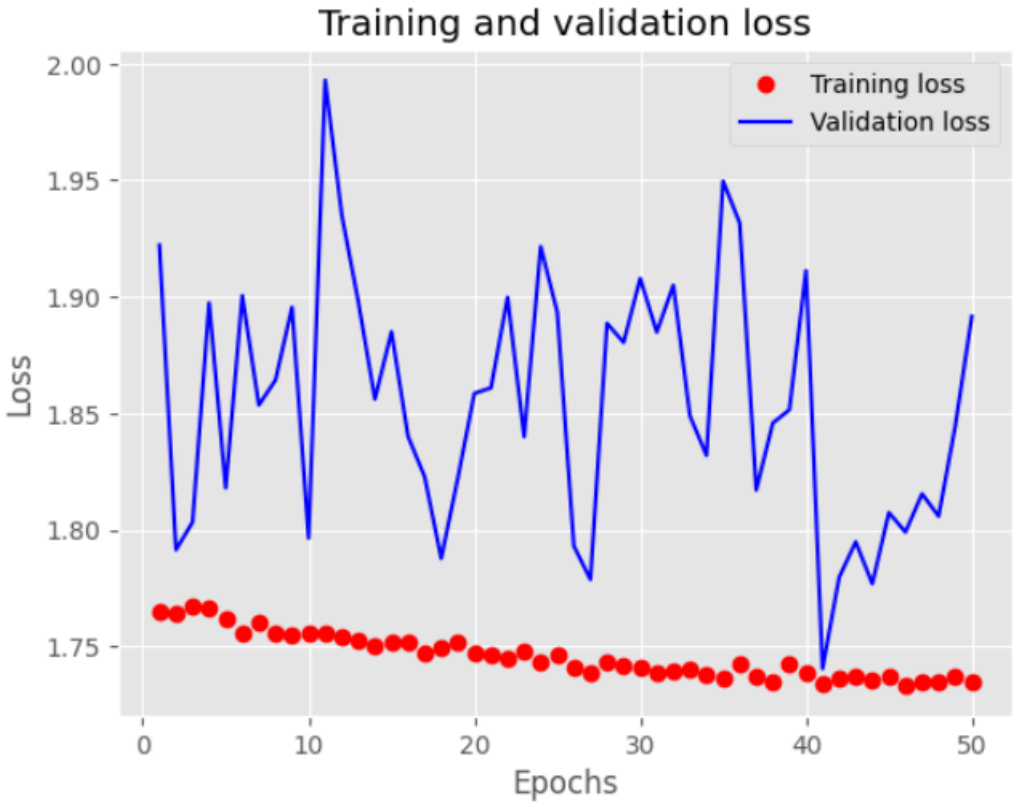


Figure 4.8 Training and validation loss comparison of GRU model.

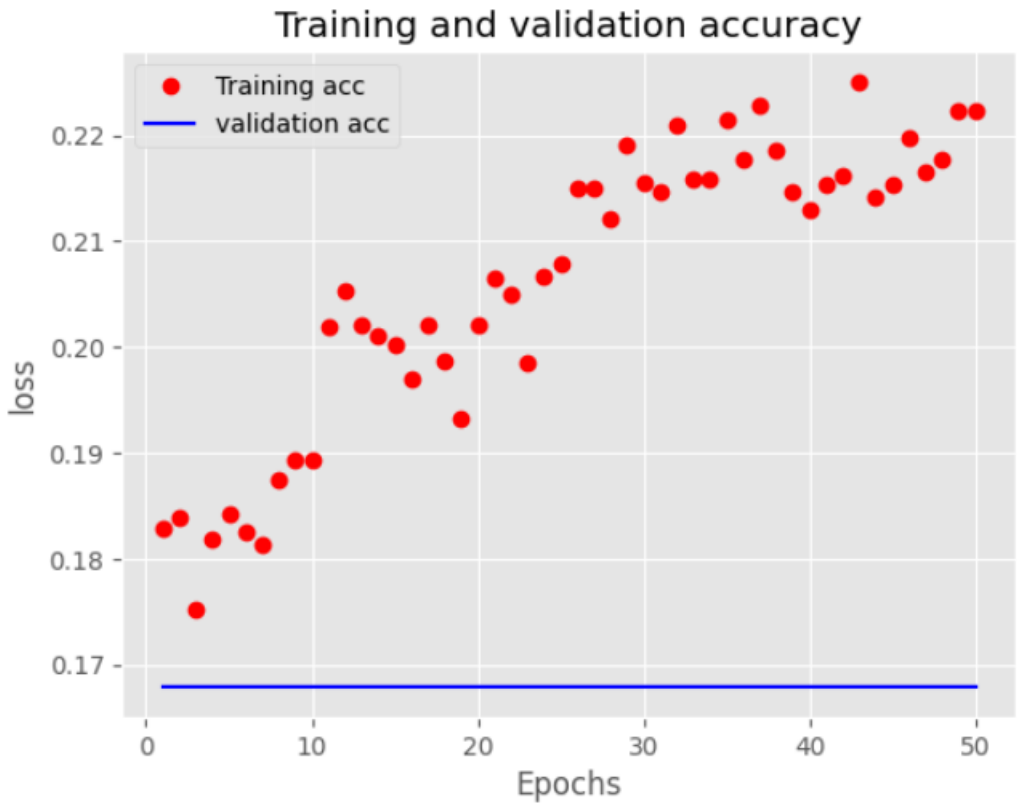


Figure 4.9 Training and validation accuracy comparison of GRU model.

4.12 Summary

In this chapter, the author tries to provide more valuable information about programming as well as a predictive model. Here the author provides complete data information as well as a summary of implementation steps means how they complete this project or the working process of this project. During this project, the author found that the entire data is audio-related. so, the author used some pre-processing techniques to properly process the dataset. After completing data processing, the author used the LSTM classifier to train the model, and after training, the model the author got 87% accuracy for getting this much accuracy the author used 200 epochs in this project. But with the help of the GRU model, the author gets only 14 percent of accuracy. That's why the author finalizes the LSTM model for this problem statement In the next chapter, the author concludes this project and also author recommends some important things about this project. Also, the author gives some future scope at the end of the chapter.

5 Conclusion & Recommendation

5.1 Conclusion

This work study is very knowledgeable and interesting because here finding the voice emotions using DL techniques. Here, after applying the different pre-processing steps author creates an appropriate model which can the ability to predict the best optimal output. But the most crucial parts of this project are the data loading and processing the data. Here the author uses CNN techniques and gives the 200 epochs while training the model. And it is observed that after

to develop computers that are highly developed.

6 References

Abdulmohsin, H.A., 2021. A new proposed statistical feature extraction method in speech emotion recognition. Computers & Electrical Engineering, 93, p.107172.

Abdel-Hamid, L., 2020. Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features. Speech Communication, 122, pp.19-30.

Al Bataineh, A. and Kaur, D., 2021. Immunocomputing-based approach for optimizing the topologies of LSTM networks. IEEE Access, 9, pp.78993-79004.

Alsabhan, W., 2023. Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. Sensors, 23(3), p.1386.

Andalibi, N. and Buss, J., 2020, April. The human in emotion recognition on social media: Attitudes, outcomes, risks. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-16).

Ashraf, H.,

Zhang, T., Shao, Y., Wu, Y., Geng, Y. and Fan, L., 2020. An overview of speech endpoint detection algorithms. Applied Acoustics, 160, p.107133.

Zhou, T., Wang, Y., Zhu, Q. and Du, J., 2022. Human hand motion prediction based on feature grouping and deep learning: Pipe skid maintenance example. Automation in Construction, 138, p.104232.