“Speech Emotion Recognition”

A Minor Project Report

Submitted in Partial Fulfillment of the Requirements for the degree

of

Bachelor of Technology

IN

INSTRUMENTATION AND CONTROL ENGINEERING

By

Swet Parekh (18BIC052)

Under the Guidance of

Prof. Harsh Kapadia

ELECTRONICS AND INSTRUMENTATION ENGINEERING DEPARTMENT
SCHOOL OF TECHNOLOGY
NIRMA UNIVERSITY
Ahmedabad 382 481

December 2022

CERTIFICATE

This is to certify that the project report entitled "Speech Emotion Recognition" submitted by Swet parekh (18bic052) & Anuj shah (19bic061) towards the partial fulfillment of the requirements for the award of the degree in bachelor of Technology (INSTRUMENTATION & CONTROL ENGINEERING) Of School of Technology is the record of work carried out by them under our supervision and guidance. The work submitted has in our opinion reached a level required for being accepted for examination. The results embodied in this project work to the best of my/our knowledge have not been submitted to any other University or Institution for award of any degree or diploma.

"Harsh Kapadia"	**"Himanshu K Patel"**
Assistant Professor	**Head of Department (EI)**

DATE:

ACKNOWLEDGEMENT

Our work would not have been possible without the support and help from many individuals and organizations. We would like to thank all of them for their help and support. We are highly indebted to Prof. Harsh Kapadia for his guidance and help throughout the project.

His guidance helped us keep motivated to complete the work. We would also express our sincere gratitude to Head of our Department Dr Himanshu K Patel and Director Dr Rajesh N Patel to provide us with such an opportunity.

ABSTRACT

Since the beginning of human settlement, spoken language has been the primary means of human interaction, which serves as the foundation for information exchange. Similar to how spoken language predates emotions, which are the initial forms of natural communication, emotions may be traced back to a primitive instinct. The purpose of the project is to identify the feelings that a speaker evokes while speaking. For instance, speech generated when feeling fearful, angry, or joyful is loud and quick, with a greater and broader range of pitch, but speech produced while feeling sad or exhausted is sluggish and low-pitched. There are several uses for the detection of human emotions using voice and speech patterns, including improving human-machine interactions. In particular, we are presenting a classification model of emotion elicited by speeches based on deep neural networks (Convolutional Neural Network), Support Vector Machine, Multilayer Perceptron Classification based on acoustic features such as Mel Frequency Cepstral Coefficient (MFCC). The model has been trained to classify eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). We have obtained the following results on our model.

1. **MLP Classifier achieved an F1 score of 0.83 over the 8 classes.**
2. **SVM Classifier achieved an F1 score of 0.82.**
3. **CNN model obtained the F1 score of 0.85.(Best choice).**

Keywords: CNN, MFCC, SER, DT, MLP, SVM

List of Figures

CONTENTS

Chapter 1 Introduction

Since its inception, speech emotion recognition has developed into a crucial part of Human-Computer Interaction (HCI). Instead of using conventional devices as input to comprehend vocal information and making it simple for human listeners to respond, these systems strive to facilitate the natural interaction between humans and machines through direct voice contact. Dialogue systems for spoken languages are used in a variety of applications, including call centre discussions, onboard car driving systems, and the use of emotion patterns from speech in medical settings. However, there are still a lot of issues with HCI systems that need to be fixed, especially as these systems transition from lab testing to real-world implementation. Therefore, efforts are needed to properly address these issues and improve machine recognition of emotions.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model . A discrete emotional approach is regarded as one of the fundamental approaches among the many models used for categorising various emotions. It makes use of a range of emotions, including grief, boredom, neutrality, surprise, disgust, and fury.

The feature extraction and features classification phases make up the majority of the speech emotion recognition (SER) technique. Feature classification using linear and non-linear classifiers is part of the second step. The Maximum Likelihood Principle (MLP) and Support Vector Machine are the most widely used linear classifiers for emotion recognition (SVM). The speech signal is typically regarded as non-stationary. As a result, it is believed that SER benefits from non-linear classifiers. Many non-linear classifiers, including as the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), are available for SER. For accurate emotion recognition from speech, energy-based features like Mel-Frequency Cepstrum Coefficients (MFCC) are frequently employed. For emotion recognition, other classifiers such as K-Nearest Neighbor (KNN), Principal Component Analysis (PCA), and Decision Trees are also used.

The three core elements of emotion detection systems based on digitised speech are signal preprocessing, feature extraction, and classification. To establish evaluation purposes of the signal, acoustic preprocessing techniques like denoising and segmentation are used. To find the pertinent features present in the signal, feature extraction is used. Finally, classifiers perform the mapping of extracted feature feature vector to relevant emotions. Speech signal processing, feature extraction, and classification are all covered in-depth in this section. Due to their importance to the subject, the distinctions between spontaneous and performed speech are also examined. A basic system for speech-based emotion recognition is shown in figure1.

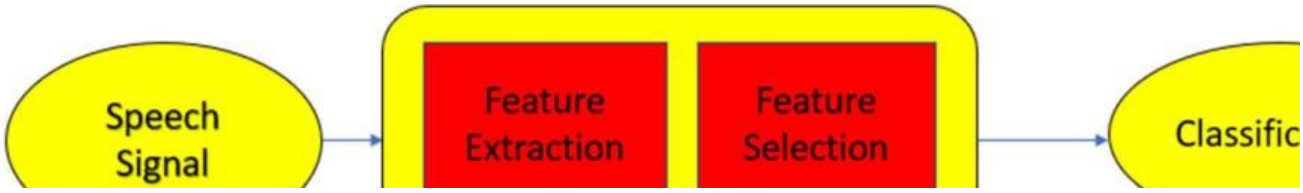


Fig 1 Basic SER Model

In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, and feature selection. The required features are extracted from the preprocessed speech signal and the selection is made from the

extracted features. Such feature extraction and selection is usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers such as GMM and HMM, etc. are utilized for classification of these features. Lastly, based on feature classification different emotions are recognized.

A. Enhancement of Speech Input Data in SER

The input data gathered for emotion recognition is frequently tainted by noise when it is being captured. The feature extraction and classification lose accuracy as a result of these flaws. This means that in order for emotion detection and recognition systems to function properly, the input data must be enhanced. The speaker and recording variation are removed during this preprocessing stage, while the emotional discrimination is retained.

B. Selection and Extraction of Features in SER

Segments are used to describe the enhanced speech signal as meaningful units. Based on the information gathered, pertinent traits are extracted and divided into several groups. Short-term classification, which is based on characteristics like energy, formants, and pitch, is one type of classification. The other is known as long term classification; mean and standard deviation are two of the often-used long term features. Among prosodic features, the intensity, pitch, rate of spoken words and variance are usually important to identify various types of emotions from the input speech signal.

Emotions	Pitch	Intensity	Spea rate
Anger	abrupt on stress	much higher	marg faste
Disgust	wide, downward inflections	lower	very faste
Fear	wide, normal	lower	mucl
Happiness	much wider, upward	higher	faste

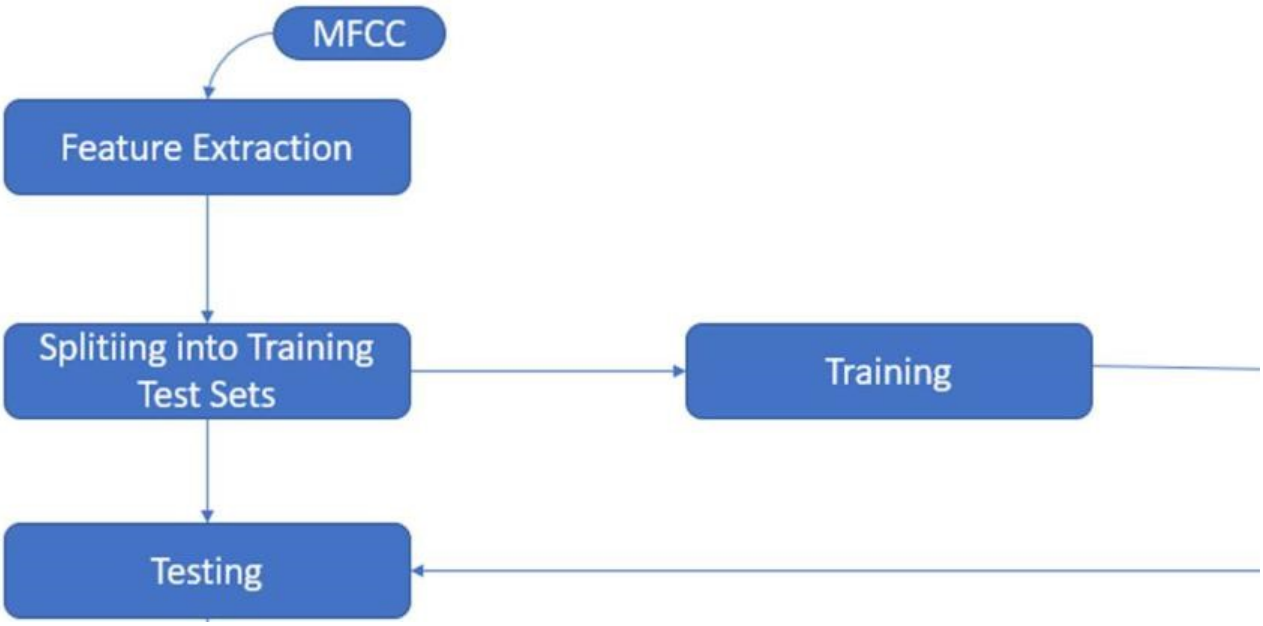
C. Measures for Acoustics in SER

Every feature of language and its variations encodes the availability of emotional information. Among the most studied cases in this area are the vocal parameters and how they relate to emotion identification. Many times, factors like voice quality, pitch, intensity, and rate of speech are taken into account. The assumption that emotions are separate categories with independent existence is a common one in the simple view of emotion. Some of these discrete emotions, as shown in Table for a subset of emotions, have rather obvious connections with acoustic characteristics. Pitch and intensity are frequently connected with activation, meaning that the intensity value rises with high pitch and falls with low pitch. Factors that affect the mapping from acoustic variables to emotion include whether the speaker is acting, there are high speaker variations, and the mood or personality of the individual.

D. Classification of Features in SER

Numerous classifiers have been researched in the literature in order to create systems like SER, speech recognition, and speaker verification, to name a few. On the other hand, the reasons for selecting a specific classifier for a given speech task are frequently left out of most applications. Typically, one of these two rules of thumb is used to choose classifiers.

Ordinarily, the two primary types of pattern recognition classifiers used for SER can be broadly divided into linear classifiers and non-linear classifiers. The most common way that linear classifiers are evaluated is as an array called a feature vector. On the other hand, non-linear classifiers are used to characterise the objects in order to create a non-linear weighted combination of them.



Dataset Used:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset the Toronto emotional speech set (TESS) dataset samples include:

1440 speech files and 1012 Song files from RAVDESS. This dataset includes recordings of 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, and song contains calm, happy, sad, angry, and fearful emotions. Each file was rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained adult research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity, interrater reliability, and test-retest intrarater reliability were reported.

2800 files from TESS. A set of 200 target words were spoken in the carrier phrase "Say the word _____' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

Chapter 2 Methodology

The classification model of emotion recognition here proposed is based on a deep learning strategy based on convolutional neural networks (CNN), Support Vector Machine (SVM) classifier, MLP Classifier. The key idea is considering the MFCC commonly referred to as the "spectrum of a spectrum", as the only feature to

train the model.

The most common machine learning application treats the MFCC itself as an 'image' and becomes feature. The benefit of treating it as an image is that it provides more information, and gives one the ability to draw on transfer learning. This is certainly legit and yields good accuracy. However, research has also shown that statistics relating to MFCCs (or any other time or frequency domain) can carry good amount of information as well.

It has been established that the MFCC, a variant of the Mel-frequency cepstrum (MFC), is the state of the art for sound formalisation in automatic speech recognition tasks. The MFC coefficients have mostly been utilised as a result of their ability to compactly and vectorially depict the amplitude spectrum of the sound wave. To obtain statistically steady waves, the audio file is segmented into frames, often using a set window size. The "Mel" frequency scale is shrunk to equalise the amplitude spectrum. This procedure is carried out to empathise the frequency more meaningfully in order to recreate the wave as accurately as the human auditory system is capable of doing. Forty features have been retrieved for each audio file. Each audio recording was converted into a floating-point time series to create the functionality. The time series was then turned into an MFCC sequence.

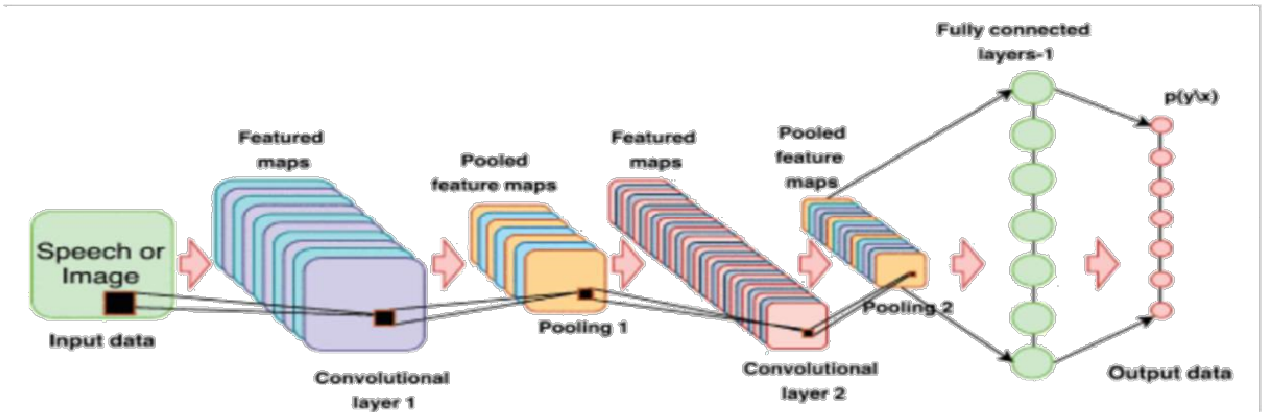
1. **CNN (Convolution neural network)**
2. **MLP(Multilayer perceptron)**
3. **SVM (support vector machine)**

2.1 Convolution Neural Network

The Convolution neural network (CNN) designed for the classification task is reported operationally in Fig. 1. The network can work on vectors of 40 features for each audio file provided as input. The 40 values represent the compact numerical form of the audio frame of 2s length. Consequently, we provide as input a of size < number of training files > x 40 x 1 on which we performed one round of a 1D CNN with a ReLu activation function, dropout of 20%, and a max-pooling function 2 x 2.

The rectified linear unit (ReLu) can be formalized as g(z) = max {0, z}, and it allows us to obtain a large value in case of activation by applying this function as a good choice to represent hidden units. Pooling can, in this case, help the model to focus only on principal characteristics of every portion of data, making them invariant by their position. We have run the process described once more by changing the kernel size. Following, we have applied another dropout and then flatten the output to make it compatible with the next layers.

Finally, we applied one Dense layer (fully connected layer) with a softmax activation function, varying the output size from 640 elements to 8 and estimating the probability distribution of each of the classes properly encoded (0=Neutral; 1= Clam; 2= Happy; Sad=3; Angry=4; Fearful= 5; Disgust=6; Surprised=7).



CNN ALGORITHM

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
conv1d_1 (Conv1D)            (None, 40, 64)
_________________________________________________________________
activation_1 (Activation)    (None, 40, 64)
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 64)
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 10, 64)
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 10, 128)
_________________________________________________________________
activation_2 (Activation)    (None, 10, 128)
_________________________________________________________________
dropout_2 (Dropout)          (None, 10, 128)
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 2, 128)
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 2, 256)
_________________________________________________________________
activation_3 (Activation)    (None, 2, 256)
```

Model Summary

Model Prediction

	precision	recall	f1
0	0.88	0.91	
1	0.77	0.76	
2	0.90	0.84	
3	0.80	0.86	
4	0.89	0.88	
5	0.88	0.80	
6	0.81	0.92	
7	0.88	0.85	

Formulas:

TP: True Positive

FP: False Positive

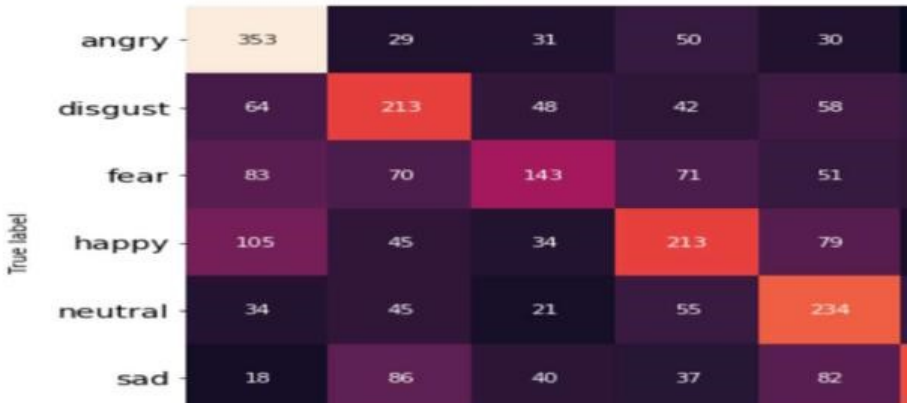
TN: True Negative

FN: False Negative

Precision: $\frac{TP}{TP+FP}$ **F1 Score:** $\frac{2*Precision*Recall}{Precision+Recall}$

Recall : $\frac{TP}{TP+FN}$ **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$

Confusion matrix



2.2 Multilayer Perceptron

A class of feedforward artificial neural network is called a multilayer perceptron (MLP) (ANN).
Backpropagation is a supervised learning method that is used by MLP during training.

MLP differs from a linear perceptron due to its numerous layers and non-linear activation. It can discriminate between data that cannot be separated linearly. A class of feedforward artificial neural network is called a multilayer perceptron (MLP) (ANN).

Backpropagation is a supervised learning method that is used by MLP during training. MLP differs from a linear perceptron due to its numerous layers and non-linear activation. It can discriminate between data that cannot be separated linearly.

Prediction

	precision	recall	f1
angry	0.94	0.84	
calm	0.72	0.72	
disgust	0.87	0.82	
fearful	0.86	0.83	
happy	0.77	0.84	
neutral	0.72	0.96	
sad	0.89	0.80	
surprised	0.89	0.79	

2.3 Support Vector Machines

A supervised machine learning approach called the Support Vector Machine (SVM) can be applied to classification or regression problems. However, classification issues are where it's most frequently employed.

The SVM algorithm plots each data point as a point in an n-dimensional space, where n is the number of features you have and each feature's value is a specific coordinate's value.

Before applying to an SVM classifier, data might be scaled to avoid attributes in larger numeric ranges. Scaling also helps to prevent various mathematical challenges that may arise throughout the calculation.

	precision	recall	f1-score
angry	0.89	0.92	0.91
calm	0.62	0.94	0.75
disgust	0.81	0.91	0.86
fear	0.76	0.80	0.78
happy	0.94	0.73	0.82
neutral	1.00	0.78	0.88
sad	0.77	0.78	0.77
surprised	0.83	0.80	0.81
accuracy			0.82
macro avg	0.83	0.83	0.82
ghted avg	0.84	0.82	0.82

0 = neutral, 1 = calm, 2 = happy, 3 = sad, 4 = angry, 5 = fearful, 6 = disgust, 7 = surprised

Implemenation and Results:

In this project we have trained our data to predict the emotions from them. After training and testing our data we even have tried to predict live data from new sources and predicted the emotion.

```
pred = livePredictions(path='testing10_model.h5',file='/content/Ac

pred.load_model()
pred.makepredictions()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 40, 64)            384

 activation (Activation)     (None, 40, 64)            0

 dropout (Dropout)           (None, 40, 64)            0

 max_pooling1d (MaxPooling1D  (None, 10, 64)           0
 )

 conv1d_1 (Conv1D)           (None, 10, 128)           41088

 activation_1 (Activation)   (None, 10, 128)           0

 dropout_1 (Dropout)         (None, 10, 128)           0

 max_pooling1d_1 (MaxPooling  (None, 2, 128)           0
 1D)

 conv1d_2 (Conv1D)           (None, 2, 256)            164096
```

```
conv1d_2 (Conv1D)           (None, 2, 256)

activation_2 (Activation)   (None, 2, 256)

dropout_2 (Dropout)         (None, 2, 256)
```

CLASS	MLP	SVM	CNN
SAD	0.81	0.82	0.80
ANGRY	0.89	0.91	0.89
HAPPY	0.82	0.84	0.90
DISGUST	0.80	0.81	0.81
SURPRISE	0.80	0.87	0.88
NEUTRAL	0.89	0.93	0.88
CALM	0.75	0.64	0.77
FEAR	0.84	0.81	0.88

Comparation of SVM,MLP and CNN models

After Running, Training and Testing Data we have come to the results mentioned above in table. As for overall emotions we found CNN model to be best fitted for the project. As per individual emotions Svm has better accuracy for neutral and angry. For emotions with lower intensity Mlp model was better.

CONCLUSION

Using audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto emotional speech collection, we proposed an architecture based on deep neural networks for the classification of emotions in this work (TESS). The model was taught to categorise seven different emotions, including neutral, calm, happy, sad, furious, afraid, disgusted, and astonished. It achieved an overall F1 score of 0.85, with the joyful class performing best (0.90) and the calm class performing worst (0.77). To achieve this, we retrieved the MFCC features (spectrum of a spectrum) from the training audio recordings.

References

[1]. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[2].https://www.kaggle.com/code/ritzing/speech-emotion-recognition-with-cnn.

[3]. Github

[4].Medium

[5].Data Camp