INDIAN INSTITUTE OF TECHNOLOGY INDORE

UNDERGRADUATE THESIS

Speech Emotion Recognition

Author:
SRIJAN SAINI 180001056
JEMIN VAGADIA 180001023

Supervisors:
Dr. ABHISHEK SRIVASTAVA

Thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Technology

in the

Department of Computer Science and Engineering

May 26, 2022

Declaration of Authorship

We, JEMIN VAGADIA and SRIJAN SAINI declare that this thesis titled, "Speech Emotion Recognition" and the work presented in it is our own. We confirm that:

- This work was done wholly or mainly while in candidature for the BTP project at IIT Indore.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.

Jemin

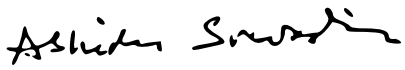
Jemin Vagadia

Srijan

Srijan Saini

Certificate

This is to certify that the thesis entitled, "*Speech Emotion Recognition*" and submitted by Jemin Vagadia ID No 180001023 and Srijan Saini ID No 180001056 in partial fulfillment of the requirements of B.Tech Project embodies the work done by them under my supervision.



Supervisor

Dr. ABHISHEK SRIVASTAVA
Professor,
Indian Institute of Technology Indore
Date:

"As for the future, your task is not to foresee it, but to enable it."

-Antoine de Saint Exupery

INDIAN INSTITUTE OF TECHNOLOGY INDORE

Abstract

Department of Computer Science and Engineering

Bachelor of Technology

Speech Emotion Recognition

Speech is one of the most natural ways for humans to express themselves. We rely on it so much that we notice its relevance while using other modes of communication, such as emails and text messages, where we frequently utilise emoticons to describe our feelings. Because emotions are so important in communication, detecting and analysing them is critical in today's digital age of remote communication. Because emotions are subjective, detecting them is a difficult task. There is no universal agreement on how to quantify or classify them. In this Project, we used the most widely used audio features like MFCC, MEL spectrogram and Chroma to classify each and every emotion. We test our algorithm's performance with different models to find the best fit for our use case. The dataset used in the Project is RAVDESS which is by far the best dataset available for the underlying problem of emotion detection. It contains in total of 1440 samples of audio detection. Atlast, we have shown our results with accuracies of different models, their confusion and classification matrix.

Acknowledgements

We would like to thank our B.Tech Project supervisor **Dr. Abhishek Srivastava** for their guidance and constant support in structuring the project and their valuable feedback throughout the course of this project. Their overseeing the project meant there was a lot that we could learn while working on it. We thank them for their time and efforts.

We are thankful to our friends and family members who were a constant source of motivation for us throughout this project.

Lastly, We offer our sincere thanks to everyone who helped us to complete this project, whose name we might have forgotten to mention.

Contents

List of Tables

List of Abbreviations

SVM	**S**upport **V**ector **M**achine
RBF	**R**adial **B**asis **F**unction
RFC	**R**andom **F**orest **C**lassifier
MLP	**M**ulti **L**ayer **P**erceptron
LBFGS	**L**imited Memory **B**royden **F**letcher **G**oldfarb **S**hanno
SGD	**S**tochastic **G**radient **D**escent
MFCC	**M**el **F**requency **C**epstral **C**oefficients

Dedicated to all the novice learners.

Chapter 1

Introduction

Speech has always been a primal way to communicate and to understand the emotions or thoughts of each other. It is a way to communicate with other human beings using their language. With the current advancement in technology, new techniques are developed on a daily basis to understand the human emotions from the audio and a lot of research is going on in this field. The basic premise underlying emotion recognition is to examine the acoustic differences that occur when pronouncing the same thing in different emotional conditions. The research looks at how human voices can be used to classify different emotions. While facial expressions and movements are the easiest way to identify one's emotions, detecting them becomes increasingly difficult as a person gets older, as people learn to regulate their expressions. Furthermore, expressions and movements only convey outer emotions like joy,anger and grief, but not emotions like disgust, boredom, and so on.

Emotionally-aware robots could deliver appropriate responses and display emotional personalities. Humans could be replaced in some situations by computer-generated characters capable of conducting a highly genuine and convincing conversations by appealing to human emotions. Speech-based emotions must be understood by machines, only with this skill a truly meaningful discourse based on mutual trust and understanding between humans and machines can be realised. Machine learning (ML) has traditionally involved calculating feature parameters from raw data. The features are then used to train a model that learns to provide the output labels. The selection of characteristics is a common challenge with this strategy. In general, it's unclear which characteristics contribute to the most effective data grouping into multiple categories (or classes). Testing a large number of different features, integrating diverse features into a single feature vector, or using various feature selection approaches can provide some insights. The quality of the self-made or self-crafted features can have a big impact on classification results.

Speech is an extremely rich data source. Depending on the sample rate the number of points sampled per second to quantify the signal one second of data could contain thousands of points. Scale this up to hours of recorded audio, and we can see how Machine Learning and Data Science nicely intertwine with signal processing techniques. Speech Emotion Recognition (SER) is an act to determine the human emotion and effective states from speech. Humans have evolved overtime to express their emotions through speech by changing the tone and pitch of their voice. Because of this phenomenon, even the animals are able to understand our emotions. SER through audio analysis is also a cost effective way compared to textual analysis and could save a lot of resources for companies. However, detecting emotions from audio is very challenging considering the wide variety of ways to express the same emotion and getting that information from audio is not always trivial.

The most important aspect of classifying emotions from audio is to create a feature vector that should be capable of storing the relevant and the essential information of the audio signal. The most predominantly used features used for this purpose are Mel-frequency Cepstral Coefficients(MFCC), MEL Spectrogram and Chroma. In our project we have used the combination of them to get the best accuracy for our model. We also did a lot of preprocessing to avoid any unnecessary information. We used the RAVDESS dataset, a multi-modal database of emotive speech and music that has been validated. The dataset includes 24 professional actors who vocalise lexically-matched sentences in a neutral North American accent.

Chapter 2

Literature Survey

Emotion plays a crucial role in everyday interpersonal human relationships. By expressing our sentiments and providing feedback to others, it helps us match and comprehend the feelings of others. Because of its many uses, such as security, audio surveillance, medical, education, assisted living environments, and so on, speech emotion recognition has become a significant problem. As a result, automatic emotion detection has emerged as a new field of study with the primary goal of comprehending and recalling desired sentiments.

Several techniques have been investigated in the past to recognise emotional states, including facial expressions [1], speech [2], physiological signals [3], and so on. Several speech features have been proposed by researchers [4] which contain emotion information like energy, pitch, Mel-frequency cepstrum coefficients (MFCC), mel, and modulation spectral features (MSFs) [5]. As a result, most researchers choose to employ a combined feature set, which is made up of a variety of features that contain more emotional data [6].

The most commonly used audio features are MFCCs, which are coefficients produced from a form of cepstral representation of an audio clip (a nonlinear "spectrum-of-a-spectrum"). [7]

Recently, researcher have proposed many classification algorithms in speech emotion recognition, such as support vector machine (SVM) [8] [9] [10] [11] [12], Gaussian mixture model (GMM) [13], neural networks (NN) [14], and recurrent neural networks (RNN) [15] [16] [17]

Convolutional networks with thick layers are also used in the majority of current sentiment analysis studies. As a result, audio files must be of a specific size. [18], [19]. Howewer, using an architecture based on Fully Convolutional Neural Networks, a method capable of analysing audio files of any size is proposed, without the need for that size to be determined beforehand (FCN). [20]

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is the most commonly used dataset in this field. It consists of 24 people's audio and visual English recordings and is gender balanced. This project focuses only the audio part of the dataset which is divided into 8 emotional classes: sad, happy, angry, calm, fearful, surprised, neutral and disgust. [21]

Chapter 3

Proposed Work

This chapter consists the methodology we used in our project.

3.1 Workflow

We first extracted the features of all the audio samples from the dataset and then we preprocessed it to make it as useful as possible by removing all the unnecessary information and null data. After that, we created the feature vectors of the audio samples which later on used to train the model. Eventually, we analyze it on our test dataset.

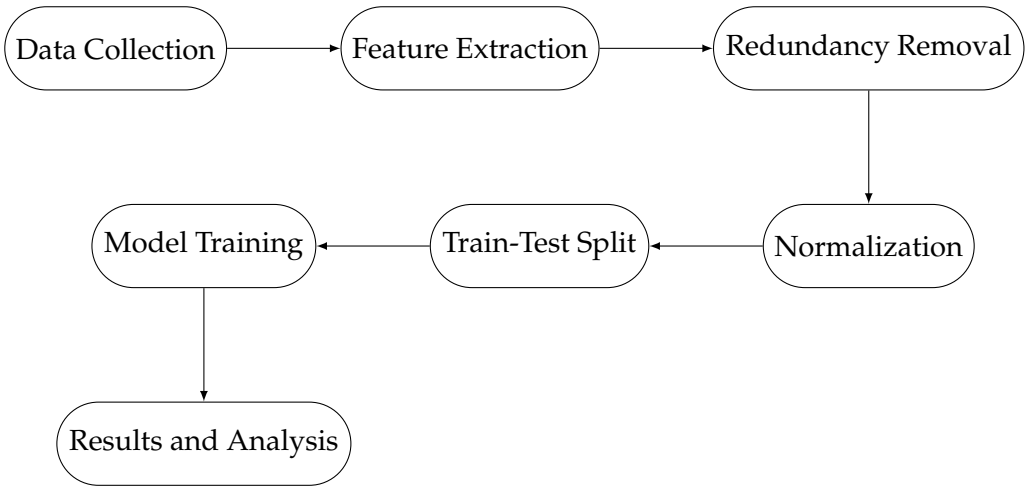
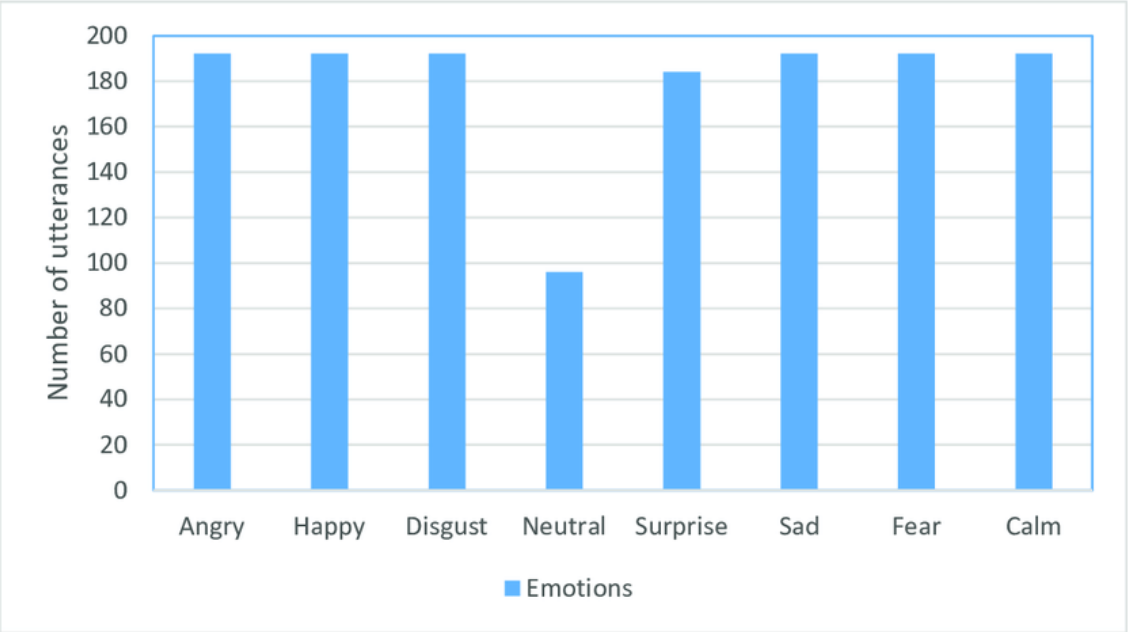


FIGURE 3.1: Proposed Workflow of Project

3.2 Dataset

The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. In total there are 1440 samples for the Speech that we could use for the domain of our problem. We did a train-test split of 80:20 and 1152 samples are used for training and 288 samples for testing. Furthermore, we randomly distributed the samples between the training and testing dataset along with maintaining the same proportions of each label as in input dataset. Distribution graph of the emotions in the input dataset is shown below:



3.3 Audio Features

As discussed in the introduction, the most predominantly used Audio features for classification are MFCC, Mel Spectrogram and Chroma. Along with that there are some more features which are very much useful to understand the waveform of the signal. For ex- Pitch, Zero-Crossing Rate etc.

3.3.1 Mel-Frequency Cepstral Coefficients

Any voice that we generate depends upon the shape of the vocal tract. If we somehow determine that shape we can also predict the emotion corresponding to that sound signal. Mel-frequency cepstral coefficients are the coefficients that make up an MFC (MFCCs). They're made from a cepstral representation of an audio clip. On the mel scale ,the MFC's frequency bands are evenly separated, which more closely approximates the human auditory system's response than the normal spectrum's linearly split frequency bands, which is the difference between the cepstrum and the mel-frequency cepstrum. For example, in audio compression, frequency warping can help to better portray sound.

MFCCs are commonly derived as follows:

- Divide the signal into windows or frames.

- Evaluate each window's Discrete Fourier Transform.
- Use triangle overlapping windows or cosine overlapping windows to map the powers of the spectrum found above onto the mel scale.
- Calculate the power's log at each of the mel frequencies.
- Calculate the DCT of the list of mel log powers, as if it were a signal.
- The MFCCs are the resulting spectrum's amplitudes.

3.3.2 Mel Spectrogram

It is made up of two words Mel and Spectrogram. They are as follows:

- The mel scale is a perceptual scale of sounds that listeners interpret to be equally spaced apart. Assigning a perceptual pitch of 1000 mels to a 1000 Hz tone defines the reference point between this scale and normal frequency measurement.
- A Spectrogram is a figure which represents the spectrum of frequencies of a recorded audio overtime.

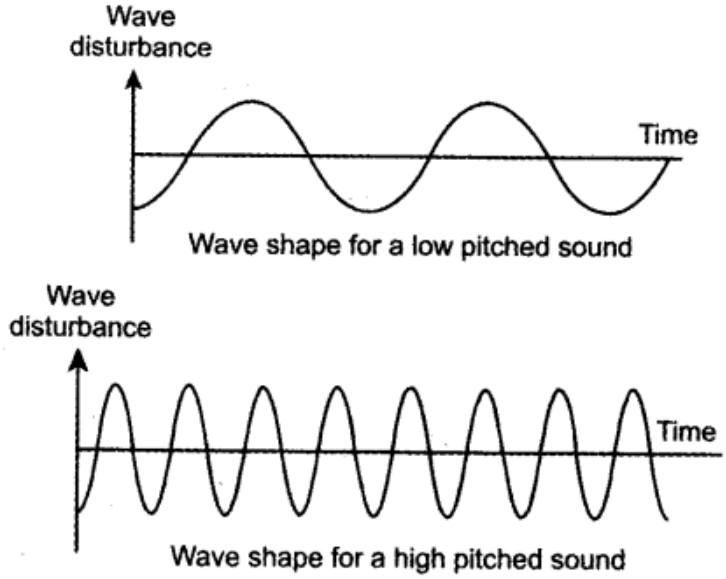
We convert the spectrogram frequency to the mel scale which is much more relevant and useful comparatively. There are around 128 mel spectrogram features in feature vector.

3.3.3 Chroma

The chroma is a condensed description that represents the tonal component of a musical audio source. The word chroma feature or chromagram refers to the twelve various pitch classes in music. Chroma-based characteristics are an effective method for assessing music with meaningfully grouped pitches. Chromatic and melodic aspects of music are captured by chroma features, which are resistant to changes in timbre and instrumentation.

3.3.4 Pitch

A listener allocates musical tones to relative positions on a musical scale based mostly on their sense of vibration frequency. Although it is linked to frequency, the two are not the same thing. Pitch is a personal preference that cannot be quantified whereas frequency can be exactly calculated.



3.3.5 Zero-Crossing Rate

It is the number of times the amplitude of an audio signal passes through a value of zero in a certain time interval frame.
Importance of this feature:

- This property is important for classifying percussive sounds and has been used extensively in speech recognition and music information retrieval.

Chapter 4

Results and Discussions

In this section, we will analyse the performance of different Machine learning models on the RAVDESS dataset. These results are based on our final combination of feature vectors: MFCC + mel (168 data columns). The total data is split in the ratio of 4:1 for training and testing. For the computation of these results, we used Jupyter notebook (Pycharm) for running the code.

4.1 Support Vector Machine

Support Vector Machine (SVM) works very well in high dimensionality problems. Hence, it gave one of the best results for these 168 dimensional data input. The following results are computed by varying the kernel function parameter:

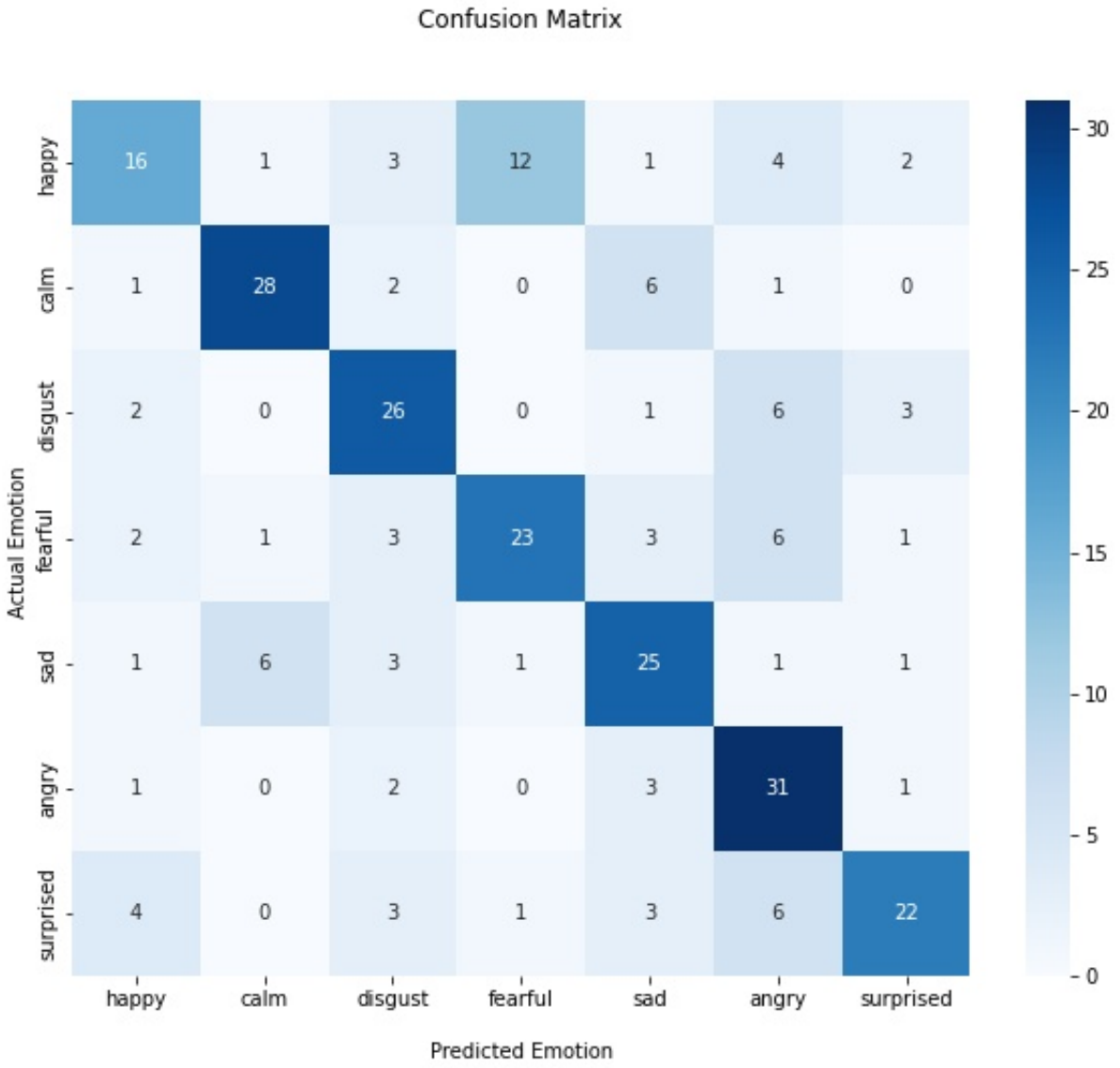
Support Vector Machine	
Kernel Name	Accuracy
Linear	45
Polynomial(Degree=2)	59
Polynomial(Degree=4)	53
Polynomial(Degree=6)	32
Polynomial(Degree=8)	25
Polynomial(Degree=10)	22
Rbf	64

TABLE 4.1: SVM kernel Results

Its clearly seen that rbf kernel is the best match for the given data. To further analyse and study the outcomes, classification report and confusion matrix are shown for the rbf kernel.

SVM - Classification report				
Emotion Class	Precision	Recall	f1-score	support
happy	0.59	0.41	0.48	39
calm	0.78	0.74	0.76	38
disgust	0.62	0.68	0.65	38
fearful	0.62	0.59	0.61	39
sad	0.60	0.66	0.62	38
angry	0.56	0.82	0.67	38
surprised	0.73	0.56	0.64	39

TABLE 4.2: SVM Classification Report



Here most of the entries are closely packed in the diagonal. However we can still detect some minor anomalies present in some of the classes. For 'Happy' class, there are 12 entries which should are actually in the happy class are predicted as fearful due to similarity in the audio wave structure. Similarly, some emotion classes are showing some overlap due to similarity in tone and pitch while expression of these emotions.

4.2 Random Forest Classifier

The RFC employs averaging to increase predicted accuracy and controls over-fitting by fitting a number of decision tree classifiers on various sub-samples of the dataset. Here, the following results are computed by varying the number of decision tress:

Random Forest Classifier	
Number of estimators (Trees)	Accuracy
100	54
200	56
300	59
400	61
500	60

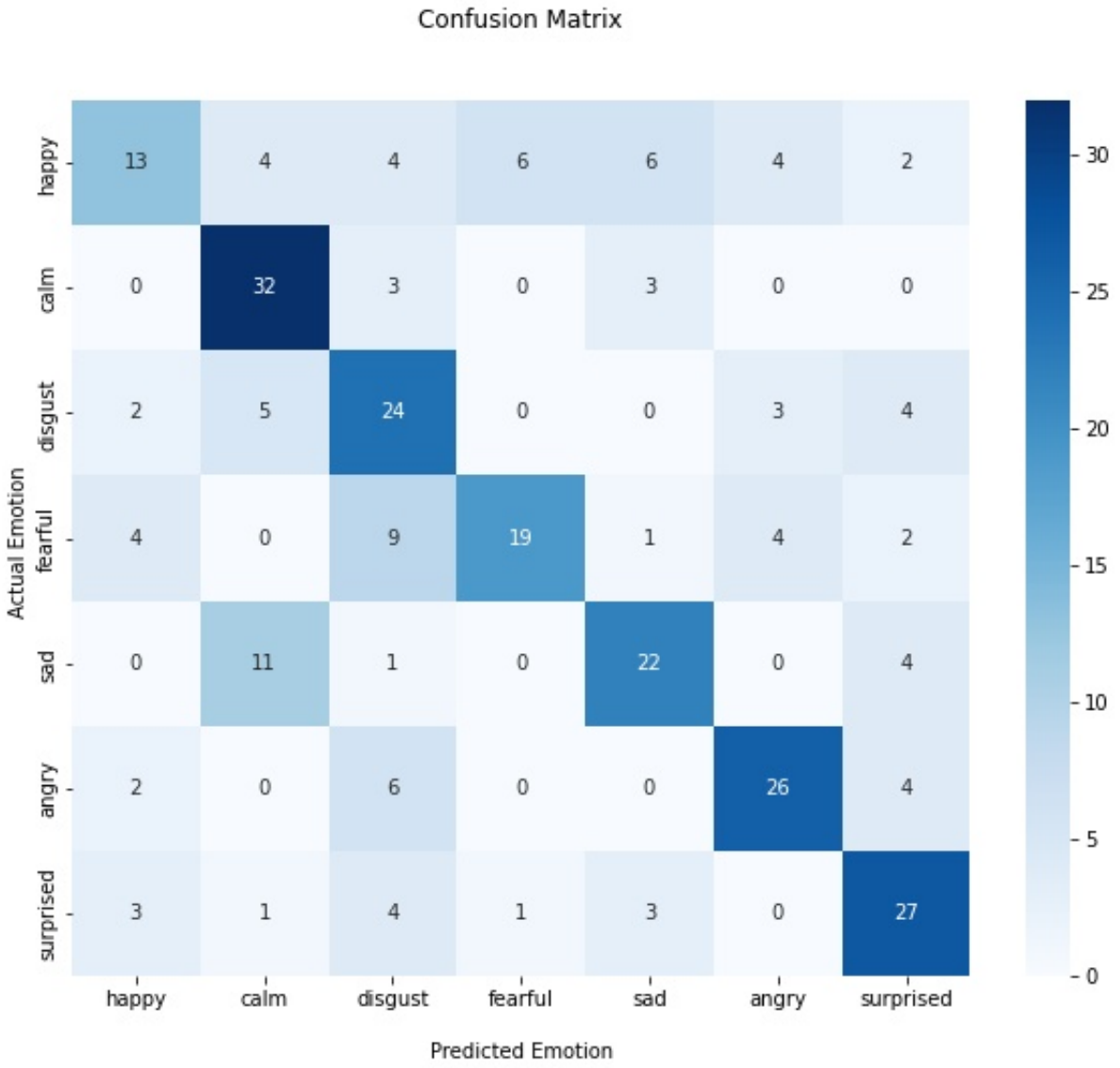
TABLE 4.3: RFC Results

Here, we are getting the best results when we use entropy criterion and 400 number of estimators. Hence, to further analyse the outcomes, the classification report and confusion matrix are computed for the best set of parameters.

SVM - Classification report				
Emotion Class	Precision	Recall	f1-score	support
happy	0.54	0.33	0.41	39
calm	0.60	0.84	0.70	38
disgust	0.47	0.63	0.54	38
fearful	0.73	0.49	0.58	39
sad	0.63	0.58	0.60	38
angry	0.70	0.68	0.69	38
surprised	0.63	0.69	0.66	39

TABLE 4.4: RFC Classification report

Here, in the confusion matrix shown below, we still have most of the entries along the diagonal which is a good performance sign. However, we can still see the similar type of anomalies as we had seen in svm. There is a significant overlap between calm and sad classes as well as disgust and fearful classes. We can also see some minor overlaps between other classes.



4.3 Multi-Layer Perceptron Classifier

MLP is a classic neural network classifier which runs iteratively to minimize the loss function according to the optimizer used. It gave the best results so far based on accuracy and confusion matrix. The following results are computed by varying the type of optimizer:

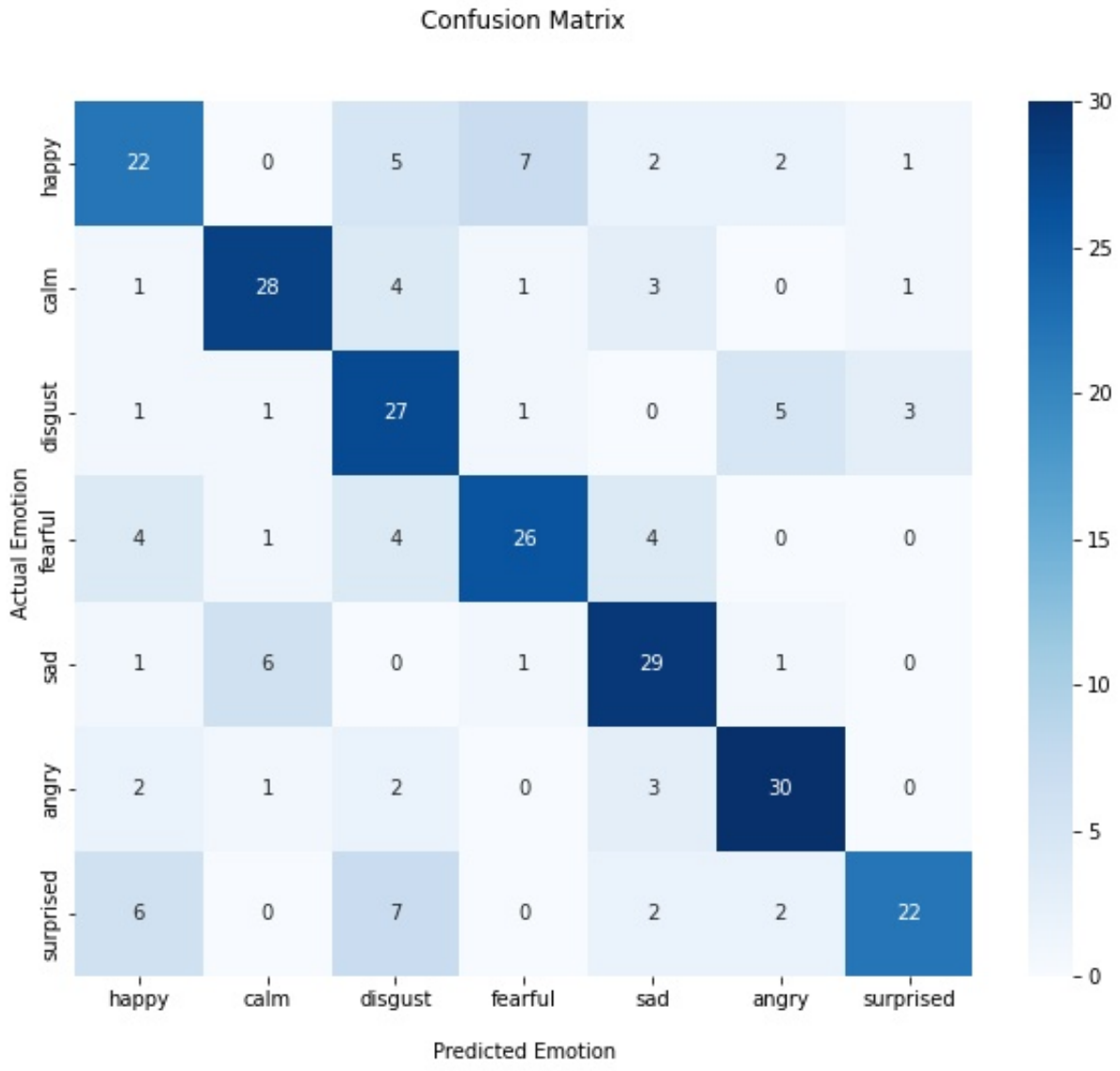
Multi-Layer Perceptron Classifier	
Optimizer	Accuracy
lbfgs	65
sgd	62
adam	70

TABLE 4.5: RFC Results

Here, adam optimizer is giving us the best results for MLP classifier. Thus, to further study the outcomes, classification report and confusion matrix are shown for the adam optimizer.

SVM - Classification report				
Emotion Class	Precision	Recall	f1-score	support
happy	0.62	0.51	0.56	39
calm	0.81	0.76	0.78	38
disgust	0.61	0.74	0.67	38
fearful	0.68	0.69	0.68	39
sad	0.70	0.79	0.74	38
angry	0.74	0.74	0.74	38
surprised	0.71	0.62	0.66	39

TABLE 4.6: SVM Classification Report



Here, there are no significant anomalies present in the confusion matrix of mlp classifier. There are still some minor overlaps present but this are the best results obtained so far. This directly increases the state of the art accuracy of audio emotion prediction on RAVDESS dataset (using only audio data) by 29%.

Chapter 5

Conclusion and Future Work

The main objectives of this project were:

- To develop fast emotion recognition models to predict emotions from audio files even with lower processing power.
- To further refine the data features for improving the accuracy.
- Evaluate the performance of these models based on different evaluation metrics.
- Analyse the significance of the produced results.

Through this project, we demonstrated how to extract underlying emotion from speech signals using machine learning algorithms. Multiple classifier models (SVM, RFC, MLP) were thoroughly analysed to get the best results. Based on their confusion matrices, the MLP classifier was able to give the best results. In the end, all these models were able to classify emotions with significant accuracy.

There is still a scope to further make the models more robust and accurate if:

- A larger dataset is available to give more accurate idea on distinguishing between emotions.
- Multiple comparable datasets are available to help generalize the models.
- Corresponding video or image data is used to supplement the output from audio analysis.
- Bigger length audio samples are available to further increase the data volume going inside the models.

Bibliography

[1] Hasimah Ali et al. "Facial emotion recognition using empirical mode decomposition". In: *Expert Systems with Applications* 42.3 (2015), pp. 1261–1277.

[2] Zhen-Tao Liu et al. "Speech emotion recognition based on feature selection and extreme learning machine decision tree". In: *Neurocomputing* 273 (2018), pp. 271–280.

[3] Martin Ragot et al. "Emotion recognition using physiological signals: laboratory vs. wearable sensors". In: *International Conference on Applied Human Factors and Ergonomics*. Springer. 2017, pp. 15–22.

[4] Meshach A Martin et al. "AUTOMATIC SPEECH EMOTION RECOGNITION USING MACHINE LEARNING." In: *International Journal of Advanced Research in Computer Science* 12 (2021).

[5] Siqing Wu, Tiago H Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features". In: *Speech communication* 53.5 (2011), pp. 768–785.

[6] Siqing Wu. *Recognition of human emotion in speech using modulation spectral features and support vector machines*. Queen's University, 2009.

[7] Anjali Bhavan, Pankaj Chauhan, Rajiv Ratn Shah, et al. "Bagged support vector machines for emotion recognition from speech". In: *Knowledge-Based Systems* 184 (2019), p. 104886.

[8] A Milton, S Sharmy Roy, and S Tamil Selvi. "SVM scheme for speech emotion recognition using MFCC feature". In: *International Journal of Computer Applications* 69.9 (2013).

[9] GS Divya Sree, P Chandrasekhar, and B Venkateshulu. "SVM based speech emotion recognition compared with GMM-UBM and NN". In: *International Journal of Engineering Science and Computing* 6.11 (2016), pp. 3293–3298.

[10] Gabriella Melki et al. "OLLAWV: online learning algorithm using worst-violators". In: *Applied Soft Computing* 66 (2018), pp. 384–393.

[11] Jeevan Singh Deusi and Elena Irena Popa. "An Investigation of the Accuracy of Real Time Speech Emotion Recognition". In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer. 2019, pp. 336–349.

[12] Peipei Shen, Zhou Changjun, and Xiong Chen. "Automatic speech emotion recognition using support vector machine". In: *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*. Vol. 2. IEEE. 2011, pp. 621–625.

[13] Martin Vondra and Robert Vích. "Recognition of emotions in german speech using gaussian mixture models". In: *Multimodal Signals: Cognitive and Algorithmic Issues*. Springer, 2009, pp. 256–263.

[14] Sathit Prasomphan. "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram". In: *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE. 2015, pp. 73–76.

[15] Alex Graves and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks". In: *International conference on machine learning*. PMLR. 2014, pp. 1764–1772.

[16] Shizhe Chen and Qin Jin. "Multi-modal dimensional emotion recognition using recurrent neural networks". In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. 2015, pp. 49–56.

[17] Wootaek Lim, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks". In: *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE. 2016, pp. 1–4.

[18] Muhammad Sajjad, Soonil Kwon, et al. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM". In: *IEEE Access* 8 (2020), pp. 79861–79875.

[19] Soonil Kwon et al. "Att-Net: Enhanced emotion recognition system using lightweight self-attention module". In: *Applied Soft Computing* 102 (2021), p. 107101.

[20] María Teresa García-Ordás et al. "Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network". In: *Biomedical Signal Processing and Control* 69 (2021), p. 102946. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.102946. URL: https://www.sciencedirect.com/science/article/pii/S1746809421005437.

[21] Steven R Livingstone and Frank A Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PloS one* 13.5 (2018), e0196391.