SPEECH EMOTION RECOGNITION USING DEEP LEARNING

A Thesis

Submitted

In Partial Fulfillment for the Degree

MASTER OF TECHNOLOGY

In

Computer Science & Engineering

Submitted by:

VANDANA SINGH

Enroll No: 1900104067

Under the Supervision of:

K C Maurya

(Assistant Professor)

Department of Computer Science and Engineering

INTEGRAL UNIVERSITY, LUCKNOW U.P, INDIA

June 2022

CERTIFICATE

This is to certify that **Ms. Vandana Singh** (Enroll No.1900104067) has carried out the research work presented in the dissertation titled "**SPEECH EMOTION RECOGNITION USING DEEP LEARNING"** submitted for partial fulfillment for the award of the **Master of Technology in Computer Science and Engineering from Integral University, Lucknow** under my supervision.

It is also certified that:

i. This dissertation embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

ii. The candidate has worked under my supervision for the prescribed period.

iii. The dissertation fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

iv. No published work (figure, data, table etc) has been reproduced in the dissertation without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

Mr K C Maurya
Supervisor
(Assistant Professor)
Department of CSE,
Integral University, Lucknow

Date: _____
Place: Lucknow

DECLARATION BY STUDENT

I **Vandana Singh**, hereby declare that the work presented herein is original work done by me and has not been published or submitted elsewhere for the requirement of a degree program. Any literature date or work done by other and cited within this thesis has given due acknowledgement and listed in the reference section

Vandana Singh

Place: Lucknow

Date:

RECOMMENDATION

On the basis of the declaration submitted by "**VANDANA SINGH**", a student of M.Tech CSE (Part Time), successful completion of Pre presentation on 02-06-2022 and the certificate issued by the supervisor K C Maurya (Assistant Professor), Computer Science and Engineering Department, Integral University, the work entitled "**SPEECH EMOTION RECOGNITION USING DEEP LEARNING**", submitted to department of CSE, in partial fulfillment of the requirement for award of the degree of Master of Technology in Computer Science & Engineering, is recommended for examination.

Program Coordinator Signature	**HOD Signature**
Dr. Faiyaz Ahmad	Mrs. Kavita Agarwal
Dept. of Computer Science & Engineering	Dept. of Computer Science & Engineering
Date:	Date:

ACKNOWLEDGEMENT

I am highly grateful to the Head of Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.

It gives me a great pleasure to express my deep sense of gratitude and indebtedness to my guide **K C Maurya (Assistant Professor), Department of Computer Science and Engineering,** for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me this opportunity to carry out the ideas and work during my project period and helping me to gain the successful completion of my Project.

I am also highly obliged to the Head of Department, **Kavita Agarwal (Associate Professor), Department of Computer Science and Engineering** and PG Program Coordinator **Dr. Faiyaz Ahamad (Assistant Professor), Department of Computer Science and Engineering,** for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly to work hard in this project. I pay my respect and love to my parents and all other my friends and supporting member for their help and encouragement throughout this course of project work.

COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: **SPEECH EMOTION RECOGNITION USING DEEP LEARNING**

Candidate Name: **VANDANA SINGH**

The undersigned hereby assigns to Integral University all rights under copyright that may exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award at the M.Tech degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University's copyright notices are indicated.

Vandana Singh

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Name
RAVDESS	Ryerson Audio-Visual Dataset of Emotions Speech and Son
SVM	Support Vector Machine
MLP	Multilayer Perceptron
MFCC	Mel Frequency Campestral Coefficient
RAVDESS	Ryerson Audio-Visual Dataset of Emotions Speech and Song
TESS	Toronto Emotional Speech Set
ML	Machine Learning
PCA	Principle Component Analysis
LFPC	Log Frequency Power Coefficients

ABSTRACT

The purpose of this study is to detect the emotions evoked by the speaker while they are speaking. Speech generated in a condition of fear, rage, or delight, for example, becomes loud and quick, with a greater and broader range of pitch, but speech produced in a state of grief or exhaustion is sluggish and low-pitched. The detection of human emotions via voice and speech patterns has a variety of applications, including improving human-machine interactions. We provide a classification model of emotions produced by speeches that uses deep neural networks (CNNs), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) Classification based on auditory data like Mel Frequency Campestral Coefficient (MFCC). The models have been taught to distinguish between seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). Using the Ryerson Audio-Visual Dataset of Emotions Speech and Song (RAVDESS) dataset and the Toronto Emotional Speech Set (TESS) dataset, we found that the suggested technique achieves accuracies of 86 percent, 84 percent, and 82 percent using CNN, MLP, and SVM, respectively, for 7 emotions.

The results indicate that the proposed architecture can produce comparable results with state-of-the-art; despite excluding data augmentation and advanced pre-processing. It was reported 3 parallel CNN pipes yielded the highest accuracy, together with a series of modified LFLBs that utilize average- pooling and ReLU activation. This shows the power of leaving the feature learning up to the network and opens up for interesting future research on time- complexity and trade-off between introducing complexity in pre-processing or in the model architecture itself.

CHAPTER 1
INTRODUCTION

The foundation for information exchange is human communication via spoken language. It is also used in a variety of practical applications in fields such as Business Process Outsourcing (BPO) Centers and Call Centers to detect emotion, which is useful for determining a customer's happiness with a product, improving speech interaction, resolving various language ambiguities, and adapting computer systems to an individual's mood and emotion. The goal of the presented models is to identify just the emotion in the audio recording that has a higher value.

In a voice-based system, a computer agent is required to completely comprehend the human's speech percept in order to accurately pick up the commands given to it. This field of study is termed as Speech Processing and consists of three components:

- Speaker Identification
- Speech Recognition
- Speech Emotion Detection

Detecting emotions is one of the most important marketing strategy in today's world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

CHAPTER 2
LITERATURE REVIEW

This section provides a background of the psychological aspects of emotions, which can be approached either from a categorical or dimensional standpoint. Further on, we go through the workings of sound and speech, their common representations, and how we can sample sound. Then we proceed into the re- cent paradigm of deep learning, going over the fundamentals, including Convolutional and Recurrent Neural Networks. Finally, we explain methods for constructing and training a network, such as regularization methods, and pro- vide an insight into related work within the field of speech emotion recognition and deep learning.

Many categorization algorithms have been proposed in this field of research throughout the years. Iqbal et al. [1] created a programme that employed Gradient Boosting, KNN, and SVM to work on granular partitioning in the RAVDESS data to find differences based on gender, with overall accuracy ranging from 40% to 80% depending on the job. Male recordings alone, female recordings only, and mixed recordings datasets were constructed. SVM and KNN have 100% recognition for all anger and neutrality in the RAVDESS (male) dataset, while Gradient Boosting outperformed SVM and KNN in excitement and melancholy. SVM obtains 100% accuracy with the same fury as half of the guys in the RAVDESS (female) dataset.

With accuracy of 87 percent and 100 percent, KNN performed well in the areas of rage and neutrality. When compared to other categories of tourists, KNN performed worse in happiness and sadness. With rage and neutrality, SVM and KNN performed much better than Gradient Boosting among the combined male and female data rates. KNN's performance was extremely depressing in terms of both happiness and grief. The classifiers' average performance in the male dataset is better than in the female dataset

without SVM. SVM is more accurate for aggregated data than gender data sets. [2]. Obtained 66.41 percent accuracy in audio data and 90 percent accuracy in blending audio and video data using another method. The scientists trained three alternative depth networks using already processed picture data, including faces and audio waveforms: one for image data only, one for fixed audio waveforms only, and one for both data and waveform data. One of the first algorithms to use the RAVDESS dataset, however it merely identified it from other emotions available [8]. Three different forms of music sharing algorithms have been proposed: a basic model, a single work area model, and a multi-task capacity model. A single, independently domain classifier was utilized in a basic model. During the training, two hierarchical kinds were employed. For each domain, the single function machine trained various classifications.

During the collecting step, noise typically corrupts the input data acquired for emotion recognition [4]. The extraction of features and categorization become less accurate as a result of these flaws [7]. This means that in emotions detection and identification systems, improving the data input is crucial. The emotional discrimination is retained in this pre - processing stage, but the speech and recording variance is removed [28].

The study will cover several deep learning algorithms in the context of SER in the next part. In comparison to traditional procedures, these methods produce more precise findings, but they are more computationally demanding. This section offers researchers and readers literature-based support for evaluating HCI(Human Computer Interaction) feasibility and analyzing the user's emotional voice in a specific scenario. Emotion identification from voice input data is a viable alternative [6], but real-time implementations of these approaches are far more challenging. Although these approaches have limitations, combining two or more of these classifiers creates a new step that may enhance emotion recognition.

2.1 METHODOLOGY

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better. The flowchart represents a pictorial overview of the process (see Figure 1). The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data. The second step, called feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address the several data representation and data quality issues. The third step is often considered the core of an ML project where an algorithmic based model is developed. This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to. The final step is to evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms. Comparison results help to choose the appropriate ML algorithm most relevant to the problem.

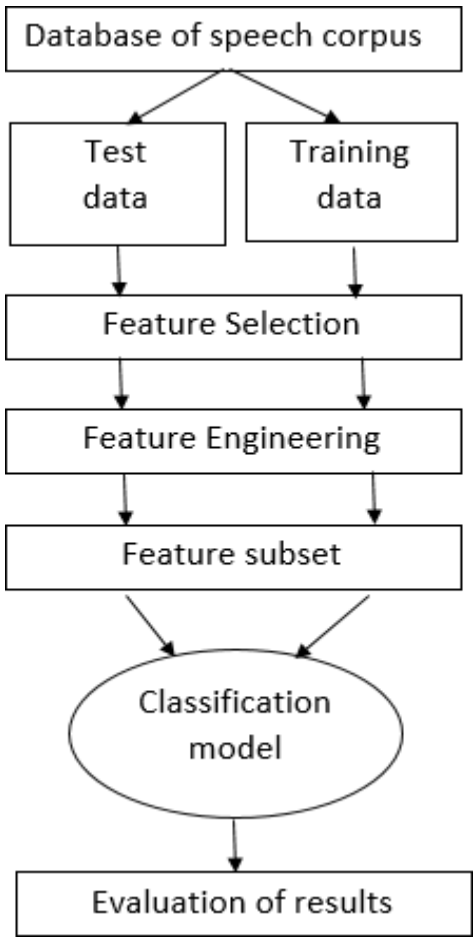

Fig.1 Flow of implementation

2.2 RANKING SVM APPROACH

Cao et al. [7] proposed a system that considered that the emotion expressed by humans is mostly a result of mixed feeling. Therefore, they suggested an improvement over the SVM algorithm thatwould consider mixed signals and choose the most dominant one. For this purpose, a ranking SVM algorithm was chosen. The ranking SVM takes all predictions from individual binary classification SVM classifiers also called as rankers, and applies it to the final multi-class problem. Using the ranking SVM algorithm, an accuracy of 44.40% was achieved in their system.

2.3 DIMENSIONALITY REDUCTION METHOD

Chen et al. [8] developed a system that had improvements in the pre-processing stage. Two pre-processing techniques, namely Fisher and Principle Component Analysis (PCA), were used in combination with two classifier algorithms, namely SVM and ANN. They carried out four experiments, each with a different combination of pre-processing and the classifier algorithm. The first experiment used Fisher method to select features for a multi-level SVM classifier (Fisher + SVM). The second experiment was to reduce feature dimensionality using Principle Component Analysis (PCA) for the SVM classifier (PCA + SVM). The third experiment used the Fisher technique over the ANN model (Fisher + ANN). Finally, PCA was applied before classification using ANN (PCA + ANN). From these experiments, two important conclusions were made. Firstly, dimensionality reduction improves the performance of the system. Secondly, SVM classifier algorithm classifies better than the ANN algorithm in the case of emotion detection. The winning experiment had an accuracy of 86.50% using Fisher for dimensionality reduction and SVM for classification.

2.4 LPC COEFFICIENT APPROACH

In the Nwe et al. [9] system, a subset of features, similar to the Mel Frequency Cepstral Coefficients (MFCC), was used. They used the Log Frequency Power Coefficients (LFPC) over a Hidden Markov Model (HMM) to classify emotions in speech. Their work is not publicly available, as they used a dataset privately available to them. However, they claim that using the LFPC coefficients over the MFFCC coefficients shows a significant improvement in terms of the accuracy of the model. The average classification accuracy in

their model is 78% and the best accuracy is even higher 96%.

2.5 EXTENDING THE FEATURE SPACE

Rong et al. [10] proposed an innovative way to improve the accuracy of existing models. Traditionally, computer scientists were using various pre-processing techniques to reduce the number of features. Contrastingly, this new system increased the number of features used for classification. They claimed to have performed classification over a small dataset containing audio percepts in the Chinese language, but do not disclose the features that they used. However, they also mentioned that none of their features are language-dependent. Using a high number of features over an ensemble random forest algorithm (ERFTrees), they achieved an accuracy of 82.54%.

Two datasets created in the English language, namely the Toronto Emotional Speech Set (TESS) and the emotional dataset from Knowledge Extraction based on Evolutionary Learning (KEEL), contain a more diverse and realistic audio. The descriptions of the dataset are as follows.

2.6 TORONTO EMOTIONAL SPEECH SET (TESS)

The researchers from the Department of Psychology at the University of Toronto have created a speech emotion based dataset in 2010, in the English language [12]. The database contains 2800 sound files of speech utterances in seven basic emotional categories, namely: Happy, Sad, Angry, Surprise, Fear, Disgust and Neutral. It is an acted recording, where actors from two age groups of Old (64-year-old) and Young (26-year-old) had performed the dictation.

CHAPTER 3
PROPOSED METHODOLOGY

A few qualities of this dataset which makes it good for this project are:

3.1 KNOWLEDGE EXTRACTION BASED ON EVOLUTIONARY LEARNING (KEEL)

KEEL is an online dataset repository contributed by machine learning researchers worldwide [13]. The emotion for speech dataset contains 72 features extracted for each of the 593 sound files. The data are labeled across six emotions, namely: Happy, Sad, Angry, Surprise, Fear and Neutral. The repository also offers data to be downloaded in 10 or 5 folds for the purpose of training and testing.

A few qualities of this dataset which makes it good for this project are:

- Data is represented as features directly, which saves conversion time and procedures.
- All basic emotional categories of data are present. A combination of these emotions can be used for further research like Sarcasm and Depression detection.

3.1.1 Supervised Learning

Supervised learning is the most common type of machine learning, where the agent learns to map between input *x* and output *y* [26]. Data labeled with outputs *y* are needed for the learning process to take place; this means such datasets have to be produced somehow, often by hand, which is costly. A classical problem suitable for supervised learning is how to predict housing prices. The agent or model is fed with a training dataset of known houses. These houses should both have data for input

features and output prices. The input x consists of features such as the number of rooms, housing area, backyard size,etc. The output y contains the price of each house. Other types of learning include semi-supervised learning and unsupervised learning (clustering), wherelabels y are either partly or entirely unknown [26].

3.1.2 Unsupervised Learning

The sudden rise of big data has become the center of attention in a lot of industries, causing a major need for establishing processes for big data management, as well as modeling and analysis to utilize the potential of the data. In the real world, structured data, i.e. data that is organized into spreadsheets or relational databases only constitute 5% of all data [30]. The remaining 95% of data is not homogeneous nor free from noise and constitute what is known as *un- structured data*. Unstructured data lack the organization that is often required by machines to learn; common examples of such data are text, images, video,and audio [27]. Approaches to learn from unstructured data, involves using in- formation extraction techniques like entity recognition that outputs structureddata, as well as deploying machine learning models such as CNNs that can automatically extract features and learn from unstructured data [31] [32] [22].

CHAPTER 4
IMPLEMENTATION PROCESS

4.1 THE PROCESS

Speech is a varying sound signal. Humans are capable of making modifications to the sound signal using their vocal tract, tongue, and teeth to pronounce the phoneme. The features are a way toquantify data. A better representation of the speech signals to get the most information from the speech is through extracting features common among speech signals. Some characteristics of good features include [14]:

- The features should be independent of each other. Most features in the feature vector are correlated to each other. Therefore, it is crucial to select a subset of features that are individualand independent of each other.
- The features should be informative to the context. Only those features that are more descriptive about the emotional content are to be selected for further analysis.
- The features should be consistent across all data samples. Features that are unique and specific to certain data samples should be avoided.

The values of the features should be processed. The initial feature selection process can resultin a raw feature vector that is unmanageable. The process of Feature Engineering will removeany outliers, missing values, and null values.

The features in a speech percept that is relevant to the emotional content can be grouped into two main categories:

1. Prosodic features
2. Phonetic features.

The prosodic features are the energy, pitch, tempo, loudness, formant, and intensity. The phonetic features are mostly related to the pronunciation of the words based on the language. Therefore, for the purpose of emotion detection, the analysis is performed on the prosodic features or a combination of them. Mostly the pitch and

loudness are the features that are very relevant to the emotional content.

4.2 LIST OF FEATURES

See Table 1 for the features that were extracted for each frame of the audio signal, along with theirdefinitions [15].

Table.1 List of features present in an audio signal

Feature ID	Feature Name	Description
1	Zero Crossing Rate	"The rate at which the signal changes its sign."
2	Energy	"The sum of the signal values squared and normalized using frame length."
3	Entropy of Energy	"The value of the change in energy."
4	Spectral Centroid	"The value at the center of the spectrum."
5	Spectral Spread	"The value of the bandwidth in the spectrum."
6	Spectral Entropy	"The value of the change in the spectral energy."
7	Spectral Flux	"The square of the difference between the spectral energies of consecutive frames."
8	Spectral Rolloff	"The value of the frequency under which 90% of the spectral distribution occurs."
9-21	MFCCs	"Mel Frequency Cepstral Coefficient values of the frequency bands distributed in the Mel-scale."
22-33	Chroma Vector	"The 12 values representing the energy belonging to each pitch class."
34	Chroma Deviation	"The value of the standard deviation of the Chroma vectors."

4.2.1 Mel Frequency Cepstrum Coefficients (MFCC) FEATURES

A subset of features that are used for speech emotion detection is grouped under a category called the Mel Frequency Cepstrum Coefficients (MFCC) [16]. It can be explained as follows:

- The word Mel represents the scale used in Frequency vs Pitch measurement (see Figure 2)[16]. The value measured in frequency scale can be converted into Mel scale using the formula **m = 2595 log10 (1 + (f/700))**
- The word Cepstrum represents the Fourier Transform of the log spectrum of the speechsignal.

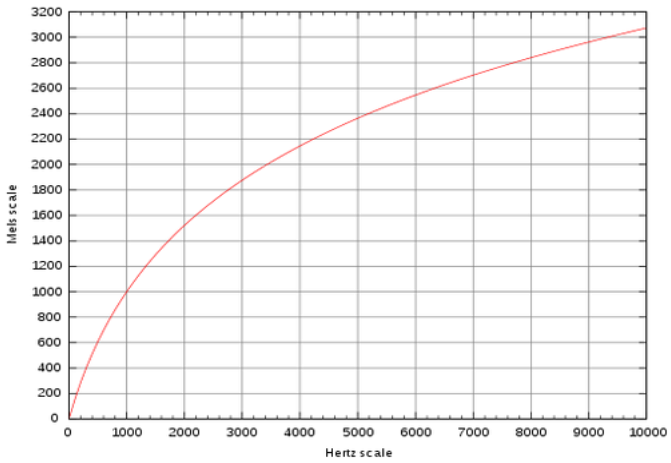

Fig.2 The Mel scale

4.2.2 COEFFICIENT COMPUTATION

Below is the mathematical approach to compute the MFCC features from a speech signal [16]:

The first step is to frame the audio signal. The method of frame blocking discussed earlier is used tosplit the audio signals into frames of an optimal length of 20ms to 30ms, with 50% overlap.

The next step is mathematical. In this step, for each frame of the signal, the power spectrum is computed. The power spectrum, also known as Periodogram, identifies the frequencies present in each frame (see Figure 3) [16]. In order to select a particular band of frequencies, the value at each frame is multiplied by a Hamming window value. Mathematically, the periodogram is the squared value of the modulus of the Discrete Fourier Transform (DFT).

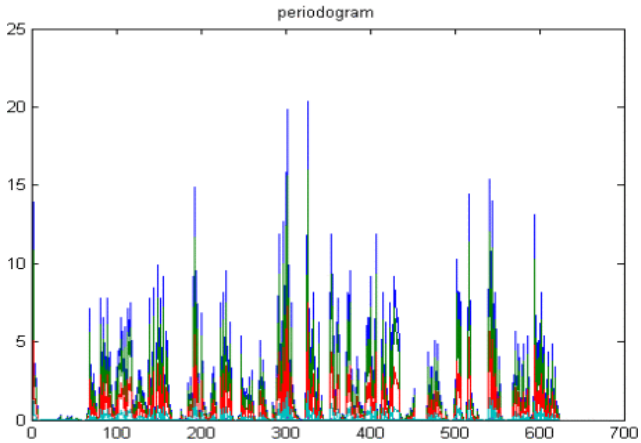

Fig.3 Periodogram

Next, the power spectra obtained can contain many closely spaced frequencies. These variations in the frequencies make it difficult to obtain the energy values present in the signal. Thus, to scale the values, a filter named Mel Filterbank is applied to the power spectrum. The Mel Filterbank is a collection of triangular filters in the frequency domain. Nearing 0Hz, the frequencies are narrow to each other; further higher, the frequencies become wider (see Figure 4) [16]. The product of the power spectrum values and the Mel Filterbank values provides the energy in each frame. However, since overlapping frames are used in the analysis, the energy values obtained for the individual frames would be correlating with each other.

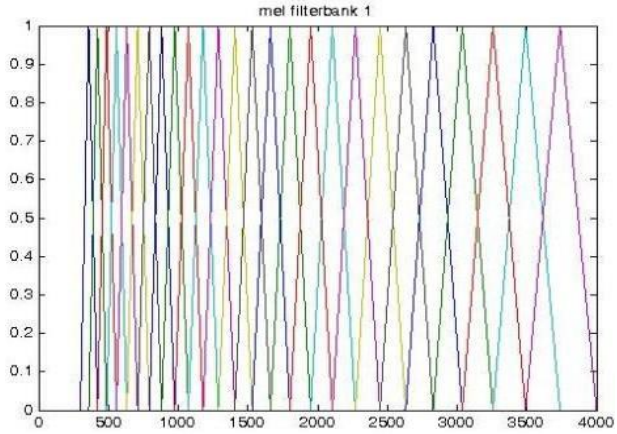

Fig.4 The Mel filterbank

The final step is to de-correlate the energies. For this purpose, the Discrete Cosine Transform (DCT) function is used. DCT outputs a list of coefficient values corresponding to the pitch and energy values obtained so far. The lower level coefficients (the first 12 to 13 coefficients) of each frame represents steady changes in the pitch and energy values, and therefore they are better for analysis. These lower-level coefficients are called the Mel Frequency cepstral coefficients.

4.3 FRAME BLOCKING

The frame blocking method is used to analyze sound signals. It is the process of dividing the sound signal into blocks known as frames and performing analysis over each block, rather than the signal at large. It is preferred to analyze individual frames because audio signals are stable within short time intervals. Several acoustic features can be interpreted from a single frame. In order to ensure the time-varying characteristics of the signal are measured accurately, some part of the neighboring frames is also analyzed at every step to identify any subtle changes in the sound signal. This value is often termed as frame overlap, indicating the amount of overlap to include from the neighboring frames. The steps of frame blocking are as follows [17]:

1. Set *frame size* value to an appropriate number.

Each frame should not be too small or too large, as this would mislead the time-varying characteristics of the features. Standard framing window size is 20ms to 30ms for audio signal processing.

2. Set *frame overlap* value.

If the overlap value is too large, more duration from the neighboring frames will need to be analyzed at each analysis step. This will increase the computation and hence is not recommended. Ideally, ½ or 1/3 of frame overlap is suggested.

3. Perform analysis on each frame

Each frame is a unit of computation of the sound signal. Feature extraction of frames will quantify the acoustic features of the audio signal.

4.4 SILENCE REMOVAL

An audio signal at the time of recording can accommodate silent regions where no utterances had been made. Such silent regions of the audio signal do not provide any useful information regarding the emotion expressed, and can be removed. A semi-supervised learning approach is used for silence detection in audio signals. In this approach, a model is initially trained with sample audio signals in order to be able to distinguish between high and low energy features [15]. Later, a percentage of high and low energy frames are used as endpoints to detect the regions of actual audio in the signal. Finally, applying the trained model over the entire signal will provide silence-free audio segments [15]. Threshold values such as frame size, frame step, and sampling frequency are tunable in order to smooth the output signals.

CHAPTER 5

RESULT

5.1 DATA QUALITY ISSUES

Data must be cleaned to perform any meaningful analysis. As a next step, the dataset thus collected had to be inspected for its quality. Some of the data quality issues addressed for this experimentation includes:

1. Missing value analysis
2. Outlier identification
3. Null value handling
4. Invalid data
5. Duplicate data

5.1.1 MISSING VALUE ANALYSIS

Due to several influencing factors, a few or more data rows can contain no values for specific features. These values are termed ‘missing’ from the dataset. A large number of missing values can provide insights into the data. For example, if a particular feature has most of its values missing for all data rows, then it can be inferred that the feature is likely uncommon and can be removed from the dataset. Contrastingly, a small number of missing values can represent data entry error. Analyzing and amending individual features over missing values will improve and fix the quality of the dataset. Some of the methods to handle missing values include; if the number of missing values is large then the corresponding data rows or features can be removed, whereas if very few values are missing for a feature it can be imputed which means replacing with the mean or most frequent value of the feature.

5.1.2 OUTLIER IDENTIFICATION

Outlier values are also considered as data modifiers because often the prediction algorithm used will be misled by the outlier values. Outliers also alter the statistics of the overall data such as mean, variance and standard deviation. The proportion of the outliers amongst the whole dataset can be used to made decisions on how to handle them. If the outliers lie within a small range of difference and contribute to a very small proportion, then no fixes will be required. Some methods to handle large proportions of outliers include, replacing outlier values with boundary, or mean, or median, or mode values.

5.1.3 NULL VALUE HANDLING

A common error that can occur in a dataset is the null value error. It is when the words ‘null’ or ‘NA’ is used in place of missing values as fillers. Null values are mostly treated and handled in a fashion similar to missing values.

5.1.4 INVALID DATA

A dataset can have values irrelevant to the data type, such as symbols and special characters. These values, despite being meaningless, can cause errors during processing. Depending on the amount of invalid data present, it can either be removed or imputed.

5.1.5 DUPLICATE DATA

Few features might be a duplicate of each other, with different names or units of measurement. Such features increase the dimensionality of the data with no further significance. Removal of duplicate features is highly recommended.

5.2 NORMALIZATION AND STANDARDIZATION

Different characteristics of the audio signal, represented by its features, are computed on different units or scales. Rescaling the values to a uniform range will ensure accurate calculations are made. Many algorithms use distance metrics for their computation. Therefore it is necessary that all the values in the dataset are normalized. Two approaches are commonly used for the purpose of rescaling, namely Normalization and Standardization. Normalization alters all numeric values to lie in the range 0 to1. For this purpose, all outliers in the data must be eliminated prior to normalizing the data. The formula for normalization is given as

$$x_{new} = (x - x_{min}) / (x_{max} - x_{min})$$

In the formula, x represents the data [18].

Standardization transforms the data to have a mean value of zero and a variance of one. However, standardizing the data provides more insights into the data than normalization. The formula for normalization is given as

$$x_{new} = (x - \mu) / \sigma$$

In the formula, x represents the data [18].

5.3 PEARSON CORRELATION COEFFICIENT

Correlation is the liner association that exists between each pair of features in the dataset and is used to identify the features that are highly associated or correlated with the decision attribute [19]. The Pearson correlation coefficient, *r* is a value that denotes the strength of the correlation. The value of *r* ranges between -1 to +1 through 0, where the negative values denote a lesser correlation between variables and the positive values denote greater correlation. A value of 0 denotes that there is no correlation between the variables (see Figure 5) [19].

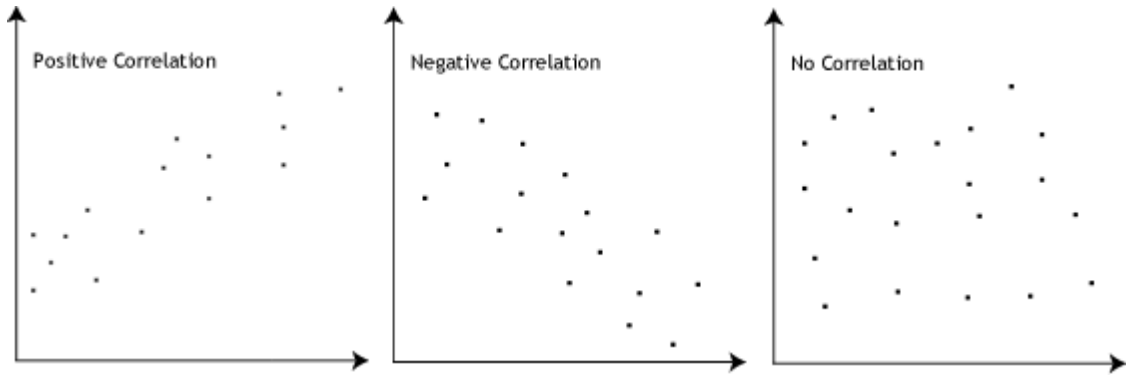

Fig.5 Pearson correlation coefficient

The strength of the association can be determined using the value of the Pearson coefficient *r*. However, the strength of the association also depends on the type of variables under measurement. The Person method of correlation computation can be used on all numeric data irrespective of whether they have been scaled or not. Additionally, this approach treats all variables equally and does not consider any proposed dependence between the variables. The following guidelines have been proposed to determine the strength of correlation (see Figure 6).

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Fig.6 Pearson correlation coefficient guidelines

5.4 CLUSTERING

K MEANS CLUSTERING ALGORITHM

K-Means is an unsupervised clustering algorithm that will form ‘k’ groups within the data based on feature similarity. It is an iterative process by which data are iteratively grouped based on the similarity between their features. Clustering the data into groups provides more insights on the distribution of the training data available, and also easily helps classify any unknown (new) data. The algorithm is a two-step iterative process in which, based on distance metrics between the data points and centroids of the cluster, groups of similar data are created. The steps of the algorithm are [20]:

Initially, random values of the centroid(s) are assumed.

1. Data Assignment

Based on the value of a distance metric (For example, Euclidean distance and Manhattan distance), data points are assigned to the closest neighboring centroids.

2. Centroid re-computation

   Centroids or mean value of all data points are calculated and updated at each step following the data assignment step.

Steps 1 and 2 are iteratively performed until the groups are distinctively classified. There are two conditions that ensure accuracy in clustering, namely: the inter-cluster distance and intra-cluster difference. The distance between the centroids of each cluster should be larger, ensuring that each group is well-separated from each other showing distinct differences. Additionally, the distance between each point within the cluster should be smaller, ensuring the similarity between data points within the group. By analyzing the final value of each centroid, the characteristics of the data belonging to the cluster can be quantitatively explained.

5.4.1 THE ELBOW METHOD-CHOOSING K-VALUE

The ‘k’ in k-means denote the number of clusters the data needed to be grouped into. Traditionally, the algorithm is repeated over different values of k and the results are compared by the average within cluster distance to the centroid. Alternatively, the elbow method can be used to depict an optimal value of k [20]. In the elbow method, the average within cluster distance to the centroid value is plotted against different values of k and the point of the curve where the distance sharply bends is the optimal value of k (see Figure 7).

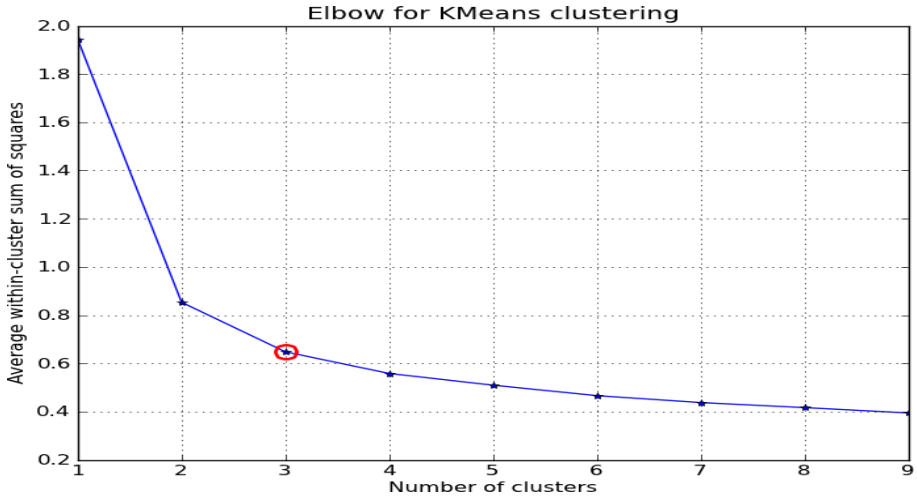

Fig.7 Elbow method for K means clustering

5.5 PRINCIPAL COMPONENT ANALYSIS

In some cases, most or all of the features might have an impact on the decision making. However,a high dimensional dataset with a large feature space could potentially slow the performance of the system in terms of space and time complexity. Choosing the right set of features for analysis can be challenging, as it requires high-level domain knowledge. A solution to this problem is the Principle Component Analysis (PCA) technique for dimensionality reduction. PCA is an approach to bring out the principal components or the important aspects of the data. By using this method, the original feature space of the data is transformed into a new set of features while retaining the variation present in the data. The technique of PCA analyses the variance of the data by measuring the covariance between the features. This is done mathematically using the concept of eigenvalues and eigenvectors. Eigenvalues are numbers denoting the value of variance in each dimension of the data, and the eigenvector is the dimension with the highest eigenvalue. This eigenvector is the principal component of the dataset [22]. Given below are the implementation steps of PCA [21].

5.5.1 DATA NORMALIZATION

PCA works with the numerical values of the dataset to compute the variance, hence it is necessary that the values are scaled and normalized. All normalized data variables will have a mean value 1 of 0.

5.5.2 COVARIANCE MATRIX COMPUTATION

An NxN covariance matrix is computed, where N is the number of features in the dataset. The elements of the covariance matrix represent the variance between each of the features in the dataset.

5.5.3 EIGENVALUES AND EIGENVECTOR COMPUTATION

Eigenvalues and eigenvectors are computed using the covariance matrix. This computation is purely mathematical and many programming libraries have built-in functions for this calculation. At the end of the computation, N eigenvalues for an N-dimensional dataset is obtained. The eigenvalues thus obtained are the components of PCA [22].

5.5.4 CHOOSING COMPONENTS

The eigenvector component with the largest eigenvalue is the 1^{st} Principal component, containing the most information about the dataset. Sorting the eigenvalues in decreasing order can give the list of principal components with the amount of variance needed. Depending on how much information is needed, programmers can choose the top P number of components needed for further analysis.

5.5.5 FORMING PRINCIPLE COMPONENTS

A new dataset is created using the principal components selected for analysis. Mathematically, left multiplication of the transposed feature vector with the scaled original dataset will produce the new dataset.

New Dataset = $(\text{Feature Vector})^T$ X $(\text{Scaled Data})^T$

5.5.6 SCREE PLOT

Scree plot is a way to select the optimal number of components to be selected such that enough information is being retained form the raw dataset. It is a curve plot having the information maintained and the number of components on the different axis. The elbow point of the curve indicates the optimal value of components to be used for further analysis.

5.6 LOGISTIC REGRESSION

Logistic Regression is a supervised classification algorithm which produces probability values of data belonging to different classes [23]. There are three types of Logistic Regression algorithms, namely Binary class, Multi-class and Ordinal class logistic algorithms depending on the type of target class. The Wikipedia definition states that "*Logistic regression computes the relationship between the target (dependent) variable and one or more independent variables using the estimated probability values through a logistic function"* [24]. The logistic function, also known as a sigmoid function, maps predicted values to probability values. The procedure of a multiclass logistic regression algorithm is as follows:

For an N class problem, divide into N pairs of binary class problems.

For each binary class problem

- For each observation of a binary class problem
- Compute probability values of the observation belonging to a class

Make the final prediction by computing the maximum probability value amongst all classes The time complexity of the algorithm is in the order of the number of data samples, represented as O (n samples).

5.7 NAÏVE BAYES

Naïve Bayes classifier is based on Bayes theorem, which determines the probability of an event based on a prior probability of events [26]. Bayes theorem is used to compute prior probability values.

This classifier algorithm assumes feature independence. No correlation between the features is considered. The algorithm is said to be Naïve because it treats all the features to independently contribute to deciding the target class. The steps of a simple Naïve Bayes

algorithm is as follows [25]:

1. Create a frequency table for all features individually. Tag the frequency of each entryagainst the target class.
2. Create a likelihood table by computing probability values for each entry in the frequencytable.
3. Calculate posterior probability for each target class using the Bayes theorem.

Declare the target class with the highest posterior probability value as the predicted outcome.

5.8 SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) are a supervised algorithm that works for both classification and regression problems. Support vectors are coordinate points in space, formed using the attributes of a data point. Briefly, for an N-dimensional dataset, each data point is plotted on an N-dimensional space using all its feature vector values as a coordinate point [27]. Classification between the classes is performed by finding a hyperplane in space that clearly separates the distinct classes. SVM works best for high dimensional data. The important aspect of implementing SVM algorithm is finding the hyperplane. Two conditions are to be met in the order given while choosing the right hyperplane.

1. The hyperplane should classify the classes most accurately

   The margin distance from the hyperplane to the nearest data point must be maximized.

5.9 EVALUATION METRICS

The most important characteristic of machine learning models is its ability to improve. Once the model is built, even before testing the model on real data, machine learning experts evaluate the performance of the model. Evaluation

metrics reveal important model parameters and provides numeric scores that will help judge the functioning of the model. The most important metric needed to evaluate the model is the confusion matrix (see Figure 8) [32].

The structure of a confusion matrix is against the actual and predicted positive and negative classes, and contains four values which are used to compute other metrics. The true positive represents the correct predictions made in the positive class, and the true negatives represent the correct predictions made in the negative class. The false positives and false negatives are the observations wrongly predicted for their respective classes.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Fig.8 Confusion Matrix

Four important metrics can be derived using the values in the confusion matrix, namely [31]:

5.9.1 ACCURACY

It is the ratio of the observations predicted correctly to the total number of observations. Accuracy works best for datasets with an equal class distribution, and hence it is not always a good measure to evaluate the model. Accuracy can be computed as,

Accuracy = (True Positives + True Negatives) / (True Positives + False positives+True Negatives + False Negatives)

5.9.2 PRECISION

It is the ratio of the positive observations predicted correctly to the total positive observations predicted. Higher the value of precision, better and more accurate the

model actually is. Precision can also work with an uneven class distribution. It can be computed as,

Precision = True Positives / (True positives + False positives)

5.9.3 RECALL OR SENSITIVITY

It is the ratio of the positive observations predicted correctly to the total positive observations. A recall score of 50% and more reveals a good performing model. Recall can also work with an uneven class distribution. It can be computed as

Recall = True positives / (True positives + False negatives)

5.9.4 F1 score

It is the weighted average value of precision and recall. The F1 score is the best metric for uneven class distribution. F1 can be computed as

F1= 2 * (Recall * Precision) / (Recall + Precision)

5.10 DATA COLLECTION

The first step in implementing the Speech Emotion Recognition system is to collect audio samples under different emotional categories which can be used to train the model. The audio samples are usually wav or mp3 files and publicly available for download. The following steps are explained relative to the experiments performed on the TESS dataset.

5.11 PYTHON LIBRARY

The next step after data collection was to represent these audio files numerically, in order to perform further analysis on them. This step is called feature extraction, where quantitative values for different features of the audio is obtained. The pyAudioAnalysis library was used for this purpose [15]. This python library provides functions for short-term feature extraction, with tunable windowing parameters such as frame size and frame step. At the end of this step,

each audio file was represented as a row in a CSV file with 34 columns representing the different features. Each feature will have a range of values for one audio file obtained over the various frames in that audio signal. The python library pyAudioAnalysis is an open Python library that provides a wide range of audio-related functionalities focusing on feature extraction, classification, segmentation, and visualization issues. The library depends on several other libraries which are:

- Numpy
- Matplotlib
- Scipy
- Sklearn
- Hmmlearn
- Simplejson
- eyeD3
- pydub

5.12 DATA VISUALIZATION

Visualizing the data gives more understanding of the problem and the type of solution to be built. The distribution of classes, the number of instances under each category, the spread of the data, the correlation between the features and clustering are a few methods to visualize the data. Python and R provide statistical functions for data visualization.

5.12.1 FEATURE ANALYSIS

Primarily, the number of rows and columns and a preview of the data is viewed (see Figure 9.A).

```
In [3]: data = pd.read_csv('data.csv') #reading the csv data

In [4]: data.shape
Out[4]: (2399, 36)

In [5]: data.head(n=3) #preview of the data
Out[5]:
```

	1	2	3	4	5	6	7	8	9	10	...	27	28	29	30	31
0	0.237025	0.001949	2.180585	0.342264	0.210608	1.761245	0.001733	0.412815	-26.535934	0.319416	...	0.000223	0.001806	0.000748	0.000370	0.005889
1	0.255306	0.019663	2.964614	0.369318	0.233425	1.677484	0.009996	0.505462	-26.642530	1.078969	...	0.002802	0.008537	0.009402	0.003874	0.002074
2	0.262975	0.004484	1.195476	0.356051	0.253092	1.473366	0.003087	0.467647	-27.314425	0.848172	...	0.002532	0.001380	0.000552	0.000348	0.000545

3 rows × 36 columns

Fig.9.A. An overview of data

Next, the number of examples under each category are counted (see Figure 9.B).

```
In [6]: data['36'].value_counts()
Out[6]: Surprise    402
        Happy       400
        Disgust     400
        Neutral     400
        Sad         399
        Fear        200
        Angry       198
        Name: 36, dtype: int64
```

Fig.9.B. Overview of data grouped by categories

Distribution of the data on each of its feature can be visualized using box plots and violin plots or using pair plots. The Seaborn package in Python provides methods to draw such plots. Shown here is the data distribution of the feature ‘Energy’ among the different categories (see Figure 10.A and Figure 10.B).

```
In [12]: import seaborn as sns
         sns.boxplot(x="36", y="2", data=data)
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x5efd3e34a8>
```

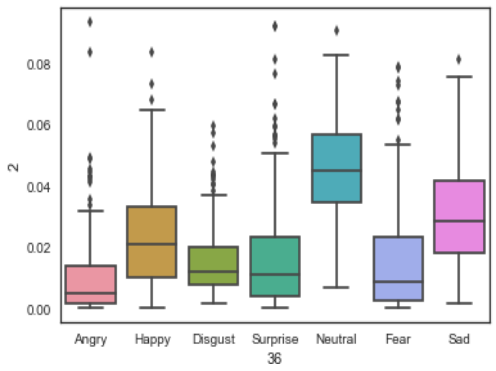

Fig.10.A. Box plot analysis of feature values

```
In [13]: sns.violinplot(x="36", y="2", data=data)
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x5efd4fb358>
```

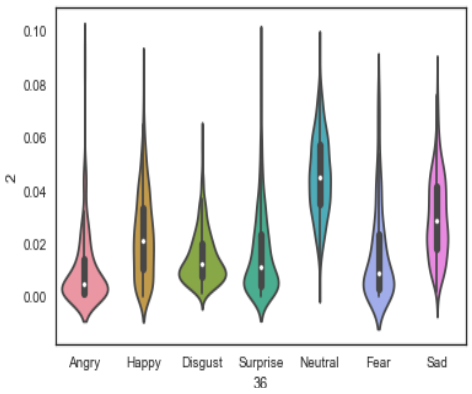

Fig.10.B. Violin plot analysis of feature values

The statistical language R provides several functions to effectively understand the statistics of the data. For each feature, the statistical values were visualized and it was observed that the raw data was not standardized (see Figure 11).

```
       X1                 X2                  X3                X4                X5                X6                 X7                  X8                X9          
 Min.   :0.01898   Min.   :0.0002479   Min.   :0.8198   Min.   :0.1076   Min.   :0.1338   Min.   :0.01319   Min.   :0.0001811   Min.   :0.02143   Min.   :-32.96  
 1st Qu.:0.15546   1st Qu.:0.0080144   1st Qu.:2.6473   1st Qu.:0.2931   1st Qu.:0.1840   1st Qu.:1.04099   1st Qu.:0.0035809   1st Qu.:0.31828   1st Qu.:-27.33  
 Median :0.22599   Median :0.0198240   Median :2.8741   Median :0.3289   Median :0.2000   Median :1.33882   Median :0.0061594   Median :0.39727   Median :-26.32  
 Mean   :0.23878   Mean   :0.0237834   Mean   :2.7632   Mean   :0.3455   Mean   :0.2023   Mean   :1.31511   Mean   :0.0073466   Mean   :0.40830   Mean   :-26.52  
 3rd Qu.:0.29751   3rd Qu.:0.0363465   3rd Qu.:2.9984   3rd Qu.:0.4005   3rd Qu.:0.2174   3rd Qu.:1.53587   3rd Qu.:0.0096574   3rd Qu.:0.48298   3rd Qu.:-25.54  
 Max.   :0.69626   Max.   :0.0933426   Max.   :3.2281   Max.   :0.7304   Max.   :0.3118   Max.   :3.00326   Max.   :0.0509154   Max.   :0.96534   Max.   :-22.13  

      X10               X11               X12               X13               X14               X15               X16               X17          
 Min.   :-1.122   Min.   :-2.1393   Min.   :-1.25896   Min.   :-1.4527   Min.   :-1.03556   Min.   :-1.04332   Min.   :-1.20529   Min.   :-1.3340107  
 1st Qu.: 1.026   1st Qu.:-0.7804   1st Qu.:-0.21547   1st Qu.:-0.6456   1st Qu.:-0.32735   1st Qu.:-0.15881   1st Qu.:-0.17463   1st Qu.:-0.1659003  
 Median : 1.498   Median :-0.3394   Median : 0.07098   Median :-0.4172   Median :-0.07785   Median : 0.04823   Median :-0.03496   Median :-0.0004108  
 Mean   : 1.495   Mean   :-0.3457   Mean   : 0.04115   Mean   :-0.4169   Mean   :-0.04215   Mean   : 0.01938   Mean   :-0.06578   Mean   :-0.0518579  
 3rd Qu.: 2.010   3rd Qu.: 0.1162   3rd Qu.: 0.32359   3rd Qu.:-0.1893   3rd Qu.: 0.19766   3rd Qu.: 0.20017   3rd Qu.: 0.08734   3rd Qu.: 0.1305859  
 Max.   : 3.780   Max.   : 1.3581   Max.   : 1.12542   Max.   : 0.5616   Max.   : 1.04685   Max.   : 0.94772   Max.   : 0.49288   Max.   : 0.7206110  

      X18               X19               X20               X21               X22               X23               X24               X25          
 Min.   :-0.96893   Min.   :-0.8357281   Min.   :-0.79690   Min.   :-0.85185   Min.   :0.0000211   Min.   :0.0000271   Min.   :0.0000623   Min.   :0.0000744  
 1st Qu.:-0.30044   1st Qu.:-0.1358281   1st Qu.:-0.29167   1st Qu.:-0.33274   1st Qu.:0.0005862   1st Qu.:0.0004490   1st Qu.:0.0009469   1st Qu.:0.0005527  
 Median :-0.15212   Median :-0.0002672   Median :-0.09269   Median :-0.17381   Median :0.0015330   Median :0.0010544   Median :0.0023683   Median :0.0012137  
 Mean   :-0.17123   Mean   :-0.0049830   Mean   :-0.06890   Mean   :-0.13697   Mean   :0.0050965   Mean   :0.0033127   Mean   :0.0047547   Mean   :0.0034065  
 3rd Qu.:-0.02544   3rd Qu.: 0.1328477   3rd Qu.: 0.11949   3rd Qu.: 0.03094   3rd Qu.:0.0048227   3rd Qu.:0.0035430   3rd Qu.:0.0056950   3rd Qu.:0.0035870  
 Max.   : 0.51956   Max.   : 0.9343624   Max.   : 1.03844   Max.   : 0.74888   Max.   :0.1042071   Max.   :0.1137400   Max.   :0.0840600   Max.   :0.0827991  

      X26               X27               X28               X29               X30               X31               X32               X33          
 Min.   :0.0000979   Min.   :0.0000397   Min.   :0.0000409   Min.   :0.0000350   Min.   :0.0000362   Min.   :0.0000343   Min.   :0.0000418   Min.   :0.0000305  
 1st Qu.:0.0009131   1st Qu.:0.0003715   1st Qu.:0.0008550   1st Qu.:0.0007248   1st Qu.:0.0005415   1st Qu.:0.0008563   1st Qu.:0.0033102   1st Qu.:0.0007767  
 Median :0.0020207   Median :0.0010257   Median :0.0022048   Median :0.0018806   Median :0.0016713   Median :0.0032559   Median :0.0096335   Median :0.0020540  
 Mean   :0.0058273   Mean   :0.0049895   Mean   :0.0046888   Mean   :0.0051188   Mean   :0.0067453   Mean   :0.0124376   Mean   :0.0161145   Mean   :0.0082052  
 3rd Qu.:0.0052759   3rd Qu.:0.0038781   3rd Qu.:0.0054777   3rd Qu.:0.0051923   3rd Qu.:0.0064854   3rd Qu.:0.0126654   3rd Qu.:0.0225707   3rd Qu.:0.0073837  
 Max.   :0.1346614   Max.   :0.1143216   Max.   :0.1299185   Max.   :0.1319882   Max.   :0.1165654   Max.   :0.1797250   Max.   :0.1867128   Max.   :0.1422646  

      X34              X35            X36       
 Min.   :0.0002408   Old  :1399   Angry   :198  
 1st Qu.:0.0062759   Young:1000   Disgust :400  
 Median :0.0105646                Fear    :200  
 Mean   :0.0128518                Happy   :400  
 3rd Qu.:0.0169517                Neutral :400
```

Fig.11 Summary of data-before standardization

5.13 CORRELATION

The Pearson correlation coefficient values were analyzed which provided insights about the features that correlate positively and negatively with the target class. In this observation, no features had a negative correlation with the target class (see Figure 12).

```
> flattenCorrMatrix(round(res2$r,2), round(res2$P,2))
    row column   cor    p
1    X1     X2 -0.19 0.00
2    X1     X3 -0.05 0.03
3    X2     X3  0.37 0.00
4    X1     X4  0.88 0.00
5    X2     X4 -0.19 0.00
6    X3     X4 -0.05 0.03
7    X1     X5 -0.10 0.00
8    X2     X5 -0.16 0.00
9    X3     X5 -0.41 0.00
10   X4     X5  0.10 0.00
11   X1     X6  0.51 0.00
12   X2     X6 -0.19 0.00
13   X3     X6 -0.02 0.24
14   X4     X6  0.72 0.00
15   X5     X6  0.25 0.00
16   X1     X7 -0.36 0.00
17   X2     X7  0.37 0.00
18   X3     X7  0.33 0.00
19   X4     X7 -0.39 0.00
20   X5     X7 -0.15 0.00
21   X6     X7 -0.49 0.00
22   X1     X8  0.80 0.00
23   X2     X8 -0.16 0.00
24   X3     X8  0.04 0.07
25   X4     X8  0.93 0.00
26   X5     X8  0.07 0.00
27   X6     X8  0.85 0.00
28   X7     X8 -0.42 0.00
29   X1     X9 -0.45 0.00
30   X2     X9  0.48 0.00
31   X3     X9  0.05 0.03
32   X4     X9 -0.62 0.00
33   X5     X9 -0.07 0.00
34   X6     X9 -0.38 0.00
35   X7     X9  0.05 0.01
36   X8     X9 -0.54 0.00
472  X7    X32  0.18 0.00
473  X8    X32 -0.31 0.00
474  X9    X32 -0.03 0.09
475 X10    X32  0.22 0.00
476 X11    X32  0.10 0.00
477 X12    X32  0.11 0.00
478 X13    X32  0.03 0.12
479 X14    X32 -0.22 0.00
480 X15    X32  0.10 0.00
481 X16    X32  0.06 0.00
482 X17    X32  0.29 0.00
483 X18    X32  0.23 0.00
484 X19    X32  0.07 0.00
485 X20    X32  0.01 0.65
486 X21    X32 -0.32 0.00
487 X22    X32 -0.03 0.15
488 X23    X32 -0.03 0.16
489 X24    X32 -0.02 0.38
490 X25    X32 -0.10 0.00
491 X26    X32 -0.13 0.00
492 X27    X32 -0.11 0.00
493 X28    X32  0.03 0.14
494 X29    X32  0.04 0.04
495 X30    X32  0.07 0.00
496 X31    X32  0.12 0.00
497  X1    X33  0.00 0.98
498  X2    X33  0.12 0.00
499  X3    X33  0.07 0.00
500  X4    X33 -0.02 0.34
501  X5    X33 -0.06 0.00
502  X6    X33 -0.17 0.00
503  X7    X33  0.35 0.00
504  X8    X33 -0.07 0.00
505  X9    X33 -0.08 0.00
506 X10    X33  0.19 0.00
507 X11    X33  0.17 0.00
508 X12    X33  0.16 0.00
509 X13    X33  0.09 0.00
```

Fig.12 Correlation values of data

5.14 CLUSTERING

Clustering the data provided a deeper understanding of the features. The k-means clustering was performed iteratively for various values of k and evaluated against the sum of squares metric. For k values less than 7, the cluster grouping was imbalanced, and for values greater than 10, the clusters were becoming too spread out. The elbow method of plotting reveals a sharp turn at k=7 which produces balanced clusters (see Figure 13). Interestingly there are 7 categories of emotions tagged in the dataset, hence making 7 distinct cluster groups of the data shows a clear separation with the feature values among the groups.

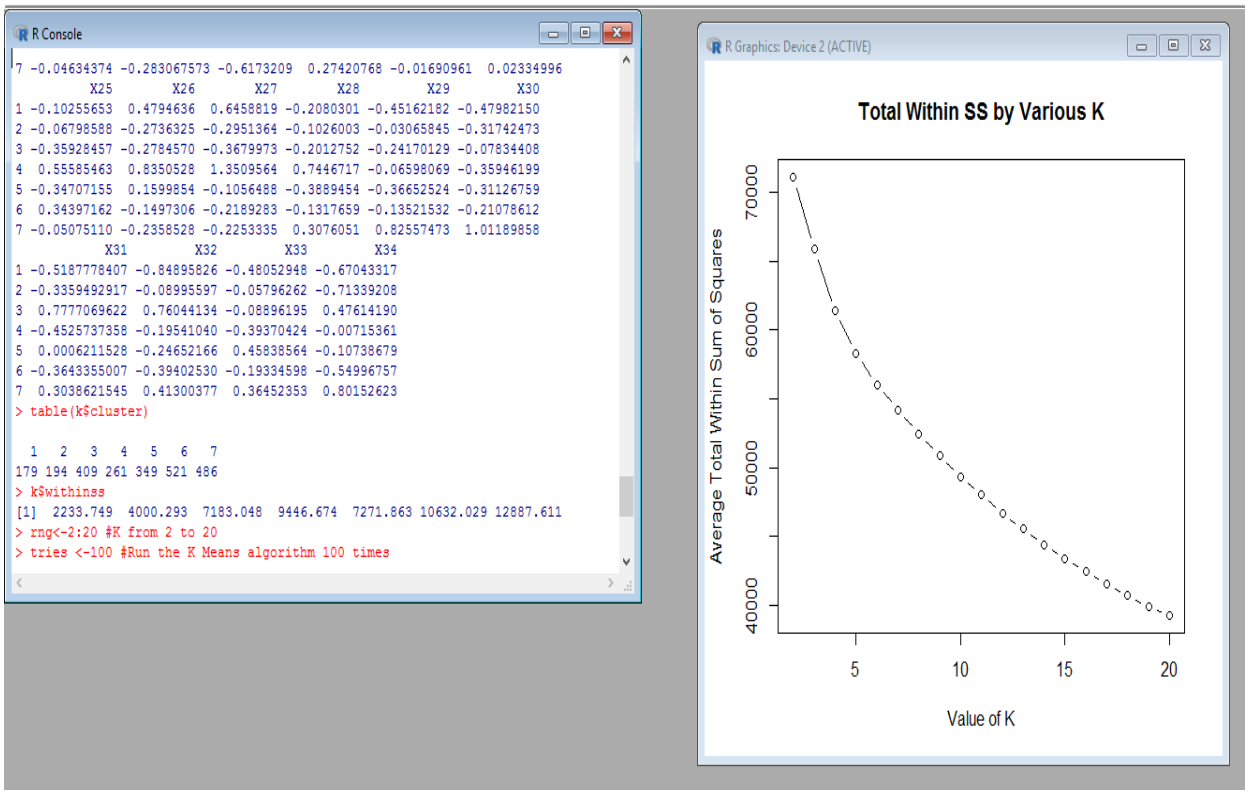

Fig.13 K-Means clustering of data and the elbow method

5.15 DATA PREPARATION

After analyzing the data through various visualizations, the next step is to prepare the data for processing. The steps of data preparation include fixing quality issues, standardization, and normalization. First, the data is checked for quality issues such as missing values (see Figure 14), outliers (see Figure 15), invalid data and duplicate data. There were no missing values, invalid or duplicate values in the dataset.

```
> sapply(data, function(x) sum(is.na(x)))
 X1  X2  X3  X4  X5  X6  X7  X8  X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 X33 X34 X35 X36
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
>
```

Fig.14 Missing value analysis

With the outlier analysis, for each feature, the proportion of the outliers are viewed along with the changes in the mean value of the feature with and without the outliers (see Figure 17). This insight will help decide if the outliers are actual outliers or if they contribute to decision making.

```
> outlierKD(data,data$X20)
Outliers identified: 39 from 2399 observations
Proportion (%) of outliers: 1.62567736556899
Mean of the outliers: 0.850866036692308
Mean without removing outliers: -0.0689005157190496
Mean if we remove outliers: -0.0841000477292373
```

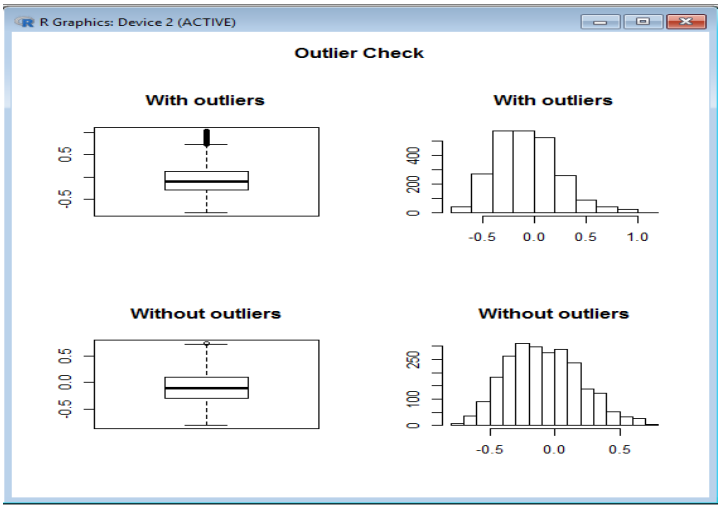

Fig.15 Outlier analysis

Next, normalization was performed on the data as the raw data was recorded on a different scale.

After standardization, all feature values now are in the range 0 to 1 (see Figure 16).

5.16 FEATURE ENGINEERING

Feature engineering is the process of transforming, reducing or constructing features for the dataset. As mentioned earlier in the raw data, each feature has multiple values for each frame of the audio signal. By the frame blocking and windowing techniques, the frame size and frame overlap values can be tuned to obtain accurate values of the audio signal. Further, using the averaging technique, average values of different features for the audio signals are obtained. Now the transformed data contains 34 discrete values representing each audio signal (see Figure 17).

```
> normalize<-function(x){
+           return ((x-min(x))/ (max(x)-min(x)))}
> data_norm <- as.data.frame(lapply(data[1:34],normalize)
+ )
> summary(data_norm)
       X1                X2                X3                X4                X5                X6                X7                X8                X9
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.2015   1st Qu.:0.08343   1st Qu.:0.7589   1st Qu.:0.2977   1st Qu.:0.2819   1st Qu.:0.3437   1st Qu.:0.06701   1st Qu.:0.3145   1st Qu.:0.5197
 Median :0.3057   Median :0.21028   Median :0.8530   Median :0.3553   Median :0.3721   Median :0.4433   Median :0.11784   Median :0.3982   Median :0.6129
 Mean   :0.3245   Mean   :0.25281   Mean   :0.8070   Mean   :0.3819   Mean   :0.3850   Mean   :0.4354   Mean   :0.14124   Mean   :0.4099   Mean   :0.5952
 3rd Qu.:0.4112   3rd Qu.:0.38776   3rd Qu.:0.9046   3rd Qu.:0.4703   3rd Qu.:0.4697   3rd Qu.:0.5092   3rd Qu.:0.18678   3rd Qu.:0.4890   3rd Qu.:0.6855
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
      X10               X11               X12               X13               X14               X15               X16               X17               X18
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.4382   1st Qu.:0.3886   1st Qu.:0.4376   1st Qu.:0.4007   1st Qu.:0.3401   1st Qu.:0.4442   1st Qu.:0.6069   1st Qu.:0.5685   1st Qu.:0.4491
 Median :0.5344   Median :0.5146   Median :0.5578   Median :0.5141   Median :0.4599   Median :0.5482   Median :0.6892   Median :0.6491   Median :0.5487
 Mean   :0.5338   Mean   :0.5128   Mean   :0.5453   Mean   :0.5142   Mean   :0.4771   Mean   :0.5337   Mean   :0.6710   Mean   :0.6240   Mean   :0.5359
 3rd Qu.:0.6389   3rd Qu.:0.6449   3rd Qu.:0.6637   3rd Qu.:0.6272   3rd Qu.:0.5922   3rd Qu.:0.6245   3rd Qu.:0.7612   3rd Qu.:0.7128   3rd Qu.:0.6339
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      X19               X20               X21               X22               X23               X24               X25               X26               X27
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
 1st Qu.:0.3954   1st Qu.:0.2753   1st Qu.:0.3243   1st Qu.:0.005424   1st Qu.:0.003710   1st Qu.:0.01053   1st Qu.:0.005781   1st Qu.:0.006058   1st Qu.:0.002903
 Median :0.4720   Median :0.3837   Median :0.4236   Median :0.014512   Median :0.009035   Median :0.02745   Median :0.013772   Median :0.014289   Median :0.008627
 Mean   :0.4693   Mean   :0.3967   Mean   :0.4466   Mean   :0.048715   Mean   :0.028893   Mean   :0.05586   Mean   :0.040280   Mean   :0.042578   Mean   :0.043312
 3rd Qu.:0.5472   3rd Qu.:0.4993   3rd Qu.:0.5515   3rd Qu.:0.046087   3rd Qu.:0.030919   3rd Qu.:0.06706   3rd Qu.:0.042461   3rd Qu.:0.038480   3rd Qu.:0.033587
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
      X28               X29               X30               X31               X32               X33               X34
 Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
 1st Qu.:0.006268   1st Qu.:0.005228   1st Qu.:0.004336   1st Qu.:0.004574   1st Qu.:0.01751   1st Qu.:0.005246   1st Qu.:0.1072
 Median :0.016661   Median :0.013987   Median :0.014032   Median :0.017929   Median :0.05138   Median :0.014227   Median :0.1834
 Mean   :0.035786   Mean   :0.038527   Mean   :0.057575   Mean   :0.069026   Mean   :0.08610   Mean   :0.057473   Mean   :0.2240
 3rd Qu.:0.041861   3rd Qu.:0.039084   3rd Qu.:0.055344   3rd Qu.:0.070294   3rd Qu.:0.12069   3rd Qu.:0.051698   3rd Qu.:0.2969
 Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
>
```

Fig.16 Summary of data-after standardization

Reducing the number of features is a crucial decision to take. Considering features to be removed is generally based on subject knowledge and hence can affect the performance of the system. Next, a series of experiments are performed with this prepared dataset in order to analyze the important features.

Several observations and conclusions can be derived from the results of the implementation. There is an overall improvement in the performance scores between the different approaches. The implementation following the first approach has a fair performance with a high score of 83% using the SVM algorithm, and the second approach worked well using the KNN algorithm for a high score of 80%, and the third approach had a 90% score using the SVM algorithm, which is the highest among all three approaches. The following observations can be made from the results:

Observation 1: Upon comparing of the results of the first and third approach, the SVM and KNN algorithm had improvements in the different performance metric scores. However, the Decision Tree, Random Forest and, Gradient Boosting Trees had diminishing scores. The Bayesian algorithm performed constantly between these approaches.

The improved scores of the SVM and KNN algorithm can be attributed to the dimensionality reduction used in the third approach. Reducing the dimensionality of the data increases the ratio of the size of the dataset to the number of dimensions, which reduces the bias of the classifier towards any particular class. Contrastingly, the performance of the tree-based algorithms improves with a larger feature set. This is because, the depth of the decision tree increases with adding more features, and thereby help making more accurate decisions. The Bayesian principle of the Naïve Bayes algorithm works on the prior probability value calculated for each data in the training

set, which remains unchanged among the two approaches, and hence the scores remain unchanged.

Observation 2: The overall performance of the second approach was lower than the other two approaches.

This is because of the selective feature approach fails to contain most of the information from the speech signal, and can be concluded that using only the MFCC values alone cannot be a good measure to classifythe emotional content of speech.

Observation 3: The classification report of the first approach (see Figure 22) shows that the misclassification is higher for the emotions Happy and Surprise.

This bias is due to the common properties of the features in these two categories. The dimensionality reduction step used in the third approach has greatly minimized this bias (see Figur 17).

1. On comparison to the baseline system by Chen et al. [8], the proposed system has improvements in the accuracy score (see Figure 24). The third experiment is the winning approach for the proposed methodology.

```
Classification Report
              precision    recall  f1-score   support

       Angry       0.80      0.81      0.80        48
     Disgust       0.80      0.94      0.86       112
        Fear       0.80      0.88      0.84        68
       Happy       0.89      0.81      0.85       135
     Neutral       0.88      0.91      0.89       118
         Sad       0.85      0.81      0.83       114
    Surprise       0.77      0.69      0.73       125

avg / total       0.83      0.83      0.83       720

Confusion Matrix
[[ 39   0   6   1   0   1   1]
 [  1 105   0   0   2   2   2]
 [  5   1  60   1   0   0   1]
 [  1   5   4 110   0   0  15]
 [  0   2   0   0 107   7   2]
 [  0   8   0   0   9  92   5]
 [  3  10   5  11   4   6  86]]
```

Fig.17 Classification report for approach-1 results

```
Classification Report
             precision    recall  f1-score   support

      Angry       0.92      0.92      0.92        24
    Disgust       0.88      0.94      0.91        52
       Fear       0.89      0.91      0.90        34
      Happy       0.98      0.94      0.96        66
    Neutral       0.95      0.98      0.96        53
        Sad       0.90      0.88      0.89        49
   Surprise       0.94      0.89      0.91        65

avg / total       0.93      0.92      0.92       343

Confusion Matrix
[[22  0  1  0  0  1  0]
 [ 0 49  1  0  0  1  1]
 [ 2  0 31  0  0  1  0]
 [ 0  1  1 62  0  0  2]
 [ 0  0  0  0 52  0  1]
 [ 0  3  1  0  2 43  0]
 [ 0  3  0  1  1  2 58]]
```

Fig.18 Classification report for approach-3 results

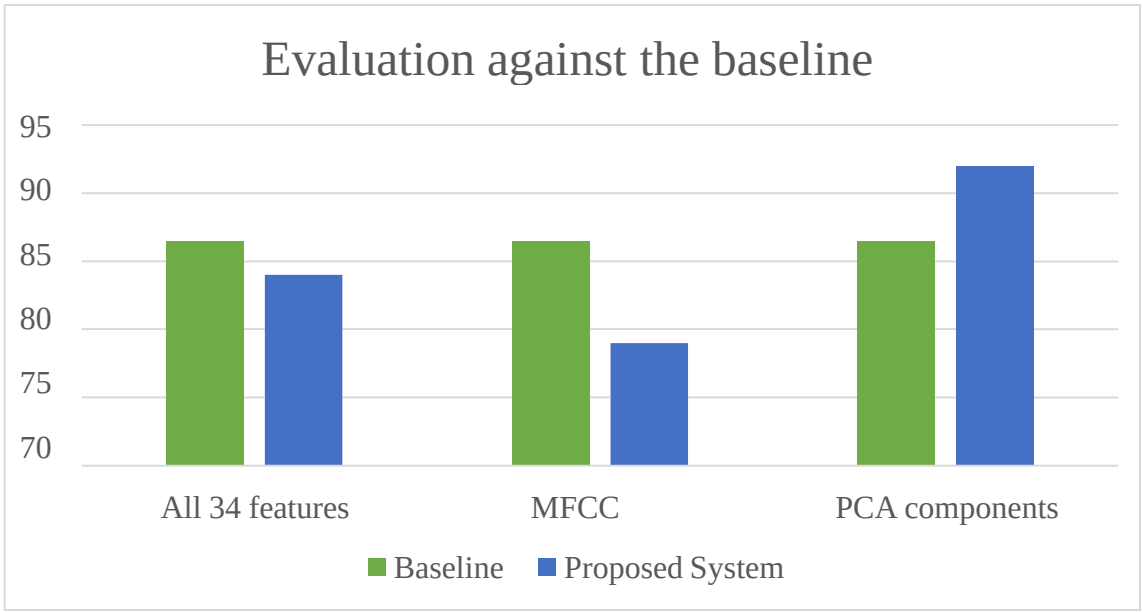

Fig.19 Evaluation against the baseline

CHAPTER 6
CONCLUSION

The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated device's work based on voice commands from the user. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system. In this thesis, the steps of building a speech emotion detection system were discussed in detail and some experiments were carried out to understand the impact of each step. Initially, the limited number of publicly available speech database made it challenging to implement a well-trained model. Next, several novel approaches to feature extraction had been proposed in the earlier works, and selecting the best approach included performing many experiments. Finally, the classifier selection involved learning about the strength and weakness of each classifying algorithm with respect to emotion recognition. At the end of the experimentation, it can be concluded that an integrated feature space will produce a better recognition rate when compared to a single feature.

For future advancements, the proposed project can be further modeled in terms of efficiency, accuracy, and usability. Additional to the emotions, the model can be extended to recognize feelings such as depression and mood changes. Such systems can be used by therapists to monitor the mood swings of the patients. A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem of emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. A sentiment detection using vocabulary, can be integrated with speech emotion detection to identify a possible sarcasm. Therefore, in the future, there would emerge many applications of a speech-based emotion recognition system.

REFRENCES

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–.

[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.

[3] LIVINGSTONE, S. R., AND RUSSO, F. A. : The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one 13, 5 (2018), e0196391.

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814

[7] Platt, J. C., Cristianini, N. and Shawe-Taylor, J. : Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, S. A. Solla,

[8] T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553

[9] Toronto emotional speech set (TESS)

(https://tspace.library.utoronto.ca/handle/1807/24487)

[10] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, “Speech emotion recognition,” in Advances in Electronics, Computers and Communica- tions (ICAECC), 2014 International Conference on. IEEE, 2014, pp. 1–4.

[11] K. R. Scherer, “What are emotions? and how can they be measured?” Social science information, vol. 44, no. 4, pp. 695–729, 2005.

[12] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, “Emotion analysis in man-machine interaction systems,” in International Workshop on Machine Learning for Multimodal Interac- tion. Springer, 2004, pp. 318–328.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias,

[14] W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[15] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech signals,” in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[16] R. W. Picard, Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[17] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[19] .-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech

signals," in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[20] A. D. Dileep and C. C. Sekhar, "Gmm-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE transactions on neural networks and learning systems, vol. 25, no. 8, pp. 1421–1432, 2014.

[21] L. Deng, D. Yu et al., "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197– 387, 2014.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," Neu- ral networks, vol. 61, pp. 85–117, 2015.

[23] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005, pp. 474–477.

[24] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155–177, 2015.

[25] [A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu,

[26] T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.

[27] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057– 1070, 2011.

[28] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in Acoustics, Speech and Signal Processing

(ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5005–5009.

[29] "Reconstruction-error-based learning for continuous emotion recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[30] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[31] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for prac- tical realisation," in Affect and emotion in human-computer interaction. Springer, 2008, pp. 75–91.

[32] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," IEEE Access, vol. 5, pp. 5235–5246, 2017.

[33] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," Journal of advances in Computer Networks, vol. 2, no. 1, pp. 28–30, 2014.

[34] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in Fourth International Conference on Spoken Language Processing, 1996.

[35] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, 2009, pp. 1–4.

ANNEXURE 1:

First Published Paper :

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Link

http://ijariie.com/FormDetails.aspx?MenuScriptId=215760

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Vandana Singh
Department of Computer Science and Engineering,
Integral University, Lucknow.
K. C. Maurya
Assistant Professor Department of Computer Science and Engineering
Integral University, Lucknow.

Abstract

The purpose of this study is to detect the emotions evoked by the speaker while they are speaking. Speech generated in a condition of fear, rage, or delight, for example, becomes loud and quick, with a greater and broader range of pitch, but speech produced in a state of grief or exhaustion is sluggish and low-pitched. The detection of human emotions via voice and speech patterns has a variety of applications, including improving human-machine interactions. We provide a classification model of emotions produced by speeches that uses deep neural networks (CNNs), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) Classification based on auditory data like Mel Frequency Campestral Coefficient (MFCC). . The models have been taught to distinguish between seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). Using the Ryerson Audio-Visual Dataset of Emotions Speech and Song (RAVDESS) dataset and the Toronto Emotional Speech Set (TESS) dataset, we found that the suggested technique achieves accuracies of 86 percent, 84 percent, and 82 percent using CNN, MLP, and SVM, respectively, for 7 emotions.

Index Terms-- Emotion detection, deep learning, machine learning, classification, mel-frequency cepstral coefficients, CNN, RAVDESS, TESS, SVM, MLP.

I. INTRODUCTION

The foundation for information exchange is human communication via spoken language. It is also used in a variety of practical applications in fields such as Business Process Outsourcing (BPO) Centers and Call Centers to detect emotion, which is useful for determining a customer's happiness with a product, improving speech interaction, resolving various language ambiguities, and adapting computer systems to an individual's mood and emotion.

The goal of the presented models is to identify just the emotion in the audio recording that has a higher value. To have a computer classify sentiments, several ways have been attempted, such as feature extraction or text analytics. The purpose of this study is use pure audio data while considering MFCC [4].

Detecting emotions is one of the most important marketing strategy in today's world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

II. LITERATURE REVIEW

Many categorization algorithms have been proposed in this field of research throughout the years. Iqbal et al. [1] created a programme that employed Gradient Boosting, KNN, and SVM to work on granular partitioning in the RAVDESS data to find differences based on gender, with overall accuracy ranging from 40% to 80% depending

on the job. Male recordings alone, female recordings only, and mixed recordings datasets were constructed. SVM and KNN have 100% recognition for all anger and neutrality in the RAVDESS (male) dataset, while Gradient Boosting outperformed SVM and KNN in excitement and melancholy. SVM obtains 100% accuracy with the same fury as half of the guys in the RAVDESS (female) dataset.

With accuracy of 87 percent and 100 percent, KNN performed well in the areas of rage and neutrality. When compared to other categories of tourists, KNN performed worse in happiness and sadness. With rage and neutrality, SVM and KNN performed much better than Gradient Boosting among the combined male and female data rates. KNN's performance was extremely depressing in terms of both happiness and grief. The classifiers' average performance in the male dataset is better than in the female dataset without SVM. SVM is more accurate for aggregated data than gender data sets. [2]. Obtained 66.41 percent accuracy in audio data and 90 percent accuracy in blending audio and video data using another method. The scientists trained three alternative depth networks using already processed picture data, including faces and audio waveforms: one for image data only, one for fixed audio waveforms only, and one for both data and waveform data. One of the first algorithms to use the RAVDESS dataset, however it merely identified it from other emotions available [8]. Three different forms of music sharing algorithms have been proposed: a basic model, a single work area model, and a multi-task capacity model. A single, independently domain classifier was utilized in a basic model. During the training, two hierarchical kinds were employed. For each domain, the single function machine trained various classifications.

III. OBJECTIVE OF THE PROJECT

During the collecting step, noise typically corrupts the input data acquired for emotion recognition [4]. The extraction of features and categorization become less accurate as a result of these flaws [7]. This means that in emotions detection and identification systems, improving the data input is crucial. The emotional discrimination is retained in this pre - processing stage, but the speech and recording variance is removed [28].

The study will cover several deep learning algorithms in the context of SER in the next part. In comparison to traditional procedures, these methods produce more precise findings, but they are more computationally demanding. This section offers researchers and readers literature-based support for evaluating HCI(Human Computer Interaction) feasibility and analyzing the user's emotional voice in a specific scenario. Emotion identification from voice input data is a viable alternative [6], but real-time implementations of these approaches are far more challenging. Although these approaches have limitations, combining two or more of these classifiers creates a new step that may enhance emotion recognition.

IV. PROBLEM DEFINITION

On the RAVDESS and TESS datasets, the assessment findings reveal that the model is effective when compared to baseline methods and the state of the art. Table I. displays the accuracy, recall, and F1 values achieved for each of the emotional classifications. These findings demonstrate that accuracy and recall are extremely well matched, allowing us to acquire F1 values for practically all classes that are spread around the value 0.85. The model's robustness is demonstrated by the limited range of F1 values, which efficiently categories emotions into eight separate categories. The model is less accurate in the classes "Calm" and "Disgust," but this is not surprising given that they are the most difficult to recognize not just by speaking but also by monitoring facial expressions or analyzing written material [15], as stated in the Introduction. We chose to examine the findings acquired from two additional methods, namely SVM and MLP classifier, in order to assess the effectiveness of the emotion classification described in this paper.

V. PROPOSED METHODOLOGY

The emotion recognition classification models given here are based on a deep learning method based on CNN, SVM, and MLP classifiers. The fundamental concept is that the MFCC [4], often known as the "spectrum of a spectrum," is the only feature used to train the model. The Mel-frequency cepstrum (MFC) is a distinct understanding of the Mel-frequency cepstrum (MFCC), and it has been shown to be the state of the art in sound formalization in voice recognition [5]. Because of its capacity to express the amplitude spectrum of a sound wave in a condensed vectorial form, the MFC coefficients have been widely employed.

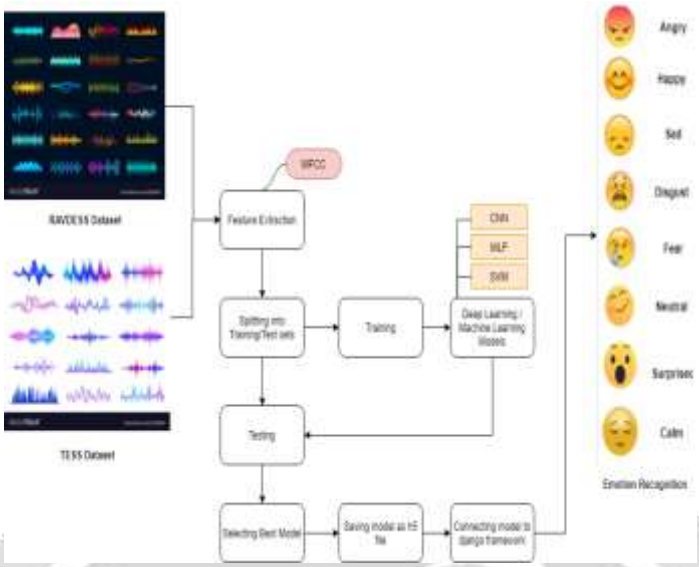

Fig-1: Proposed System Architecture.

Fig-2: Proposed CNN Layer Discription.

Method:

Figure 2 shows the operational results of the deep neural network (CNN) constructed for the classification job. For each audio file supplied as input, the network is capable of working on 40 feature vectors. The 40 values reflect the two-second audio frame's condensed numerical representation. As a result, we give a set of training data (40 x 1) on which we ran one cycle of a 1D CNN with a ReLU activation function [6, a 20% dropout, and a 2 x 2 max-pooling function. The rectified linear unit (ReLU) is defined as g(z) = max0, z, and it allows us to acquire a big value in the event of activation by using this function to represent hidden units. In this situation, pooling can assist the model in focusing solely on the most important aspects of each segment of input, rendering them position invariant. By adjusting the kernel size, we repeated the method outlined above. We then applied another washout and flattened the result to ensure compatibility with the next layers.

VI. REFERENCE

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–.

[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.

[3] LIVINGSTONE, S. R., AND RUSSO, F. A. : The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one 13, 5 (2018), e0196391.

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814

[7] Platt, J. C., Cristianini, N. and Shawe-Taylor, J. : Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, S. A. Solla,

[8] T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553

[9] Toronto emotional speech set (TESS) (https://tspace.library.utoronto.ca/handle/1807/24487)

[10] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in Advances in Electronics, Computers and Communica- tions (ICAECC), 2014 International Conference on. IEEE, 2014, pp. 1–4.

[11] K. R. Scherer, "What are emotions? and how can they be measured?" Social science information, vol. 44, no. 4, pp. 695–729, 2005.

[12] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in International Workshop on Machine Learning for Multimodal Interac- tion. Springer, 2004, pp. 318–328.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias,

[14] W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[15] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[16] R. W. Picard, Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[17] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

[18] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[19] .-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[20] A. D. Dileep and C. C. Sekhar, "Gmm-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE transactions on neural networks and learning systems, vol. 25, no. 8, pp. 1421–1432, 2014.

[21] L. Deng, D. Yu et al., "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197– 387, 2014.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," Neu- ral networks, vol. 61, pp. 85–117, 2015.

[23] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005, pp. 474–477.

[24] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155–177, 2015.

[25] [A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu,

[26] T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.

[27] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057– 1070, 2011.

[28] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion

recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5005–5009.

[29] "Reconstruction-error-based learning for continuous emotion recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[30] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[31] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for prac- tical realisation," in Affect and emotion in human-computer interaction. Springer, 2008, pp. 75–91.

[32] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," IEEE Access, vol. 5, pp. 5235–5246, 2017.

[33] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," Journal of advances in Computer Networks, vol. 2, no. 1, pp. 28–30, 2014.

[34] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in Fourth International Conference on Spoken Language Processing, 1996.

[35] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, 2009, pp. 1–4.

ANNEXURE

ANNEXURE 2: PUBLISHED PAPER

ADCIS 2022 (author)

ADCIS 2022 Submission 62

Update information
Update authors
Update file

The submission has been saved!

Submission 62	
Title	SPEECH EMOTION RECOGNITION USING DEEP LEARNING
Paper:	(Jun 04, 12:04 GMT)
Author keywords	Emotion detection deep learning machine learning classification mel-frequency cepstral coefficients CNN RAVDESS TESS SVM MLP
Abstract	The purpose of this study is to detect the emotions evoked by the speaker while they are speaking. Speech generated in a condition of fear, rage, or delight, for example, becomes loud and quick, with a greater and broader range of pitch, but speech produced in a state of grief or exhaustion is sluggish and low-pitched. The detection of human emotions via voice and speech patterns has a variety of applications, including improving human-machine interactions. We provide a classification model of emotions produced by speeches that uses deep neural networks (CNNs), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) Classification based on auditory data like Mel Frequency Campestral Coefficient (MFCC). . The models have been taught to distinguish between seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). Using the Ryerson Audio-Visual Dataset of Emotions Speech and Song (RAVDESS) dataset and the Toronto Emotional Speech Set (TESS) dataset, we found that the suggested technique achieves accuracies of 86 percent, 84 percent, and 82 percent using CNN, MLP, and SVM, respectively, for 7 emotions.
Submitted	Jun 04, 12:04 GMT
Last update	Jun 04, 12:04 GMT
Status of using third-party	I am not using third-party material for which formal permission is

material in your article	required

Authors

first name	last name	email	country	affiliation	Web page	corresponding?
Vandana	Singh	vandanasingh586@gmail.com	India	Punjab National Bank		✓
K C	Maurya	kcmaurya@iul.ac.in	India	Integral University, Lucknow, Uttar Pradesh		✓

ANNEXURE

ANNEXURE 3: PLAGIARISM REPORT

RE-2022-22293 (2)-plag-report

ORIGINALITY REPORT

12%	17%	8%	28%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to CSU, San Jose State University Student Paper	4%
2	scholarworks.sjsu.edu Internet Source	2%
3	www.diva-portal.org Internet Source	2%
4	www.coursehero.com Internet Source	1%
5	Submitted to Bennett University Student Paper	1%
6	kth.diva-portal.org Internet Source	1%
7	Submitted to Manipal University Student Paper	<1%
8	github.com Internet Source	<1%
9	Marco Giuseppe de Pinto, Marco Polignano, Pasquale Lops, Giovanni Semeraro. "Emotions Understanding Model from Spoken Language	<1%

using Deep Neural Networks and Mel-Frequency Cepstral Coefficients", 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2020

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		