Abnormal Activity Detection

MINI PROJECT

Submitted in partial fulfillment of the requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS & COMMUNICATION ENGINEERING

BY

VRITTIK SHARMA
05611502817

VAIBHAV SURI
05511502817

UJJWAL GABA
05311502817

PRATYUSH RAJ
03411502817

Guided by

Dr.q Manoj Sharma

CANDIDATE'S DECLARATION

It is hereby certified that the work which is being presented in the B.Tech **Mini Project** entitled **"Abnormal Activity Detection "** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology and submitted in the Department of **Electronics and Communication Engineering** of **BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING, New Delhi**

(Affiliated to Guru Gobind Singh Indraprastha University, Delhi) is an authentic record of our own work carried out during a period under the guidance of **Dr. Manoj Sharma, Professor.**

VRITTIK SHARMA
05611502817

VAIBHAV SURI
05511502817

UJJWAL GABA
05311502817

PRATYUSH RAJ
03411502817

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Dr. Manoj Sharma)

Professor

CONTENTS:

INTRODUCTION TO PROJECT:

Abnormal Activity Detection

1. Definition

1.1. Project Overview

The project mainly focuses on identification of normal and abnormal activities in a setting like a household, classroom etc. The classification will be done on basis of Speech Emotion. The Speech is an extremely important indicator of the type of activity going on. Speech can be used to indicate abnormal activities on basis of emotions like fear, anger etc. There are other factors like Pitch, Frequency, Voice tone, Modulation etc which play an important role. The amalgamation of the aspects will be used to determine abnormal and normal activities. The analysis will be done on Ravdess Dataset and an Indian Speech dataset. The results from the 2 datasets will be then compared and analysed.

1.2. Problem Statement

Speech plays an important role in identifying the tone of a person, and tone in turn helps in identifying the behaviour of a person. The problem is to make a machine classify human behaviour as normal and abnormal based on speech produced by the person.

The problem is divided into 2 subtasks, one is training the machine with Indian Dataset and the other is to train the machine with a foreign dataset of speech and then comparing both in terms of accuracy.

2. Methodology

2.1 Preprocessing

Preprocessing is the very first step after collecting data that will be used to train the classifier in a SER system. Some of these preprocessing techniques are used for feature extraction, while others are used to normalize the features so that variations of speakers and recordings would not affect the recognition process.

Noise Reduction and Mel Frequency Cepstral Coefficients

When sound is produced by a person, it is filtered by the shape of the vocal tract. The sound that comes out is determined by this shape. An accurately simulated shape may result in an accurate representation of the vocal tract and the sound produced. Characteristics of the vocal tract are well represented in the frequency domain . Spectral features are obtained by transforming the time domain signal into the frequency domain signal using the Fourier transform. They are extracted from speech segments of length 20 to 30 milliseconds that are partitioned by a windowing method.Mel Frequency Cepstral Coefficients (MFCC) feature represents the short term power spectrum of the speech signal. To obtain MFCC, utterances are divided into segments, then each segment is converted into the frequency domain using a short time discrete Fourier transform. A number of sub-band energies are calculated using a Mel filter bank. Then, the logarithm of those sub-bands is calculated. Finally, an inverse Fourier transform is applied to obtain MFCC. It is the most widely used spectral feature .

Basic concept of feature extraction :

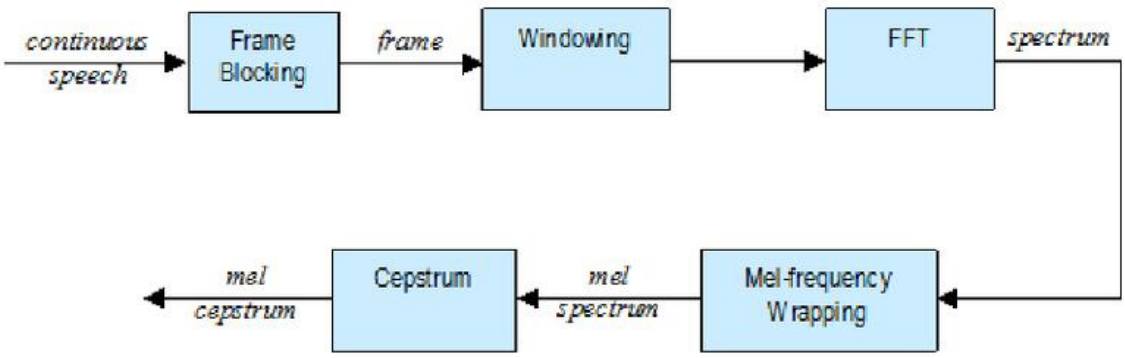


Fig 1: Mel Frequency Cepstral Coefficients Calculation

Steps involved in MFCC Feature Extraction:

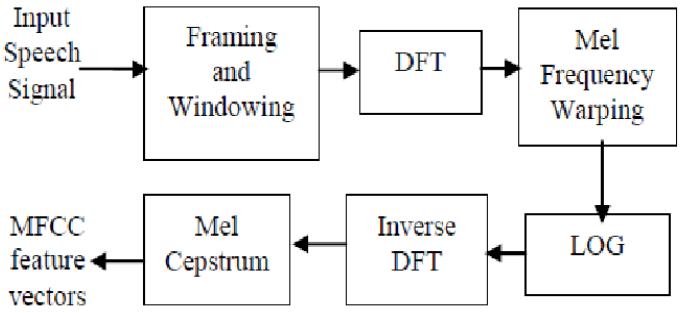


Fig 2: DFT method of calculation mfccs

3. Convolutional Neural Networks

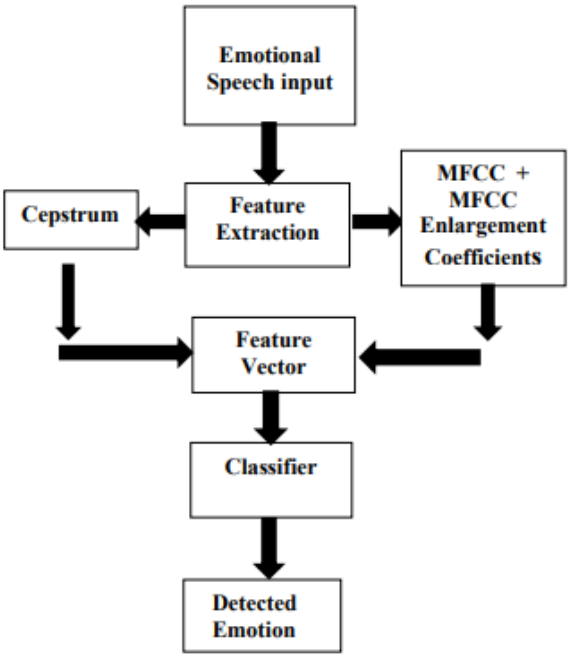
Convolutional Neural Networks (CNNs) are types of neural networks which are designed to process data that has a grid-like topology, such as images. Through applications of several relevant filters, CNN can successfully capture temporal and spatial dependencies from an input source. The inputs are reduced into a form without loss of feature so that computational complexity decreases, and the success rate of the algorithm is increased. A CNN is composed of several layers: convolution layer, polling layer, and Fully Connected layer.

A convolution layer is used to extract high-level features from the input. Mathematically a convolution means combining two functions to obtain a third one. In CNN, the input is taken and then a kernel is applied to it. The resulting output is a feature map.

Polling layer is used to reduce the size of convoluted features to decrease computational complexity through dimensionality reduction. It is useful for extracting dominant features of the input data.

After passing input from several convolution and pooling layers and extracting the high-level features, the resulting features are used as an input to a fully connected layer by flattening the 2D data to a column array and feeding it to a feed-forward network that operates as an ordinary neural network.

Block Diagram of the proposed algorithm:



4. Classification, Training and Testing of the Data

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modelling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

The 2 categories of the classification were **Normal** and **Abnormal** Sounds

There are 2 datasets that were worked on. The first Dataset is the The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the other one is the Indian EmoSpeech Command Dataset.

The RAVDESS Dataset consists of 8 classes of human emotion. This dataset contains labelled files in 3 modality (full AV, video-only and audio-only) and 2 vocal channels (speech and song) from male and female actors. Since our focus was on speech sentiment analysis, and also due to file size constraints, our models were trained on audio-only speech samples which consist of Labels:

The neutral, calm ,happy and surprised were put under the Normal Category whereas the Sad,Angry,Fearful,Disgust and Surprised were put under the Abnormal Category

Emotion	**Speech Sample Count**	**Speech+Video Sample Count**	**Summed Count**
Normal(Neutral,Calm, Happy)	96+192+192	92+184+184	940
Abnormal(Sad,Angry,Fearful,Disgust,Surprised)	192+192+192+192+192	184+184+184+0+0	1512
Total	1440	1012	2452

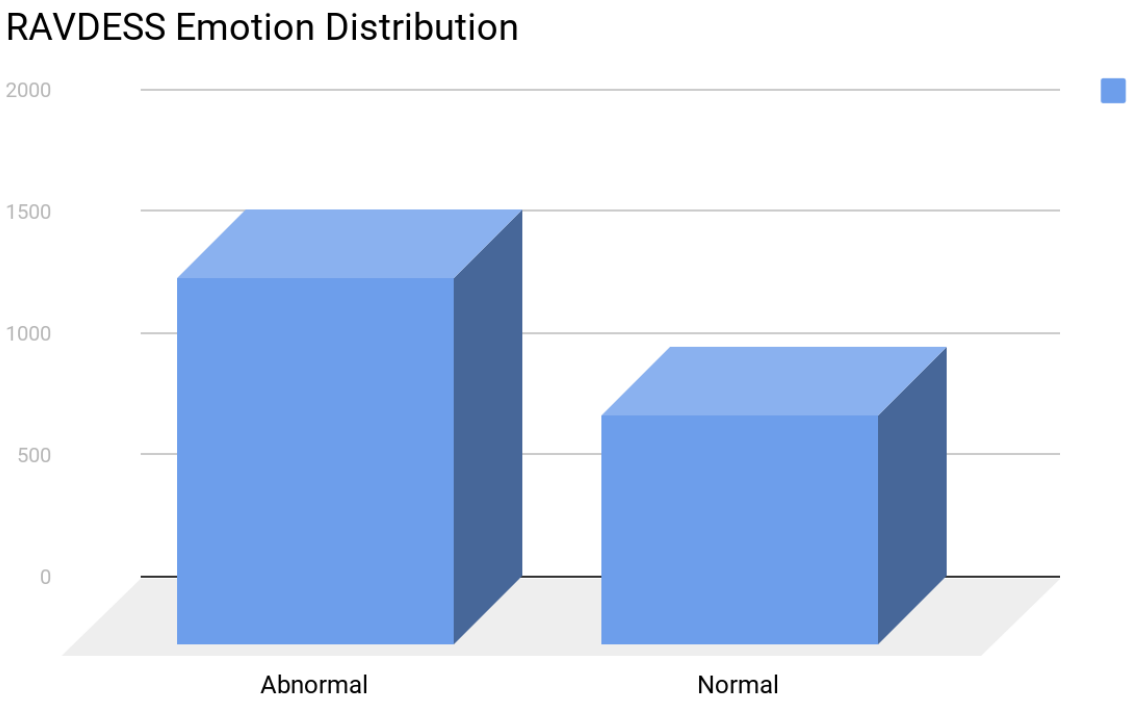


Fig 3 : RAVDESS Emotion Distribution

The other dataset was the Indian EmoSpeech Command Dataset. There are 4 emotions present in the dataset:Happy,Calm,Fearful and Angry. The Happy and Calm emotions were put under NORMAL category and the Fearful and Angry emotions were put on the ABNORMAL category.

Emotion	Speech Sample Count
Normal(Happy, Calm)	834
Abnormal(Angry, Fearful)	723
Total	1557

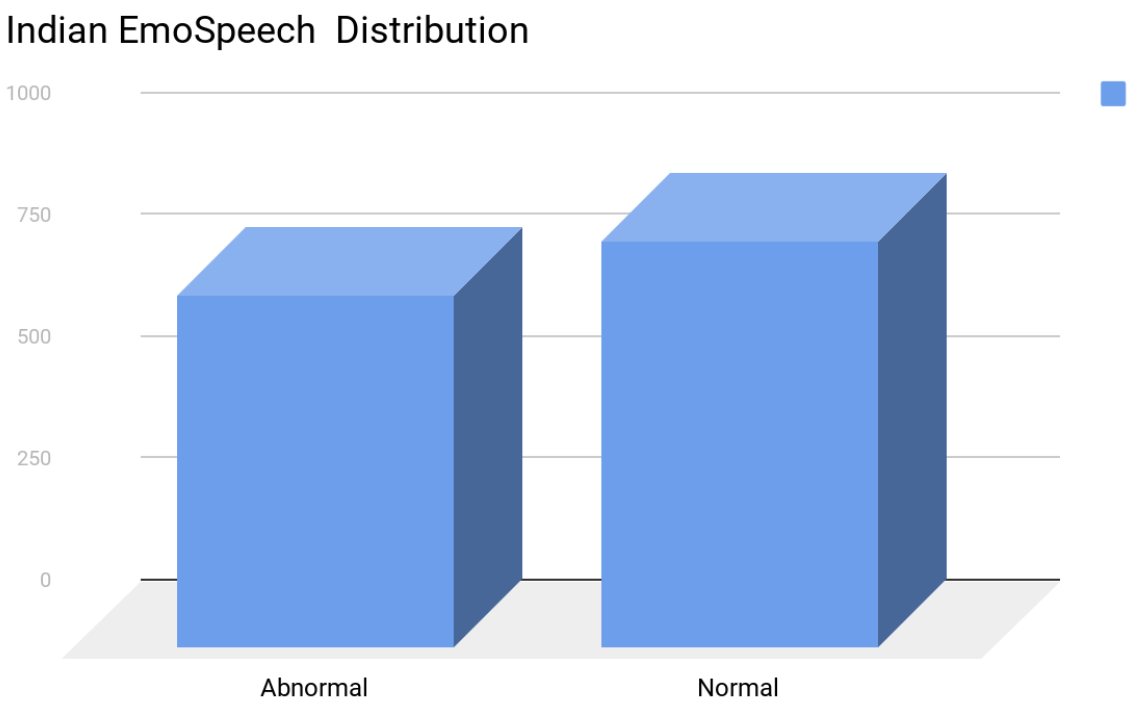


Fig 4 : Indian EmoSpeech Distribution

5. RESULT AND ANALYSIS

5.1 Ravdess model results

The emotion samples from the database are divided into training and testing samples. The training data consists of 2305 samples (384 samples of each emotion) and the testing data consists of 576 samples (83 samples of each emotion). In each of the training and testing data, signals are arranged in sets, each set having one signal of each emotion in the order calm, happy, sad, anger, fearful, disgust, surprise, neutral . Corresponding target matrices are constructed for training and testing data.

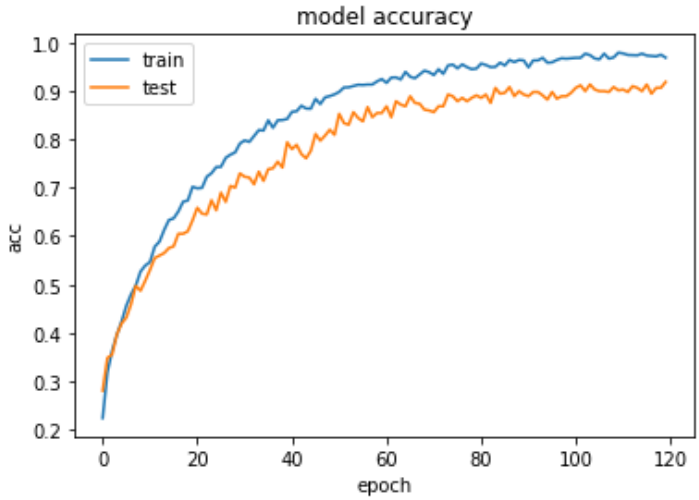


Fig 5 : Ravdess Model Accuracy vs Epoch curve

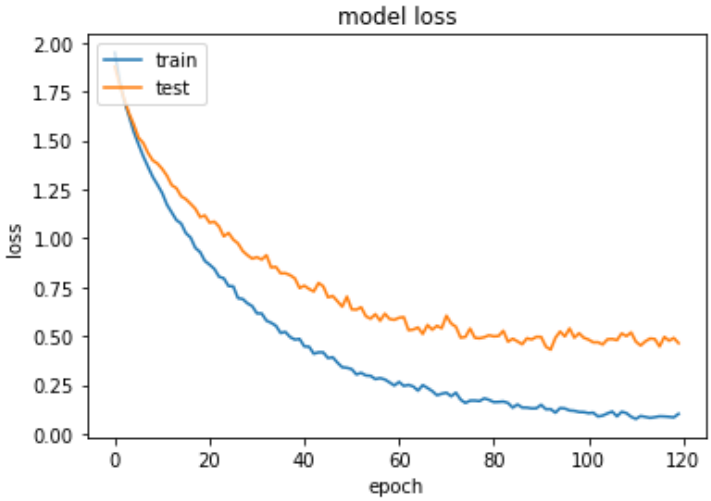


Fig 6: Ravdess Model Loss vs Epoch curve

Neural networks classifier performs supervised training and testing. The amount and extent of training depends upon the number of iterations which in turn depend upon the number of hidden neurons and number of layers in the network. The training of the classifier consists of three stages: training, validation and testing.

CONFUSION MATRIX

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. The sum of the values in the main diagonal of the matrix gives the total correct predictions and the sum of other values gives the total incorrect predictions.

Actual(row)/ Predicted(column)	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Neutral	40	1	0	4	0	0	2	2
calm	0	71	2	2	0	0	0	0
Happy	0	0	70	4	0	3	0	0
Sad	0	2	0	62	2	4	0	0
Angry	0	2	2	0	82	0	0	0
Fearful	0	0	4	0	1	71	0	0
Disgust	0	0	2	2	0	0	68	0
Surprise	0	0	0	0	2	4	0	66

TABLE 1 : Confusion Matrix for Ravdess Model

5.2 Indian model results

The emotion samples from the database are divided into training and testing samples. The training data consists of 1246 samples (638 samples of normal & 608 samples of abnormal) and

the testing data consists of 311 samples. In each of the training and testing data, signals are arranged in sets, each set having one signal of each emotion in the order happy, calm, angry and fearful . Corresponding target matrices are constructed for training and testing data.

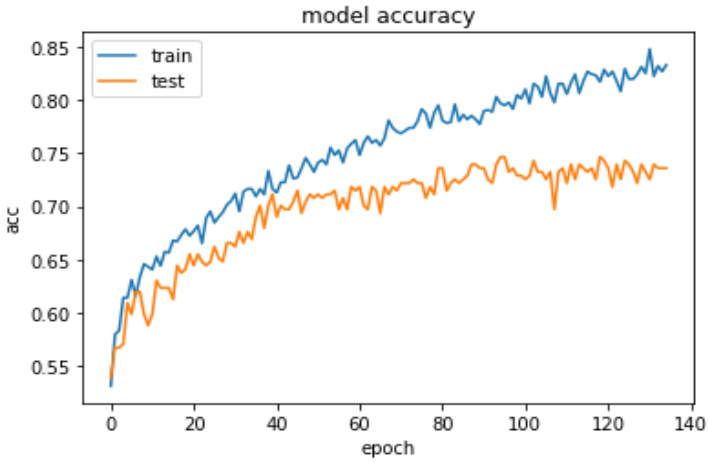


Fig 7: Indian Model Accuracy vs Epoch curve

Fig 8: Indian Model Loss vs Epoch curve

CONFUSION MATRIX :

Actual(row)/ Predicted(column)	Normal	Abnormal
Normal	84	50
Abnormal	25	125

TABLE 2 : Indian Model Confusion Matrix

6. CONCLUSION

Human speech is a strong factor in analysing the emotions of a person and emotions in turn helps in finding out the behaviour of a person. The main aim of this project is to automate the current surveillance techniques to the point where they can be used to identify abnormalities in human behaviour , whether the person is in the metro ,home ,school ,college etc.
We used two speech corpuses for this project namely RAVDESS DATASET ,INDIAN EMOSPEECH DATASET
The deep learning model trained on the Ravdess dataset is showing a higher value of accuracy (94.6%) in comparison to that of the model trained on Indian dataset (73.49%). This is since there are more noise factors in Indian dataset in comparison to Ravdess dataset that in turn reduces the valuable pitch and tone features in the audio. The calibration of the model trained on Indian dataset can be done by adding more audio files so that the model becomes more accurate in predicting abnormal and normal speech.

7. References:

Reza Chu **, "**Speech Emotion Recognition with Convolutional Neural Network**",**Section **:** Default ModelArchitecture,blog@https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3 , Jun 1, 2019 [Accessed Jan 5, 2020]

Omar Raghib, Eshita Sharma, Tameem Ahmad, Faisal Alam, "Emotion analysis And speech signal processing", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Sept 21, 2017

Aniket Sharma, Piyush Aggarwal, " Indian EmoSpeech Command Dataset : A dataset for emotion based speech recognition in the wild", @https://emo-speech.web.app [Accessed April 5, 2020]

George Georgoulas, Voula C. Georgopoulos, Chrysostomos D. Stylios, "Speech Sound Classification and Detection of Articulation Disorders with Support Vector Machines and Wavelets", 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, Aug 30, 2006