

Sonargaon University

Department of Electrical & Electronic Engineering

B.Sc. Dissertation

On

Speech Emotion Recognition Using Machine Learning Technique

A thesis submitted in partial fulfillment of the requirement of the degree of Bachelor of Science in Electrical and Electronic Engineering.

Supervised By

Syed Raihan Hossain

Lecturer, EEE, SU

Submitted By

MD. ENAMUL HAQUE

ID: EEE 1601007140

MD. SHAKIRUL ISLAM

ID: EEE 1601007129

KAWSER AHMED

ID: EEE 1601007137

February 2020

Candidate's Declaration

It is declared hereby that this thesis paper or any part of it has not been submitted to anywhere else for the award of any degree.

…….....

Md. Enamul Haque

…….....

Md. Shakirul Islam

…….....

Kawser Ahmed

Certification

I certify that I have read this thesis work and that in my opinion it is fully adequate in scope & quality as a dissertation for the degree of Bachelor of Science in engineering.

Supervised By:

……..

Syed Raihan Hossain

Lecturer, EEE,

Sonargaon University

Department of Electrical and Electronic Engineering, Sonargaon University

ACKNOWLEDGEMENT

First of all, we would like to express our cordial gratefulness to almighty Allah for His kindness, for which we successfully completed our thesis within time and we also apologize to him for our any kind of mistakes.

We would like to express our boundless honor and respect to our thesis Supervisor **Syed Raihan Hossain,** Lecturer, Department of EEE, and Sonargaon University for his encouragement and for giving us permission to get involved with this thesis work. We were realy able to learn and grow under your guidance and advice during the project and throughout our thesis work.

We would like to thank ours parents ,Teachers and all the staffs of the department of EEE who actively supported us time to time by providing valuable resources and their valuable time.

ABSTRACT

Speech emotion recognition is a challenging problem partly because it is not clear what features are effective for the task. In this thesis, we propose a comparative study of speech emotion recognition (SER) systems. Theoretical definition, categorization of effective state and the modalities of emotion expression are presented. To achieve this study, we performed pre-processing necessary for emotion recognition from speech data on SER system, based on Multi-layer Perceptron (MLP) classifiers and MLP methods for features extraction, and this generates the training and testing datasets that contain the emotions of Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprised. The MLP classifiers are then used for the classification stage in order to predict the emotion. Mel-frequency Cepstrum coefficients (MFCC), Chroma, and Mel features are extracted from the speech signals and used to train MLP classified with the Mel-frequency cepstral coefficient (MFCC) feature extraction algorithms & Stochastic L-BFGS algorithm. Bangla and RAVDESS databases are used as the experimental data set. This study shows that for RAVDESS database all classifiers achieve an accuracy of 53.89% and for Bangla database 45.83% when a speaker normalization (SN) and a feature selection are applied to the features. The demand for machines that can interact with its users through speech is growing. For example, four of the world's largest IT companies; Amazon, Apple, Google and Microsoft, are developing intelligent personal assistants who are able to communicate through speech. In this thesis, we have investigated the effect of feature extraction when classifying emotions in speech, using a Artificial neural network (ANN). We used the "kernels" on Kaggle to extract sets of features from recorded audio, and compared the MLP classification accuracy of the sets with eight classes of emotions. We used one architecture of the ANN, to be fair when comparing each feature set. The ANN architecture was developed by an experimental approach. Using python in recent years, the working which requires human-machine interaction such as speech recognition, emotion recognition from speech recognition is increasing. Not only the speech recognition also the features during the conversation is studied like Melody, emotion, chunk, etc. It has been proven with the research that it can be reached meaningful results using prosodic features of speech. In addition, a Confusion Matrix (CM) technique is used to evaluate the performance of these classifiers. The proposed system is tested on RAVDESS and Bangla databases and has achieved a prediction rate of 70.89%.

Keywords: Speech processing, speech emotion recognition (ESR), feature extraction, MFCC, Chroma, Mel, Artificial neural network, MLP Classifier machine learning.

CONTENTS

CHAPTER Page

ACKNOWLEDGEMENT ..... **iii**

ABSTRACT ..... **iv**

ACRONYMS ..... **viii**

LIST OF TABLES ..... **x**

LIST OF FIGURES ..... **xi**

1 INTRODUCTION **1**

1.1 Background Study 1

1.2 Speech Emotion Recognition 2

1.3 Speech Emotion Recognition System 3

1.4 Feature Extraction and Selection 4

1.5 Applications 5

2 LITERATURE REVIEW **7**

2.1 Introduction 7

2.2 Review 7

2.3 Comparison Study and Feature Direction 11

3 DATASET DESCRIPTION **14**

3.1 Introduction 14

3.2 RAVDESS Database 14

3.3 Bangla Database 14

3.4 File Description 14

3.4.1 Audio-only file 15

3.4.2 File Summary 15

3.4.3 File naming convention 16

3.5 Data Field Descriptions 18

3.5.1 Data fields 19

3.6 Summary 19

4 METHODOLOGY **20**

4.1 Introduction 20
4.2 Feature Extraction & Selection 20
4.2.1 Mel Frequency Cepstral Coefficients (MFCC) 22
4.2.2 Chroma 24
4.2.3 MEL ..... **Error! Bookmark not defined.**
4.3 Neural Network Model 26
4.4 Artificial neural networks (ANN) 27
4.5 Multi-layer Perceptron (MLP) 28
4.5.1 Activated Function: 30
4.6 MLP Classifier 30
4.7 Training Multi-layer Perceptron Network 31
4.7.1 Selecting the Number of Hidden Layers 31
4.7.2 Deciding how many neurons to use in the hidden layers 31
4.7.3 Finding a globally optimal solution 32
4.8 The Algorithm 33
4.8.1 Construction of the Inverse Hessian Approximation $\boldsymbol{Hr}$ 35
4.8.2 Gradient descent 36
4.8.3 Stochastic gradient descent (SGD) 36
4.9 Evaluation metric 37
4.10 Confusion Matrix 38
4.11 Kaggle 41
4.11.1 Kaggle's Services 41
4.12 Summary 42
5 RESULT AND DISCUSSION 43
5.1 Result 43
5.1.1 Calculation for RAVDESS dataset 44
5.1.2 Calculation for Bangla dataset 48
5.1.3 Key results 52
5.2 Discussion 53
5.2.1 Interpretation and Analysis of results 53
5.2.2 Answer of our research question 54
5.2.3 Justification of our approach 54

5.2.4 Critical evaluation of our study 55
6 CONCLUSION AND FUTURE WORK 56
6.1 Conclusions 56
6.2 Future work 58
REFERENCES 59

ACRONYMS

SER	**: Speech Emotion Recognition**
MLP	**: Multi-layer Perceptron**
SN	**: Speaker Normalization**
MFCC	**: Mel-Frequency Cepstrum Coefficients**
ANN	**: Artificial Neural Network**
CM	**: Confusion Matrix**
HMI	**: Human-Machine Interaction**
LPCC	**: Linear Prediction Cepstrum Coefficient**
KNN	**: K-Nearest Neighbors**
HMM	**: Hidden Markov Model**
SVM	**: Support Vector Machine**
GMM	**: Gaussian Mixtures Model**
FS	**: Forward Selection**
PCA	**: Principal Component Analysis**
LFPC	**: Log Frequency Power Coefficients**
MSF	**: Modulation Spectral Features**
RBF	**: Radial Basis Function**
VAM	**: Vera am Mittag**
MDS	**: Multi-Dimensional Scaling**
MDT	**: Meta Decision Tree**
SBS	**: Sequential Backward Selection**
PAD	**: Position- Arousal-Dominance**
GMVAR	**: Gaussian Mixture Vector Autoregressive**

FDA	**: Functional Data Analysis**
QDC	**: Quadratic Discriminant Analysis**
RAVDESS	**: Ryerson Audio-Visual Database of Emotional Speech and Song**
BD	**: Bangla Database**
LPC	**: Linear Prediction Coefficients**
DCT	**: Discrete Cosine Transform**
CQT	**: Constant-Q Transform**
MFC	**: Mel-Frequency Cepstrum**
CNN	**: Convolutional Neural Networks**
RNN	**: Recurrent Neural Networks**
L-BFGS	**: Limited-Memory Broyden Fletcher Goldfarb Shanno**
GD	**: Gradient Descent**
SGD	**: Stochastic Gradient Descent**
FS	**: Feature Selection**

LIST OF TABLES

TABLE Page

Table 2.1: Automatic Speech Emotion Recognition Background Work 12
Table 3.1 Description of factor-level coding of RAVDESS filenames. 17
Table 3.2 Description of factor-level coding of Bangla filenames. 18
Table 4.1 Recognition performances of the MLP with different numbers of neurons in the hidden layers 31
Table 5.1 Confusion matrix 40
Table 5.2 Recognition accuracies (%) with different observed emotions of the training data used to calculate the confusion matrices of the dataset 52

LIST OF FIGURES

FIGURE Page

Figure 1.1 Structure of Speech Emotion Recognition System 3
Figure 4.1 (a) Waveform 21
Figure 4.2 (b) Spectrogram 21
Figure 4.3 Schema of MFCC extraction (Srinivasan et al., 2014) 22
Figure 4.4 Chroma Feature 24
Figure 4.5 Short-term power Mel Spectrogram 25
Figure 4.6 A multilayer perceptron neural network. 26
Figure 4.7 Neuron and myelinated axon, with signal flow from inputs at dendrites to outputs at axon terminals. 27
Figure 4.8 A block diagram of an MLP shown as a feed forward neural network. 29
Figure 4.9 Plot of error as a function of the weights. 32
Figure 4.10 Different combinations of predicted and actual values. 38
Figure 5.1 Performance Confusion Matrix 44
Figure 5.2 Performance Confusion Matrix For (RAVDESS) 47
Figure 5.3 Performance Confusion Matrix For (Bangla) 51

CHAPTER

1 INTRODUCTION

1.1 Background Study

The speech signal is one of the fastest, efficient and most natural methods of communications between humans and between humans and machines also. The emotional detection is very natural for humans but it is very difficult task for machines. Therefore, the purpose of emotion recognition system is to use emotion related knowledge in such a way that human machine communication will be improved. In recent time, Human-machine interaction (HMI) has become a growing area of innovation in industry as well as academic field [1]. A speech signal is basically a logical arrangement of sounds. Our brain performs a complex set of analysis of auditory input (i.e. sounds). It converts the sounds into some conceptual ideas and thoughts which forms the basis of instructions, commands, information & entertainment.

Speech emotion recognition is a kind of analyzing vocal behavior. The speech processing involves three main steps i.e. preprocessing, feature extraction and pattern recognition. In case of speech signal, vowels carry the most of the informative part. Vowels are mainly voiced part of the spoken words. Therefore, it is customary to separate out voiced part from unvoiced part of the information spoken and proceed further with signal processing on only voiced part.

For an effective and natural HMI, emotion recognition plays a vital role. Emotions reflect the mental state of the person through speech, facial expressions, body postures and gestures and also other physical parameters like body temperature, blood pressure, muscle action, etc. The mental state of the person indirectly affects the speech produced by the person.

In speech emotion recognition, the emotions from the speech of male or female speakers are found out [2] . In the past century some speech features were studied which involved the fundamental frequencies, Mel frequency Cepstrum coefficient (MFCC), linear prediction Cepstrum coefficient (LPCC), etc., which form the basis for speech processing even today. In one of the research the spectrograms of real and acted emotional speech were studied and found similar recognition rate for both, which recommend that later one can be used for the speech emotion recognition system. In another research a correlation between emotion and speech features were present. Further humans and machine emotion recognition rate was compared, in which same recognition rates were found for both.

Emotion recognition from the speaker's speech is very difficult because of the following reasons: In differentiating between various emotions which particular speech features are more useful is not clear. Because of the existence of the different sentences, speakers, speaking styles, speaking rates accosting variability was introduced, because of which speech features get directly affected. The same utterance may show different emotions. Each emotion may

correspond to the different portions of the spoken utterance. Therefore, it is very difficult to differentiate these portions of utterance. Another problem is that emotion expression is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system. There may be two or more types of emotions, long term emotion and transient one, so it is not clear which type of emotion the recognizer will detect [2].

Emotion recognition from the speech information may be the speaker dependent or speaker independent. The different classifiers available are k-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM). The paper reviews the mentioned classifiers [3].

1.2 Speech Emotion Recognition

Interpersonal communication is an interaction which involves the exchange of reciprocal ideas and emotions. Gestures and sound are a way of conveying information in a human-to-human interaction. Speech, a special form of sound, is one of the fundamental ways of conveying information between people. Words are not sole component of speech. Acoustic properties of speech also carry important effective features. Emotions exist in every part of the speech. Emotions in speech are transmitted from one communicator to another during an interaction. As a result of exchange of emotions during an ongoing conversation, emotional state of a speaker may easily trigger an interlocutor emotional state resulting in a change in the speech style or tone. Emotional states are transferred and mutually shaped through this process. Communication involves a source and a receiver. Speaker is the source and listener is the receiver. In a dyadic conversation, participants are both source and receiver in turns. On the source side, vocal track plays important roles. In the vocal fold, speech is shaped in a way which reflects the emotional state of the speaker. On the receiver side, heard speech signal is exposed to a series of transformations in the auditory system. Auditory system converts the voice in a way, which allows us to perceive content and context. Speech emotion recognition resides on the receiver side which aims at recognizing the underlying emotional state of a speaker.

Acoustic part of speech carries important cues about emotions. Each emotion has unique properties that make us recognize them. Main task of a speech emotion recognition algorithm is to detect the emotional state of a speaker from speech. General automatic speech emotion recognition algorithm composed of two parts which are feature extraction and classification stages. Prosodic and spectral features of speech are the most popular features that are used in speech emotion recognition algorithm. Intonation, pause, stress, pitch and rhythm are prosodic feature examples. Spectral features investigate frequency components of speech signal. Used classification algorithm varies from algorithm to algorithm. Support vector machines, Gaussian mixture models, hidden Markov model and neural networks are the most popular classification algorithms used in speech emotion recognition task. Automatic speech emotion

research generally focuses on the selection of right feature set and to detect in which emotion which features are more effective. Different from these studies, in this study, human auditory system is investigated. Our brain investigates the input from our auditory system. There are many auditory models that can simulate the process in our ear. In this study one of the models is investigated and its output is used to extract the emotional state of a speaker. The output of the auditory model is named as modulation transfer function. Yet, there is not any model that is able to tell how our brain evaluates the data from the auditory system and extracts the emotions, this study aims to constitute a machine learning algorithm which recognize emotions using the features extracted from a computational model of human auditory system. In the scope of this study, selected Bengali and English databases are applied to the machines. Besides machines, to make comparison, Bengali database is applied to human listeners who do not know any Bengali to measure human emotion recognition rate using only acoustic cues. Comparison results are going to verify the success rate of a computational model of human-like speech emotion recognition algorithm.

1.3 Speech Emotion Recognition System

Speech emotion recognition is nothing but the pattern recognition system. This shows that the stages that are present in the pattern recognition system are also present in the speech emotion recognition system. The structure of the speech emotion recognition is as shown in Figure 1.1.The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output [1].

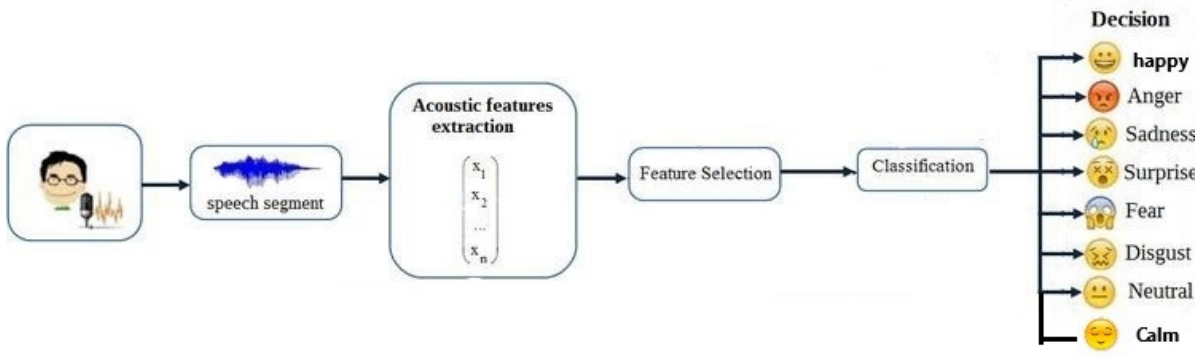


Figure 1.1 Structure of Speech Emotion Recognition System

The need to find out a set of the significant emotions to be classified by an automatic emotion recognizer is a main concern in speech emotion recognition system. A typical set of emotions contains 300 emotional states. Therefore, to classify such a great number of emotions is very complicated. According to “Palette theory” any emotion can be decomposed into primary

emotions similar to the way that any color is a combination of some basic colors. Primary emotions are anger, disgust, fear, happiness, sadness and surprise [2] .

The evaluation of the speech emotion recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. If the inferior database is used as an input to the system then incorrect conclusion may be drawn. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations [2].

1.4 Feature Extraction and Selection

Any emotion from the speaker's speech is represented by the large number of parameters which is contained in the speech and the changes in these parameters will result in corresponding change in emotions. Therefore, an extraction of these speech features which represents emotions is an important factor in speech emotion recognition system [4]. The speech features can be divided into two main categories that is long term and short-term features. The region of analysis of the speech signal used for the feature extraction is an important issue which is to be considered in the feature extraction. The speech signal is divided into the small intervals which are referred as a frame [2].

The prosodic features are known as the primary indicator of the speakers' emotional states. Research on emotion of speech indicates that pitch, energy, duration, formant, Mel frequency Cepstrum coefficient (MFCC), and linear prediction Cepstrum coefficient (LPCC) are the important features [4] [5]. With the different emotional state, corresponding changes occur in the speak rate, pitch, energy, and spectrum. Typically anger has a higher mean value and variance of pitch and mean value of energy. In the happy state there is an improvement in mean value, variation range and variance of pitch and mean value of energy. On the other hand, the mean value, variation range and variance of pitch decrease in sadness, also the energy is weak, speak rate is slow and decrease in spectrum of high frequency components. The feature of fear has a high mean value and variation range of pitch, improvement of spectrum in high frequency components. Therefore, statistics of pitch, energy and some spectrum feature can be extracted to recognize emotions from speech [4,5] .

One of the main speech features which indicate emotion is energy and the study of energy depends on short-term energy and short-term average amplitude [4]. As the arousal level of emotions is associated with the short term speech energy therefore it can be used in the field of emotion recognition. The pitch signal which is also referred as the glottal wave form is one more main feature which indicates emotion in speech. The pitch signal depends on the tension of the vocal folds and sub glottal air pressure, and it is produced from the vibration rate of the vocal cord. The pitch signal is characterized by the two features that is pitch frequency, and glottal air velocity at the vocal fold opening time instant. Number of harmonics present in the spectrum is directly get affected by the pitch frequency [6].

Linear Prediction Cepstrum coefficient (LPCC) gives the details about the characteristics of particular channel of any individual person and this channel characteristic will get change in accordance with the different emotions, so by using these features one can extract the emotions in speech. The merits of using the LPCC is that it involves less computation, its algorithm is more efficient and it could describe the vowels in better manner. Mel frequency Cepstrum coefficient (MFCC) is extensively used in speech recognition and speech emotion recognition systems and the recognition rate of the MFCC is very good. In the low frequency region better frequency resolution and robustness to noise could be achieved with the help of MFCC rather than that for high frequency region [4]. Mel frequency Cepstrum is an illustration of short term power spectrum of sound [3].

In feature extraction all of the basic speech feature extracted may not be helpful and essential for speech emotion recognition system. If all the extracted features give as an input to the classifier this would not guarantee the best system performance which shows that there is a need to remove such useless features from the base features. Therefore, there is a need of systematic feature selection to reduce these features. Forward selection (FS) feature selection method could be used to select the best feature subset. In the initial stage forward selection initializes with the single best feature out of the whole feature set. The remaining features are further added which increases the classification accuracy. If the added number of features attained the preset number, the selection process should stop [7].

The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. The average accuracy of the most of the classifiers for speaker independent system is less than that for the speaker dependent. Automatic emotion recognitions from the human speech are increasing now a day because it results in the better interactions between human and machine. To improve the emotion recognition process, combinations of the given methods can be derived. Also by extracting more effective features of speech, accuracy of the speech emotion recognition system can be enhanced.

1.5 Applications

A new wave of interest has recently risen attracting both psychologists and artificial intelligence specialists. There are several reasons for this renewed interest such as: technological progress in recording, storing, and processing audio and visual information; the development of non-intrusive sensors; the advent of wearable computers; the urge to enrich human-computer interface from point-and-click to sense-and-feel; and the invasion on our computers of lifelike agents who supposed to be able express, have and understand emotions (Elliot and Brzezinski, 1998). A new field of research in AI known as affective computing has recently been identified (Picard, 1997). As to research on recognizing emotions in speech, on one hand, psychologists have done many experiments and suggested theories (reviews of about 60 years of research can be found in (van Bezooijen 1984; Scherer et al., 1991)). On the other

hand, AI researchers made contributions in the following areas: emotional speech synthesis (Canh, 1989; Murray and Arnott, 1993), recognition of emotions (Dellaert et al., 1996), and using agents for decoding and expressing emotions (Tosa and Nakatsu, 1996). The following pieces of software have been developed: Emotion Recognition Game (ERG), Emotion Recognition software for call centers (ER), and a dialog emotion recognition program (SpeakSoftly). No humans are ever non emotional. We speak emotionally, perceive others emotions and communicate emotionally. Despite this, contemporary human machine dialog systems always speak with the same unmoved voice and ignore customer's irony, anger or elation. This is partly due to insufficient technological performance with respect to recognition and simulation, and partly to a gap with respect to the necessary artificial intelligence to support emotional behaviour. In Figure 1, we display some possibilities of emotional processing in human machine interaction [8] . Emotional awareness can be included in several places of an information-processing system [9].

In telecommunication, it might be desirable to provide a special channel for emotional communication. A popular example are the so-called 'emoticons' used in e-mail communication. The human emotional expression can be analysed in different modalities, and this knowledge is used to alter the system reaction of recognition emotional expression can be mimicked by the system in order to enhance a natural interface or to access further channels of communication, like uttering urgent messages in an agitated speech style. Internal models of emotional representations can be used to represent user or system states or as models for artificial intelligence, e.g. influence decision making.

On the whole, the speech emotion recognition techniques are vastly applied in the field of medicine, E-learning, monitoring, environment, law, marketing etc.

• SER has many hot applications, for example, human robotic interfaces, smart call centers, intelligent spoken tutoring systems and spoken dialog system, to mention just a few.

• SER is handled as a pattern recognition task, which implies that the solution normally includes: feature selection and extraction, classifier choice, training and testing. And all the results of PR hold for SER.

• Databases of acted or authentic emotions can be used. For any applicable research, authentic data is more suitable. Yet we concern ourselves with a number of theoretical problems. And for now, the six databases of emotional speech used in this study contain only acted speech.

• Signal features included into the vector can be short-time or long-time. Long-time features have been shown to be more discriminative, although there are good systems, based on short-time features.

• Artificial intelligence to optimize speech recognition functions. Examples include intelligent personal assistants, e.g. Amazon's "Alexa" or Apple's "Siri".

CHAPTER

2 LITERATURE REVIEW

2.1 Introduction

Speech is the fast and best normal way of communicating amongst human. This reality motivates many researchers to consider speech signal as a quick and effective process to interact between computer and human. It means the computer should have enough knowledge to identify human voice and speech. Although, there is a significant improvement in speech recognition but still researcher is away from natural interplay between computer and human, since computer is not capable of understanding human emotional state. The recognition of emotional speech aims to recognize the emotional condition of individual utterer by applying his/her voice automatically. Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis [10] . Recognizing of emotional conditions in speech signals are so challengeable area for several reasons. First issue of all speech emotional methods is selecting the best features, which is powerful enough to distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features include pitch, energy [11]. Furthermore, it is possible to have a more than one specific emotion at the same time in the same speech signal, each emotion correlate with a different part of speech signals. Therefore, defining the boundaries between parts of emotion is very challenging task. The majority of works are concentrated on monolingual emotion recognition, and making a presumption that there is no cultural diversity between utterers. However, the multi-lingual emotion classification process has been considered in some research [12] .

2.2 Review

Over the last years, an excessive investigation has been completed to recognize emotions by using speech statistics.

In [13] the author proposed a ranking SVM (support vector machine) method for synthesize information about emotion recognition to solve the problem of binary classification. This ranking method, instruct SVM algorithms for particular emotions, treating data from every utterer as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM achieves two advantage, first, for training and testing steps in speaker- independent it obtains speaker specific data. Second, it considers the intuition that each speaker may express mixed of emotion to recognize the dominant emotion. Ranking

approaches achieves substantial gain in terms of accuracy compare to conventional SVM in two public datasets of acted emotional speech, Berlin and LDC. In both acted data and the spontaneous data, which comprises neutral intense emotional utterances, ranking-based SVM achieved higher accuracy in recognizing emotional utterances than conventional SVM methods. Unweight average (UA) or Balance accuracy achieved 44.4% [13].

In [14] the author proposed aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method classifies different emotions from coarse to fine then select appropriate feature by using Fisher rate. The output of Fisher rate is an input parameter 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 level are 86.5%, 68.5% and 50.2% separately in Beijing university database of emotional speech (BHUDES).

In [15] the author 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 demonstrate 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.

In [16] the author proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filter bank and an auditory filter bank 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.

In [17] the author presented an ensemble random forest to trees (ERFTrees) 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.

In [18] the author proposed a fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs). In this fusion method, first AP features are extracted then three different types of base-level classifier include GMMs, SVMs, MLP and Meta decision tree (MDT) are used. The maximum entropy model (MaxEnt) in the semantic labels method are applied. MaxEnt modeled the association between emotion association rules (EARs) and emotion states in emotion recognition. In the final state to define the emotion recognition outcome, the integrated consequence from the SL-based and AS-based are utilized. The experimental result on private dataset shows the performance based on MDT archives 80%, SL-based recognition archives 80.92, and mixture of AP and SL archives 83.55%.

In [19] the Author proposed domain-specific emotion recognition by utilizing speech signals from call center application. Detecting negative and non-negative emotion (e.g. anger and happy) are the main focus of this research. Different types of information include acoustic, lexical, and discourse are used for emotion recognition. In addition, information-theoretic contents of emotional salience are presented to obtain data at emotion information at the language level. Both k-NN and linear discriminant classifier are used to work with different types of features. Experimental result confirms that the best results are achieved by combination of acoustic and language data. Outcomes demonstrates by combining three information sources instead of one source, classification accuracy increases by 40.7% for males and 36.4% for females. Compare to pervious work improvement range in accuracy is from 1.4% to 6.75% for male and 0.75% to 3.96% for female.

In [20] the author presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory. First, beginning from predicted pitch of a speech signals, then computing spherical autocorrelation of pitch histogram. It calculates the incidence of dissimilar two pitch duration, which cause a harmonic or inharmonic impression. In Classification step, Bayesian classifier plays an important rule with a Gaussian class-conditional likelihood. Experimental result in Berlin emotion database by using harmony features indicate an improvement in recognition performance. Recognition rate improved by 2% in average.

In [21] the author investigate a new spectral feature in order to determine emotions and to characterize groups. In this study based on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design a novel hierarchical technique for classification of emotions. The innovation of the proposed method is two things, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Experimental result in Berlin dataset demonstrates the hierarchical approach achieves the better performance compare to

best standard classifier, with decuple cross-validation. For example, performance of standard HMM method reached 68.57% and the hierarchical model reached 71.75%.

In [22] the author represent a hierarchical computational structure to identify emotions. This method via following layers of binary classifications, maps input speech signal in one of the corresponding emotion classes. The main concept of different level in tree is to solve the classification task in easiest way to diminish error propagation. AIBO and USC IEMOCAP databases are employed to evaluate the classification method. Over the baseline SVM, the absolute result improve the accuracy archives an absolute improvement of 72.44%- 89.58%. The consequence proves the reported hierarchical method is efficient for classifying emotional speech in various databases.

In [23] the author proposed hierarchical structure for binary decision tree in emotion recognition fields. This method concentrate on locates the simpler classification obstacle at top level of tree to diminish agglomeration of error. This structural method also maps input speech data into one the emotion classes via following layer of binary classification. The result on AIBO database shows 3.3% absolute over baseline model in [17], which archives 70.1% and 65.1% for two-class and five-class problem respectively. As an alternative solution instead of out- putting hard labels at every step, measuring of probability as a soft label can makes the powerful framework modeling. Bayesian Logistic Regression and SVM were employed as a binary classifier. Outcome were not significantly higher than Bjorn et al. [14, 15].

In [25] the author proposed a segment based method for recognition of emotion in Mandarin speech. This approach is containing the following process. First, define the k parameter in weighted discrete k-NN classifier, the experimental testing of different k shows the best performance for k-NN is when k sets to 10. For selecting the foremost feature set, sequential forward selection (SFS) and sequential backward selection (SBS) are employed. SFS and SBS improves feature accuracy to 84% and 82% respectively. The highest accuracy in segment-based method achieves 86%. The experimental result build on private corpus by inviting 18 males and 16 females. It is essential to gather more expressive speech to explore the extensive emotional investigation, in the future.

In [26] the author proposed a computational approach for recognition of emotion and analysis the specifications of emotion in voiced social media such as WeChat. This approach approximates the mixed emotion and dynamic fluctuations in position- arousal-dominance (PAD) by extracting 25 acoustic features of speech signals and employing trained least squares-support vector regression (LV-SVR) model as well. The experimental results demonstrate the recognition rate for different emotion are different and the average rate of recognition achieves 82.43%, which is the best existing result by similar examination.

In [27] the author proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is mixture of GMM with vector autoregressive for classification problem of speech emotion recognition. The key idea of GMVAR is its capability to multi-modality in their dissemination and design the dependency between speech feature set. Berlin emotional dataset was used for evaluation of GMVAR. The experimental result shows classification accuracy

achieves 76% when for HMM reached 71%, for k-NN 67% and 55% for feed-forward neural networks. The advantage of this method better differentiation amongst high and low arousal with neutral emotions compare to HMM.

In [28] the author proposed a novel shape based method by using neutral model to recognize emotional salience in the basic frequency. The new method supported by functional data analysis (FDA) that target to obtain the natural changeability of F0 contours. For a certain F0 contour, PCA are calculated to use as feature for speech emotion recognition. The empirical result indicate the proposed approach accuracy achieved 75.8% in binary classification. It means 6.2% higher than trained benchmark system with over-all F0 statics. The approach is evaluated by the SEMAINE dataset. The results designate that to recognize speech emotion, utilizing the shape-based method in real application can be effective.

In [29] the author proposed a multi-dimensional model by utilizing emotion primitives for speech emotion recognition. Three dimension were made by composing of three different value of emotion primitives, which is called valence, activation, and dominance. The value of these factors assumed to be in the range of [-1, +1]. A text free, image-based method was introduced to assess the emotion primitives and achieves best inter-evaluator agreement. For extracting acoustic feature such as energy, pith and spectral specifications, both fuzzy logic and rule based estimator are employed. The approached are validate by testing two EMA and VAM datasets, which are acted emotion and spontaneous speech emotion. Both dataset is recorded form talk-show in German TV. Finally, for mapping the emotion primitives to certain emotion category, k-NN was employed as a classifier. K-NN achieves total recognition rate up to 83.5%.

2.3 Comparison Study and Feature Direction

Generally, analysis on speech emotion recognition aims to increase the recognition rate as well as accuracy. Table-2.1 presents the current available methods, which targeted the speech emotion recognition systems, and these are evaluated with its classifier, features set, and recognition rate and on different dataset levels. As Table-2.1 depicts that SVM has many motivating properties such as easy to fulfilment because of mathematical basis compare to simple classifier such as quadratic discriminant analysis (QDC) [22]. Normally, SVM is employed as alone and also with combination of other classifier such as ANN and RBF to reduce the dimensionality. It shows most accuracy on according to table-2, where it uses the log frequency power coefficient (LEPC) and MFCC features set and achieve more than 95% in best case. The main limitation author just tested on private dataset. It should be tested with Berlin and VAM dataset or some other to verify the accuracy. Besides SVM, k-nearest neighbor (k-NN) is one of the well-known classifier in speech emotion recognition. Yeh et al. [14] have proven in the best situation can achieve 86% of recognition rate using k-NN method. Nwe et al. [12] by using HMM and adopting LFPC as feature parameters achieves the highest recognition rate in this research, which is 96% in the best situation, and his average rate is 78%.

Table 2.1: Automatic Speech Emotion Recognition Background Work

Reference	Corpus	Feature Set	Classification Method	Emotions	Accuracy
(Petrushin, 2000)	140 utterances per emotional State at 22-kHz/16 bit.	Pitch, vocal energy, frequency spectral features, formants, speech rate and pausing	Neural Network	Happiness, anger, sadness, fear, and neutral	70%
(Schuller, Rigoll, & Lang, 2003)	From five speaker 5250 samples In German and English	Low level contours	Hidden Markov Models	Surprise, joy, anger, fear, disgust, sadness, neutral	77.8%
(M. M. H. E. Ayadi, KaMel, & Karray, 2007)	Berlin emotional speech database	12 Mel frequency Cepstrum Coefficient MFCC, 12 delta coefficients, 0th cepstral coefficient, and the speech energy.	Gaussian mixture vector autoregressive model	Anger, boredom, fear, happiness, sadness and neutral	76%
(Nogueiras, Moreno, Bona-fonte, & no, 2001)	Spanish INTERFACE Database	Instant values and contours of Pitch and energy.	Hidden semi continuous Markov models	Anger, disgust, fear, joy, sadness and surprise, neural	70%
(Nwe et al., 2003)	Six Burmese and six Mandarin speakers generated 720 utterances	Logarithmic Frequency Power Coefficients	Hidden Markov Model	Anger, Disgust, Fear, Joy, Sadness and Surprise	77.1%
(Tawari & Trivedi,2010)	Berlin emotional speech database	Speech intensity, pitch and speaking rate, MFCC	Support Vector Machine	Surprise, joy, anger, fear, disgust, sadness, neutral	84%

Continued on next page

Table 2.1 – continued from previous page

(Ang et al., 2002)	Collected over the telephone and Sampled at 8 kHz.	Duration and speaking rate features, pause features, pitch features, energy features, and spectral tilt features	Decision tree	Neutral, annoyed, frustrated, tired, amused	71.7%
(S. Wu, Falk, & Chan, 2009)	Berlin emotional speech database	Spectro-temporal features, trajectories of pitch and intensity,	Support Vector Machine	Surprise, joy, anger, fear, disgust, sadness, neutral	88.6%
(Iliou & Anagnos-topoulos, 2010)	Berlin emotional speech database speaker dependent	133 features from pitch, Mel frequency cepstral coefficients, energy and formants.	probabilistic neural Network	Surprise, joy, anger, fear, disgust, sadness, neutral	94%
(Cichosz & Slot, 2007)	Berlin emotional speech database (independent)	Regression parameters of pitch and mean energy in low frequency sub-bands	Decision tree	Surprise, joy, anger, fear, sadness, neutral	72%
(Casale, Russo, Scebba, & Serrano, 2008)	Berlin Database of Emotional Speech	15 log energy coefficient, the 12 cepstral coefficients C1 C12, the pitch period, and the voicing class.	Support Vector Machine	Surprise, joy, anger, fear, disgust, sadness, neutral	92%

CHAPTER

3 DATASET DESCRIPTION

3.1 Introduction

In our Research, Dataset description we take two kind of language dataset. First North American Accent data and second Sonargaon University campus Students own voice Bangla. On first, these are two datasets originally made use in the repository RAVDESS and SAVEE, and we only adopted RAVDESS, and second took Bangla in our model. The RAVDESS and Bangla, there are two types of data having speech and song but we research here only RAVDESS and Bangla data of speech.

3.2 RAVDESS Database

The RAVDESS [30] is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest inter-rater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. All recordings are made freely available under a Creative Commons license.

3.3 Bangla Database

Similarly, Bangla Database was recorded from 24 non-professional Bangladeshi actors (both male and female). The audio-speeches include the same above mentioned 8 emotions. The recordings were converted from mp3 to wav to chunk to stereo to mono format. Finally, the database was given as input to the Kaggle-kernel python program and a satisfactory output accuracy was obtained.

3.4 File Description

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 1440 files (total size: 163 MB). Bangla contains 1536 files (total size: 141 MB).The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry,

fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in single modality formats: Audio-only (mono, 16bit, 16 kHz .wav).

3.4.1 Audio-only file

Audio-only files of all actors (01-24) are available on this file as these are shown on the speech bellow:

Speech file:

(Audio_Speech_Actors_01-24.zip, 74.8 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440 files for RAVDESS dataset.

(Audio_Speech_Actors_01-24.zip, 123.45MB) contains 1536 files: 64 trials per actor x 24 actors =1536 files for Bangla Database.

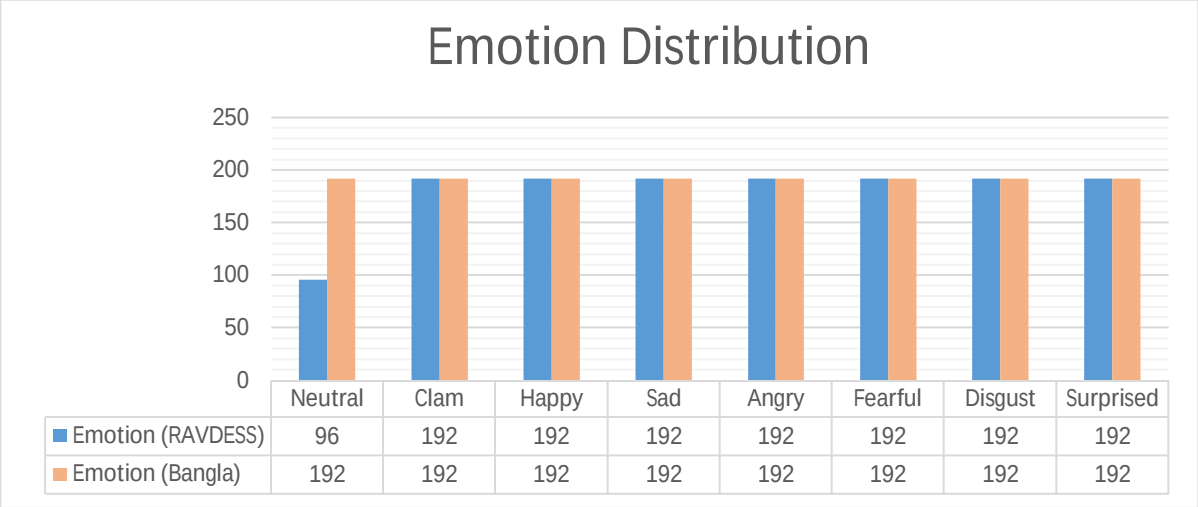
3.4.2 File Summary

In total, the RAVDESS collection includes 1440 files and Bangla 1536 files.

Total Class:

Emotion	Speech Sample Count (**RAVDESS**)	Speech Sample Count **(Bangla)**
Neutral	96	192
Calm	192	192
Happy	192	192
Sad	192	192
Angry	192	192
Fearful	192	192
Disgust	192	192
Surprised	192	192
Total	1440	1536

Here is the emotion class distribution bar chart (RAVDESS and Bangla):



3.4.3 File naming convention

Each **RAVDESS** and **Bangla** file has a unique filename. The filename consists of seven two-digit numerical identifiers, these identifiers define the stimulus characteristics and they are separated by hyphens (**e.g., 03-01-06-01-02-01-12.mp3**). Each two-digit numerical identifier defines the level of a different experimental factor. The identifiers are ordered: Modality–Channel–Emotion–Intensity–Statement–Repetition–Actor.mp4 or .wav. The numerical coding of levels is described in Table 3.1 and Table 3.2. For example, the Filename example: "03-01-06-01-02-01-12.mp4" refers to –

- Audio-only (03)
- Speech (01)
- Fearful (06)
- Intensity normal (01)
- Statement "dogs" (02)
- First Repetition (01)
- Twelfth Actor (12)
- The gender of the actor is coded by the actor's number, where odd numbered actors are male, even numbered actors are female.

Table 3.1 Description of factor-level coding of RAVDESS filenames.

Identifier	Coding Description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06=Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, ...…, 24 = Twenty-fourth actor

Table 3.2 Description of factor-level coding of Bangla filenames.

Identifier	Coding Description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06=Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 & 02 both are Taking from different chunk of "**Bangla language**" >"স্বপ্ন সেটা নয় যা আমরা সকালে ঘুমিয়ে থাকা অবস্থায় দেখি স্বপ্ন হলো সেটাই, যা আমাদের ঘুমাতে দেয় না " >"মানুষ মরে গেলে পচে যায়, বেঁচে থাকলে বদলায়, কারণে-অকারণে বদলায়" >"পরিচয় দ্বারা পাওয়া কাজ কিছু সময়ের জন্যে থাকে কিন্তু কাজ থেকে প্রাপ্ত পরিচয় সারাজীবন থাকে।" >"মানুষ মানুষের জন্যে, জীবন জীবনের জন্য একটু সহানুভূতি কি মানুষ পেতে পারে না।" >"যাঁরা বলেন বাংলায় বিজ্ঞান চর্চা সম্ভব নয় তারা হয় বাংলা জানেনা অথবা বাংলা বোঝেনা "
Repetition	01 & 02 = Number of repetition in the above statement.
Actor	01 = First actor, ...…, 24 = Twenty-fourth actor

3.5 Data Field Descriptions

Train - Training set. 10 days of click-through data, ordered chronologically. Non-clicks and clicks are subsampled according to different strategies.

Test - Test set. 1 day of ads to for testing our model predictions.

SampleSubmission.csv - Sample submission file in the correct format, corresponds to the All-0.5 Benchmark.

3.5.1 Data fields

- id: ad identifier
- click: 0/1 for non-click/click
- hour: format is YYMMDDHH, so 19091123 means 23:00 on Sept. 11, 2019 UTC.
- C1 -- anonymized categorical variable
- banner_pos
- site_id
- site_domain
- site_category
- app_id
- app_domain
- app_category
- device_id
- device_ip
- device_model
- device_type
- device_conn_type
- C14-C21 -- anonymized categorical variables

3.6 Summary

This chapter presented Bangla and RAVDEES database of emotional speech recognition were introduced. Among both of these database contains acted emotions, whereas the other two databases contain spontaneous or elicited emotions. Several non-linguistic and linguistically motivated units that can possibly be used for emotion recognition from speech were presented.

CHAPTER

4 METHODOLOGY

4.1 Introduction

Emotion recognition can have interesting applications in human-robot interaction, thus paving the way for a scenario where human-robot interaction will normally take place in the real world. When a speaker expresses an emotion while adhering to an inconspicuous intonation pattern, human listeners can nevertheless perceive the emotional information through the pitch and intensity of speech. An experimental methodology was set up, consisting of two different databases built from speakers of different areas of the world. The First database contains recordings from American English speakers, The second database contains recordings from Bangladeshi Bangla speakers includes Sonargaon University students in the age group from 21 to 26 years (12 women and 12 men).

4.2 Feature Extraction & Selection

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC) and Modulation spectral features. In this work, we have selected MFCC, Chroma and Mel to extract the emotional features.

Define a function extract feature to extract the MFCC, Chroma, Mel features from a sound file. This function takes 4 parameters- the file name and three Boolean parameters for the three features:

- MFCC: Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound.
- Chroma: Pertains to the 12 different pitch classes.
- Mel: Mel Spectrogram Frequency.

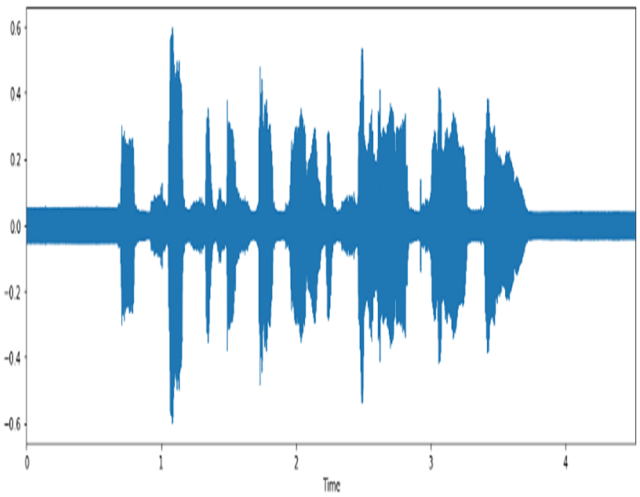


Figure 4.1 (a) Waveform

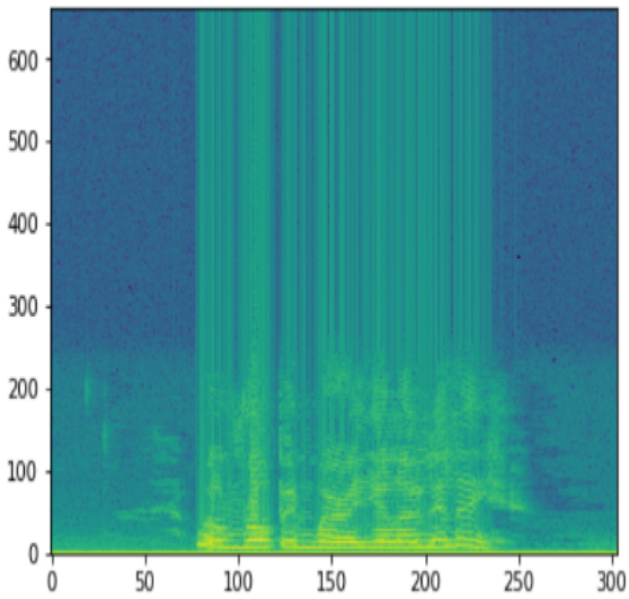


Figure 4.2 (b) Spectrogram

Figure 4.1 & 4.2 Audio waveform and it's converting Spectrogram

Loading audio data and converting it to MFCCs format can be easily done by the Python package librosa.

Advantages of feature selection:

- It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed.
- It removes redundant, irrelevant or noisy data.
- The immediate effects of data analysis tasks are speeding up the running time of the learning algorithms.
- Improving the data quality.
- Increasing the accuracy of the resulting model.
- The feature set reduction, to save resources in the next round of data collection or during utilization.
- Performance improvement, to gain in predictive accuracy.
- Data understanding to gain knowledge about the process that generated the data or simply visualizes the data.

4.2.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel-frequency Cepstrum coefficient is the most used representation of spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration.

MFCCs are calculated by applying a Mel-scale filter bank to the Fourier transform of a windowed signal. Subsequently, a DCT (discrete cosine transform) transforms the logarithmized spectrum into a Cepstrum. The MFCCs are then the amplitudes of the Cepstrum. Usually, only the first 12 coefficients are used. Through the mapping onto the Mel-scale, which is an adaptation of the Hertz-scale the frequency to the human sense of hearing, MFCCs enable a signal representation that is closer to human perception. The MFCC values used in the classification process. Usually, the process of calculating MFCC is shown in Figure 4.3.

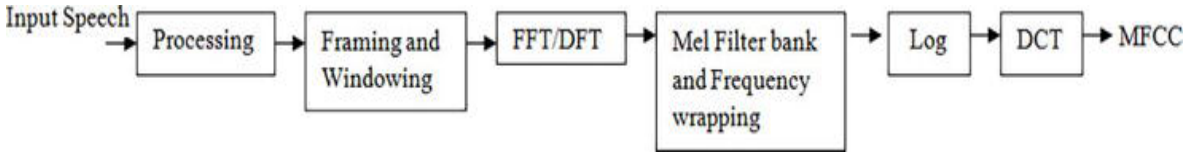


Figure 4.3 Schema of MFCC extraction (Srinivasan et al., 2014)

The calculation of MFCC involves computing the cosine transform of the real logarithm of the short-time power spectrum on a Mel-warped frequency scale, which can be broken down into the following steps as shown in Figure 4.3.

(a) DFT is applied in each speech frame as

$$X[k] = \sum_{N=0}^{N-1} x[n] e^{-\frac{j\omega k}{N}}, \quad 0 \le k \le N-1$$

(b) Mel-scale filter bank

The Fourier spectrum is non-uniformly quantized to conduct Mel-filter bank analysis. The window functions that are first uniformly spaced on the Mel-scale and then transformed back to the Hertz scale are multiplied with the Fourier power spectrum and accumulated to achieve the Mel-spectrum filter-bank coefficients. A Mel-filter bank has filters linearly spaced at low frequencies and approximately logarithmically spaced at high frequencies, which can capture the phonetically important characteristics of the speech signal while suppressing insignificant spectral variation in the higher frequency bands [31].

(c) The Mel-spectrum filter-bank coefficients is calculated as

$$F[m] = \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right) \qquad 0 \le m \le M$$

(d) The Discrete Cosine Transform (DCT) of the log filter bank energies is calculated to find the MFCC as

$$c[n] = \sum_{m=0}^{M} F[M] \cos\left(\frac{\pi n(m-1)}{2M}\right) \qquad 0 \leq n \leq M$$

Where c [n] is the nth coefficient.

In our research, we extract the first 12-order of the MFCC coefficients where the speech signals are sampled at 16 kHz. For each order coefficients, we compute the mean, standard deviation, Kurtosis, and Skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional.

Though the argumentation for the MFCCs is not without problems, it has become the most used feature in speech and audio recognition applications. It is used because it works and because it has relatively low complexity and it is straightforward to implement. Simply stated,

If we are unsure which inputs to give to a speech and audio recognition engine, we should try first the MFCCs.

The beneficial properties of the MFCCs include:

- Quantifies the gross-shape of the spectrum (the spectral envelope), which is important in, for example, the identification of vowels. At the same time, it removes the fine spectral structure (micro-level structure), which is often less important. It thus focuses on that part of the signal which is typically most informative.
- Straightforward and computationally reasonably efficient calculation.
- Their performance is well-tested and -understood.

Some of the issues with the MFCC include:

- The choice of a perceptual scale is not well-motivated. Scales such as the ERB or gamma-tone filter banks might be better suited. However, these alternative filter banks have not demonstrated consistent benefit, whereby the Mel-scale has persisted.
- MFCCs are not robust to noise. That is, the performance of MFCCs in the presence of additive noise, in comparison to other features, has not always been good.
- The MFCCs work well in the analysis but for synthesis, they are problematic. NaMely, it is difficult to find an inverse transform (from MFCCs to power spectra) which is simultaneously unbiased (=accurate) and congruent with its physical representation (=power spectrum must be positive).

4.2.2 Chroma

The Chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore Chroma features can be considered as an important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. A better quality of the extracted Chroma feature enables much better results in these high-level tasks. Short-Time Fourier Transforms and Constant Q Transforms are used for Chroma feature extraction.

In music, the term Chroma feature [32] or Chroma-gram closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale. One main property of Chroma features is that they capture harmonic and Melodic Characteristics of music while being robust to changes in timbre and instrumentation.

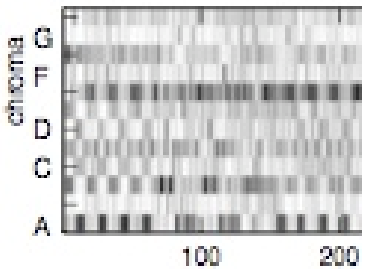


Figure 4.4 Chroma Feature

The human perception of pitch is periodic in the sense that two pitches are perceived as similar in “color” (playing a similar harmonic role) if they differ by one or several octaves (where, in our scale, an octave is defined as the distance of 12 pitches). For example, the pitches p = 60 and p = 72 are one octave apart, and the pitches p = 57 and p = 71 are two octaves apart. A pitch can be separated into two components, which are referred to as tone height and CHROMA. The tone height refers to the octave number and the Chroma to the respective pitch spelling attribute. In Western music notation, the 12 pitch attributes are given by the set

{C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B}

Extracting harmonic information directly from a spectrogram comes with a lot of problems, such as the fact that the logarithmic human perception of sound collides with the linear frequency resolution of the Fourier Transform. We, therefore, need to adapt the result of the STFT to a non-linear scale that matches the human perception of music. An appropriate auditory model to achieve this is the Constant-Q Transform (CQT) as presented in [33].

The CQT of a signal can be seen as a filter bank in which every single frequency band k_{cq} represents one semitone of the equally tempered scale. Theoretically, this could be extended to quarter or 1/8-tone resolution. Using the semitone instead as rasterization seems more appropriate as occidental music was developed upon this measure (even though there do occur notes like the harmonic seventh that are derived from the overtone scale and would fit a quarter-tone scale more tightly [34]).

To avoid misinterpretation of songs not played according to the standard pitch of 440Hz we perform a tuning step by detecting the center of energy within ±1 quarter-tone around the

standard pitch. The energy of the corresponding STFT bins is summarized over time and the position of the peak determines the most probable tuning frequency.

To receive the Chroma vector we summarize the energy that is distributed over all the bins that belong to one tone - those can be addressed as c (k_{cq}) = mod (k_{cq}, 12).

Thus, Chroma is a measure for the spectral energy that is divided into the 12 semitones of the well-tempered scale despite of their real-world frequencies as utilized by Bartsch and Wakefield in [35].

4.2.3 MEL

A value in Mels. This can also be a numpy array, conversion proceeds element-wise. In sound processing, the Mel-frequency cestrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency.

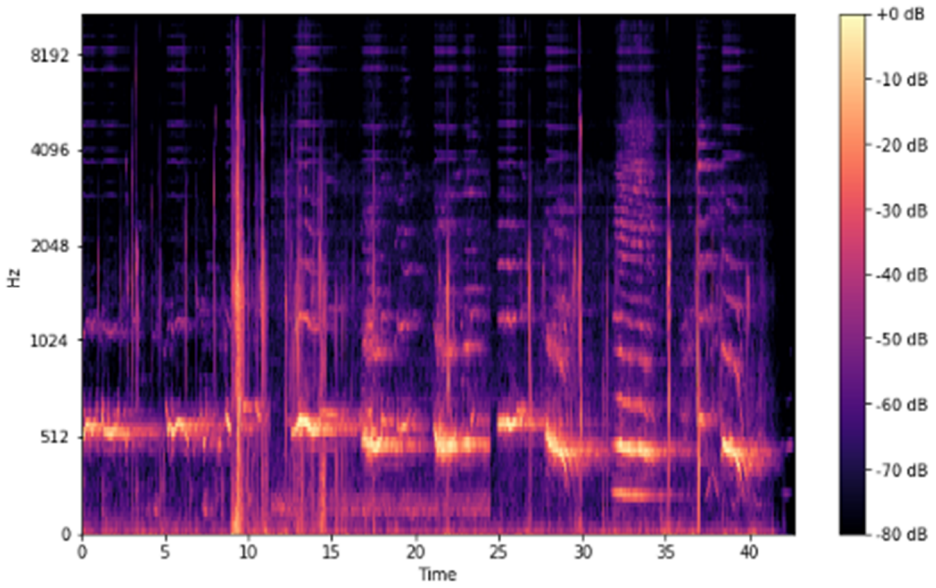


Figure 4.5 Short-term power Mel Spectrogram

Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear (humans are better at identifying small changes in a speech at lower frequencies). This scale has been derived from sets of experiments on human subjects. Let me give you an intuitive explanation of what the Mel scale captures.

The range of human hearing is 20Hz to 20 kHz. Imagine a tune at 300 Hz. This would sound something like the standard dialer tone of a land-line phone. Now imagine a tune at 400 Hz (a little higher pitched dialer tone). Now compare the distance between these two howsoever this may be perceived by your brain. Now imagine a 900 Hz signal (similar to a microphone feedback sound) and a 1 kHz sound. The perceived distance between these two sounds may

seem greater than the first two although the actual difference is the same (100Hz). The Mel scale tries to capture such differences. A frequency measured in Hertz (f) can be converted to the Mel scale using the following formula:

$$Mel\,(f) = 2595 \log(1 + \frac{f}{700})$$

The inverse transform can be readily derived as

$$f = 700\,(10^{(\frac{m}{2595})} - 1)$$

Any sound generated by humans is determined by the shape of their vocal tract (including the tongue, teeth, etc.). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the time power spectrum of the speech signal is representative of the vocal tract and MFCC (which is nothing but the coefficients that make up the Mel-frequency Cepstrum) accurately represents this envelope.

4.3 Neural Network Model

ANNs are characterized by their architecture. As stated in Section 4.6, neural network architecture refers to the number of layers, nodes in each layer and the number of arcs. Networks with one hidden layer are generally sufficient for most problems including classification. All networks used in this study will have one hidden layer. For this classification problem, the number of input nodes is the number of predictor variables. For example, for the prediction of the HIV status, the networks will have six input nodes in the first layer corresponding to six predictor variables. Node biases will be used in the output nodes and the logistic activation function will be specified in the networks.

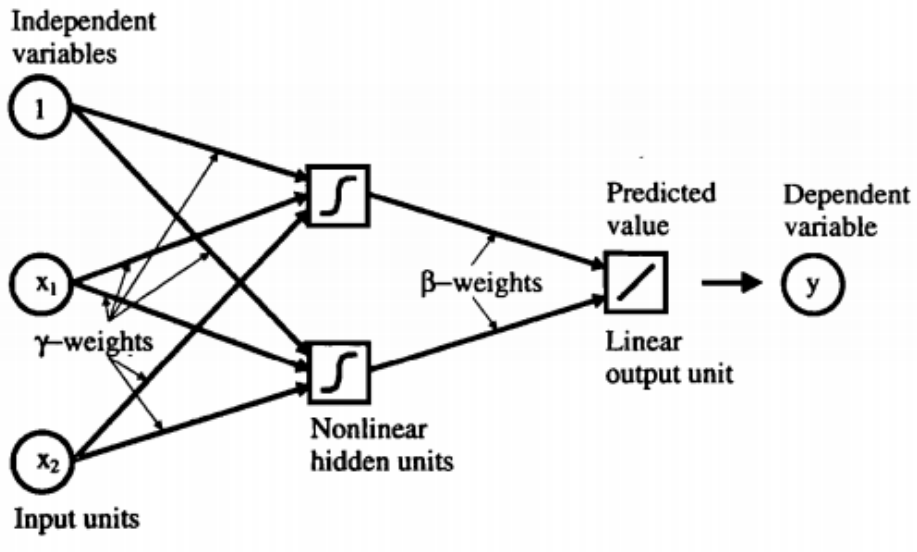


Figure 4.6 A multilayer perceptron neural network.

For this study data preprocessing is conducted, though some studies suggest that data [36] preprocessing is not beneficial for a classification exercise. Neural network training is a nonlinear nonconvex minimization problem and hence global solutions cannot be guaranteed. To reduce the likelihood of being trapped in bad local minima, the ANNs were each trained 50 times by using 50 sets of randomly selected initial weights and the best solution of weights among the 50 runs is retained for particular network architecture.

4.4 Artificial neural networks (ANN)

There are three classes of artificial neural networks that we recommend to focus on in general. They are:

- Multilayer Perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)

Artificial neural network (ANN) models based on static features vector as well as normalized temporal features vector, were used to recognize emotion state from speech. Moreover, relative features obtained by computing the changes of acoustic features of emotional speech relative to those of neutral speech were adopted to weaken the influence of the individual difference.

ANNs began as an attempt to exploit the architecture of the human brain to perform tasks that conventional algorithms had little success with. They soon reoriented towards improving empirical results, mostly abandoning attempts to remain true to their biological precursors. Neurons are connected to each other in various patterns, to allow the output of some neurons to become the input of others. The network forms a directed.

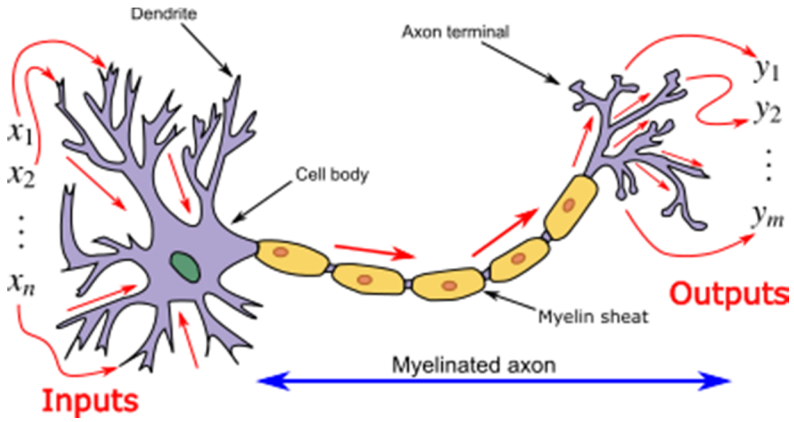


Figure 4.7 Neuron and myelinated axon, with signal flow from inputs at dendrites to outputs at axon terminals.

An artificial neural network consists of a collection of simulated neurons. Each neuron is a node that is connected to other nodes via links that correspond to biological axon-synapse-dendrite connections. Each link has a weight, which determines the strength of one node's influence on another.

An artificial neural network (ANN) combines biological principles with advanced statistics to solve problems in domains such as pattern recognition and game-play. ANNs adopt the basic model of neuron analogues connected to each other in a variety of ways.

Artificial neural networks can be employed to solve a wide spectrum of problems in optimization, parallel computing, and matrix algebra and signal processing. Taking a computational approach, this book explains how ANNs provide solutions in real time, and allow the visualization and development of new techniques and architectures. Features include a guide to the fundamental mathematics of neuro-computing.

4.5 Multi-layer Perceptron (MLP)

The Multi-Layer Perceptron (MLP) [26] is an ANN (Artificial Neural Network) which has its fundamentals in the human brain, where each neuron (here perceptron or node) fires an output depending on the input and its internal weights and then squashing it through a function which contains the output range.

Organizing these neurons in layers permits us to determine non-linear relationships between inputs.

The MLP is a sophisticated classifier that has been shown to work well on a number of problems. However, depending on the complexity of the network, training, as well as testing, can take very long. The MLP is a special kind of Artificial Neural Network (ANN). ANNs are developed to replicate the learning and generalization abilities of human behavior with an attempt to model the functions of biological neural networks of the human brain.

Architecturally, an MLP is a feed-forward layered network of artificial neurons. Each artificial neuron in the MLP computes a sigmoid function of the weighted sum of all its inputs. An MLP consists of one input layer, one output layer and a number of hidden or intermediate layers, as shown in Figure 4.8. The output from every neuron in a layer of the MLP is connected to all inputs of each neuron in the immediate next layer of the same. Neurons in the input layer of the MLP are all basically dummy neurons as they are used simply to pass on the input to the next layer just by computing an identity function each. The numbers of neurons in the input and the output layers of an MLP are chosen depending on the problem to be solved. The number of neurons in other layers and the number of layers in the MLP are all determined by a trial and error method at the time of its training. An ANN requires training to learn an unknown input-output relationship to solve a problem.

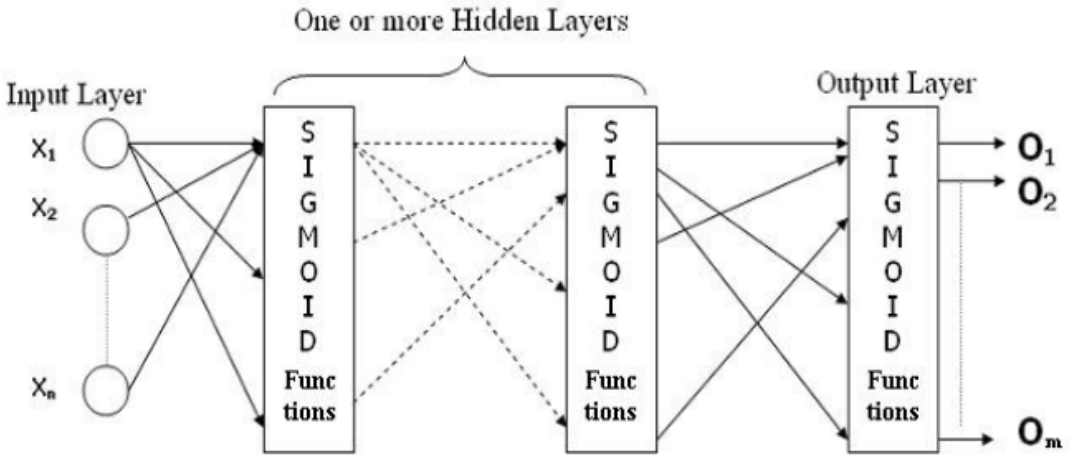


Figure 4.8 A block diagram of an MLP shown as a feed forward neural network.

Depending on the models of ANNs, training is performed either under the supervision of some teacher (i.e., with labeled data of known input-output responses) or without supervision. The MLP to be used for the present work requires supervised training. During the training of MLP weights or strengths of neuron-to-neuron connections, also called synapses, are iteratively tuned so that it can respond appropriately to all training data and also to other data, not considered at the time of training. The learning and generalization abilities of an ANN are determined on the basis of how best it can respond under these two respective situations.

In the Multilayer perceptron, there can more than one linear layer (combinations of ***neurons***). If we take the simple example of the three-layer network, the first layer will be the input layer and the last will be output layer and the middle layer will be called a hidden layer. We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task. Feed-forward network, is the most typical neural network model. Its goal is to approximate some function **f ().** Given, for example, a classifier **y = f ∗ (x)** that maps an input x to an output Class, y, the MLP finds the best approximation to that classifier by defining a mapping, **y = f(x; θ)** and learning the best parameters θ for it. The MLP networks are composed of many functions that are chained together. A network with three functions or layers would form **f(x) = f (3)(f (2)(f (1)(x)))** .Each of these layers is composed of units that perform an affine transformation of a linear sum of inputs. Each layer is represented as **y = f(WxT + b)**. Where f is the activation function (covered below), W is the set of parameters, or weights, in the layer, x is the input vector, which can also be the output of the previous layer, and b is the bias vector. The layers of an MLP consists of several fully connected layers because each unit in a layer is connected to all the units in the previous layer. In a fully connected layer, the parameters of each unit are independent of the rest of the units in the layer that means each unit possesses a unique set of weights.

In a supervised classification system, each input vector is associated with a label, or ground truth, defining its class or class label is given with the data. The output of the network gives a class score, or prediction, for each input. To measure the performance of the classifier, the loss function is defined. The loss will be high if the predicted class does not correspond to the true class, it will be low otherwise. Sometimes the problem of overfitting and under-fitting occurs at the time of training the model. In this case, our model performs very well on training data but not on testing data. In order to train the network, an optimization procedure is required for this we need loss function and an optimizer. This procedure will find the values for the set of weights, W that minimizes the loss function.

A popular strategy is to initialize the weights to random values and refine them iteratively to get a lower loss. This refinement is achieved by moving on the direction defined by the gradient of the loss function. And it is important to set a learning rate defining the amount in which the algorithm is moving in every iteration.

4.5.1 Activated Function:

Activation functions also known non- linearity, describe the input-output relations in a non-linear way. This gives the model power to be more flexible in describing arbitrary relations. Here are some popular activation functions Sigmoid, Relu, and TanH. In this MLP hidden layer we shown sigmoid function.

$$Sigmoid\ (x) = \frac{1}{1 + e^{-x}}$$

4.6 MLP Classifier

MLP classifier designed for the present work is trained with the Back Propagation (BP) algorithm. It minimizes the sum of the squared errors for the training samples by conducting a gradient descent search in the weight space. The number of neurons in a hidden layer in the same is also adjusted during its training.

The problem of pattern classification involves two successive transformations as follows: M→F→D

Where, M, F and D stand for the measurement space, the feature space and the decision space respectively. Once a feature set is fixed up, it is left with the design of a mapping (δ) as follows:

δ: F→D

ANNs with their learning and generalization abilities can approximate a general class of functions given below.

$f : R^n \rightarrow R$

Pattern classification with ANNs requires approximating δ as a discrete-valued function shown below.

$\delta: R^n \rightarrow \{1.2 \ldots m\}$

Where n and m denote the number of features and the number of pattern classes respectively. So an ANN-based pattern classifier requires n number of neurons in the input layer and m number of neurons in the output layer. Conventionally 1-out-of-m representation is used for its output.

Table 4.1 Recognition performances of the MLP with different numbers of neurons in the hidden layers

No of Hidden neurons	35	40	45	50	55	60	65	70	75
Percentage recognition rate the training samples	80.35	80.93	82.8	84.7	85.70	86.46	87.36	86.69	85.65
Percentage recognition rate test samples	70.35	70.70	71.6	73.15	74.2	75.05	73.65	74.7	72.00

4.7 Training Multi-layer Perceptron Network

The goal of the training process is to find the set of weight values that will cause the output from the neural network to match the actual target values as closely as possible. There are several issues involved in designing and training an MLP perceptron network such as:

- Selecting the number of hidden layers to use in the network.
- Deciding how many neurons to use in each hidden layer.
- Finding a globally optimal solution that avoids local minima.
- Converging to an optimal solution in a reasonable period of time.
- Validating the neural network to test for overfitting.

4.7.1 Selecting the Number of Hidden Layers

For nearly all problems, one hidden layer is sufficient. Using two hidden layers rarely improves the model, and it may introduce a greater risk of converging to local minima. There is no theoretical reason for using more than two hidden layers. Three-layer models with one hidden layer are recommended.

4.7.2 Deciding how many neurons to use in the hidden layers

One of the most important characteristics of a perceptron network is the number of neurons in the hidden layer(s). If an inadequate number of neurons are used, the network will be unable to model complex data, and the resulting fit will be poor.

If too many neurons are used, the training time may become excessively long, and, worse, the network may over-fit the data. When overfitting occurs, the network will begin to model random noise in the data. The result is that the model fits the training data extreMely well, but it generalizes poorly to new, unseen data. Validation must be used to test for this.

4.7.3 Finding a globally optimal solution

A typical neural network [37] might have a couple of hundredweights whose values must produce an optimal solution. The output of a neural network as a function of the inputs is often highly nonlinear, making the optimization process complex. The plot of error as a function of the weights for neural networks will most likely appear as a rough surface with many local minima such as shown in Figure 4.9.

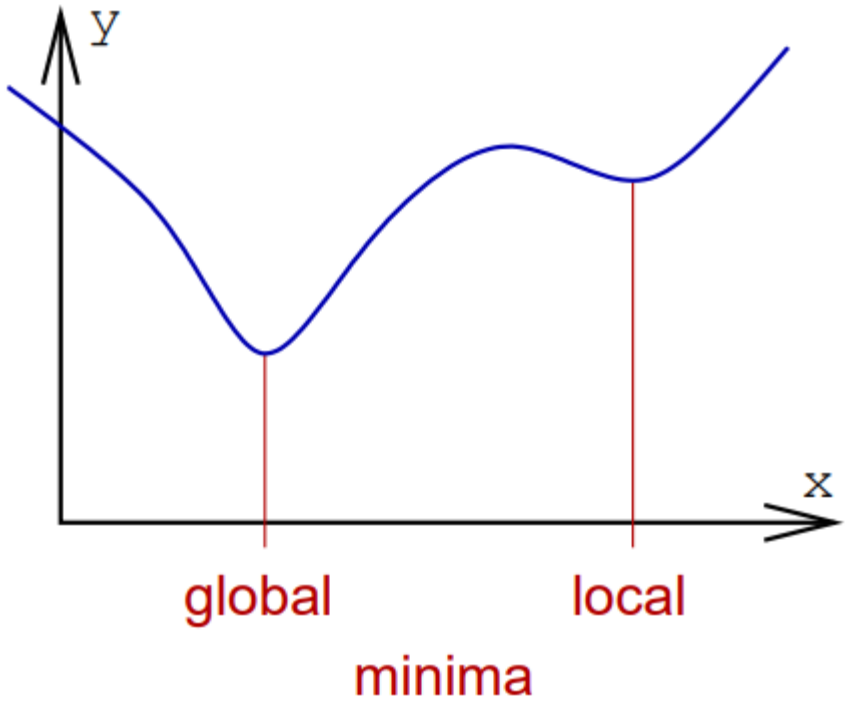


Figure 4.9 Plot of error as a function of the weights.

Optimization methods such as steepest descent and conjugate gradient are highly susceptible to finding local minima if they begin the search in a valley near a local minimum. They have no ability to see the big picture and find the global minimum. Several methods have been tried to avoid local minima. The simplest is just to try a number of random starting points and use the one with the best value. A more sophisticated technique called simulated annealing improves on this by trying widely separated random values and then gradually reducing the random jumps in the hope that the location is getting closer to the global minimum. Conjugate gradient usually finds the optimum weights quickly, but there is no guarantee that the weight values it finds are globally optimal.

4.8 The Algorithm

We introduce a stochastic variant of L-BFGS that incorporates the idea of variance reduction and has two desirable features. First, it obtains a guaranteed linear rate of convergence in the strongly-convex case. In particular, it does not require a diminishing step size in order to guarantee convergence (as partially evidenced by the fact that if our algorithm is initialized at the optimum it will stay there). Second, it performs very well on large-scale optimization problems, exhibiting a qualitatively linear rate of convergence in practice.

L-BFGS (**Limited-memory B**royden **F**letcher **G**oldfarb **S**hanno) is a numeric optimization **method** that has been effectively used for parameter estimation to train various machine learning models [40].

We consider the problem of minimizing the function

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} f_i(w) \dots\dots\dots\dots (4.1)$$

Over $w \in \mathbb{R}^d$. For a subset $S \subseteq \{1, \dots . N\}$, we define the subsampled function f_S by

$$fs(w) = \frac{1}{|S|} \sum_{i \epsilon S} f_i(w). \dots\dots\dots\dots (4.2)$$

Our updates will use stochastic estimates of the gradient ∇fs as well as stochastic approximations to the inverse Hessian $\nabla^2 f\tau$ Following Byrd et al. (2014), we use distinct subsets $S, \tau \subseteq \{1, \dots . N\}$ in order to decouple the estimation of the gradient from the estimation of the Hessian. We let $b = |s|$ and $b_H = |\tau|$.

Following Johnson and Zhang (2013), we occasionally compute full gradients, which we use to reduce the variance of our stochastic gradient estimates.

The update rule for our algorithm will take the form $W_{k+1} = W_k - \eta_k H_k V_k$.

In the gradient method, H_k is the identity matrix. In Newton's method, it is the inverse Hessian $(\nabla^2 f(w))^{-1}$. In our algorithm, as in LBFGS, H_k will be an approximation to the inverse Hessian. Instead of the usual stochastic estimate of the gradient, V_k will be a stochastic estimate of the gradient with reduced variance.

The code for our algorithm is given in **Algorithm 1**. Our algorithm is specified by several parameters. It requires a step size η, a memory size M, and positive integers m and L. Every m iterations, the algorithm performs a full gradient computation, which it uses to reduce the variance of the stochastic gradient estimates. Every L iterations, the algorithm updates the inverse Hessian approximation. The vector S_r records the average direction in which the algorithm has made progress over the past 2L iterations. The vector y_r is obtained by multiplying S_r by a stochastic estimate of the Hessian. Note that this differs from the usual

L-BFGS algorithm, which produces y_r by taking the difference between successive gradients. We find that this approach works better in the stochastic setting. The inverse Hessian approximation H_r is defined from the pairs (S_j, y_j) for $r - M + 1 \leq j \leq r$ using the standard L-BFGS update rule, which is described in **Section 4.8.1**. The user must also choose batch sizes b and b_H from which to construct the stochastic gradient and stochastic Hessian estimates.

In Algorithm 1 and below, we use I to refer to the identity matrix. We use $\mathcal{F}_{k,t}$ to denote the sigma algebra generated by the random variables introduced up to the time when the iteration counters k and t have the specified values. That is

$$\mathcal{F}_{k,t} = \sigma(\{S_{\dot{k},\dot{t}} : k' < k \; or \; k' = k \; and \; t' < t\} \cup \{\tau_r : \text{rL} \leq \text{mk+t}\})$$

We will use $\mathbb{E}_{k,t}$ to denote the conditional expectation with respect to $\mathcal{F}_{k,t}$. We define the inverse Hessian approximation H_r in **Section 4.8.1**. Note that we do not actually construct

Algorithm 1 Stochastic L-BFGS

Input: initial state ω_0 , parameters m, M, and L, batch sizes b and b_H, and step size η
Initialize r = 0
Initialize H_0 = I
for k = 0, . . . do
    Compute a full gradient $\mu_k = \Delta f(W_k)$
    Set $x_0 = \omega_k$
    for t = 0, . . . ,m − 1 do
        Sample a minibatch $S_{k,t} \subseteq \{1, \ldots, N\}$
        Compute a stochastic gradient $\nabla f S_{k,t}(x_t)$
        Compute a variance reduced gradient $V_t = \nabla f S_{k,t}(x_t) - \nabla f S_{k,t}(\omega_k) + \mu_k$
        Set $X_{t+1} = X_t - \eta H_r V_t$
        if $t \equiv 0 \bmod L$ then
            Increment r $\leftarrow$r + 1
            Set $U_r = \frac{1}{L}\sum_{j=t-L}^{t-1} xj$
            Sample $T_r \subseteq \{1, \ldots, N\}$ to define the stochastic approximation $\nabla^2 f T_r(U_r)$
            Compute $S_r = U_r - U_{r-1}$
            Compute $y_r = \nabla^2 f T_r(U_r) S_r$
            Define H_r as in Section 4.8.1
    Set $\omega_{k+1} = x_i$ for randomly chosen i ϵ {0, |. ,m − 1}

the matrix H_r because doing so would require $0(d^2)$ computation. In practice, we directly compute products of the form $H_r v$ using the two-loop recursion [41] .

Algorithm (forward pass)
Require: pattern ~x, MLP, enumeration of all neurons in topological order **Ensure: calculate output of MLP** **1: for all input neurons i do** **2: set a i ← x i** **3: end for** **4: for all hidden and output neurons i in topological order do** **5: set net i ← w i 0 + P j ∈Pred(i) wij a j** **6: set a i ← flog (net i)** **7: end for** **8: for all output neurons i do** **9: assemble a i in output vector ~y** **10: end for** **11: return ~**

Generic MLP learning algorithm:	
1: choose an initial weight vector ~w **2: intialize minimization approach** **3: while error did not converge do** **4: for all (~x, ~d) ∈ D do** **5: apply ~x to network and calculate the network output** **6: calculate ∂e(~x) ∂wij for all weights** **7: end for** **8: calculate ∂E(D) ∂wij for all weights suming over all training patterns** **9: perform one update step of the minimization approach** **10: end while**	**1: choose an initial weight vector ~w** **2: intialize minimization approach** **3: while error did not converge do** **4: for all (~x, ~d) ∈ D do** **5: apply ~x to network and calculate the network output** **6: calculate ∂e(~x) ∂wij for all weights** **7: perform one update step of the minimization approach** **8: end for** **9: end while**

4.8.1 Construction of the Inverse Hessian Approximation H_r

To define the inverse Hessian approximation H_r from the pairs (S_j, y_j) , we follow the usual **LBFGS** method.

Let $1/s_j^T$ y$_j$ and recursively define

$$H_r^{(j)} = (I - \rho_j s_j y_j^T)^T H_r^{(j-1)} (I - \rho_j s_j y_j^T)^T + \rho_j s_j s_j^T)^T \quad (4)$$

for $r - M + 1 \leq j \leq r$. Initialize $H_r^{(r-M)} = (S_r^T y_r / ||y_r||^2)$I and set $H_r = H_r^{(r)}$

Note that the update in Equation 4 preserves positive definiteness (note that $\rho_j > 0$), which implies that H_r and each $H_r^{(j)}$ will be positive definite, as will their inverses.

4.8.2 Gradient descent

It has often been proposed (e.g., [42]) to minimize the empirical risk En(fw) using gradient descent (GD). Each iteration updates the weights w on the basis of the gradient of En(fw),

$$w_t = w_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q(z_i, w_t), \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.3$$

Where γ is an adequately chosen learning rate. Under sufficient regularity assumptions, when the initial estimate w0 is close enough to the optimum, and when the learning rate γ is sufficiently small, this algorithm achieves linear convergence [6], that is, − log ρ ∼ t, where ρ represents the residual error.1 Much better optimization algorithms can be designed by replacing the scalar learning rate γ by a positive definite matrix Γt that approaches the inverse of the Hessian of the cost at the optimum:

$$w_{t+1} = w_t - \Gamma_t \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q(z_i, w_t) \ \ldots\ldots\ldots\ldots\ldots 4.4$$

This second order gradient descent (2GD) is a variant of the well-known Newton algorithm. Under sufficiently optimistic regularity assumptions, and provided that w_0 is sufficiently close to the optimum, second order gradient descent achieves quadratic convergence. When the cost is quadratic and the scaling matrix Γ is exact, the algorithm reaches the optimum after a single iteration. Otherwise, assuming sufficient smoothness, we have − log log ρ ∼ t.

4.8.3 Stochastic gradient descent (SGD)

Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties (e.g. differentiable or sub differentiable). It can be regarded as a stochastic approximation of gradient descent optimization since it replaces the actual gradient (calculated from the entire data set) by an estimate thereof (calculated from a randomly selected subset of the data).

The stochastic gradient descent (SGD) algorithm is a drastic simplification. Instead of computing the gradient of $E_n(f_w)$ exactly, each iteration estimates this gradient on the basis of a single randomly picked example z_t :

$$w_{t+1} = w_t - \gamma_t \nabla_w (Z_t, w_t) \ldots\ldots\ldots\ldots\ldots . (4.5)$$

The stochastic process $\{w_t,\ t = 1...\}$ depends on the examples randomly picked at each iteration. It is hoped that (4.5) behaves like its expectation (4.3) despite the noise introduced by this simplified procedure.

Since the stochastic algorithm does not need to remember which examples were visited during the previous iterations, it can process examples on the fly in a deployed system. In such a situation, the stochastic gradient descent directly optimizes the expected risk, since the examples are randomly drawn from the ground truth distribution.

Stochastic approximation methods are a family of iterative methods typically used for root-finding problems or for optimization problems. The recursive update rules of stochastic approximation methods can be used, among other things, for solving linear systems when the collected data is corrupted by noise, or for approximating extreme values of functions which cannot be computed directly but only estimated via noisy observations.

Stochastic approximation algorithms deal with a function of form $f(\theta) = E_{\xi}[F(\theta, \xi)]$ which is the expected value of a function depending on a random variable ξ. The goal is to recover the properties of such a function f without evaluating it directly. Instead, stochastic approximation algorithms use random samples of $F(\theta, \xi)$ to efficiently approximate properties of f such as zeros or extrema.

4.9 Evaluation metric

Evaluation metrics are used to measure the quality of the statistical or machine learning model. Evaluating machine learning models or algorithms is essential for any project. There are many different types of evaluation metrics available to test a model. These include classification accuracy, logarithmic loss, confusion matrix, and others. Classification accuracy is the ratio of the number of correct predictions to the total number of input samples, which is usually what we refer to when we use the term accuracy. Logarithmic loss, also called log loss, works by penalizing the false classifications. A confusion matrix gives us a matrix as output and describes the complete performance of the model. There are other evaluation metrics that can be used that have not been listed. Evaluation metrics involves using a combination of these individual evaluation metrics to test a model or algorithm.

There are various evaluation metrics applicable to this emotion prediction task, we report several evaluation metrics. Firstly, we use a relatively simple evaluation metric that measures the absolute difference between the predicted output and the reference input (also used in [13]):

$$e_i = |x_i^{pred} - x_i^{ref}|.$$

We report the e_{avg} which is obtained by averaging over N segments:

$$e_{avg} = \frac{1}{N}\sum_i^N e_i .$$

The lower e_{avg}, the better the performance. Secondly, as human-machine agreement measures, we report Krippendorff's $\alpha_{ord,5}$ to allow comparison with human performances. Finally, Pearson's ρ is reported.

4.10 Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification. When we get the data, after data cleaning, pre-processing and wrangling, the first step we do is to feed it to an outstanding model and of course, get output in probabilities. But hold on! How in the hell can we measure the effectiveness of our model? Better the effectiveness, better the performance and that are exactly what we want. And it is where the Confusion matrix comes into the liMelight.

Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. We can measure Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

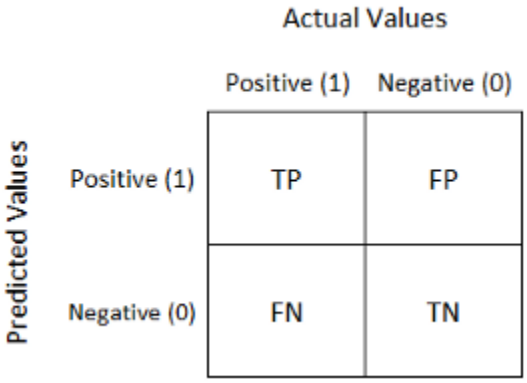
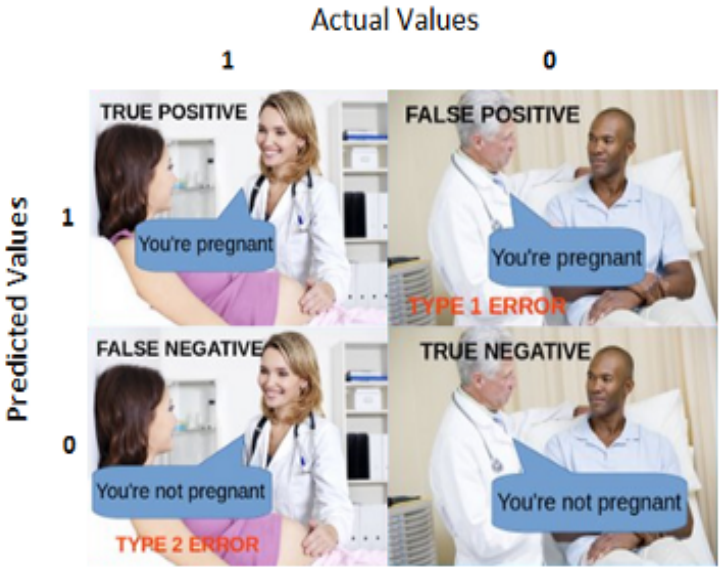


Figure 4.10 Different combinations of predicted and actual values.

Let’s understand TP, FP, FN, TN in terms of pregnancy analogy example



True Positive: *Interpretation:* You predicted positive and it's true. You predicted that a woman is pregnant and she actually is.	**False Positive: (Type 1 Error**) *Interpretation*: You predicted positive and it's false. You predicted that a man is pregnant but he actually is not.
True Negative: *Interpretation:* You predicted negative and it's true. You predicted that a man is not pregnant and he actually is not.	**False Negative: (Type 2 Error)** *Interpretation*: You predicted negative and it's false. You predicted that a woman is not pregnant but she actually is.

Just Remember, We describe predicted values as Positive and Negative and actual values as True and False.



Calculate Confusion Matrix for a 2-class classification

Table 5.1 Confusion matrix

y	Y pred	Output for threshold 0.6	Recall	Precision	**Accuracy**
0	0.5	0	1/2	2/3	4/7
1	0.9	1			
0	0.7	1			
1	0.7	1			
1	0.3	0			
0	0.4	0			
1	0.5	0			

Recall:

$$Recall = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 4.6$$

Out of all the positive classes, how much we predicted correctly. It should be high as possible.

Precision:

$$Precision = \frac{TP}{TP + FP} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4.7$$

Out of all the positive classes we have predicted correctly, how many are actually positive. And Accuracy will be Out of all the classes, how much we predicted correctly, which will be, in this case 4/7. It should be high as possible.

F-measure:

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \ldots\ldots\ldots\ldots\ldots \quad 4.8$$

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more

Accuracy:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ prediction}{Total\ number\ of\ predictios} \ldots\ldots\ldots\ldots\ldots\ 4.9$$

Accuracy above is also used to compare how well the model finds the correct predictions; however, the accuracy formula is not differentiable so it cannot be used for backpropagation. The accuracy is still a good metric to evaluate the performance of the models. One thing to note that is higher accuracy does not always mean that the model performance is better if the dataset is skewed. For instance, for the dataset which has 99% of instances in one class and only 1 % in the other, predicting that every instance belongs to the majority class will result in an accuracy of 99%. But this is not a problem in this work because the data is re-sampled to be balanced, and the accuracy is utilized for model evaluations.

4.11 Kaggle

Kaggle is the world's largest community of data scientists, statisticians, and machine learning engineers with powerful tools. It is a cloud-based workbench for data science and Artificial Intelligence education.

It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars from the likes of Hal Varian (Chief Economist at Google), Max Levchin (Paypal), Index and Khosla Ventures and then ultimately been acquired by the Google in March of 2017.

Kaggle enables data scientists and other developers to engage in running machine learning contests, write and share code, and to host datasets. The types of data science problems posted on Kaggle can be anything from attempting to predict cancer occurrence by examining patient records for analyzing sentiment to evoke by movie reviews and how this affects audience reaction.

4.11.1 Kaggle's Services

- **Kaggle Kernels:** A cloud-based workbench for data science and machine learning. Allows data scientists to share code and analysis in Python and R. Over 150K "kernels" (code snippets) have been shared on Kaggle covering everything from sentiment analysis to object detection.
- **Public Datasets Platform:** Here community members can share datasets publicly and among themselves.

- **Machine Learning Competitions:** This was Kaggle's first product. Companies post problems and machine learners compete to build the best algorithm.
- **Kaggle Learn:** In this way, we can learn Data Science education, python, Machine learning, Dataset visualization, pandas, Feature Engineering Deep Learning, SQL, Micro challenges, etc.
- **Jobs Board:** Employers post machine learning and AI jobs.

4.12 Summary

This chapter presented the methodology for speech emotion recognition system. Including two kinds of datasets. For feature extraction, an approach computing global statistics from basic feature types for each unit is chosen. Finally, the classification algorithms for a classifier comparison experiment were presented. The calculation of the basic feature type's MFCCs, Chroma, Mel, duration, spectral and a Confusion Matrix (CM) technique is used to evaluate the performance of these MLP classifiers. This database was given as input to the Kaggle-kernel python program and a satisfactory output accuracy described in the next chapter.

CHAPTER

5 RESULT AND DISCUSSION

In the results section, subjective speech emotion recognition performance is presented. Besides, subjective emotion recognition performance, developed automatic speech emotion recognition algorithms rates are exhibited. The developed algorithm is tested on two different databases. Algorithms' speaker and language dependence/independence cases are measured and exhibited in this section. In the comparison part, subjective performance and automatic recognition performance are compared.

To consider the different characteristics of speech among the speakers, the speech data define a function extract feature to extract the **MFCC**, **Chroma**, and **Mel** features from a sound file. This function takes 4 parameters- the file name and three Boolean parameters. Now, Load the data and extract features for each sound file that are going time to split the dataset into training and testing sets.

Let's keep the test set 25% of everything and use the load data function for this and observe the shape of the training and testing datasets and get feature extracted numbers.

5.1 Result

We can find result on two ways, such as

1. Calculate the accuracy by MLP model.
2. Calculate the accuracy by formula. [Equation 4.6, 4.7, 4.9]

5.1.1 Calculation for RAVDESS dataset

Calculate the accuracy to observe 2 emotions on confusion matrix

$$Recall = \frac{TP}{TP + FN}$$

Recall= 41 / (41+10) = 0 .80392

$$Precision = \frac{TP}{TP + FP}$$

Precision =41 / (41+7) = 0.85417

$$Accuracy = \frac{\sum TP}{Total}$$

= (41+38)/ (41+38+7+10)

= 0.8229

= 82.29 %

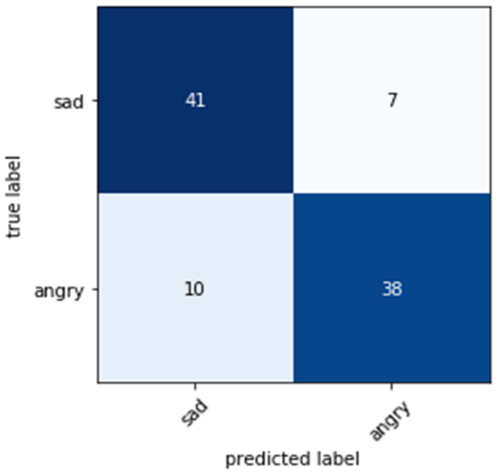
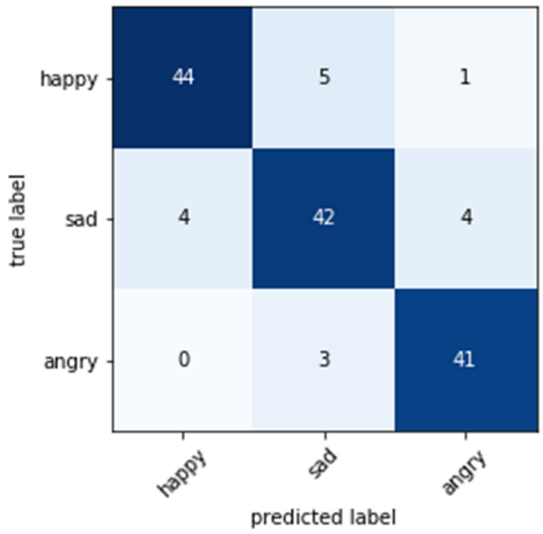


Figure 5.1 Performance Confusion Matrix

Now all class of emotion varying result are shown given below.

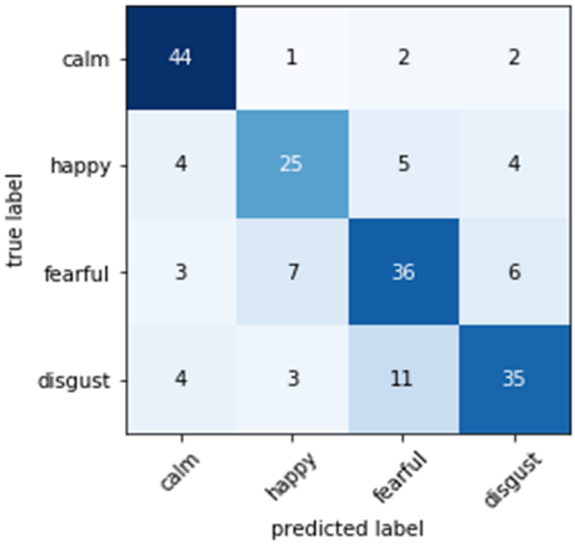
Observation of 3 emotions data (happy, sad, angry):

`( x_train.shape , x_test.shape )`	(432, 144)
`Features extracted`	180
`Accuracy`	88.19 %



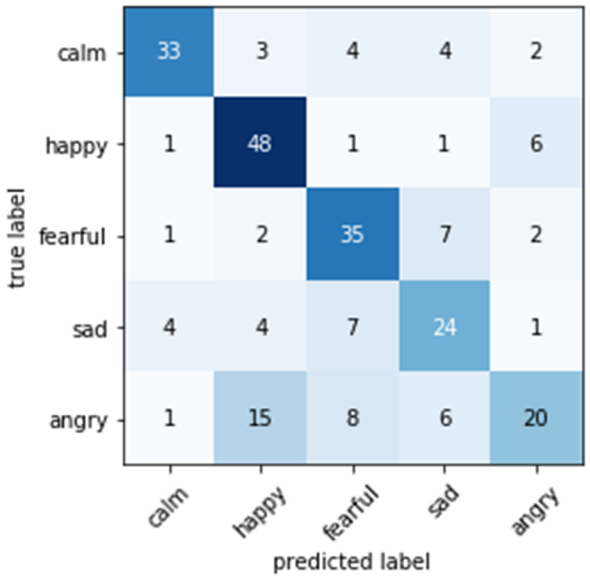
Observation of 4 emotions data (calm, happy, fearful and disgust):

`( x_train.shape , x_test.shape )`	(576, 192)
`Features extracted`	180
`Accuracy`	72.92 %



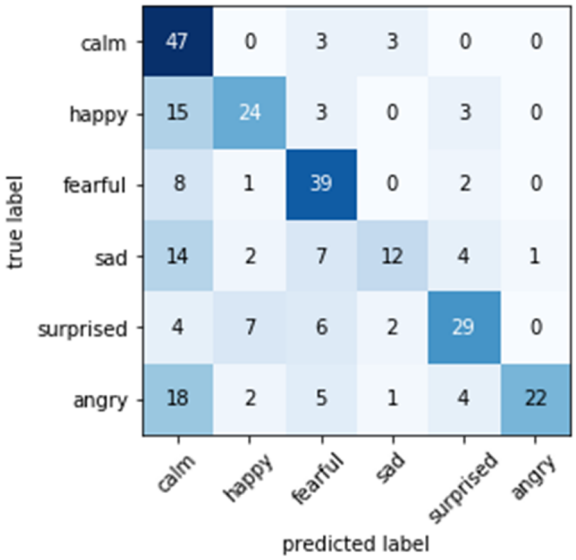
Observation of 5 emotions data (calm, happy, fearful, sad, and angry):

`( x_train.shape , x_test.shape )`	(720, 240)
`Features extracted`	180
`Accuracy`	66.67 %



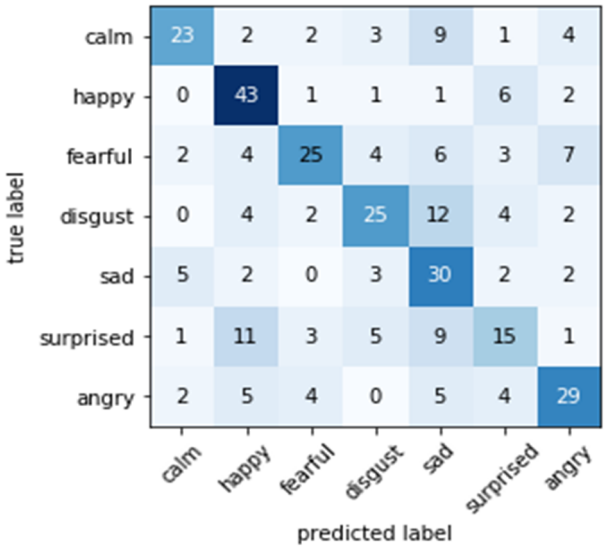
Observation of 6 emotions data (calm, happy, fearful, sad, surprised and angry):

`( x_train.shape , x_test.shape )`	(864, 288)
`Features extracted`	180
`Accuracy`	60.07 %



Observation of 7 emotions data (calm, happy, fearful, disgust, sad, surprised, and angry):

`( x_train.shape , x_test.shape )`	(1008, 336)
`Features extracted`	180
`Accuracy`	56.55 %



Observed 8 emotions data (calm, happy, fearful, disgust, sad, neutral, surprised and angry):

`( x_train.shape , x_test.shape )`	(1080, 360)
`Features extracted`	180
`Accuracy`	53.89 %

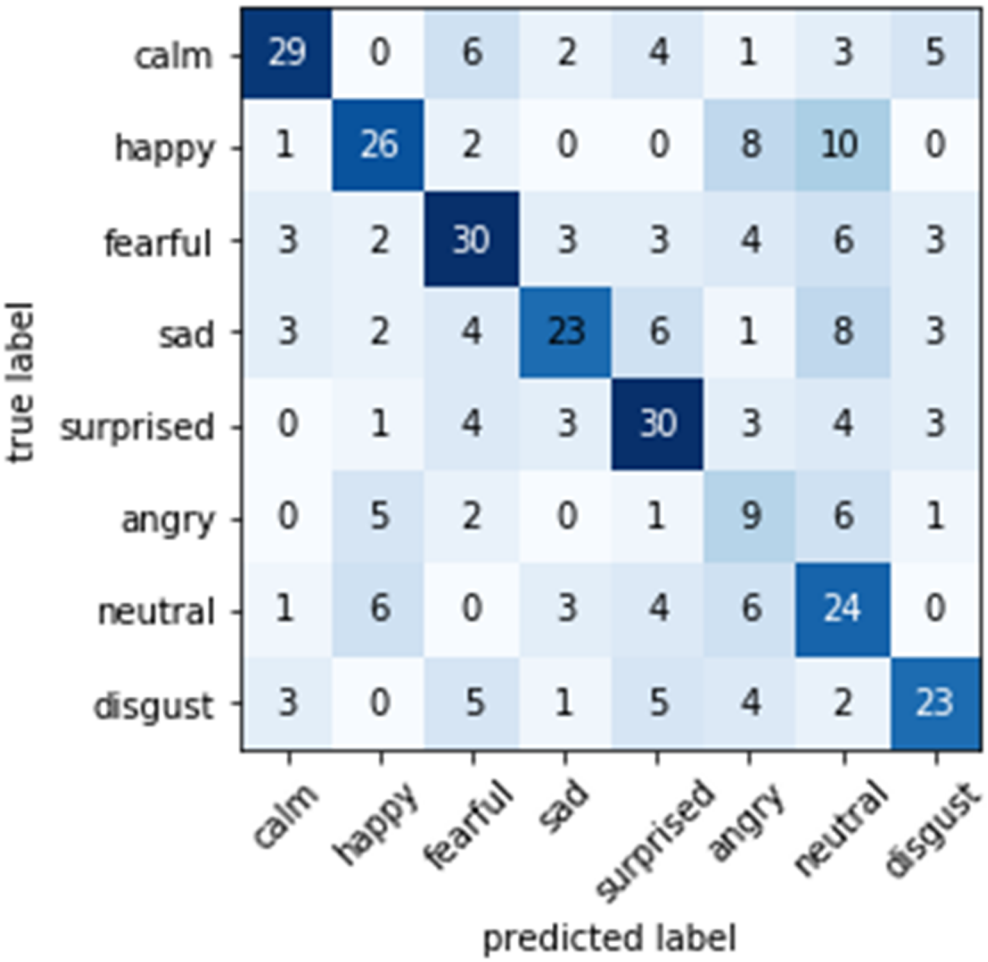


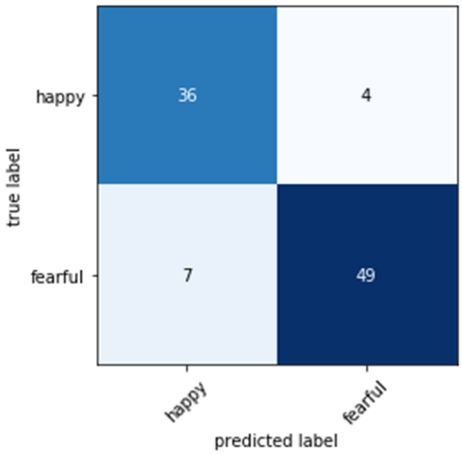
Figure 5.2 Performance Confusion Matrix For (RAVDESS)

5.1.2 Calculation for Bangla dataset

Now showing all class of emotion varying result are shown given below

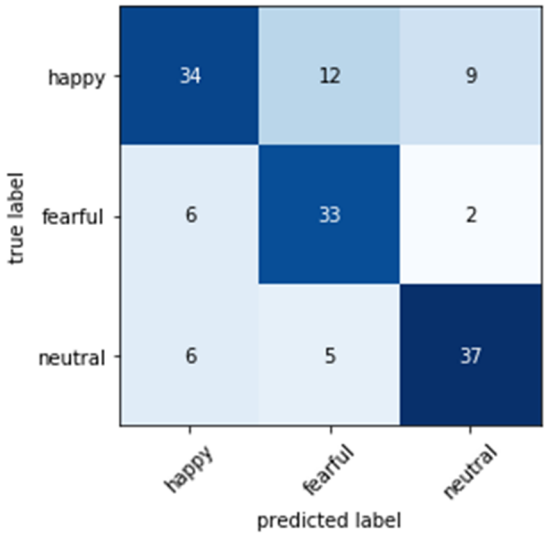
Observation of 2 emotions data (happy, fearful):

`( x_train.shape , x_test.shape )`	(288, 96)
`Features extracted`	180
`Accuracy`	88.54 %



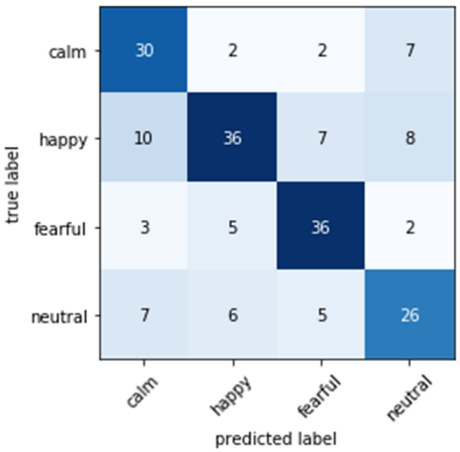
Observation of 3 emotions data (happy, fearful, neutral):

`( x_train.shape , x_test.shape )`	(432, 144)
`Features extracted`	180
`Accuracy`	72.22 %



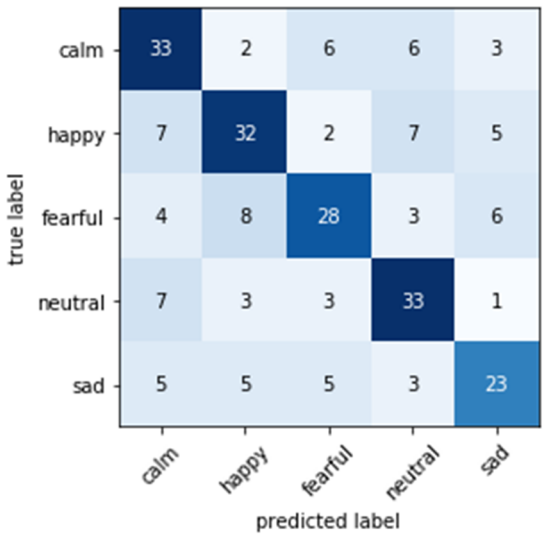
Observation of 4 emotions data (calm, happy, fearful, and neutral):

`( x_train.shape , x_test.shape )`	(576, 192)
`Features extracted`	180
`Accuracy`	66.67 %



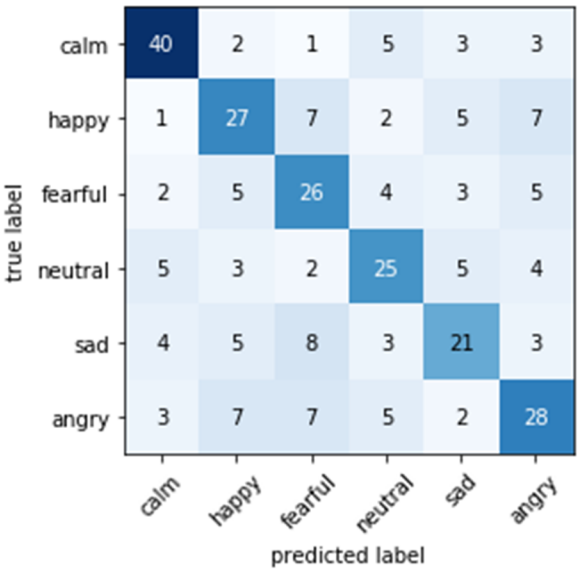
Observation of 5 emotions data (calm, happy, fearful, neutral and sad):

`( x_train.shape , x_test.shape )`	(720, 240)
`Features extracted`	180
`Accuracy`	62.08 %



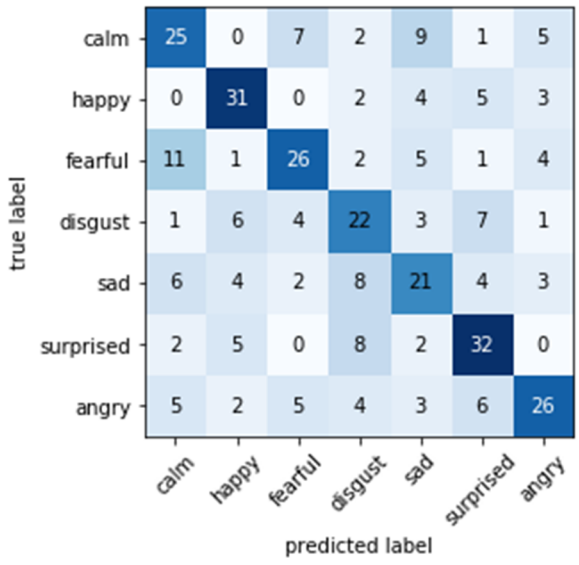
Observation of 6 emotions data (Calm, happy, fearful, neutral, sad, and angry):

`( x_train.shape , x_test.shape )`	(864, 288)
`Features extracted`	180
`Accuracy`	57.99 %



Observation of 7 emotions data (calm, happy, fearful, disgust, sad, surprised, and angry):

`( x_train.shape , x_test.shape )`	(1008, 336)
`Features extracted`	180
`Accuracy`	54.46 %



Observation of 8 emotions data (calm, happy, fearful, sad, surprised, angry, neutral and disgust):

`( x_train.shape , x_test.shape )`	(1152, 384)
`Features extracted`	180
`Accuracy`	45.83 %

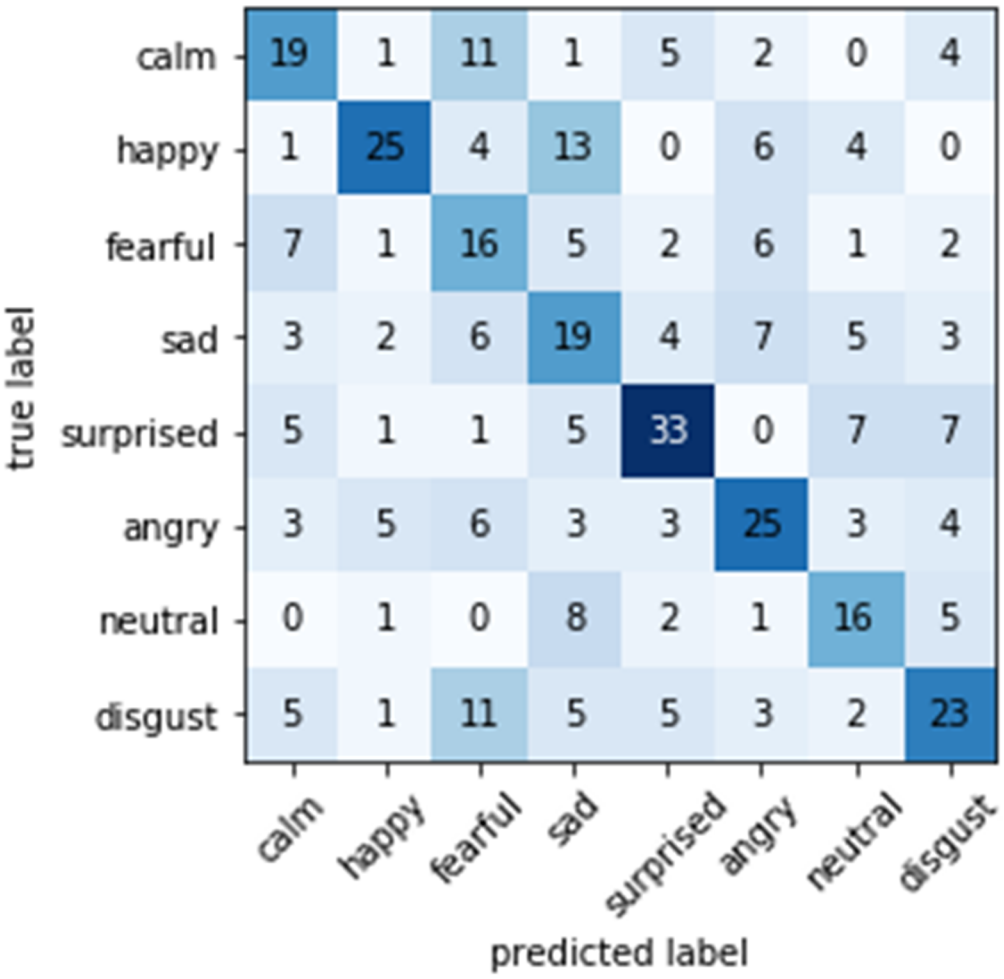
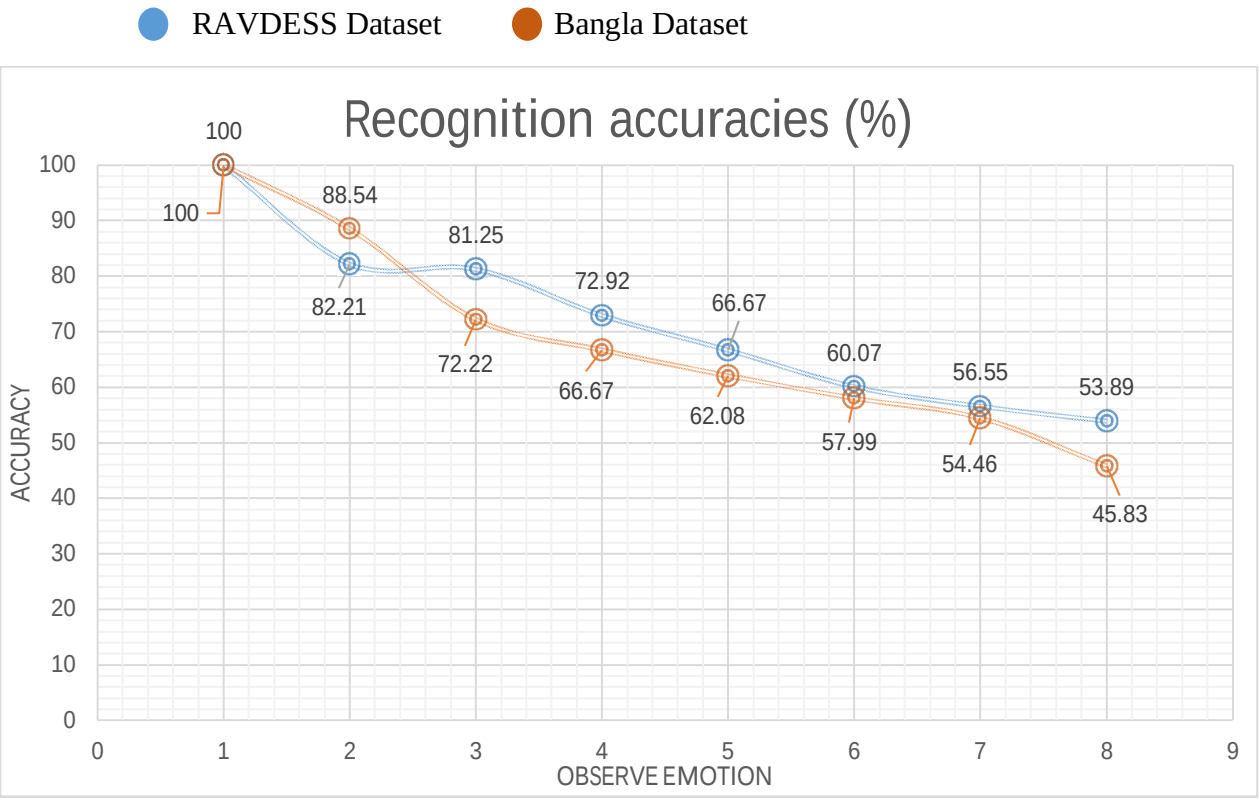


Figure 5.3 Performance Confusion Matrix For (Bangla)

Table 5.2 Recognition accuracies (%) with different observed emotions of the training data used to calculate the confusion matrices of the dataset.

Observed Emotion	Accuracy (RAVDESS)	Accuracy (Bangla)
8	53.89	45.83
7	56.55	54.46
6	60.07	57.99
5	66.67	62.08
4	72.92	66.67
3	81.25	72.22
2	82.21	88.54
1	100	100



5.1.3 Key results

An accuracy rate of about 53.89% for RAVDESS and 45.83% for Bangla is achieved from the data model for predicting the emotions. It is evident from the above figure that taken 8 emotions is the evaluated accuracy rate achieved during validation of data on all sides of emotion.

5.2 Discussion

In this paper, we described the construction and validation of the RAVDESS and Bangla, a set of emotional expressions that are dynamic and multimodal. The RAVDESS has several important features that lend itself for use by scientists, engineers, and clinicians: it is large in number, it contains visual and auditory depictions of spoken and sung expressions, and it consists of professional actors from North America, on other side Bangla was taken from the local area of Bangladesh. It consists of non-professional actors. Thus it has a variety of emotional expressions at two levels of emotional intensity, and it is made freely available under a Creative Commons non-commercial license. Validation of the RAVDESS was performed with 247 raters from North America whereas Bangla is new to be introduced here. Validity referred to the accuracy with which participants correctly identified the actors' intended emotions. We examined proportion correct scores, as is commonly reported in the literature.

5.2.1 Interpretation and Analysis of results

For RAVDESS dataset

Emotions are subjective, people would interpret it differently. It is hard to define the notion of emotions. Collecting data is complex. There are lots of audio data that can be achieved from films or news. However, both of them are biased since news reporting has to be neutral and actors' emotions are imitated. It is hard to look for neutral audio recordings without any bias.

After we selected 1 actor and 1 actress's dataset and listened to all of them. We found out male and female are expressing their emotions in a different way. Here are some findings:

- Male's *Angry* is simply increased in volume.
- Male's *Happy* and *Sad* significant features were laughing and crying tone.
- Female's *Happy, Angry* and *Sad* are increased in volume.

According to various intra observe emotions, their accuracies are different. If we take four emotions (calm, happy, fearful, disgust) the accuracy is 72.92 %. And again we take 4 emotions (calm, happy, sad, angry) the accuracy is 67.19 %.

In all emotions, we see some observed emotions' accuracy is significantly high. Like

Disgust, Clam, happy, fearful, surprised.

Sad, angry- those kinds of emotional accuracies are low performance. And Neutral emotions are quit average performance.

Some of the reasons for less accuracy rate are, Transfer Learning is used to train the model, there could've been fewer spectrograms used for training, which leads to less accuracy. There are also less data set used for the training process which also leads to the case.

For Bangla dataset

The recording-process of Bangla audio speech is highly difficult. We were able to overcome these difficulties by taking recordings from 24 actors directly by employing immense labor. All male and female actors express their emotions in a very different way. But in Bangla, though all the 8 emotions are co-related in presentation, they can also be distinguished or differentiated subsequently. The accuracies obtained due to the Bangla dataset is quite praiseworthy. The first footprints on the sands of time regarding the Bangla audio speech dataset claim great success. Thus, we can train and test the model using the Bangla dataset and hope to have good accuracies.

After we selected 1 actor and 1 actress's dataset and listened to all of them. We found out male and female are expressing their emotions in a different way. Here are some findings:

- Male's *Angry* is simply increased in volume.
- Male's *Happy* and *Sad* significant features were laughing and crying tone.
- Females' *Happy, Angry* and *Surprised* are increased in volume.
- Some Class of emotion we get proper and a few samples of females can't get proper as well as male.
- Neutral emotion is quite better than the RAVDESS data set.

According to various intra observe emotions, their accuracies are different. If we take four emotions (calm, happy, fearful, neutral) the accuracy is 66.67 % and again we take 4 emotions (calm, happy, fearful, sad) the accuracy is 61.99 %. Here the quality of neutral emotion is more prominent than sad emotion.

Overall we see that all emotion classes provide reasonable accuracy in both cases.

5.2.2 Answer of our research question

Annotating an audio collection and recording is challenging. We should label a single word, sentence or a whole conversation. Our all desired emotions are defined to recognize them from their signals.

5.2.3 Justification of our approach

Recognizing human emotion has always been a fascinating task for data scientists. Lately, we are working on an experimental Speech Emotion Recognition (SER) project to explore its potential. In this paper (Python mini project), we learned to recognize emotions from speech. We used an MLP Classifier for this and made use of the sound file library to read the sound file, and the librosa library to extract features from it. As you'll see, the model delivered an accuracy of 53.89 % (RAVDESS) and 45.83% (Bangla) of 8 emotions that's good enough for us yet.

5.2.4 Critical evaluation of our study

In the RADVESS dataset, each actor has to perform 8 emotions by saying and singing two sentences and two times for each. As a result, each actor would induce 4 samples for each emotion except neutral, disgust and surprised since there is no singing data for these emotions. Each audio wave is around 3 second, the first and last second are most likely silenced.

We have presented a novel system for RAVDESS dataset for tone expression feature extraction and Speech for recognition. We have illustrated the performance of the proposed method applied on sequential emotion expression datasets for eight emotion expressions. The result shows that it can improve the emotion expression feature extraction task. Furthermore, MLP dealing with MFCC processed sequential data can provide the higher recognition rate of 53.89% in comparison to the results from the conventional feature extracting methods with MFCC. We conclude that the holistic features can be used for the MLP method rather than usage of action units for any to recognize the partial activities.

In our Bangla Dataset we explored how well people and computers recognize emotions in speech. Several conclusions can be drawn from the above results. For Bangla dataset, each actor has to perform 8 emotions by saying and singing five sentences and eight times for each. There is no singing data for these emotions. Each audio wave is around 3 seconds after making chunk of 24 seconds, the first and last second are most likely silenced. Furthermore, MLP dealing with MFCC processed sequential data can provide the higher recognition rate of 45.83% in comparison to the results from the conventional feature extracting methods with MFCC.

Other physiological signals may be helpful in emotion recognition. Finally, some studies (Parameswaran, 2002; Winston, O'Doherty & Dolan, 2003) indicated that multiple emotions may be co-occurring in the same person. For example, when a person is angry about one thing and sad about another, his/her face shows one emotion and the voice shows another. However, this study mainly focuses on identifying single emotion during learning processes, but the phenomenon of multiple emotions is valuable for further investigation.

CHAPTER

6 CONCLUSION AND FUTURE WORK

6.1 Conclusions

In this current study, we presented a speech emotion recognition (SER) system using machine learning algorithms, MLP to classify eight emotions. Thus, three types of features (MFCC, MEL, and Chroma) were extracted from two different acted databases (RAVDESS and Bangla databases), and a combination of these features was presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from these features. SER reported the best recognition rate of 53.89% on the RAVDESS database using MLP classifier without speaker normalization (SN) and with feature selection (FS). For the Bangla database, the classifiers achieve an accuracy of 45.83% when a speaker normalization (SN) and a feature selection (FS) are applied to the features. From this result, we can see that MLP often performs better with more data and it suffers from the problem of very long training times. Therefore, we proposed that the MLP models have good potential for practical usage for limited data in comparison with MLP Classifier.

In our research, we treated all stages of emotion analysis, from data collection to classification using short-term observations, and evaluated several fusion methods as well as a hybrid decision scheme. We also compared the results from multimodal classification with the unimodal results. As in our earlier work (Kim et al. 2005) where we relied on longer observation phases and a different set of features, the best results were obtained by feature level fusion method in combination with the feature selection stage. In this case, not only user-dependent, but also user-independent emotion classification could be improved compared to the unimodal methods. We did not achieve the same high gains that were achieved for audio data which seems to indicate that speech and physiological data contain less complementary information. Furthermore, in a natural setting like ours, we cannot exclude that the subjects are inconsistent in their emotional expression. Inconsistencies are less likely to occur in scenarios where actors are asked to deliberately express emotions via speech and mimics.

This study integrates sensor, signal processing, wireless communication, system-on-chip and machine-learning technologies to construct an embedded human emotion recognition system that supports communication tailored to emotions in web-based learning environments. The forecasting accuracy rate of the proposed human emotion recognition system evaluated by cross-validation is 61.98% based on the proposed emotion feature extraction scheme. The forecasting accuracy rate is sufficient to support teachers in immediately understanding the emotions of individual learners during learning processes. Additionally, this study also applied the proposed embedded human emotion recognition system to support teachers in reducing the

anxiety of English learners while speaking in a web-based one-to-one synchronous learning environment. Experimental results demonstrate that the embedded human emotion recognition system provides benefits such as reducing language-learning anxiety and increasing the effectiveness of English conversation training.

In the past decade, a lot of research has gone into Automatic Speech Emotion Recognition (SER). The primary objective of SER is to improve the man-machine interface. It can also be used to monitor the psychophysiological state of a person in lie detectors. In recent times, speech emotion recognition also finds its applications in medicine and forensics. In this paper, 8 emotions are recognized using MFCC, Chroma and MEL features. The majority of the speech features used in this work are in time domain. Multi-Layer Perceptron (MLP) classifier has been used for classifying the emotions. RAVDESS emotional database is chosen for the task. A good recognition rate of 53.89% was obtained. The paper that was considered as the reference for our work recognized 4 emotions and obtained a recognition rate of 67.19%.

Furthermore, some important problems are pointed out, such as the use of posterior probabilities when fusing information with a high disparity accuracy. Most of the existing classifiers used in the literature are generalized methods based on statistics or estimating linear regression of given data. Such classifiers may not be able to capture emotion-specific features and to apply self-adapting decision rules that consider contextual information, for instance.

The Bangla emotional database was introduced as additional work in order to widen our research horizon in accordance with the RAVDESS database. Almost similar output achievements from both the databases ensure the correctness of the classifiers of our models and thereby keeps the future hope that database from any linguistic accent can be recognized and be implemented for the betterment of the total mankind.

6.2 Future work

We aim to use other feature selection methods because the quality of the feature selection affects the emotion recognition rate: a good emotion feature selection method can select features reflecting emotion state quickly. The overall aim of our work is to develop a system that will be used in a pedagogical interaction in classrooms, in order to help the teacher to orchestrate his class. For achieving this goal, we aim to test the system proposed in this work. Although the proposed embedded human emotion recognition system is helpful in conveying individual learner emotions to teachers who can then reduce language learning anxiety, several issues warrant further investigation. First, developing non-invasive physiological signal detection technologies, such as wearable physiological clothes and optical sensors, to support context-aware learning is necessary because using a pulse sensor worn on a finger to determine the emotions of students during learning processes is inconvenient. Moreover, In [43] author claimed that the human pulse shape varied with age. Therefore, investigating how human emotions are affected by variability such as gender, personality and health statuses has been considered as our future work. Additionally, the physiological signal in this study is the pulse; however, other physiological signals associated with human emotion, such as blood pressure, ECG and EEG, should be considered in future research. More important issue in future work would be how to complementarily combine the multiple modalities, since it is obvious that combining modalities by equally weighting them does not always guarantee improving recognition accuracy. Toward the human-like analysis and finer resolution of recognizable emotion classes, an essential step would be therefore to find innate priority among the modalities to be preferred for each emotional state. A considerable scheme might be to decompose an emotion recognition problem into several refining processes using additional modalities, for example, arousal recognition through physiological channels, valence recognition by using audiovisual channels, and then resolving subtle uncertainty between adjacent emotion classes, or predicting even the “stance” in 3D emotion model, by cumulative analysis of user’s context information. Therefore, the design of an emotion-specific classification scheme is one of the most important issues for the future, and this issue becomes even more critical when classifying combined multimodal observations. To overcome these problems, we need to develop a multilayer fusion scheme with parametric refinement stages in each decision layer. Besides, we plan to carry out a larger research on low level feature extraction, alternative acoustic modeling techniques, and multi-modal emotion recognition.

REFERENCES

[1] K. Soltani and R. N. Ainon, "Speech emotion detection based on neural networks," in *2007 9th international symposium on signal processing and its applications*, 2007, pp. 1–3.

[2] M. Srivastava and A. Agarwal, "Classification of emotions from speech using implicit features," in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, 2014, pp. 1–6.

[3] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in *2014 International Conference on Advances in Electronics Computers and Communications*, 2014, pp. 1–4.

[4] J. Pohjalainen and P. Alku, "Multi-scale modulation filtering in automatic detection of emotions in telephone speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 980–984.

[5] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2014, pp. 3724–3728.

[6] M. Gilke, P. Kachare, R. Kothalikar, V. P. Rodrigues, and M. Pednekar, "MFCC-based vocal emotion recognition using ANN," in *International Conference on Electronics Engineering and Informatics (ICEEI 2012) IPCSIT*, 2012, vol. 49.

[7] G. E. Peterson, "Basic physical systems for communication between two individuals," *J. Speech Hear. Disord.*, vol. 18, no. 2, pp. 116–120, 1953.

[8] "A taxonomy of applications that utilize emotional." .

[9] "R. Picard. Affective computing. MIT Press, 1997 - Google Scholar." [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=R.+Picard.+Affective+computing.+MIT+Press%2C+1997&btnG=. [Accessed: 20-Jan-2020].

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

[11] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, p. 614, 1996.

[12] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *Int. J. Speech Technol.*, vol. 6, no. 3, pp. 311–320, 2003.

[13] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, 2015.

[14] 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, 2012.

[15] 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, 2003.

[16] 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, 2011.

[17] 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, 2009.

[18] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, 2010.

[19] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, 2005.
[20] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” *Signal Process.*, vol. 90, no. 5, pp. 1415–1423, 2010.
[21] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, 2011.
[22] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1162–1171, 2011.
[23] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *interspeech*, vol. 53, no. 9–10, pp. 320–323, 2009.
[24] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
[25] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, “Segment-based emotion recognition from continuous Mandarin Chinese speech,” *Comput. Hum. Behav.*, vol. 27, no. 5, pp. 1545–1552, 2011.
[26] W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion recognition and affective computing on vocal social media,” *Inf. Manage.*, vol. 52, no. 7, pp. 777–788, 2015.
[27] M. M. El Ayadi, M. S. KaMel, and F. Karray, “Speech emotion recognition using Gaussian mixture vector autoregressive models,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, 2007, vol. 4, pp. IV–957.
[28] J. P. Arias, C. Busso, and N. B. Yoma, “Shape-based modeling of the fundamental frequency contour for emotion detection in speech,” *Comput. Speech Lang.*, vol. 28, no. 1, pp. 278–294, 2014.
[29] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, 2007.
[30] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.
[31] S. Davis and P. MerMelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.
[32] “Chroma feature", En.wikipedia.org.” .
[33] J. C. Brown, “Calculation of a constant Q spectral transform,” *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.
[34] R. H. M. Bosanquet, *An Elementary Treatise on Musical Intervals and Temperament: With an Account of an Enharmonic Harmonium Exhibited in the Loan Collection of Scientific Instruments, South Kensington, 1876, Also of an Enharmonic Organ Exhibited to the Musical Association of London, May, 1875*. London: Macmillan, 1876.
[35] M. A. Bartsch and G. H. Wakefield, “To catch a chorus: Using Chroma-based representations for audio thumbnailing,” in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, 2001, pp. 15–18.

[36] J. Eliashberg, V. R. Rao, and S. Shanker, “A Market Demand Model for a Preannounced Product,” Working Paper, 1996.
[37] W. Sibanda and P. Pretorius, “Novel application of Multi-Layer Perceptrons (MLP) neural networks to model HIV in South Africa using Seroprevalence data from antenatal clinics,” *Int. J. Comput. Appl.*, vol. 35, no. 5, pp. 26–31, 2011.
[38] S. Parthasarathy and I. Tashev, “Convolutional neural network techniques for speech emotion recognition,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 121–125.
[39] S. Chen and Q. Jin, “Multi-modal dimensional emotion recognition using recurrent neural networks,” in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 49–56.
[40] “web link : https://en.wikipedia.org/wiki/Limited-memory_BFGS.” .
[41] “A Linearly-Convergent Stochastic L-BFGS Algorithm - Google Scholar.” [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+Linearly-Convergent+Stochastic+L-BFGS+Algorithm&btnG=. [Accessed: 13-Jan-2020].
[42] “RuMelhart: Learning internal representations by error... - Google Scholar.” [Online]. Available: https://scholar.google.com/scholar_lookup?title=Learning%20internal%20representations%20by%20error%20propagation&author=D.E..%20RuMelhart&author=G.E..%20Hinton&author=R.J..%20Williams&pages=318-362&publication_year=1986. [Accessed: 13-Jan-2020].
[43] M. H. Sherebrin and R. Z. Sherebrin, “Frequency analysis of the peripheral pulse wave detected in the finger with a photoplethysmograph,” *IEEE Trans. Biomed. Eng.*, vol. 37, no. 3, pp. 313–317, 1990.