Speech Emotion Detection Using Supervised, Unsupervised And Feature Selection Algorithms

By

Abu Nuraiya Mahfuza Yesmin Rifat
15101048
Aditi Biswas
16301135
Nadia Farhin Chowdhury
15301087

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
April, 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name Signature:

Abu Nuraiya Mahfuza Yesmin
Rifat
15101048

Aditi Biswas
16301135

Nadia Farhin Chowdhury
15301087

Approval

The thesis titled "Speech Emotion Detection Using Supervised, Unsupervised And Feature Selection Algorithms" submitted by

1. Abu Nuraiya Mahfuza Yesmin Rifat (15101048)
2. Aditi Biswas (16301135)
3. Nadia Farhin Chowdhury (15301087)

of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on 25th April, 2019.

Examining Committee:

Supervisor:
(Member)

Dr Jia Uddin
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Amitabha Chakrabarty
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Md. Abdul Mottalib
Professor
Department of Computer Science and Engineering
Brac University

Abstract

A tremendous research is being done on Speech Emotion Recognition (SER) in the recent years with its main motto to improve human machine interaction. In this thesis work,we have introduced a scheme for emotion recognition from speech. We have classified three emotions (happy, angry and sad) for both male and female. Recognition task has been done using Mel-frequency Cepstrum Coefficient (MFCC) based features.Four classifiers are used for the purpose of classification. They are Random Forest, Gradient Boosting, SVMand CNN. Among them, CNN has shown the best accuracy of 71.17%. Random Forest has shown an accuracy of 61.26%, Gradient Boosting 60.36% and SVM60 36%. After using RFE method, PCA and P-Valuefor less significant feature reduction the accuracy improved to 62.16% for Random Forest, 62.16% for Gradient Boostingand 61.26% for SVM.

Keywords: SER; MFCC; Random Forest; Gradient Boosting; SVM; CNN; RFE; P-Value; PCA.

Dedicated to

Our beloved Parents and honourable Supervisor Dr. Jia Uddin for their endless support and patience

Acknowledgement

Firstly, we are grateful to Almighty for the good health and well being that were necessary to complete this thesis and for always guiding us to the right path. There are also many people we would like to acknowledge for their support during this journey.

We would like to thank our thesis Supervisor Dr. Jia Uddin for always having his door open for us whenever we ran into a trouble. He consistently allowed this paper to be our own work, but steered us in the right direction whenever he thought we needed it.

Finally, we must express our very profound gratitude to our parents and siblings for providing us with unfailing support and continuous encouragement throughout our years of study and while writing this thesis paper.

This accomplishment would not have been possible without any of them.

Thank you.

Table of Contents

List of Tables

List of Figures

List of Acronyms

CNN	Convolutional Neural Network
MFCC	Mel-Frequency Cepstral Coefficient
RFE	Recursive Feature Elimination
PCA	Principal Component Analysis
SER	Speech Emotion Recognition
SVM	Support Vector Machine

Chapter 1

INTRODUCTION

1.1 Introduction

In our daily lives, emotions accompany us, playing a key role in non-verbal communication. Emotions can also influence the part of the brain which controls thought, reason and judgment. It is very difficult to predict human emotions quantitatively. Though facial expressions and gestures are the best ways to figure out one's emotions, it becomes difficult to identify them as the age of a person increases, because people learn to control their expressions with age and experience. To overcome this, different methods are discovered for emotion recognition. Speech Emotion Recognition(SER) is one of them. Speech signal can be used to articulate various kinds of emotions. SER system identifies emotions on paralinguistic basis. It also plays an important role in finding out the psychological state of a person.The urge to improvise the efficiency and naturalness of Human Machine Interaction (HMI) derives major motivation for the work.

MFCC coefficients derived from human speech samples play a vital role in the field of speech signal processing. They are used in applications including speaker verification, speaker recognition, emotion detection etc.The mel-frequency scale is a quasi-logarithmic spacing roughly resembling the resolution of the human auditory system.This makes the MFCC features "biologically inspired." So MFCC co-efficient or MFCC based features can effectively detect emotion of human from speech.

The task of emotion classification involves two stages. The first stage is feature extraction followed by classification. Here MFCC coefficients are considered. From coefficients nine features are developed for the purpose of supervised machine learning classifiers. Features are mean, median, standard deviation, amplitude, pitch, variance, mode, kurtosis, skew-ness. The effect of these features and their possible combinations on SER is analyzed. As supervised classifier, we used SVM, Random Forest classifier, Gradient Boosting classifier. CNN has been used as Neural Network classification. For CNN, MFCC coefficients are directly used as input as it is not a feature based classifier.

1.2 Motivation

Main thought of this thesis work is to detect emotions based on MFCC coefficients. Moreover, we have suggested improvements of the detection work. For improvement work, feature reduction (PCA) and feature selection methods (RFE, P-Value calculation) have been used. The hope of this thesis work was to find a new way of emotions and using algorithms for feature selection or reduction, enhancing the model performance.

1.3 Objectives

Main objectives of this thesis are summarized as follows:

- Design a system based on MFCC coefficients that can detect basic emotions of human
- Choose suitable features to design the system based on MFCC coefficients
- For the purpose of classifying with Neural Network, choose CNN
- Use ofextracted features to detect the emotions
- Apply feature reduction or selection algorithms on results to introduce further improvement on classifier performance
- Compare results among different classifier

1.4 Literature review

Earlier researchers have incorporated MFCC coefficients in the feature vector for identifying the paralinguistic content but could recognize few numbers of emotions. Only three emotions [1] were detected by B. Jaramillo. A. FirozShah detected four emotions with poor recognition accuracy [2]. The feature vector consisted of first 19 MFCC coefficients [2] and a total of 63 MFCC features [3]. The available raw signal is too small to use it both for training and testing. Hence, the speech signals are synthetically enlarged so that the signal will be sufficient both for training and testing. K.V.Krishna compared the MFCC methods with the use of Sub band based Cepstral parameters increases the classification efficiency by 19% [3].Using Synthetic enlargement of MFCC's, the emotion misclassification efficiency reduces [4]. This work is done by I. M. Herranz1. He introduced enlargement on MFCC. A. V. Vidyapeetham did emotion classification using MFCC and Cepstrum Features [5]. They classified seven emotions.Chen et al. [29] aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method classify different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is input parameters for multi- level SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments include Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence indicates in dimension reduction Fisher is better than PCA and for classification, SVM is more expansible than ANN for emotion recognition in speaker independent is. The recognition rates for three levels are 86.5%, 68.5% and 50.2% separately in Beihang university database of emotional speech (BHUDES) [29]. Nwe et al. [30] proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficients (LFPC) and discrete HMM to characterize the speech signals and classifier respectively. This method classified the emotion into six different categories then used the private dataset to train and test the new system. In order to evaluate the performance of the proposed method, LFPC is compared with the mel-frequency Cepstral coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Result demonstrates the average and best classification accuracy achieved 78% and 96% respectively. Furthermore, results expose that LFPC is a better option as feature for emotion classification than the standard features [30]. Wu et al. [31] proposed new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for convey important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) are adopted. Berlin and Vera am Mittag (VAM) are employed to evaluate MSFs. In experimental result, the MSFs display capable performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC). When MSFs utilized augment prosodic features, there is a considerable improvement in performance of recognition. Furthermore overall recognition rate of 91.6% is achieved for classification [9]. Rong et al. [32] presented an ensemble random forest to trees (ERFTress) method with a high number of features for emotion recognition without referring any language or linguistic information remains an unclosed problem. This method is applied on small size of data with high number of features. In order to evaluate the proposed method an experiment results on a Chinese emotional speech dataset designates, this method achieved improvement on emotion recognition rate. Furthermore, ERFTrees performs better than popular dimension reduction methods such as PCA and multi-dimensional scaling (MDS) and recently developed ISOMap. The best accuracy with 16 features for female dataset achieved the maximum correct rate of 82.54%, while the worst is only 16% on 84 features with natural data set.

1.5 Proposed System

Most of the early researchers detected a few number of emotions on their research work. In our case we have detected three emotions forboth male and female totalling to six emotions. Previously, they worked with only neural networks or only supervised algorithms but here we proposed a model combining both and showed comparisons. Moreover, performance improvementshave been introduced here. The system will work in the following process,

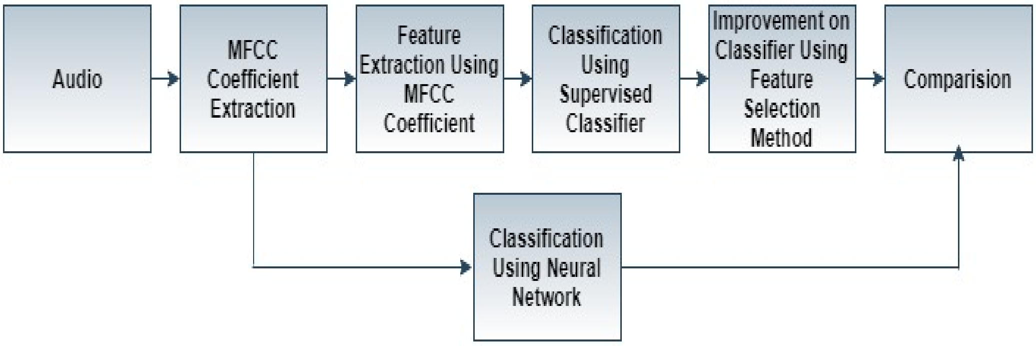
Audio
MFCC Coefficient Extraction
Feature Extraction Using MFCC Coefficient
Classification Using Supervised Classifier
Improvement on Classifier Using Feature Selection Method
Comparision
Classification Using Neural Network

Figure 1.1: Block Diagram of Proposed System

Chapter 2

DATASET DESCRIPTION

2.1 Introduction

Dataset was collected from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6]. It is a validated multimodal database of emotional speech and song. Description on this dataset is given below.

2.2 Database Properties

The database is gender balanced, consisting of 24 professional actors, in a neutral North American accent vocalizing lexically matched statements. Speech includes expressions of calm, happiness, sadness, anger, fear, surprise and disgust, and song contains emotions of calm, happiness, sadness, anger, and fear. Eight emotions were selected for speech: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Calm and neutral were selected as baseline conditions, while the remaining states constitute the set of six basic or fundamental emotions that are thought to be culturally universal. Each expression is produced with a further neutral expression at two levels of emotional intensity. Face-and-voice, face-only, and voice-only formats are available for all conditions. Twenty-four professional actors who are working in Toronto, Ontario, Canada were hired for stimulus creation (M = 26.0 years; SD = 3.75; age range = 21–33; 12 males and 12 females). Actors self-identified as Caucasian (N = 20), East-Asian (N = 2), and Mixed (N = 2, East-Asian Caucasian, and Black-Canadian First nations Caucasian). To be eligible, actors needed to have English as their first language, speak with a neutral North American accent, and to not possess any distinctive features (e.g., beards, facial tattoos, hair colourings, facial piercings). Participants were also required to identify text presented at 1.5 m distance without wearing glasses.
The set of 7356 recordings was rated on emotional validity, intensity, and authenticity 10 times each. 247 individuals who were characteristic of North America's untrained research participants provided ratings. Test-retest data was provided by another set of 72 participants. High emotional validity levels and reliability of test-retest interpreters have been reported. Corrected accuracy and composite ”goodness” measures are presented to support the selection of stimuli by researchers. All recordings are made freely available under a Creative Commons license and can be downloaded at https://doi.org/10.5281/zenodo.1188976.

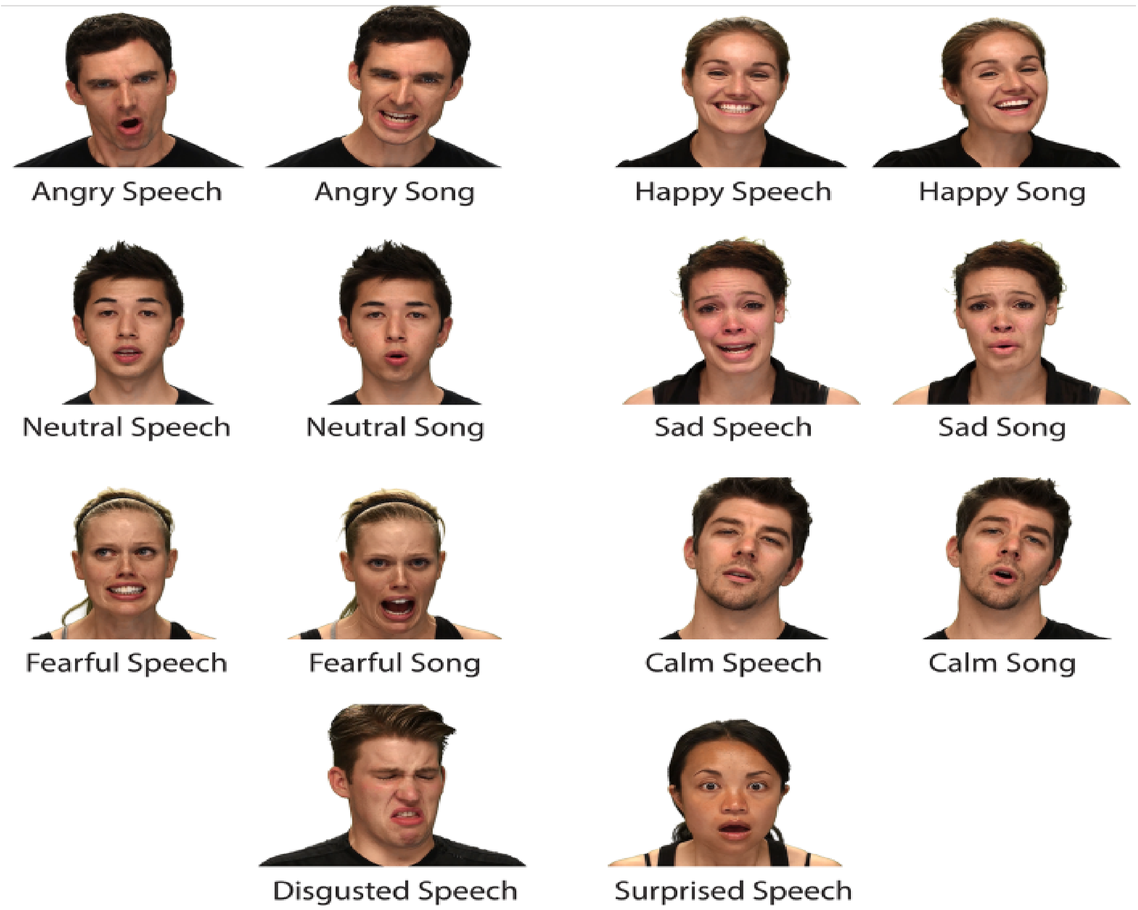

Figure 2.1: Example of RAVDESS emotion acquisition

2.3 Data Acquisition Process

Procedural steps of data acquisition are given below:

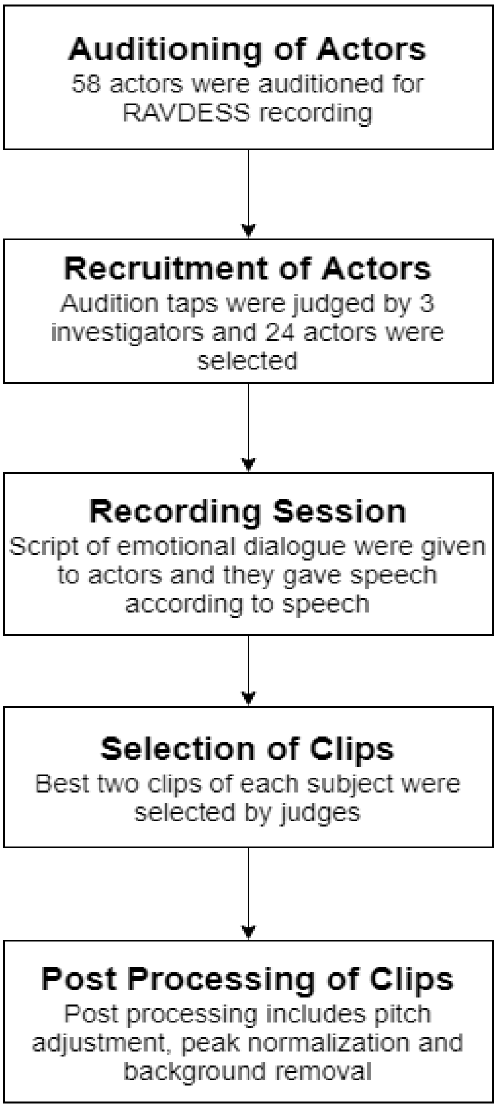
Auditioning of Actors
58 actors were auditioned for RAVDESS recording
Recruitment of Actors
Audition taps were judged by 3 investigators and 24 actors were selected
Recording Session
Script of emotional dialogue were given to actors and they gave speech according to speech
Selection of Clips
Best two clips of each subject were selected by judges
Post Processing of Clips
Post processing includes pitch adjustment, peak normalization and background removal

Figure 2.2: Data acquisition process

2.4 Dataset used in the Study

Total database of RAVDESS was not used in this study. Only 552 trails with three emotions for both male and female are considered for this study. Description of emotions is given below:

Types	Emotions	Trail Number
Male	Happy	96
	Sad	96
	Angry	96
Female	Happy	88
	Sad	88
	Angry	88
Total Trails		552

Table 2.1: Emotions with trail number

2.5 Summary

Considered database properties, acquisition and contribution of considered database in our work are described properly. Considered database is free, so anybody can access it. Considered portion of database will be used for dataset here and further MFCC will be measured using that dataset.

Chapter 3

MFCC AND FEATURE EXTRACTION

3.1 MFCC

3.1.1 Introduction

The Mel-frequency cepstrum (MFC) in sound processing is a representation of a sound's short-term power spectrum based on a linear cosine transformation of a log power spectrum on a nonlinear frequency Mel scale.
Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC [7]. They are derived from an audio clip (a nonlinear "spectrum-of - a-spectrum") type of cepstral representation. The difference between the cepstrum and the cepstrum of the Mel-frequency is that the frequency bands are spaced equally on the Mel scale in the MFC, which approximates the response of the human auditory system more closely than the linear spaced frequency bands used in the normal cepstrum. That's right.

3.1.2 Derivation of MFCC

MFCCs are generally derived as follows [8]:

- Take the Fourier transformation of a signal (the windowed excerpt of).
- Use triangular overlapping windows to map the spectrum power obtained above to the Mel scale.
- Take the power logs for each Mel frequency.
- Take the discrete cosine transformation of the Mel log power list, as though it were a signal.
- The MFCCs are the resulting spectrum amplitudes.

3.1.3 Advantages and Disadvantages of MFCC

Early researches proved that MFCC is much effective in case of audio emotion detection. It has a number of advantages in case of detection. But it has also some disadvantages. Advantages and disadvantages of MFCC are given below,

Advantages

MFCC coefficients derived from human speech samples play a vital role in the field of speech signal processing. The mel-frequency scale is a quasi-logarithmic spacing roughly resembling the resolution of the human auditory system. This makes the

MFCC features “biologically inspired.” So MFCC co-efficient or MFCC based features can effectively detect emotion of human from speech.

Disadvantages

In the presence of additive noise, MFCC values are not very robust, so it is common to standardize their values in speech recognition systems to reduce noise influence. Some researchers are proposing modifications to the basic MFCC algorithm to improve robustness, for example by raising the log-mel amplitudes to an appropriate power (around 2 or 3) before taking the DCT (Discrete Cosine Transform), which reduces the influence of low-energy components.

3.1.4 Application of MFCC

Application of MFCC are given below,

- MFCCs are commonly used as voice recognition features [9].
- MFCCs are also increasingly finding use in music-recovery applications such as genre classification, audio-like measurements, etc. [10]
- Speaker verification
- Speaker recognition
- Emotion detection etc.

3.1.5 Early researches using MFCC

MFCC has been used in many researches in early days. It is used as an important feature extraction methodology now-a-days. L. Muda [11] used MFCC to develop a voice recognition algorithm. Voice recognition is important in digital signal processing. F Zheng [12] described the effect of MFCC in case of filter designing when working with signal processing. MFCC effectively affects (1) the number of filters, (2) the shape of filters, (3) the way in which filters are spaced, and (4) the way in which the power spectrum is warped. MFCC features have also been used in speaker recognition [13]. Research on detection of emotionally abused woman [14] has also been done with the help of MFCC. So MFCC has a huge effect on research line.

3.1.6 Summary

In this section, we briefly discussed about MFCC, the definition of MFCC, its derivation, advantages and disadvantages, application of MFCC and influence of MFCC in research work. So now it is clear, why MFCC has been used in speech emotion recognition.

3.2 FEATURE EXTRACTION

3.2.1 Introduction

Using MFCC co-efficient total nine features are extracted for this study. These features have been used in many emotion detection research works. Considered features are given below,

- Mean
- Median
- Standard Deviation
- Amplitude
- Pitch
- Variance
- Mode
- Kurtosis
- Skewness

All of these features are statistical features. Description of them are given below,

3.2.2 Features

Mean

In different mathematics branches there are several types of means (especially statistics). They may be mean arithmetic, mean geometric, mean harmonic, etc. The arithmetic mean, also called the mathematical expectation or average, is the central value of a discrete set of numbers for a data set: specifically, the sum of the values divided by the number of values. [15]. If we consider a sample, x1,x2,x3. . . .,xn, usually arithmetic meanx will be the sum of the sampled values divided by the number of items. So,

$$\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} xi) = \frac{(x1+x2+x3+...+xn)}{n} \text{ .... (1)}$$

Median

Median is the middle number in a number list that has been sorted. The numbers must first be arranged in value order from the lowest to the highest in order to determine the median value in a sequence of numbers. If there is an odd number amount, the median value is the number in the middle, with the same number below and above. If there is an even amount of numbers in the list, the middle pair must be determined, added together and divided by two to find the median value. The median can be used to determine an approximate average, or mean. The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values. Outliers can affect the median of a sequence less than the mean.

For example, if given list x1,x2,x3,x4,x5 where, x1$<$ x2$<$ x3$<$ x4$<$ x5, then x3 will be the median of this sequence.

Standard Deviation

The standard deviation (SD, also represented by the lower case Greek letter sigma or the Latin letter s) in statistics is a measure used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the set mean (also known as the expected value), whereas a high standard deviation indicates that the data points are spread across a wider range of values [16].

The formula for the sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{N}(xi-\bar{x})^2}{(N-1)}}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Where, x1,x2,x3,....., xN are the observed values of the sample item, x is the mean value of these observations, N is the number of observations in the sample.

Amplitude

The amplitude of a fluctuation in relation to a time series, is the ordinate's value at its peak or trough from some mean value or trend line. The difference between peak and trough values is sometimes referred to as "amplitude."

In a time series, let consider, some mean values are x1,x2,x3,x4,x5.If among them x2 has the peak value, then x2 will be the amplitude of the sequence.

Pitch

Pitch is a perceptual property of sounds that enables them to be ordered on a frequency-related scale, or more commonly, pitch is the quality that enables sounds to be judged as "higher" and "lower" in the sense associated with musical melodies. Only sounds with a frequency that is clear and stable enough to distinguish between noise can determine pitch. Pitch is a major musical tone auditory attribute, along with duration, loudness, and timbre. Pitch can be quantified as a frequency, but pitch is not a purely objective physical property; it is a subjective psychological acoustic sound attribute. Historically, pitch and pitch perception study has been a central problem in psycho-acoustics and has been instrumental in the development and testing of sound representation, processing and perception theories in the auditory system [17].

Variance

Variance is the expectation of the square deviation of a random variable from its mean in probability theory and statistics. Informally, it measures how far from their average value a set of (random) numbers is spread. Variance plays a central role in statistics, including descriptive statistics, statistical inference, hypothesis testing, fitness goodness, and Monte Carlo sampling. In sciences where statistical analysis of data is common, variance is an important tool. The variance is the square of the standard deviation, the second central moment of a distribution, and the co-variance of the random variable with itself, and it is often represented by σ^2, s^2 or var(X) [18]. The variance of a random variable X is the expected value of the squared deviation from the mean of X,μ=E(x):

$$var(x) = E[(X{-}\mu)^2] \text{ ....(3)}$$

Mode

A set of data values mode is the value that appears most often. If X is a discrete random variable, the mode is the valuex (i.e. X=x) where the maximum value is taken from the probability mass function. In other words, the most likely value to be sampled is the value. Like the statistical mean and average, the mode is a way to express important information about a random variable or population in a (usually) single number. The mode's numerical value is the same as the mean and median in a normal distribution, and in highly skewed distributions it can be very different. The mode is not necessarily unique to a given discrete distribution because at several points x1, x2, etc., the probability mass function can take the same maximum value. The most extreme case occurs in uniform distributions, where all values occur equally frequently [19].
If we consider a sequence like, x1, x2, x3, x1, x3, x1, x1; then x1 will be mode of this sequence as it is occurring most frequently.

Kurtosis

Kurtosis is a measure of the "tailedness" of the probability distribution of a real-evaluated random variable in probability theory and statistics. Similar to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and there are different ways to quantify it for a theoretical distribution as well as corresponding ways to estimate it from a population sample. There are different interpretations of kurtosis, depending on the specific measure of kurtosis used, and how specific measures should be interpreted [20].
The kurtosis is the fourth standardized moment, defined as

$$kurt(x) = E[(X - \mu)\,/\sigma]^4 = \mu^4\,/\sigma^4 \text{ ....} \quad (4)$$

Where, 4 is the fourth central moment and is the standard deviation

Skewness

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.
In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal.
● In the normal peRFEct distribution (green solid curve in the illustration below), the exact mirror images of each side of the curve.
● The tail on the left side of the curve is longer than the tail on the right side when a distribution is skewed to the left (red dashed curve), and the mean is less than the mode. This situation is also referred to as negative skews.
● The tail on the right side of the curve is longer than the tail on the left side when a distribution is skewed to the right (blue dotted curve), and the mean is greater than the mode. Positive skewness is also called this situation.

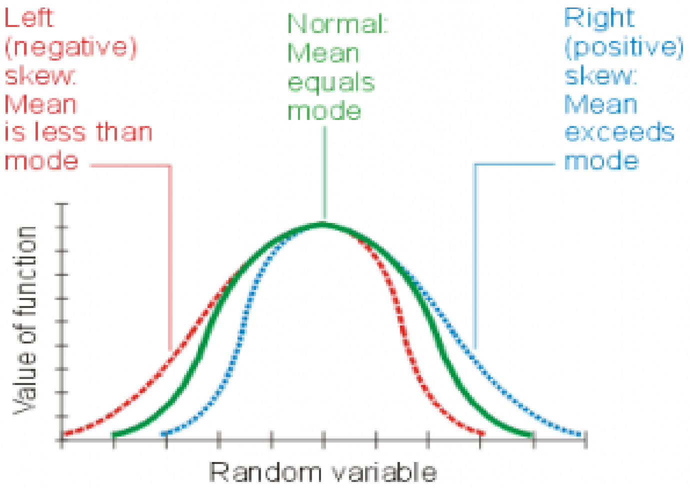

Figure 3.1: Skewness

3.2.3 Summary

This section contains working flow chart for feature extraction and classification. Different features are extracted from MFCC coefficients. These features can be used for analysingemotions of human. Classification methodology based on these features will be explained in next chapter.

4 CLASSIFIERS AND FEATURE SELECTION

4.1 CLASSIFIERS

4.1.1 Introduction

Classification is the process whereby the class of data points is predicted. Sometimes classes are referred to as targets / labels or categories. Predictive modeling classification is the task of approximating a mapping function (f) from variables of input (X) to discrete variables of output (y). In classification as lazy learners and eager learners, there are two types of learners.

i. Lazy learners

Simply store the training data for lazy learners and wait until test data appears. Classification is performed when it does, based on the most related data in the training data stored. Lazy learners have less time to train but more time to predict compared to eager learners, e.g.: k-nearest neighbour, Case-based reasoning.

ii. Eager learners

Before receiving classification data, eager learners build a classification model based on the given training data. It must be capable of committing to a single hypothesis covering the entire space of the instance. Eager learners take a long time to train and less time to predict because of the model construction, i.e..: Decision Tree, Naive Bayes, Artificial Neural Networks.

In this study four classifiers are considered for classification. They are Support Vector Machine (SVM), Random Forest Classifier, Gradient Boosting classifier and Convolutional Neural Network. Among them first three are supervised and last one is neural network.

4.1.2 Classification Algorithms

A lot of classification algorithms are available now, but it can not be concluded which one is superior to the other. It depends on the application and nature of the data set available. Each classifier used here are described below.

Support Vector Machine

Support Vector Machine (SVM) have good performances in many applications like bioinformatics because of their accuracy and their ability to deal with a large number of predictors solving highly nonlinear problems with datasets of small number.
An SVM performs classification by building a N-dimensional hyper-plane that optimally divides data into two categories [26]. The standard SVM takes a set of input data and predicts a model that assigns new examples to one or the other category for each input, each marked as belonging to one of the two categories. An SVM model is a portrayal of examples as space points.
The support vectors are the closest data points to the separating hyper-plane; these points are on the slab's boundary. Margin means the maximum slab width parallel to the hyperplane which does not have any internal data point. Some binary classification problems as a useful separating criterion do not have a simple hyper-plane. There is a variant of the mathematical approach for these problems, which retains almost all the simplicity of a separating hyper-plane SVM. Nonlinear SVM then operates to find a separate hyper-plane in the transformed predictor space. This approach uses the results of the kernel reproduction theory.
Generally, There are two categories of SVM:-
i. Linear SVM and
ii. Non-linear SVM

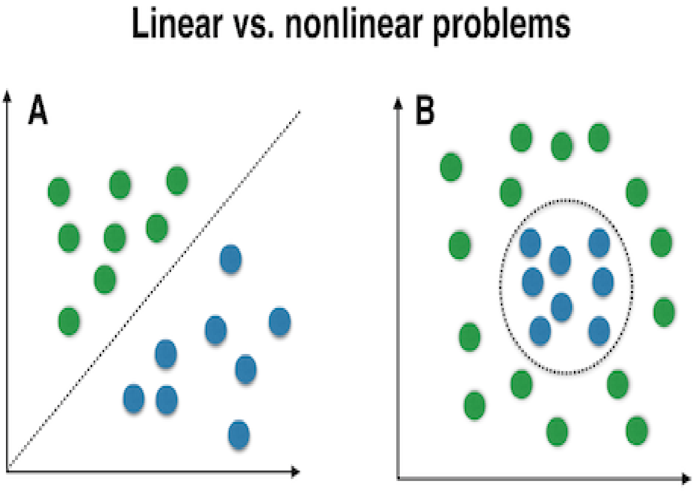

Figure 4.1: Linear and Non-linear SVM

i. Linear SVM

We are given a training dataset of n points of the form
$(x1$, y1) ....$(xn$, yn) ,
Where, the yi are either 1 or 1, each of which indicates the class to which point xi , belongs. Each xi is a real vector of p-dimension. We want to find the "hyper-plane of the maximum margin" that divides the group points xi for which yi= 1 from the points group for which yi = 1, the distance between the hyperplane and the nearest point xi is defined from either group is maximized.
Any hyper-plane can be described as the set of points xi which satisfies,
w .x b = 0, ...(10)
Where, w is the (not essentially normalized) normal vector to the hyperplan. This is similar to the normal form of Hesse, except that w is not necessarily a unit vector. The parameter $b\backslash$ || x || determines the hyperplane offset from the source along the normal vector w . Linear SVM can be used both on soft and hard margin.

ii. Non-Linear SVM

In non-linear SVM, train data are distributed non-linearly. This type of SVM works in soft margin. Moreover, kernel trick is needed in this type of classification. Basic kernels are given below,
Polynomial (Homogenous): $k($xi,xj$) = ($xi,xj$)^d$
$Polynomial(Non - homogenous)$:k (xi,xj) = (xi,xj + 1) ^{d}
$GaussianRadialBasisfunction$:k (xi,xj) = exp($-\gamma$ || xi - xj ||)^2

Soft Margin

We introduce the hinge loss function to extend SVM to cases where the data is not linearly separable, max $(0, 1 - yi($w. x- b))
.Note that yi is the i-th target (i.e., in this case, 1 or -1), and w. x- bis the current output.

Hard Margin

If the training data is linearly separable, two parallel hyper-plans can be selected that separate the two data classes so that the distance between them is as large as possible. The region bounded by these two hyper-plans is called the "margin," and the hyper-plane of the maximum margin is the hyper-plane between them. With a standardized or standardized dataset, the equations can describe these hyper-planes,

w. x- b = 1, (anything on or above this boundary is of one class, with label 1)

w. x- b = -1, (anything on or below this boundary is of other class, with label -1) The distance between these two hyperplanes is geometrically, $2/\mid x \mid$, So, to maximize the distance we want to minimize between the planes x. The distance is calculated from a point to a plane equation using the distance. We must also prevent data points from falling into the margin, adding the following restriction: for each i either

w . x - b$\geq$1, if yi=1 ....(11)

w . x - b$\leq$1, if yi=-1 .... (12)

Random Forest Algorithm

Random Forest or Random Decision Forests are an ensemble learning method for classification, regression and other tasks that works by building a multitude of decision trees at the time of training and outputting the class which is the class mode (classification) or mean prediction (regression) of the individual tree. It is a flexible, easy-to-use machine learning algorithm that generates a great result most of the time, even without hyper-parameter tuning. It is also one of the most frequently used algorithms due to its simplicity and the fact that it can be used for classification and regression tasks. Further information on how they are calculated is useful for understanding and using the different options. Most of the options are based on two Random Forest data objects. Random Forest is a supervised learning algorithm[27]. It creates a forest, making it random somehow. When sampling with replacement draws the training set for the current tree, approximately one-third of the cases are left out of the sample. This oob (out - of-bag) data is used as trees are added to the forest to get a running unbiased estimate of the classification error. It is also used to obtain variable importance estimates.

After building the tree, all data will run down tree and for each pair of cases, proximities will be calculated. If two cases have the same terminal node, they will increase their proximity by one. At the end of the run, by dividing by the number of trees, the proximities are normalized. Proximities are used to replace missing data, locate outliers, and generate low-dimensional data views. The forest it builds is an ensemble of Decision trees, mostly trained with the method of "bagging." The general idea of the bagging method is that the overall result is increased by a combination of learning models. You can see below in figure 4.2 how two trees would look like a Random Forest.

Random Forest's big advantage is that it can be used for both classification and regression issues, which make up the majority of current machine learning systems. I'm going to talk about Random Forest as classification is sometimes considered the machine learning building block. The Random Forest Training Algorithm applies the general bootstrap aggregation technique, or bagging, to tree learners. Given a training set X = x1,...,xn with responses Y = y1,...,yn, repeated bagging (B times) selects a random sample to replace the training set and fits trees to the samples:

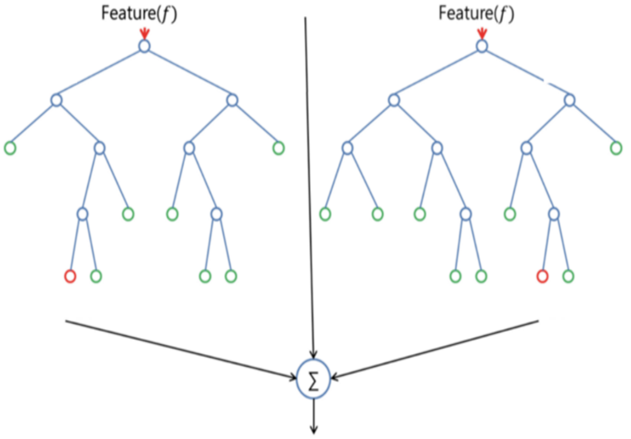

Figure 4.2: Random Forest classifier with two trees

For b = 1,...,B
1.Sample, with substitute, n training examples from X, Y; call these Xb,Yb.
2.Training a classifier or regression tree fb on Xb,Yb.
Predictions of unseen samples after training x' can be made by averaging the predictions from all the individual regression trees on x':

$$\sigma=\sqrt{\frac{\sum_{b=1}^{B}(fb(x^1)-f^1)^2}{(B-1)}}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(13)$$

It could also be done by taking the majority vote in the case of classification trees. This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single workout set would give strongly correlated trees (or even the same tree many times if the workout algorithm is deterministic); bootstrap sampling is a way to de-correlate trees by showing them different workouts. In addition, an estimate of the prediction uncertainty can be made as the standard deviation of the predictions from all the regression trees. on x':

$$f' = \frac{1}{B}\sum_{b=1}^{B} fb(x^1)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(14)$$

A free parameter is the number of samples / trees, B. Depending on the size and nature of the training set, a few hundred to several thousand trees are typically used. An optimal number of trees B can be detected by cross-validation or by observing the error outside the bag: the mean prediction error on each training sample x, using only the trees that did not have x in their bootstrap sample. Training and testing error tend to level off after fitting a number of trees.

Gradient Boosting Algorithm

Gradient Boosting is a machine learning technique for regression and classification issues that generates a predictive model in the form of a set of weak predictive models, typically decision trees [33]. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Gradient Boosting is typically used with decision trees. Like other methods of boosting, Gradient Boosting in an iterative fashion combines weak "learners" into a single strong learner. In the least square regression setting, it is easiest to explain where the goal is to "teach" a model F to predict shape values y'= F(x)by minimizing the mean squared error1 $/n\sum_i (y'i - yi)^2$ where i indexes over some training set of size nof actual values of the output variable p. It can be assumed that there is some imperfect model at each stage of m, 1¡m ¡ M, of Gradient Boosting Fm. The algorithm for gradient boosting improves Fm by building a new model adding an estimator h to provide a better model: Fm+1(x)=Fm(x)+h(x). To find h, the Gradient Boosting solution starts with the observation that a perfect h would imply,
Fm+1(x)=Fm(x)+h(x)=y(x) ....(15)

Or equivalently,
h(x)=y-Fm(x) ....(16)

Gradient boosting will therefore fit h to the residual, y-Fm(x). As in other variants of boosting, each Fm+1(x) attempts to correct the errors of its predecessor, Fm. A generalization of this idea to loss functions other than square error and problems of classification and ranking, follows from the observation that residuals y-Fm(x)for a given model are the negative gradients of the squared error loss function, $1/2\ (y - f(x))^2$. So, Gradient Boosting is an algorithm of gradient descent, and generalizing it involves "plugging in" a different loss and gradient.

Convolutional Neural Network

CNN is a class of deep neural networks in deep learning, most frequently used to analyze visual imagery[28]. CNNs use a multilayer perception variation designed to require minimal pre-processing. A convolutionary neural network consists of a layer of input and output, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, RELU layer i.e. activation function, pooling layers, fully connected layers and normalization layers. The process is described by convention as a convolution in neural networks. Mathematically it is a cross-correlation rather than a convolution (although cross-correlation is a related operation). This only has significance for the indices in the matrix, and thus which weights are placed at which index.

• Convolution

Convolutionary layers apply an operation of convolution to the input and pass the result to the next layer. The convolution emulates the visual stimuli response of an individual neuron. For its receptive field, each convolutionary neuron processes data only. Although it is possible to use fully connected feedforward neural networks to learn features and classify data, applying this architecture to images is not practical. Due to the very large input sizes associated with images, where each pixel is a relevant variable, a very high number of neurons would be needed, even in shallow (opposite to deep) architecture. A fully connected layer, for example, for a (small) size image 100 x 100 has 10000 weights for each neuron in the second layer. This problem is solved by the convolution operation as it reduces the number of free parameters, allowing the network to be deeper with fewer parameters. For example, irrespective of the image size, tiling size regions 5 x 5, each with the same shared weights, requires only 25 learnable parameters. This way, by using back propagation, it solves the problem of disappearing or exploding gradients in training traditional multi-layer neural networks with many layers.

• Pooling

Convolutionary networks may include layers of pooling locally or globally. Pooling layers reduce data dimensions by combining neuron cluster outputs at one layer in the next layer into a single neuron. Local pooling, typically 2x 2, combines small clusters. Global pooling acts on the convolutionary layer's neurons. Pooling can also calculate a max or an average. The maximum value of each cluster of neurons in the previous layer is used by Max pooling. Average pooling uses the average value of each of the previous layer's clusters of neurons.

- **Fully Connected**

Fully connected layers connect each neuron to each neuron in another layer in one layer. It is basically the same as the traditional neural perceptron multi-layer (MLP) network. To classify the images, the flattened matrix passes through a fully connected layer.

- **Receptive Field**

Each neural network receives input from a number of preceding layer locations. Each neuron receives input from each element of the previous layer in a fully connected layer. Neurons receive input from only a limited sub-area of the previous layer in a convolutionary layer. The sub-area is typically square in shape (e.g. size 5 by 5). A neuron's input area is called its receptive field. Thus, the receptive field is the entire previous layer in a fully connected layer. The receptive area is smaller in a convolutionary layer than the entire previous layer.

- **Weight**

Each neuron in a neural network calculates an output value by applying a certain function in the previous layer to the input values from the receptive field. A weight vector and a bias (typically real numbers) specify the function that is applied to the input values. Neural network progress can be made through incremental adjustments to biases and weights. The weights vector and bias are called a filter and represent some input feature (e.g., a specific shape). A distinctive feature of CNNs is that the same filter is shared by many neurons. This reduces memory footprint as a single bias and a single weight vector is used across all receptive fields sharing that filter instead of each receptive field having its own bias and weight vector.

4.1.3 Summary

Total four classifiers have been described comprehensively in this section as it is going to be performed on different features of this thesis work. A step by step classification procedure has been given. The mechanisms of the algorithms are also discussed briefly with mathematical equations.

4.2 FEATURE SELECTION

4.2.1 Introduction

In machine learning, we tend to add as many features as possible at first to capture useful indicators and get a more accurate result. However, the model's performance will decrease after a certain point with the increasing number of elements. This phenomenon is often called "Dimensionality's Curse" [34].

The curse of dimensionality occurs because with the increase in dimensionality, the sample density decreases exponentially. If we continue to add features without also increasing the number of training samples, the dimensionality of the feature space will increase and become sparser and sparser. Because of this sparsity, finding a "perfect" solution for the machine learning model that is highly likely to result in overfitting becomes much easier.

Overfitting occurs when the model is too close to a specific data set and is not well generalized. In the training dataset, an overfitted model would work too well to fail future data and make the prediction unreliable.

So how can we overcome the dimensionality curse and avoid overfitting especially when we have many features and relatively few samples of training? One popular approach is the reduction of dimensionality. Reduction of dimensionality is the process of reducing with consideration the dimensionality of the feature space by obtaining a set of main features. Reduction of dimensionality can also be broken into selection of features and extraction of features.

Selection of features attempts to select a subset of the original features to be used in the model of machine learning. We could remove redundant and irrelevant features in this way without incurring a lot of information loss.

Extraction of features is also called projection of features. While selection of features returns a subset of the original features, extraction of features creates new features by projecting the data to a smaller space in the high-dimensional space. Informative and non-redundant features can also be derived from this approach.

We can use selection of features and extraction of features together. Extraction of features can be done on selected elements that contain relevant information instead of using the original features. Besides avoiding overfitting and redundancy, reducing dimensionality also leads to better human interpretations and lower computational costs with model simplification.

In our study, we used principle component analysis (PCA) for dimension reduction and Recursive Feature Elimination (RFE) and P-Value calculation for feature selection.

4.2.2 Principle Component Analysis

Principal Component Analysis (PCA) is a statistical procedure using an orthogonal transformation to convert a set of observations of potentially correlated variables (entities each assuming different numerical values) into a set of values of linearly un-

correlated variables called main components [35]. If n observations are made with p, the number of separate main components is min (n 1, p). This transformation is defined in such a way that the first main component has the greatest possible variance (i.e. accounts for as much data variability as possible), and each successor component has the highest possible variance under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set (each being a linear combination of variables and containing n observations). PCA is sensitive to the original variables ' relative scaling.
Suppose we measured and plotted two variables, length and width, as shown below. Both variables have about the same variance and are highly correlated. We could pass a vector from right angles to the first through the long axis of the point cloud and a second vector, with both vectors passing through the data centroid.

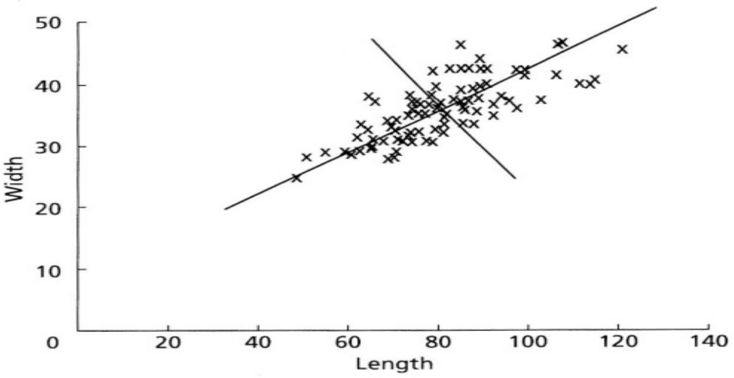

Figure 4.3: Plot of two variables (features- length and width) with n number of samples

Once we made these vectors, as shown here, we could find the coordinates of all the data points related to these two perpendicular vectors and re-plot the data.

Note in this new frame of reference that the variance on axis 1 is greater than on axis 2. Note also that the points ' spatial relationships are unchanged; this process has only rotated the data. Finally, note that there is no correlation between our new vectors or axes. The new axes could have specific explanations by performing such a rotation. Axis 1 could be considered as a measure of size in this case, with samples on the left having both small length and width and samples on the right having large length and width. Axis 2 could be considered as a measure of shape, with samples having different length to width ratios at any axis 1 position (i.e., of a given size).

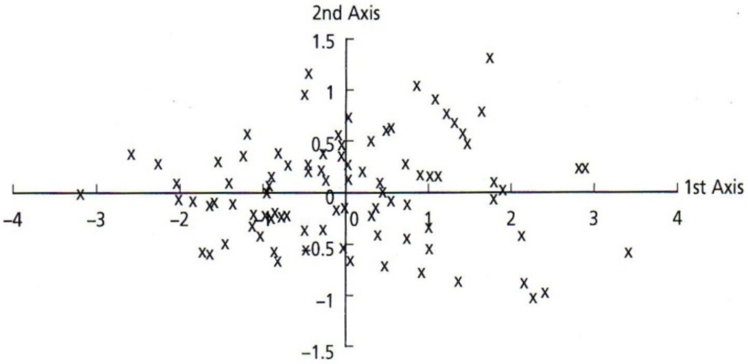

Figure 4.4: New reference frame with n samples

Generally speaking, PC axes will not exactly match any of the original variables. Although these relationships may seem obvious, this process allows one to evaluate any relationships between variables much faster when dealing with many variables. The variance of some axes may be great for data sets with many variables, while others may be so small that they may be ignored. This is known as reducing a data set's dimensionality, so you could start with thirty original variables, but end up with just two or three meaningful axes. The formal name for this rotating data approach is known as Principal Components Analysis, or PCA, so that each successive axis shows a decrease in variance. PCA produces linear combinations of the original axis variables, also known as main components, or PCs.

Computation of PCA

Given a data matrix with p variables and n samples, the data are first centred on the means of each variable. This will insure that the cloud of data is centred on the origin of our principal components, but does not aect the spatial relationships of neither the data nor the variances along our variables. The first principal components (Y1) is given by the linear combination of the variables X1, X2,Xp which is-
Y1 = a11X1 + a12X2 + ...+ a1pXp . . . .(5)

Or, in matrix notation
Y1 = aT1 X . . . .(6)

The first main component is calculated to account for the maximum variance in the data set. Of course, one could make the variance of Y1 as large as possible by choosing large values for the weights a11, a12,... a1p. To prevent this, weights are calculated with the constraint that their sum of squares is 1.
a211 + a212 + ...+ a21p = 1 . . . .(7)

The second main component is calculated in the same way, provided that it is uncorrelated with (i.e., perpendicular to) the first main component and the next highest variance is accounted for.
Y2 = a21X1 + a22X2 + ...+ a2pXp . . . .(8)

This goes on until a total of p main components, equal to the original number of variables, have been calculated. At this point, the sum of the variances of all the main components will be equal to the sum of the variances of all variables, i.e. all the original information was explained or accounted for. Collectively, all these original variables are transformed into the main components i.
Y = XA . . . .(9)

4.2.3 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a method of selecting features that fits a model and removes the weakest feature (or features) until reaching the specified number of features [27]. Features are ranked by the coefficient or feature attributes of importance of the model, and by repeatedly eliminating a small number of features per loop, RFE attempts to eliminate the dependencies and collinearity that may exist in the model. RFE requires a specified number of features to be maintained, but how many features are valid is often not known in advance. To find the optimum number of cross-validation features, RFE is used to score various subsets of features and select the best collection of features to score.
RFE is a type of method of reverse selection. RFE works on the ranking system of features, however. First model fits on all variables based on linear regression. It then calculates the coefficients of the variable and their significance. On the basis of linear regression fit, it then ranks the variable and then removes the low ranking variable in each iteration.

4.2.4 P-Value Calculation

A P-Value helps you determine the meaning of your results when performing a statistical hypothesis test. To test the validity of a claim made about a population, hypothesis tests are used. In essence, this claim on trial is called the null hypothesis [25].
The alternative hypothesis if the null hypothesis is concluded to be untrue is the one that would be believed. Your data and the statistics that accompany it are the evidence in the trial. Ultimately, all hypothesis tests use a P-value to weigh evidence strength (what the data tells you about the population). The P-value ranges from 0 to 1 and is interpreted as follows:
● A small P-value (typically 0.05) shows strong evidence against the null hypothesis, so the null hypothesis is rejected.

• A large P-value (> 0.05) indicates weak evidence against the null hypothesis, so the null hypothesis is not rejected.
• P-Values that are very close to the cutoff (0.05) are considered marginal (may go either way).
Generally, P-Values ≤ 0.05 indicates the significant feature. The less P-Value indicates the more significance of a feature. So, for feature selection, P-Value with $\leq$ 0.05 is selected. P-Value can be calculated in many ways, for example, using t-test or anova test or etc. For this study t test has been applied.

4.2.5 Summary

In this chapter, basic concepts of dimension reduction and feature selection have been discussed. The reasons of applying dimension reduction and feature selection is discussed properly here. Moreover, as dimension reduction algorithm how PCA works, its computational process are also described here. As feature selection algorithm, RFE and P-Value has been chosen.

5 RESULTS AND DISCUSSION

5.1 Introduction

In this chapter, speech emotions are classified in accordance with different features that are used in this thesis work. A final overall result is also presented in this chapter. The results are compared against different performance parameters.

5.2 MFCC Extraction

First 14 coefficients are considered for extraction process. Time considered for each trial is 2.15 second. So, for each trail a matrix of 14215 dimension was achieved in this study. For each trail, only first row was considered. That means the first coefficients with 215 components are considered for each trail. Sampling frequency was 512 Hz. Using those coefficients, total seven features were extracted. They are mean, standard deviation, median, variance, mode, kurtosis and skewness. Amplitude and pitch- these features are developed from raw audio. Value of all features are given below,

Emotion	Mean	Standard Deviation	Median	Variance	Mode	Kurtosis	Skewness	Amplitude	Pitch
Female_angry	-34.3821	5.346704	-35.0646	45.48376	-45.2776	3.88519	0.734483	-1.37E-05	-0.45038
Female_angry	-28.5147	9.129523	-27.0051	95.5266	-46.5939	2.611645	-0.64786	4.88E-05	1.598211
Female_angry	-40.3449	7.937005	-40.1681	86.61492	-53.458	2.410482	0.183788	-4.51E-06	0
Female_angry	-35.2793	8.687238	-32.7979	93.60074	-52.2576	2.228528	-0.45834	-9.93E-05	0
Female_angry	-46.3319	6.596312	-44.7893	74.33651	-55.711	2.270145	0.064612	-4.62E-07	0
Female_angry	-38.1508	7.605945	-38.2202	78.78612	-52.849	2.842288	0.45179	-1.73E-05	-0.56551
Female_angry	-38.4427	7.263208	-37.0982	74.17801	-50.0937	3.563157	0.304459	7.11E-07	0
Female_angry	-34.9621	5.540055	-35.2018	48.33695	-44.3384	2.62335	0.244592	2.24E-06	0.073337
Female_angry	-39.2322	7.778736	-38.6512	82.84626	-55.5329	3.081797	0.01892	-3.96E-07	0
Female_angry	-36.0194	7.782884	-35.416	79.43503	-48.3727	2.353787	-0.11012	3.00E-07	0
Female_angry	-36.4734	7.804929	-36.1298	80.26296	-48.077	3.291108	0.462055	8.26E-07	0.027059
Female_angry	-37.022	6.902982	-36.895	67.5117	-50.7622	2.723007	0.003464	1.43E-06	0

Table 5.1: Sample feature extraction result

5.3 Classification Result

Using extracted features, four classifiers are used for classification. Among them CNN shows the best accuracy. Table 5.2 describes the result.

Classifier Name	Accuracy
Random Forest	61.26%
SVM	60.36%
Gradient Boosting	60.36%
CNN	71.17%

Table 5.2: Accuracy comparison of different classifiers

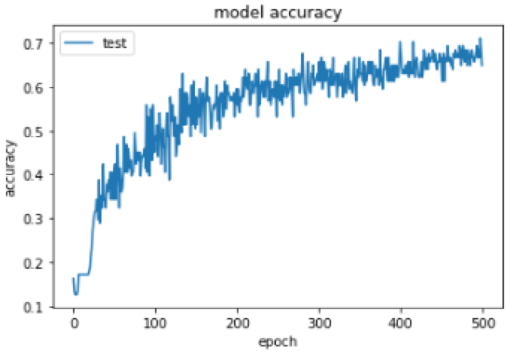

Figure 5.1: Epoch wise accuracy improvement of CNN (500 Epoch)

Now this time we wanted to improve accuracy of supervised learning classifiers using feature selection and dimension reduction algorithms. We used the following techniques-
a. RFE
b. P-Value Calculation
c. PCA

a. RFE

As feature selection algorithm, firstly we tried RFE. RFE ranks the features according to their importance.

i. RFE on Random Forest

For Random Forest, feature rank using RFE is,

Varying number of features according to rank the accuracies we get,

Name of Features	Mean	STDEV	MEDIAN	Amplitude	Pitch	Variance	Mode	kurtosis	skewness
Rank	1	4	3	7	9	5	6	8	2

Table 5.3: Feature ranking of Random Forest classifier using RFE

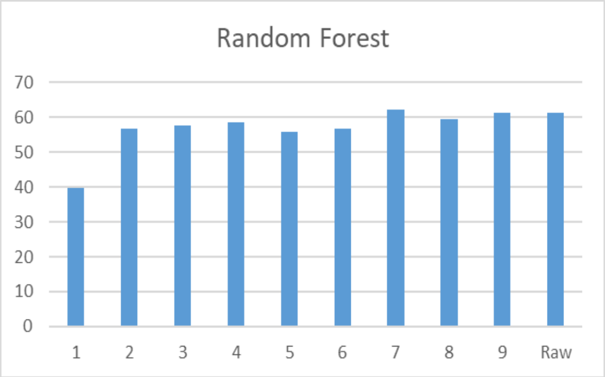

Figure 5.2: Bar chart showing accuracies using RFE on Random Forest

Features	1	2	3	4	5	6	7	8	9
Result(%)	39.634	56.75	57.65	58.55	55.85	56.75	62.16	59.46	61.26

Table 5.4: Accuracy using Random Forest with different combination of features

Using 7 features according to rank shows best accuracy of 62.16%.

ii. RFE on SVM

SVM has no effect on RFE. It shows same importance on all features (Table 5.5). For SVM, accuracy is same after applying RFE (60.36%).

Varying number of features according to rank accuracy we get,

Name of Features	Mean	STDEV	MEDIAN	Amplitude	Pitch	Variance	Mode	Kurtosis	skewness
Rank	1	1	1	1	1	1	1	1	1

Table 5.5: Feature ranking of SVM classifier using RFE

Features	1	2	3	4	5	6	7	8	9
Result(%)	60.36	60.36	60.36	60.36	60.36	60.36	60.36	60.36	60.36

Table 5.6: Accuracy using SVM with different combination of features

iii. RFE on Gradient Boosting

For Gradient Boosting, feature rank using RFE is,

Name of Features	Mean	STDEV	MEDIAN	Amplitude	Pitch	Variance	Mode	kurtosis	skewness
Rank	4	3	2	8	9	5	7	6	1

Table 5.7: Feature ranking of Gradient Boostingclassifier using RFE

Varying number of features according to rank the accuracies we get,

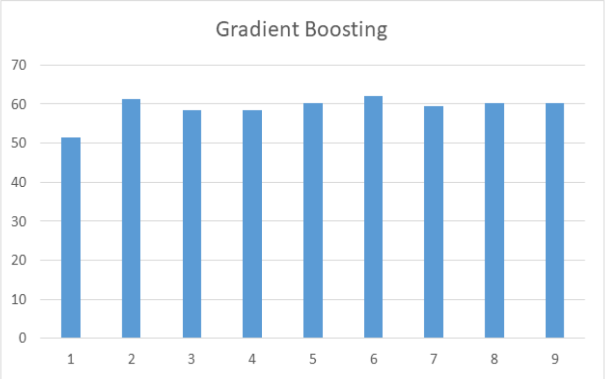

Figure 5.3: Bar chart showing accuracies using RFE on Gradient Boosting

Features	1	2	3	4	5	6	7	8	9
Result(%)	51.35	61.26	58.55	58.55	60.36	62.16	59.45	60.36	60.36

Table 5.8: : Accuracy using Gradient Boosting with different combination of features

Using 6 features according to rank shows best accuracy of 62.16%.

b. P-Value Calculation

Now we calculated P-Valuefor each features. It is said that, if P-Value for a feature is <0.05 then that feature is significant. From Figure: 5.4 it is clearly noticed that amplitude and pitch have higher value which greater than 0.05. So we avoided them.

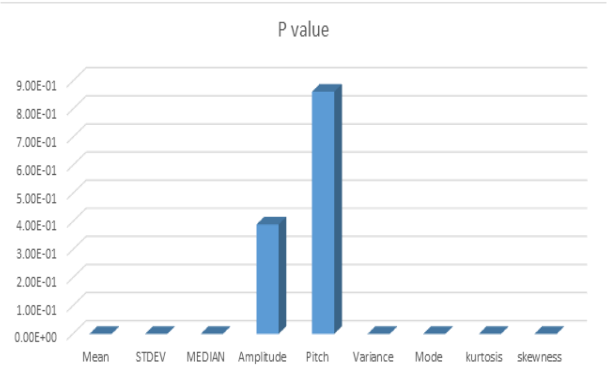

Figure 5.4: Bar chart showing Pitch and Amplitude has a higher value than others

Name of Features	Mean	STDEV	Median	Amplitude	Pitch	Variance	Mode	Kurtosis	Skewness
P-Value	1.05E-131	1.06E-155	7.11E-169	0.389255041	0.862034339	9.71E-113	2.87E-45	4.19E-10	4.56E-167

Table 5.9: P-Values of all features

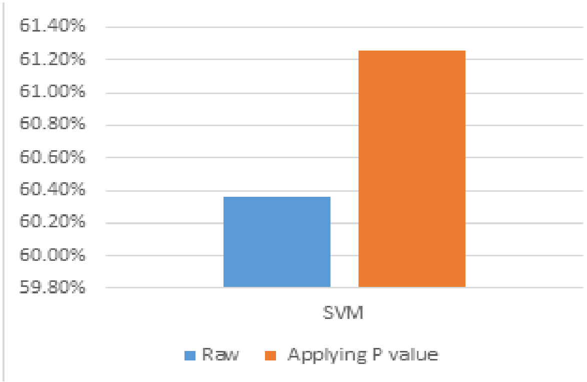

Figure 5.5: Bar Chart showing improvement on SVM

After discarding pitch and amplitude from the feature list SVM has some improvement than before. The improvement is shown in Figure: 5.5

c. PCA

The last task is to use dimension reduction algorithm to see if it can give any improvement on accuracy. For that purpose,PCA was used. But PCA didn't give any improvement as the data was uncorrelated.Figure 5.6 and Table 5.10 shows the result after applying PCA.

PCA features	1	2	3	4	5	6	7	8	9
Random Forest	35.13%	52.25%	36.93%	36.93%	35.13%	39.64%	38.74%	40.54%	40.54%
SVM	27.92%	49.54%	43.24%	37.83%	37.83%	36.93%	36.94%	39.64%	39.63%
Gradient Boosting	27.92%	45.94%	43.24%	36.93%	39.63%	41.44%	44.14%	43.24%	42.34%

Table 5.10: Accuracy after using PCA

Lastly, all the interpretations on the supervised algorithms show that applying RFE on Gradient Boosting gives the highest accuracy (62.16% in 6 dimensions). The overall accuracy comparison of each classifier is given on Table 5.11.

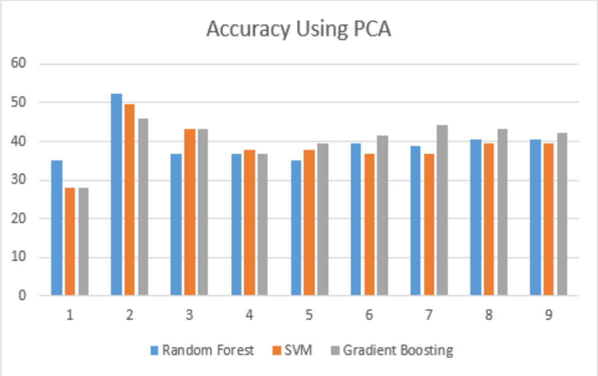

Figure 5.6: Bar Chart showing the comparison of PCA on all three classifiers

Name of Classifiers	Raw Accuracy	Applying RFE	Applying P-Value result	Applying PCA	Best Accuracy After Applying RFE,PCA and P-Value
Random Forest	61.26%	**62.16%**	57.65%	52.25%	**62.16%**
SVM	60.36%	60.36%	**61.26%**	49.54%	**61.26%**
Gradient Boosting	60.36%	**62.16%**	59.46%	45.94%	**62.16%**
CNN	**71.17%**				**71.17%**

Table 5.11: Accuracy comparison of each classifier

5.4 Discussion

We have conducted a comparative analysis on several techniques to show which method works the best. Future work can be done on these methods. From the results it can be concluded that PCA is not a good method when the data is uncorrelated like in our case, RFEseems like the most useful method out of the three feature selection and dimension reduction methods because it works well for smaller data points and for most classifiers. So, for any further work we can safely say that not all features are important for emotion detection, so before using classifiers the feature set can undergo RFE on Gradient Boosting. However, at any given dataset CNN works the best, so further works or improvements can be done using CNN or any other neural nets if work is done using unsupervised algorithms and if supervised classifiers are to be used then RFE can used on the features before moving on to the classification stage.

The proposed method is implemented in MATLAB and Python. MATLAB is used for the purpose of P-Value calculation and Python has been used for the purpose of classification and other algorithms. Device specification: Intel Core i3-4030U, CPU 1.90 GHz, RAM 12 GB, 64 bit operating system with x64 based processor. Classification and performance analysis have been done using python class scikit-learn and keras. They are for machine learning class in python. All the classifier performance has been measured properly. In spite of parameter tuning, this procedure can also be used for enhancing performance of a classifier. To establish the procedure in practical case, we need more data for research. More experimental analysis will provide strong evidence that this methodology can be used for emotion detection. In future, we would like to extend the work with more data, more data will make this analysis strong.

5.5 Summary

Result of feature extraction, classification performance has been described properly. Comparison among classifiers and comparison within classifier (Before and after using feature selection and dimension reduction algorithms) has been done properly. All the results are shown on table and figure. Finally, a discussion on this thesis work has been made.

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influence the recognition of emotion in speech.
In our research a model for successful detection of emotion was proposed. Study was done using MFCC coefficient. Using MFCC coefficients, nine features were extracted. Classification was done with four classifiers.CNN shows the best accuracy. PCA, RFE and P-Value were used for dimension reduction or feature selection which enhances classification performance of supervised classifiers. Three supervised classifiers, SVM,Random Forest and Gradient Boostinghave shown improvement on RFE and P-Value, but PCA couldn't give any improvement.And after the comparative analysis it has been shown that even for smaller or lighter data points, RFE shows most accuracy. This comparative analysis was done on three different feature selection or dimension reduction methods as mentioned, which have not been done earlier. Literature review shows that others who have worked on this field used any one of the methods or none and they have given more focus on the better use of classifiers.

6.2 Future Work

Dataset used for classification purpose is not so heavy for which the accuracy was not very much. So, the work will be done with heavy dataset in future. Accuracy, for supervised classifiers are not so high. In future, accuracy improvement task will be done by changing features. Moreover using any other scheme instead of MFCC, effect can be observed. May be, any other scheme can give better result such as emotion detection by speech spectrogram.
Moreover, we would like to extend our work to creating a Bengali dataset ourselves and then conduct the same process and then compare the two languages and emotions.

References

1. B. Jaramillo, E. B. Bolanos, T. V. Canas, J.R. Orozco and J. D. AriasLondono, J. F. V. Bonnilla “ Automatic Emotion detection in Speech using Mel frequency Cepstral Coefficients”, XV Simposio De Tratamiento De Senales, Images, Vision, Artificial-STSIVA, 2012.

2. A. FirozShah, V. Krishnan, A. RajiSukumar and P. BabuAnto, “Speaker Independent Automatic Emotion Recognitionfrom Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms,” 2009 International Conference on Advances in Recent Technologies in Communication and Computing, pp. 528-531,2009.

3. K.V. Krishna and P. K. Satish “Emotion Recognition in Speech using MFCC and Wavelet Features,” 3rd IEEE International Advance Computing Conference, 2013.

4. I. Mohino, R. G. Pita1, S. A. Diaz and M. R. Zureral, “MFCC Based Enlargement of the Trainingset for Emotion Recognition in Speech,” International Journal (SIPIJ), Vol.5, No.1, February 2014.

5. S Lalithaa, D Geyasrutia, R Narayanana and Shravani M, ”Emotion Detection using MFCC and Cepstrum Features,” 4th International Conference on Ecofriendly Computing and Communication Systems, 2015.

6. Steven R. Livingstone and Frank A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLOS One, 2018.

7. M. Xu, “HMM-based audio keyword generation,” 2004.

8. Sahidullah M. and S. Goutam, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,” Speech Communication, pp. 543–565, May, 2012.

9. T. Ganchev, N. Fakotakis and G. Kokkinakis, “Comparative evaluation of various MFCC implementations on the speaker verification task Archived 2011-07-17 at the Wayback Machine,” 10th International Conference on Speech and Computer SPECOM, Vol. 1, pp. 191–194, 2005.

10. M. Müller, “Information Retrieval for Music and Motion,” Springer, pp. 65, 2007.

11. Muda, Lindasalwa, M. Begam, and I. Elamvazuthi, “Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques,” arXiv preprint arXiv: 1003. 2010.

12. Zheng, Fang, G. Zhang and Z. Song, “Comparison of different implementations of MFCC,” Journal of Computer science and Technology, pp. 582-589, 2001.

13. Murty, K. Sri Rama and B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition,” IEEE signal processing letters 13.1, pp. 52-55, 2006.

14. B. Engel, “M. F. C. C. The emotionally abused woman: Overcoming destructive patterns and reclaiming yourself,” Ballantine Books, 2017.

15. Underhill L.G., Bradfield D., Introstat, Juta and Company Ltd, 1998.

16. Bland J.M., Altman D.G., “Statistics notes: measurement error,” 1996.

17. AnssiKlapuri, “Introduction to Music Transcription”, Signal Processing Methods for Music Transcription, Springer, p.p. 08, 2006.

18. Yuli Zhang, Huaiyu Wu, Lei Cheng, “Some new deformation formulas about variance and covariance,” Proceedings of 4th International Conference on Modelling, Identification and Control, ICMIC, pp. 987–992, June, 2012.

19. Damodar, N. Gujarati, “Econometrics. McGraw-Hill Irwin”, 3rd edition, pp. 110, 2006

20. Joanes D. N., Gill C. A., “Comparing measures of sample skewness and kurtosis,” Journal of the Royal Statistical Society, pp. 183–189, 1998.

21. M. S. Park; J. H. Na; J. Y. Choi, “PCA-based feature extraction using class information,”2005 IEEE International Conference on Systems, Man and Cybernetics, 2005.

22. Y. Sun ; J. Liu ; H. Zhang ; T. Wang ; L. Zheng, “On the algorithm of analyzing the features of magnetic flux leakage signal for pipeline defect based on PCA,” 36th Chinese Control Conference (CCC), 2017

23. B. Scholkopf, A. Smola and K.-R.Muller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem.” Neural Computation, vol. 10, no. 5, pp. 1299-1319, 1998.

24. https://www.scikit-yb.org/en/latest/api/features/RFEcv.html

25. https://www.dummies.com/education/math/statistics/what-a-P-Value-tells-you-about-statistical-data/

26. W. Jamal, S. Das, IA Oprescu, K. Maharatna, F. Apicella, F. Sicca “Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates,” J. Neural Eng. 11(4):046019, 2014.

27. G. Wei, J. Zhao, Z. Yu, Y. Feng, G. Li and X. Sun, ”An Effective Gas Sensor Array Optimization Method Based on Random Forest*,” 2018 IEEE SENSORS, New Delhi, 2018, pp. 1-4.

28. https://en.wikipedia.org/wiki/Convolutional_neural_network

29. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process, vol. 22, no. 6, pp. 1154–1160, Dec. 2012. 1

30. T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003. 1

31. S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

32. J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009. 1

33. http://everything.explained.today/Gradient_boosting

34. https://medium.com/@cxu24/why-dimensionality-reduction-is-important-dd6 0b5611543

35. https://en.wikipedia.org/wiki/Principal_component_analysis