Speaker Emotion Recognition Based On Speech Features And Classification Techniques

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted By
HIMANSHU BANSAL
2K17/SWE/10

Under the supervision of

Mr. RAHUL
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Bawana Road,
Delhi-110042

JUNE, 2019

Department of Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Himanshu Bansal, 2K17/SWE/10, student of Master of Technology (Software Engineering), hereby declare that the Major Project-II Dissertation titled **"Speaker Emotion Recognition Based On Speech Features And Classification Techniques"** which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of requirement for the award of degree of Master Of Technology (Software Engineering) is original and not copied from any source without proper citation. This work has not been previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi HIMANSHU BANSAL

Date: 2K17/SWE/10

Department of Computer Science and Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Speaker Emotion Recognition Based On Speech Features And Classification Techniques"** which is submitted by Himanshu Bansal, (2K17/SWE/10) to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of requirement for the award of the degree of Master of Technology, is a record of project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi Mr. Rahul

Date: (Supervisor)

Assistant Professor

CSE Department

Delhi Technological University

(Formerly Delhi College of Engineering)

Shahbad, Daulatpur, Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without accomplishing the people who made it all possible and whose constant guidance and encouragement secured us the success.

First of all, I would like to thank the Almighty, who has always guided me to follow the right path of the life. My greatest thanks are to my parents who bestowed the ability and strength in me to complete this work.

My thanks is addressed to my mentor **Mr. Rahul**, Department of Computer Science and Engineering who gave me this opportunity to work in a project under her supervision. It was her enigmatic supervision, unwavering support and expert guidance which has allowed me to complete this work in due time. I humbly take this opportunity to express my deepest gratitude to her.

Date: Himanshu Bansal M.Tech (SWE)-4thSem
2K17/SWE/10

ABSTRACT

The main principle behind the speaker recognition system is that of feature extraction and matching. Feature extraction deals with the extraction or reading of important characteristics or feature from a human speech signal. Those characteristics might be pitch, or frequency, which are unique to different persons. Speaker recognition methods can also be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc. All speaker recognition systems have to serve two phases. The first one is referred to the enrolment session or training phase while the second one is referred to as the operation session or testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. It consists of two main parts. The first part consists of processing each persons input voice sample to condense and summarize the characteristics of their vocal tracts. The second part involves pulling each person's data together into a single, easily manipulated matrix. The testing system mirrors the training system architecture. First the input signal is analyzed, and then it is compared to the data stored in the codebook. The difference is used to make recognition decision.

Keyword: Speaker recognition, Manipulated Matrix

LIST OF CONTENTS

Contents **Page No.**

Declaration Ii
Certificate Iii
Acknowledgement Iv
Abstract V
Contents vi-vii
List of Figures Viii
List of Tables Ix
List of Acronyms X
Research Publication

CHAPTER 1
INTRODUCTION 1
1.1 Speech emotion recognition system 2
1.2 Motivation 3
1.3 Challenges 4
1.4 Problem statement of speech emotion recognition 4
1.5 Software used 5-6
1.6 Types of emotion in speech signal 6-7
1.7 Acoustic feature used for speech emotion recognition 7-9
1.8 Objective of research work 9
1.9 Table on previous work on speech emotion recognition 9
1.10 Application of Speech Emotion Recognition 10
1.11 Thesis organisation 10

CHAPTER 2
Brief Literature Survey 11-19

CHAPTER 3
Contribution of Cognitive Science 20-25

CHAPTER 4
Methodology used for speech emotion recognition
4.1 Speech 26
4.2 Emotion Recognition 27
4.3 Speech emotion recognition 28-29
4.4 Speech Classification 30
4.5 Machine Learning Algorithm using for speech emotion recognition 31
4.6 Database for emotional speech 31
4.7 Feature selection 32
4.8 Feature Discretization 32-33
4.9 Feature normalization 33

4.10 Emotional training set

4.11 Speech emotion recognition using SVM algorithm

4.11.1 SVM algorithm 34-35

CHAPTER 5

Simulation results and Discussion
5.1 Emotion recognition based on speech analysis 36
5.2 Results of experiments 36
5.3 k-NN algorithm 37-39

5.4 SVM algorithm 40-41

5.5 Discussion of results 42-43

5.6 Emotion Classification 44

5.6.1 Anger emotion 45
5.6.2 Fear emotion 45
5.6.3 Disgust emotion 46
5.6.4 Happiness emotion 47
5.6.5 Sadness emotion 47
5.6.6 Boredom emotion
5.6.7Neutral emotion 48
CHAPTER 6
Conclusion and Future scope
6.1 Conclusion 49
6.2 Future scope 49-50

REFERENCE 51-53

LIST OF FIGURES

Sr. No. Title of Figure Page No.

1.1 Block diagram of speech emotion recognition system 2
4.1 Speech emotion recognition pipeline 27
4.2 Segmental and Suprasegmental features in a speech signal 28
4.3 Typical architecture for a combined classifier or a voting scheme 31
4.4 Classification schema for speech emotion recognition 35

5.1 Emotion classification error for various k for the k-NN classifier 37
For speaker dependent and speaker independent configurations.
5.2 Emotion classification error for various k for the k-NN classifier. 37
5.3 Emotion classification error for various k for the ANN classifier. 38
5.4 Emotion classification error for various k for the SVM classifier. 40
5.5 Comparison of speaker dependent and speaker independent 43
configurations for all three classifier tested.
5.6 Anger Emotions 44
5.7 Fear Emotions 45
5.8 Disgust Emotions 45
5.9 Happiness Emotion 46
5.10 Sadness Emotion 46
5.11 Boredom Emotion 46
5.12 Neutral Emotion 47

LIST OF TABLES

Sr. No. Title of Tables Page No.

1 Previous work of speech emotion recognition 8-9

2 Confusion matrix (%)for the SVM with Gaussian Radial 35
Basis function kernel (σ =1)

3 Confusion matrix (%)for linear SVM 35

4 Classification results (in percentages) for various number of 37
speaker samples (i) in the training set for the k-NN classifier.

5. Confusion matrix for the SI-I(upper) and SD-I (lower) configurations 38
for k-NN (in percentages) of correctly recognized emotions .

6 Confusion matrix for the SI-I(upper) and SD-I (lower) configurations for 39
ANN (in percentages) of correctly recognized emotions. The diagonals
show the percentages of correctly recognized emotions.

7. Confusion matrix for the SI-I(upper) and SD-I (lower) configurations for 40
SVM (in percentages) of correctly recognized emotions. The diagonals
shows the percentages of correctly recognized emotions.

8. Emotion classifications 46

LIST OF ACRONYMS

ABBREVIATOR ABBREVIATION

SVM Support vector machine
GMM Gaussian mixture Model
HMM Hidden Markov Model
MFCC Mel frequency cepstral coefficients
ASR Acoustic source localization
ANN Artificial neural network
LFPC Log frequency power coefficients
MLNN Multilayer neural network
RBF Radial basis function
HCI Human Computer Interaction
SNR Signal Noise Ratio
ELM Extreme Learning Machine
SER Speech Emotion Recognition
LIF Local invariant Features
SAE Sparse Auto Encoder
CNN Convolutional Neural Network
ZCR Zero Crossing Rate
DCT Discrete Cosine Transform
DTW Dynamic Time Wraping
MLP Multi Layer Perception

CHAPTER1

INTRODUCTION

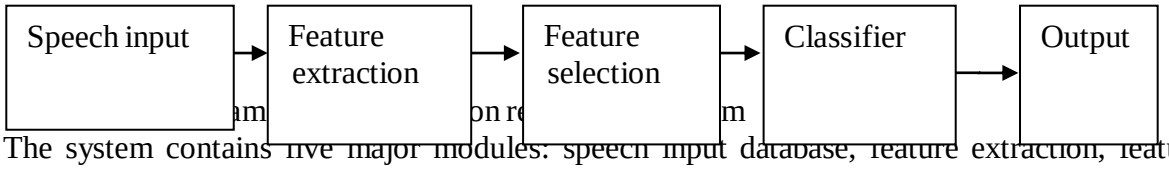
Interpersonal communication is an interaction which involves the exchange of reciprocal ideas and emotions. Gestures and sound are a way of conveying information in a human-to-human interaction. Speech, a special form of sound, is one of the fundamental ways of conveying information between people. Words are not sole component of speech. Acoustic properties of speech also carry important affective 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 the 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 clues 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 affective. 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 remove the emotional condition 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 evaluate 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 German, Polish and English databases are applied machines. Besides machines, to make comparison, German database is applied to human listeners who do not know any German 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.

Emotions play a crucial role in modulating how humans experience and interact with the outside world and have a huge effect on the human decision making process. They are an essential part of human social relations and take role in important life decisions. Therefore detection of emotions is crucial in high level interactions. Each emotion has unique properties that make us recognize them. Acoustic signal generated for the same utterance or sentence changes primarily due to biophysical changes triggered by emotions. This relation between acoustic cues and emotions made speech emotion recognition one of the trending topics of the affective computing domain. The main reason of a speech emotion recognition algorithm is to identify the emotional condition of a speaker from recorded speech signals.

1.1Speech emotion recognition system:

Speech emotion recognition is nothing but an application of the pattern recognition system in which patterns of derived speech features such as Pitch, Energy, MFCC are mapped using classifier like ANN, SVM, HMM etc.



am on re m

The system contains five major modules: speech input database, feature extraction, feature selection, classifier & recognized output as illustrated in figure – 1 above. Overall, the system is based on deep analysis of the generation mechanism of speech signal, extracting some of features which contain information about speaker's emotion & taking appropriate pattern recognition model to identify states of emotion. Typically, a set of emotion having 300 emotional states. Whenever, signal is passed to the feature extraction & selection process, the extracted speech features are selected in terms of emotion relevance. All over procedure revolves around the speech signal for extraction to the selection of speech features corresponding to emotions. Forward step is generation of database for training as well as testing

of extracted speech features. At the end, detection of emotions has been done using classifier with the usage of pattern recognition algorithm. The Speech emotion recognition is similar to the speaker recognition system but different types of approach to detect emotions make it secure & intelligent. The evaluation of the system is depending on naturalness of the input databa**se.**

1.2 Motivation

Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues. Speech emotion recognition refers to analysing vocal behaviour as a marker of affect, with focus on the nonverbal aspects of speech. Its basic assumption is that there is a set of impartially measurable parameters in voice that replicate the affective state a person is presently expressing. This assumption is supported by the fact that most affective states involve physiological reactions which in turn modify the process by which voice is produced. For example, anger often produces changes in respiration and increases muscle tension, influencing the vibration of the vocal folds and vocal tract shape and affecting the acoustic characteristics of the speech. So far, vocal emotion expression has received less attention than the facial equivalent, mirroring the relative emphasis by pioneers such as Charles Darwin. In the past, emotions were considered to be hard to measure and were consequently not studied by computer scientists. Although the field has recently received an increase in contributions, it remains a new area of study with a number of potential applications. These include emotional hearing aids for people with autism; detection of an angry caller at an automated call centre to transfer to a human; or presentation style adjustment of a computerised e-learning tutor if the student is bored. A new application of emotion detection proposed in this dissertation is speech tutoring. Especially in persuasive communication, special attention is required to what non-verbal clues the speaker conveys. Untrained speakers often come across as bland, lifeless and colourless. Precisely measuring and analysing the voice is a difficult task and has in the past been entirely subjective. By using a similar approach as for detecting emotions, this report shows that such judgements can be made objective.

1.3 Challenges

This section describes some of the expected challenges in implementing a real time speech emotion detector. Firstly, discovering which features are indicative of emotion classes is a difficult task. The key challenge, in emotion detection and in pattern recognition in general, is to maximise the between-class variability whilst minimising the with in class variability so that classes are well separated. However, features indicating different emotional states may be

overlapping, and there may be multiple ways of expressing the same emotional state. One strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others, creating a compact emotion code that can be used for classification. This avoids making difficult a priori assumptions about which features may be relevant. Secondly, previous studies indicate that several emotions can occur simultaneously [14]. For example, co-occurring emotions could include being happy at the same time as being tired, or feeling excited touched and surprised, when enquiry good news. This requires a classifier that can infer multiple temporally co-occurring emotions. Thirdly, real-time classification will require choosing and implementing efficient algorithms and data structures. Despite there existing some working systems, implementations are still seen as challenging and are generally expected to be imperfect and imprecise.

1.4 Problem statement of speech emotion recognition:

Emotion Recognition using speech started from affective computing by Picard's. It has big area of applications such as in Robotics, expert System, and psychology and for autistic people. Various classifications and clustering mechanism are there which work efficiently in the fields of speech processing. The problem statement can be more rectified and said as:

In digital world, emotion recognition is one of the recent topics. Many research work has been done into emotion recognition from speech but problem is to have high correctness rate. Detecting the motion of speech is not the easy as it seems to be. We tried to improve the accuracy of the system with noisy signals. This work has been done to categorize three emotion namely Happy, Fear and Sad using SVM algorithms. Noise levels are taken so that the emotion can be identified even though if the voice signal is highly noised or other emotion.

1.5 SOFTWARE USED

1.5.1 MATLAB:

MATLAB stands for matrix laboratory. It is a high recital language for technical computing. It involves computation, visualization and programming language environment. Furthermore, MATLAB is a contemporary programming language environment: it has classy data structures, contains built-in editing and debugging tools, and supports object - oriented programming. These influences make MATLAB an excellent tool for teaching and research.

1.5.2 WHY USE MATLAB

MATLAB may behave as a calculator or as a programming language.
MATLAB association nicely calculation and graphic plotting.
MATLAB is comparatively easy to learn.
MATLAB is not accumulated, errors are easy to fix.
MATLAB is improved to be relatively fast when performing matrix operations.

MATLAB does have approximately object oriented elements.

1.5.3 PROCEDURE OF SOFTWARE INSTALLATION

STEP 1: Open the setup.exe application file, if you can see the user control notice click Yes. first setup windows see image, wait for 1-2 minutes automatically open the install windows. We are just display one by one step found image see this image.

STEP 2: After open the math works installer window. Choose the option Install without using the internet, and click **Next.**

STEP 3: After clicked next see the license terms agreement of Matlab click Yes and Next.

STEP 4: Display the File installation key window, then enter the serial number or licence key and click Next.

STEP 5: Choose the installation type. Recommended select the **Typical** and click **Next.**

STEP 6: Folder Selection window set your installation path, if don't have space in your c: then choose the another drive of path to Browse to set new path. Recommended default path is automatically generated click **Next**.

STEP 7: Confirmation check the list of products and click **Next. S**tart the installing process.

STEP 8: Start the install process 0%…. 10%…43%… wait 10-15 minutes.. 93% complete the process wait 3-5 minutes more..93% complete the process wait 3-5 minutes more.

STEP 9: Finish...MATLAB installed successfully.

1.6Types of emotion in speech signal :

1.6.1 Anger
1.6.2 Fear
1.6.3 Sadness
1.6.4 Happiness
1.6.5 Digust
1.6.6 Boredom

1.6.1Anger

Anger requires high energy to be expressed. Definition meaning of the anger is simple extreme displeasure. In case of anger, aggression increases in which control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, joy and sadness (Ververidis & Koropoulos, 2006). The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions. Besides a faster speech rate is observed in angry speeches (Burkhardt & Sendlmeier, 2000)

1.6.2 Fear

In emotional dimension, fear has similar features to anger. High pitch level and raised intensity level are correlated with fear (Ververidis & Koropoulos, 2006). It is stated that fear has a wide pitch range. Highest speech rate is observed in fear speeches (Murray & Arnott, 1993). The pitch contour trend separates fear from joy. Although the pitch contour of fear resembles the sadness having an almost downwards slope ,emotion of joy have a rising slope(Ververidis & Koropoulos, 2006).

1.6.3. Sadness

In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo (Murray & Arnott, 1993). Speech rate of a sad person is lower than the neutral one (Ververidis & Koropoulos, 2006).

1.6.4. Happiness

Happiness exhibit a pattern with a high activation energy, and positive valence. Strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range and variance increases (Ververidis & Koropoulos, 2006). In (Murray & Arnott, 1993), it is stated that fundamental and formant frequencies increases in case of smile. Moreover, amplitude and duration also increase for some speakers.

1.6.5 Disgust

In (Ververidis & Koropoulos, 2006) a low intensity level, low mean pitch level and a slower speech rate is observed when disgust is compared with the neutral state. Disgust is stated the lowest observed speech rate and increased pause length (Murray & Arnott, 1993).

1.6.6. Boredom

Boredom is a negative emotion with negative valence and low activation level same as sad. A lowered mean pitch and a narrow pitch range with a slow speech rate are defined as the properties of a bored expression (Burkhardt & Sendlmeier, 2000).

1.7 Acoustic Features Widely Used In Emotion Recognition:

As mentioned previously, extraction of speech features is a very important process in speech emotion recognition. Speech features can be separated into numerous categories. In (E. Ayadi et 13 al., 2011), speech features are divided into 4 categories; spectral continuous, qualitative and TEO-based. Continuous features are pitch, energy and formants. Quantitative features are described as voice quality features which are harsh, tense and breathy voices. Most popular acoustic features used in emotion recognition process are outlined below.

1.7.1 Pitch features

Pitch is the fundamental frequency of the glottal excitation. Pitch depends on the tension of the vocal folds and subglottal air pressure. Pitch frequency is one of the widely used features in emotion from speech applications. Pitch frequency is also known as the fundamental frequency. The time elapsed between successive vocal fond openings determine the fundamental frequency (Ververidis & Koropoulos, 2006). From pitch features given features could be extracted which are min, max, mean, standard deviation, range at the turn level, slope (mean and max) in the voiced segments, regression coefficient and its mean square error and maximum cross-variation of F0 between two adjoining voiced segments (inter-segment) and with each voiced segment(intra-segment) (Vidrascu & Devillers, 2005).

1.7.2Teager energy operator

Produced number of harmonics due to the non-linear air flow in the vocal tract is another useful acoustic feature. In case of anger, the fast air flow causes nonlinear stress (Teager, 1990). In (E. Ayadi et al., 2011), it was stated that TEO-based features can be used to detect stress in speech.

1.7.3 Vocal tract features

Formants are a vocal tract feature. Each formant has its own bandwidth and center frequency (Ververidis & Koropoulos, 2006). Slackened speech can be distinguished from an articulated speech using formant features. Other widely used feature is the energy of a certain frequency which corresponds to the critical bands of the human ear (Ververidis & Koropoulos, 2006).

1.7.4 Spectral features

Mel-frequency cepstrum coefficients, linear predictive coding and log frequency power coefficients are the most popular spectral features. Mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC) are set as discriminating features in many studies. (D. Wu, Parsons, & Narayanan, 2010).

1.7.5 Duration features

Mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments are (D. Wu et al., 2010) duration features.

1.7.6 Energy features

Energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter quantile distance are the popular energy based features used in speech emotion recognition task (D. Wu et al., 2010).

1.8 Table on previous work on Speech emotion recognition:

Reference	Corpus	Feature set	Classification method	Emotion	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%
(Nogueiras, Moreno, Bonafonte, & 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%
(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%
(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%
(Schuller, Rigoll, & Lang, 2003)	Six Burmese and six Mandarin speakers generated 720 utterances	Logarithmic Frequency Power Coefficients	Hidden Markov Model	Anger, Disgust, Fear, Joy, Sadness and Surprise	77.8%
(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%

(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%
(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 & Anagnostopoulos, 2010)	Berlin emotional speech database speaker dependent	133 features from pitch, mel fre-quency cepstral coefficients, energy and formants.	probabilistic neural Net- work	Surprise, joy,anger, fear, disgust, sadness, neutra	94%

1.9 Objective of research work:

The objective of research work are:

1. To perform extensive literature review of speech emotion recognition**.**
2. To study and analysis the various feature, database, machine learning algorithm of speech emotion recognition.
3. The main objective of this thesis is to use Support Vector Machine (SVM) classifier to classify seven different emotions happiness, anger, sadness, boredom, disgust, neutral, fear.

1.10 Application of speech emotion recognition

Applications of emotion classification based on speech have already been used to facilitate interactions in our daily lives.

For example

1. In call centres apply emotion classification to prioritize impatient customers.

2. A warning system has been developed to detect if a driver exhibits anger or aggressive emotions.

3. For distance learning, to indentifying students emotion timely and taking appropriate action can improve the quality of teaching.

4. Emotion sensing has also been used in behaviour studies sound features have been widely explored in both the time domain and the frequency domain.

1.11 THESIS ORGANIZATION

The thesis is organised as follows:

- **Chapter 1** Discuss the introduction and overview of the entire thesis, introduction of speech emotion recognition system, type of emotion, various feature of it, database and machine algorithm of speech emotion recognition.
- **Chapter 2** This chapter explains the literature review of speech emotion algorithm using SVM algorithm.
- **Chapter 3** This chapter explains the contribution of cognitive science in speech emotion recognition.
- **Chapter 4** This chapter explains the methodology used for speech emotion recognition.
- **Chapter 5** This chapter explains the simulation results of speech emotion recognition.
- **Chapter 6** This chapter explains the conclusions of work done and future scope of the work.

CHAPTER 2

LITERATURE REVIEW

To carry out this work, extensive literature has been gone through. In this section, we present in detail different papers which have been taken into consideration. They provide a clear picture of the research which is going on emphasis has been put on literature related to speech emotion recognition using SVM.

Wootaek Lim et al., (2017): investigated that with rapid developments in the design of deep architecture models and learning algorithms, methods referred to as deep learning have come to be widely used in a variety of research areas such as pattern recognition, classification, and signal processing. Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In addition, Recurrent Neural Networks (RNNs) show significant success in various sequential data processing tasks. In this study, we explore the result of the Speech Emotion Recognition (SER) algorithm based on RNNs and CNNs trained using an emotional speech database. The main goal of our work is to suggest a SER method based on concatenated CNNs and RNNs without using any fixed hand-crafted features. By applying the proposed methods to an emotional speech database, the classification result was verified to have better accuracy than that achieved using conventional classification methods.

Pavitra Patel et al., (2017): investigated that speech has several characteristic features such as naturalness and efficient, which makes it as attractive interface medium. It is possible to express emotions and attitudes through speech. In human machine interface application emotion recognition from the speech signal has been current topic of research. Speech emotion recognition is an important issue which affects the human machine interaction. Automatic recognition of human emotion in speech purposes at identifying the basic emotional state of a speaker from the speech signal. The minimum error rate classifier (i.e. Bayesian optimal classifier) and Gaussian mixture models (GMMs) and are popular and active tools for speech emotion recognition. Normally, GMMs are used to perfect the class-conditional distributions of acoustic features and their issues are assessed by the expectation maximization (EM) algorithm formed on a training data set. Then, classification is performed to minimize the categorization error w.r.t the estimated class conditional distributions. We call this method the EM-GMM algorithm. In this paper, we establish a boosting algorithm for constantly and accurately estimating the class-conditional GMMs. The resulting algorithm is

named the Boosted-GMM algorithm. Our speech emotion recognition experiments display that the emotion recognition rates are efficiently and expressively increased by the Boosted-GMM algorithm as related to the EM-GMM algorithm.

Prajakta P. Dahake et al., (2016): investigated that in human computer interaction, speech emotion recognition is playing a pivotal part in the field of research. Human emotions consist of being angry, happy, sad, disgust, neutral. In this paper the features are extracted with hybrid of pitch, formants, zero crossing, MFCC and its statistical parameters. The pitch detection is done by cepstral algorithm after comparing it with autocorrelation and AMDF. The training and testing part of the SVM classifier is compared with different kernel function like linear, polynomial, quadratic and RBF. The polish database is used for the classification. The comparison between the different kernels is obtained for the corresponding feature vector.

Ritu D.Shah and Dr. Anil. C.Suthar (2016): In this paper methodology for emotion recognition from speech signal is presented. Some of sound features are removed from speech signal to analyze the features and behaviour of speech. The system is used to identify the basic emotions: Anger, Neutral, Happiness and Sadness. It can provide as a basis for advance designing an application for human like interaction with machines through and improving the efficiency of emotion and natural language processing. In this, formant, energy, Mel Frequency Cepstral Coefficients (MFCC) has been used for feature extraction from the speech signal. Support Vector Machine (SVM) are used for recognition of emotional states. English datasets are used for study of emotions with SVM Kernel functions. Using this analysis the machine is trained and designed for detecting emotions in real time speech. Finally results for various combination of the features and on various databases are compared and explained.

Kunxia Wang et al., (2015): Recently, studies have been performed on harmony characteristics for speech emotion recognition. It is found in our study that the first- and second-order differences of harmony features also play an important role in speech emotion recognition. Investigational results show that the offered Fourier parameter (FP) features are effective in recognising various emotional states in speech signals. They improve the recognition rates over the methods using Mel frequency cepstral coefficient (MFCC) features by 16.2, 6.8 and 16.6 points on the German database (EMODB), Chinese elderly emotion database (EESDB) and Chinese language database (CASIA). In particular, when combining FP with MFCC, the

recognition rates can be further improved on the aforementioned databases by 17.5, 10 and 10.5 points, respectively.

Rahul B.Lanjewar et al., (2015): investigated that the kinship between man and machines has become a new trend of technology such that machines now have to respond by considering the human emotional levels. The signal processing and machine learning technologies have boosted the machine intelligence that it gained the capability to understand human emotions. Incorporating the aspects of speech processing and pattern recognition algorithms an intelligent and emotions specific man-machine interaction can be achieved which can be harnessed to design a smart and secure automated home as well as commercial application. This paper emphasizes on implementation of speech emotion recognition system by utilizing the spectral components of Mel Frequency Cepstrum Coefficients (MFCC), wavelet features of speech and the pitch of vocal traces. The different machine learning algorithms used for the classification are Gaussian Mixture Model (GMM) and K- Nearest Neighbour (K-NN) models for the recognition of six emotional categories namely neutral, angry, fearful, sad, surprised and happy, from the standard speech database Berlin emotion database (BES) followed by the comparison of the two algorithms for performance analysis which is supported by the confusion matrix.

S. Lalitha et al., (2015): investigated that emotion recognition from speech helps us in improving the effectiveness of human-machine interaction. This paper presents a method to identify suitable features in DWT domain and improve good accuracy. In this work, 7 emotions (Berlin Database) are recognized using Support Vector Machine (SVM) classifier. Entropy of Teager Energy operated Discrete Wavelet Transform (DWT) coefficients, Linear Predictive Cepstral Coefficients (LPCC), Mel Energy Spectral Dynamic Coefficients (MEDC), Zero Crossing Rate (ZCR), shimmer, spectral roll off, spectral flux, spectral centroid, pitch, short time energy and Harmonic to Noise Ratio (HNR) are considered as features. The obtained average accuracy is 82.14 %. Earlier work done on emotion recognition using DWT coefficients yielded an accuracy of 63.63 % and 68.5% for 4 emotions on Berlin and Malayalam databases respectively. The proposed algorithm shows a significant increase in accuracy of about 15% to 20% for 7 emotions on Berlin database. Also, 100% efficiency has been achieved for four emotions with Simple Logistic classifier of WEKA 3.6 tool.

Qirong Mao et al., (2014): investigated that an essential way of human emotional behaviour understanding, speech emotion recognition (SER) has attracted a great deal of attention in human-centered signal processing. Accuracy in SER heavily depends on finding good affect-

related, discriminative features. In this paper, we propose to learn affect-salient features for SER using convolutional neural networks (CNN). The training of CNN involves two stages. In the first stage, unlabeled samples are used to learn local invariant features (LIF) using a variant of sparse auto-encoder (SAE) with reconstruction penalization. In the second step, LIF is used as the input to a feature extractor, salient discriminative feature analysis (SDFA), to learn affect-salient, discriminative features using a novel objective function that encourages feature saliency, orthogonality, and discrimination for SER. Our experimental results on standard datasets show that our advance leads to stable and strong recognition performance in complex scenes (e.g. and environment distortion and with speaker and language variation) and outperforms numerous well-established SER features.

Monica Feraru and Marius Zbancioc (2013): investigated an improved version of the classical KNN algorithm which associates to each parameter from the features vectors weights according to their performance in the classification process. We obtained the recognition percents of emotions around 65-67%, for the Romanian language, on the SROL database, which are comparable with the results for other languages, with non-professional voice database. This is the first study when the parameters are extracted on the sentence level. Until now, the analysis was made on the phoneme level.

Thapanee Seehapoch and Sartra Wongthanavasu (2013): studied that automatic recognition of emotional states from human speech is a current research topic with a wide range. In this paper an attempt has been made to recognize and classify the speech emotion from three language databases, namely, Berlin, Japan and Thai emotion databases. Speech features consisting of Fundamental Frequency (F0), Energy, Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) from short-time wavelet signals are comprehensively investigated. In this regard, Support Vector Machines (SVM) is utilized as the classification model. Empirical experimentation shows that the combined features of F0, Energy and MFCC provide the highest accuracy on all databases provided using the linear kernel. It gives 89.80%, 93.57% and 98.00% classification accuracy for Berlin, Japan and Thai emotions databases, respectively.

Yixiong Pan et al., (2012): studied that Speech Emotion Recognition is a recent research area in the field of Human Computer Interaction (HCI). Here we recognize three emotional states: happy, sad and neutral. The discovered features contain: linear predictive spectrum coding (LPCC), mel-energy spectrum dynamic coefficients (MEDC), energy, pitch mel-frequency spectrum coefficients (MFCC). Our self built Chinese emotional databases and a German

Corpus (Berlin Database of Emotional Speech) are used for training the Support Vector Machine (SVM) classifier. Finally results for different combination of the features and on different databases are compared and explained. The overall experimental results reveal that the feature combination of MFCC+MEDC+ Energy has the highest accuracy rate on both Chinese emotional database (91.3%) and Berlin emotional database (95.1%).

Mansour Sheikhan et al., (2012): investigated that recognition of emotional speech, the performance of automatic speech recognition (ASR) systems is degraded significantly. To improve the recognition rate of ASR systems, we can neutralize the Mel-frequency cepstral coefficients (MFCCs) of emotional speech as the most frequently used features in ASR. In this way, the neutralized MFCCs are used in a hidden Markov model (HMM)-based ASR system that has been trained by non emotional speech. In this paper, the frequency range that is most affected by emotion is determined, and the frequency warping is applied in the calculation process of MFCCs. This warping is performed in Mel filter bank module and/or discrete cosine transform (DCT) module in the process of MFCCs' calculation. To determine the warping factor, a combined structure using dynamic time warping (DTW) technique and multi-layer perception (MLP) neural network is used. Experimental results show that the recognition rate in anger and happiness emotional states is improved when the warping is performed in each of the mentioned modules when the MFCCs are calculated. Also, when the warping is performed in both the Mel filter bank and the DCT modules, the recognition rate of speech in anger and happiness emotional states is improved by 6.4 and 3.0%, respectively.

Simina Emerich and Eugen Lupu (2011): studied the recognition of the internal emotional state of a person plays an important role in several human-related fields. The present approach proposes the classification of 7 emotions (happiness, anger, fear, boredom, sadness, disgust and neutral) by using the speech signal. Different wavelet decomposition structures are used for feature vector extraction. The models were trained and tested with a Support Vectors Machine classifier. The extracted features were estimated over the whole utterance. This choice was made, due to the fact that in the literature global statistics is generally thought to be more suitable for this area of applications. In future, we intend to introduce, other voice parameterization in order to minimize the confusion between states. We also plan to investigate other wavelet techniques that can be used to overcome some of the deficiencies in the methods presented.

MoatazEl Ayadi and Mohamed S.Kamel (2011): investigated that increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. This paper is a

survey of speech emotion classification addressing three important aspects of the design of a speech emotion recognition system. The first one is that while high classification accuracies have been obtained for classification between high-arousal and low arousal emotions, N-way classification is still challenging. Moreover, the performance of current stress detectors still needs significant improvement. The average classification accuracy of speaker-independent speech emotion recognition systems is less than 80% in most of the proposed techniques. In some cases, such as, it is as low as 50%. For speaker-dependent classification, the recognition accuracy exceeded 90% only in few studies. Many classifiers have been tried for speech emotion recognition such as the HMM, the GMM, the ANN, and the SVM. However, it is hard to decide which classifier performs best for this task because different emotional corpora with different experimental setups were applied. Most of the current body of research focuses on studying their relations to the emotional content of the speech utterance and many speech features. New features have also been developed such as the TEO-based features. There are also attempts to employ different feature selection techniques in order to find the best features for this task. However, the conclusions obtained from different studies are not consistent. The main reason may be attributed to the fact that only one emotional speech database is investigated in each study. Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. In many databases, it is difficult even for human subjects to determine the emotion of some recorded utterances; e.g. the human recognition accuracy was 67% for DED, 80% for Berlin, and 65% in. There are some other problems for some databases such as the low quality of the recorded utterances, the small number of available utterances, and the unavailability of phonetic transcriptions. Therefore, it is likely that some of the conclusions established in some studies cannot be generalized to other databases. To address this problem, more cooperation across research institutes in developing benchmark emotional speech databases is necessary. In order to improve the performance of current speech emotion recognition systems, the following possible extensions are proposed. The first extension relies on the fact that speaker-dependent classification is generally easier than speaker-independent classification. At the same time, there exist speaker identification techniques with high recognition performance such as the GMM-based text-independent speaker identification system proposed by Reynolds. Thus, a speaker-independent emotion recognition system may be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system.

Iulia Lefter et al., (2010): studied possibilities for enhancing the generality, portability and robustness of emotion recognition systems by combining databases and by fusion of classifiers.

In a first experiment, we investigate the performance of an emotion detection system tested on a certain database given that it is trained on speech from either the same database, a different database or a mix of both. We observe that generally there is a drop in performance when the test database does not match the training material, but there are a few exceptions. Likewise, the performance dribs when a mixed quantity of acted databases is used for training and testing is carried out on real-life recordings. In a second experiment we study the effect of training several emotion detectors, and fusing these into a single detection system. We observe a drop in the Equal Error Rate (EER) from 19.0 % on average for 4 individual detectors to 4.2 % when fused using Focal.

Aditya Bihar Kandali et al., (2008): investigated a method based on Gaussian mixture model (GMM) classifier and Mel-frequency cepstral coefficients (MFCC) as features for emotion recognition from Assamese speeches. For training and testing of the method, data collection is carried out in format (Assam, India), which consisted of acted speeches of one short emotionally biased sentence repeated 5 times with different styles by 27 speakers (14 Male and 13 Female) for training and one long emotional speech by each speaker for testing. The experiments are performed for the cases of (i) text-independent but speaker-dependent and (ii) text-independent and speaker-independent. From observations of the results presented in the tables we find that the surprise emotion is the most difficult one to disambiguate from other emotions, since surprise may be expressed along with any other emotion such as angry-surprise, fear-surprise, happy-surprise, etc.

Hao Hu et al., (2007): investigated that speech emotion recognition is a challenging yet important speech technology. In this paper, the GMM super vector based SVM is applied to this field with spectral features. A GMM is skilled for each emotional utterance, and the equivalent GMM super vector is used as the input feature for SVM. Experimental results on an emotional speech database show that the GMM super vector based SVM outperforms standard GMM on speech emotion recognition. GMM KL divergence kernel was shown to produce better performance than other commonly used kernels in the suggested system. The results propose that the gender information should be considered in speech emotion recognition, and determine that the GMM super vector based SVM system significantly outperforms standard GMM system. For the commonly confused emotional states, other type of features, such as prosodic and voice value features can be merged with our proposed method to improve the emotion recognition performance in future work.

Suja P et al., (2005): Human face communicates important information about a person's emotional condition. In this paper an approach for facial expression recognition using wavelet transform for feature extraction and neural network classifier for five basic emotions is proposed. The strength of the algorithm is the reduction in feature size and use of less number of images for training the network, compared to existing similar approaches. Static images of the Cohn-Kanade Face Expression image database have been used for experimentation. The facial expression information that are mostly concentrated on mouth, eye, eyebrow and mouth regions are segmented from the face. Then the low-dimension features are acquired using 2-level Karhunen Loeve transform and Discrete Wavelet Transform. A neural network classifier is constructed to categorize the emotions. The neural network based classifier yielded an average accuracy of 96.4%. The expressions that are recognized are anger, surprise, disgust, happiness and sadness.

B. Schuller et al., (2004): They introduce a novel approach to the combination of acoustic features and language information for a most robust automatic recognition of a speaker's emotion. Seven emotional states are classified during the work. Firstly a model for the recognition of emotion by sound features is presented. The resulting features of the signal, spectral contours, energy and pitch are ranked by their quantitative contribution to the estimate of an emotion. Numerous different classification techniques containing linear classifiers, neural nets, support vector machines and Gaussian mixture models are compared by their performance within this task. Moreover an method to emotion recognition by the spoken content is introduced applying belief network based spotting for emotional key-phrases. Finally the two information sources are integrated in a soft decision fusion by using a neural net. The gain is evaluated and compared to other advances. Two emotional speech corpora used for evaluation and training are explained in detail and the results achieved applying the propagated novel advance to speaker emotion recognition are obtainable and discussed.

Tin Lay Nwe et al., (2003): investigated that emotion classification of speech signals, the popular features employed are statistics of fundamental frequency, duration of silence, energy contour and voice quality. Conversely, the performance of systems employing these features degrades significantly when more than two categories of emotion are to be classified. A text independent method of emotion classification of speech is purposed. The proposed method makes use of a discrete hidden Markov model (HMM) as the classifier and short time log frequency power coefficients (LFPC) to characterize the speech signals. The emotions are classified into six categories. The category labels used are, the typical emotions of Fear, Joy,

Sadness, Anger, Disgust, Surprise Fear, Joy, Sadness and Anger. A database consisting of 60 emotional utterances, each from twelve speakers is constructed and used to test and train the proposed system. Performance of the LFPC feature factors is compared with that of the linear prediction Cepstral coefficients (LPCC) and the mel-frequency Cepstral coefficients (MFCC) feature parameters usually used in speech recognition systems. Results show that the proposed system yields an average accuracy of 77% and the best accuracy of 95.7% in the classification of six emotions.

CHAPTER 3

CONTRIBUTION OF COGNITIVE SCIENCE

3.1 comprehensive definition for emotion

Providing one comprehensive definition for emotion is a hard task. My favourite definition for emotion is that, emotion is a way of representing ones circumstances, mood or relations to others. Emotion is a way of characterizing states of mind like joy, anger, love, hate derived from

natural instinctive state of mind (Gordon, 1990). Emotion has many forms and many different representations. Recent brain research shows that, there is a part, which plays an important role in expressed and embodied emotions (Gordon, 1990). The amygdale is a small structure in the limbic system, which resides on brains medial temporal lobe. It is revealed that when pictures of treating faces are shown to an individual, some neurons are activated from this cluster (Scheingold, 2010). Moreover, this region has a major role in recognizing emotional responses like facial expressions. In the literature, there are some different perspectives about theories about emotions. Evolutionary theory is in the middle in Darwinian perspective (Izard, 1984). It is stated that without evolutionary history, emotions cannot be understood. As an example, innate need for survival forces the emotions when we see a bear. Second perspective is Jamessian perspective. It claims that, a bodily change which caused by an outside stimuli is required in order to experience emotions (James, 1994). Our adrenaline production increases involuntarily when we see a bear, which results in fear as an emotion. In cognitivists aspect, physiological changes are experienced immediately and imperceptivity when a bear is seen. Socio-constructivist perspective states that emotions are products of culture (Parkinson, 1996). In this aspect, culture imposes us that bear is frightening and when

3.2 Interesting approach to emotions

you saw a bear, you should fear. An interesting approach to emotions is the 3.2dimensions of emotions. Emotional space is constituted by two dimensions namely; valence and arousal. Valence represents the level of pleasantness or unpleasantness. Arousal or with its other known name activation is the level of bodily energy. Neutral remains in the middle of two dimensional space. Happy and anger both have high activation level. However, happy has a positive valence value and anger has a negative valence value. Emotion space is important in embodied emotions since, emotional space has a representation in physical level at the human body. In high activation level and negative valence level emotions such as fear and anger, we feel shortness of breath and the trembling (Hatsimoysis, 2003). When one is fearful, the eye brow muscles grow tense, whereas if one feels joyful, the muscles in the check will form a smile position. In this section, arguments about embodied emotions will be discussed. Mind-body relation is induced to body-emotion in this case. Main arguments in embodied emotions are generally about Jamessian perspective. He claims that when we perceive some bodily changes because 5 of outside stimuli, emotions occur. Our feelings for the same changes are emotions. On the other hand, there are many opposing ideas. They stated that, emotions are more like judgments or thoughts, than perceptions (Hatsimoysis, 2003). On my viewpoint this view is true if we are

trained for that case. Such that, in a clash, soldiers emotions are related with their body status. When a body reaction occurs, unconsciously our brain forces us into an emotional state. It is hard to judge emotions in most of the cases. I defend the idea that emotions are based on perceptions of patterned changes in the body. Emotional dimension is a good proof for these patterned changes. Six basic emotions (anger, disgust, fear, joy, sadness, surprise) have unique body patterns. Anger, disgust, fear and surprise reside on the upper half of the emotional space. All requires a high activation level which means, our heart pulse rate increases. Since all basic emotions have a unique body pattern embodied emotion thesis is supported. Yet, there are some opposing ideas to this approach. In (Hatsimoysis, 2003), some opposing arguments are provided. I have briefly discussed the most interesting ones. Critics point out that some emotions do not involve bodily change at all. Does guilt or lonelyness has any relation with body movement (Hatsimoysis, 2003)? Long standing emotions correlation with the body changes over time. If someone is in love for a long time, does he always have a high heart rate? Valence level of the lover changes over time. Other critics come from another point. All perceived bodily changes are not emotions. Sport makes people happy. Exercise causes a change in the arousal level, which make them to perceive happier. In the literature, many emotional modes have been developed. The aim of these models is to generate an artificial agent, who has embodied emotions. These models try to generate emotions using environment and body. These artificial characters should be upset when they lost money just like real humans. In (Bartneck, 2002), success condition of an emotion model is defined such that, generated model should show the right emotion in the right time with right intensity. The OCC (Ortony, Clore, & Collins, 1990) is a complex yet well-known emotional model. The OCC model is able to distinguish 22 different emotional categories. In this model, each emotion has a weight. It is expected that weights should not change rapidly, since a regular persons emotional state do not change rapidly. The OCC model has five phases. The name of first phase is classification. In the classification phase, event or an action is evaluated by the character. Each object or an event has a relation with particular emotions. Next phase is the quantification phase. Every event or action requires different emotions to be activated with different intensity levels. In this phase, intensity levels of the activated emotions are determined. Following phases name is interaction. In the interaction phase, shift from one emotional state to other one is modelled. For a person who is surfing on the internet, speed of the internet is important. If the speed of the internet drops this people get angry.

3.3 OCC model

If you give this a food that he loves, his anger will fade slowly. Interaction phase models this smooth transition. Next phase is mapping. As mentioned OCC model could distinguish 22 different emotions. In the classification phase, activated emotions were selected. In this phase, first physical state is determined, and then appropriate emotions are selected from the activated emotions. Facial expression could not expose all emotions but some of them. Final phase is the expression phase. In the expression phase, emotions are exposed with all possible channels. The OCC model is designed to mimic human like emotions. Each OCC model has a character. Consistency is important to generate a stable character. If banana make a character happy, then in the next time should character required to be happy? If a banana is given after one another, then happiness level will drop to neutral level. On the other hand, if banana is given after a certain time again a high level of happiness should be generated. Distributed emotion is a framework which inherits all aspects of emotion and investigates a 6 full picture of emotion among people and environment (Hollan, Hutchins, & Kirsh, 2000). We live in a very dynamic society and emotion plays a crucial role to shape these relations. Instead of resizing the concept of emotion to individuals, it must be related with the interactions between individuals (Parkinson, 1996). Emotion is distributed among people. One group member can fire the emotions of all group members or one members sadness makes other members to feel sad. Mimicking of emotions is observed very frequently in intimate groups. By the time, close friends smile even look similar. Different from mimicking emotions, same event may cause different emotions on different people. One event may make a people feel happy on the other hand, made others sad. Rain is a good example. Slight rain make a farmer happy, on the other hand, made a basketball player unhappy. In (Glazer, 2003), distributed emotion is segmented into 3 titles. Emotion is distributed across members of social groups. Emotions are shared socially. One individuals emotion may distribute among the social group. Emotion is coordinated between external material or environmental and internal structures. People can load their emotions into physical structures and can load them back. Photographs are good examples for this case. When you look at your photographs, you memorize the past event, past emotions you have felt at this event. Emotion is distributed though time. Emotions are time-varying perceptions. When an event occurred, your emotions are generally sharper. People may respond some events overreactions at the beginning of the event. Years later, when you remember the same event, you could assess event as an adventure and you could feel happy. Other tendency is that generally grief, fear turns to rage and anger. Embodied mind thesis claims the argument that nature of our mind is largely determined by the form of the human body. Embodied emotion argument claims that when we perceive some bodily changes because of outside stimuli,

emotions occur. This argument generalizes the embodied emotion argument. We can resize the argument into the form that, emotions are only perceptions. In the big picture, most of it is true. It is fact that, we react unconsciously to the most of the events. This reaction gives birth to some emotions. Some scientific studies also showed that, there is a region in our brain, which controls embodied emotions. Embodied emotion framework is important in level of designing a human-like robot. Without emotions, it is hard to call a robot humanoid. Embodied emotions framework offers a basic model. Embodiment thesis enables the transformation of environmental inputs on body to emotions. We survive in a social society. These social relations are generally shaped by emotional state of ours. In a society, emotions flows though interaction. Emotions mix with each other and turn into other emotional states. Emotions are just like colors. Different from colors, emotions have only two primary units, valence and activation. In a happy society, valence flows one individual to others. Since emotions are distributed over time, valence level always changes. Embodied emotions are directly related with distributed emotions. Each social interaction has a pattern in our body. The concept of bed, have some effects on our body which results in some emotions. Just like this, social interaction generates some concepts and these concepts have body patterns in our body. In a cognitive system, embodied and distributed cognition complement each other. In a telephone call of a two person, generated emotion model is going to be in closed loop form. What you hear from the speaker has some effects on your body, which generate an emotional state. With this emotional state, you generate a speech. Speech emotion recognition algorithms take part in this task an important role. When you hear something you must convert speech signal into an emotion. In this closed loop cycle, firstly, detected emotion generates a body pattern in your body. Next time you speak, and other listener extracts the emotion.

3.4Distributed emotions

Distributed emotions take part in the transformation of extracted emotions into body patterns. Embodied emotions are results of body patterns. As seen, all concepts are related to 7 each other. Although the relationship between emotion and cognition is omitted for a long time, recent studies especially on magnetic resonance imaging (MRI), has provided important cues. Besides being dependent, both have many commonalities as being embodied and distributed. Yet, there are not any common view about emotions, there is a relative agreement upon cognition. Processes related with memory, attention, language and problem solving is known as cognition (Pessoa, 2008). When the distinction between emotion and cognition is projected into brain, affect is being related with unconscious processing and sub cortical activity. On the other hand,

cognition is related with the conscious processing and cortical involvement. Emotional visual expressions such as watching a war scene results in an increased activity in visual cortex (Pessoa, 2008). In addition to that, emotional content has an effect on attention. Detection of happy and anger faces are more rapid when compared with neutral ones (Eastwood, 2001). These results could be evaluated such that, amygdala has some effects on cognitive process. It arises two possibilities. First is that, excitation in amygdala has an enhancement effect on visual processing module. Second possibility is that, amygdala results in enhancement in the part of brain which are responsible for attention. Another research has provided interesting information. Emotional faces evoke responses in the amygdala although attention has major on another stimuli (Anderson, Christoff, Panitz, Rosa, & Gabrieli, 2003). This shows that, cognitive and emotional processes occurs independently. Besides attention and perception, in memory event, emotional content has an effect. Studies have shown that humans are better at remembering emotionally arousing information. In a study, subjects are exposed to two videos. First video composed of neutral film clips and the second one contains emotional content (Cahill et al., 1996). After 3 weeks of watching these clips, subjects were better at remembering emotional ones although both clips were taken from the same source. At recent years, human to human interaction is being replaced by human machine interaction. Such shift forces machines to behave like humans. In order to accomplish such a task, emotion recognition and generation of a counter emotional state are the crucial part of such devices. Such a system requires an interdisciplinary field which fuses artificial intelligence, psychology, philosophy. Design and development of human like machines are possible with

3.5 Design and development

cognitive systems. In the figure 2.1, a human machine dialog flowchart is provided. Given cognitive system is based on an interaction with speech. Speech is one the main source of communication. Speech both carries content and non-verbal elements such as emotions. Therefore, a human like machine should both have speech recognition algorithm and speech emotion recognition algorithm. In the given model, information from speech recognition and speech emotion recognition blocks is combined in artificial intelligence chat box. . In the speech synthesizer, generated sentence will be converted to emotional speech. In the scope of this thesis, speech emotion recognition task is accomplished. Recognition of emotion is crucial in an interaction. Different from many other speech emotion recognition algorithms, a human auditory model is used. Auditory model outputs 8 are transformed into feature sets. Auditory

model is a computation model of human ear. Yet, the process of emotion recognition in the brain has no model. In order to measure similarity of results, listening test is applied on humans. A German speech dataset is applied both to computers and humans. In the developed computational model is content free. Therefore, human evaluators are selected from the people who do not know any German. Output results of subjective and automatic speech emotion recognition tasks are compared with each other. Comparison of results have provided important information both on the source and receiver side.

CHAPTER 4

METHODOLOGY USED FOR SPEECH EMOTION RECOGNITION

4.1 SPEECH

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech recognition. It is evident that research into emotional speech recognition is limited to certain emotions, because the majority of emotional speech data collections encompass 5 or 6 emotions, although there are many more emotion categories in real life. Two

speech corpora were used in this investigation: the first Database of Emotional Speech and contains semantic units made up of sentences; the second, in English, is called Speech Under Simulated and Actual Stress and comprises semantic units made up of single words.

4.2 Emotion recognition

Emotion detection from the speech signal is a relatively new field of research, that has many potential applications in the current scenario. In human we can easily identify the emotion of a person by interacting with each other. But in human-computer interaction systems, emotion recognition is not as easy task. Emotion system could provide users with improved services by being versatile to their emotions. In virtual world, emotion recognition could help in simulating more realistic interaction. Emotion in speech is quite limited we are not one per cent sure that always we can find the emotion of a person. Currently, researchers are still debating regarding the recognition of emotion in speech for accuracy. There is also considerable apprehension as to the best algorithm for classifying emotion, and which emotions to class together or what can be their feature. In this project we use K-Means and Support Vector Machines (SVM) algorithm to classify opposing emotions. We separate the speech by speaker gender to investigate the relationship between gender and emotional content of speech. Here we extract a variety of temporal and spectral features from human speech. In this We use pitch relating statistics, Mel Frequency Cepstral Coefficients (MFCC) algorithms. The emotion recognition accuracy of these experiments is quite high accuracy as comparison to other algorithm. This process also allows us to de develop criteria to class emotions together. In this algorithm our system accuracy is more high.

4.3 Emotion classification

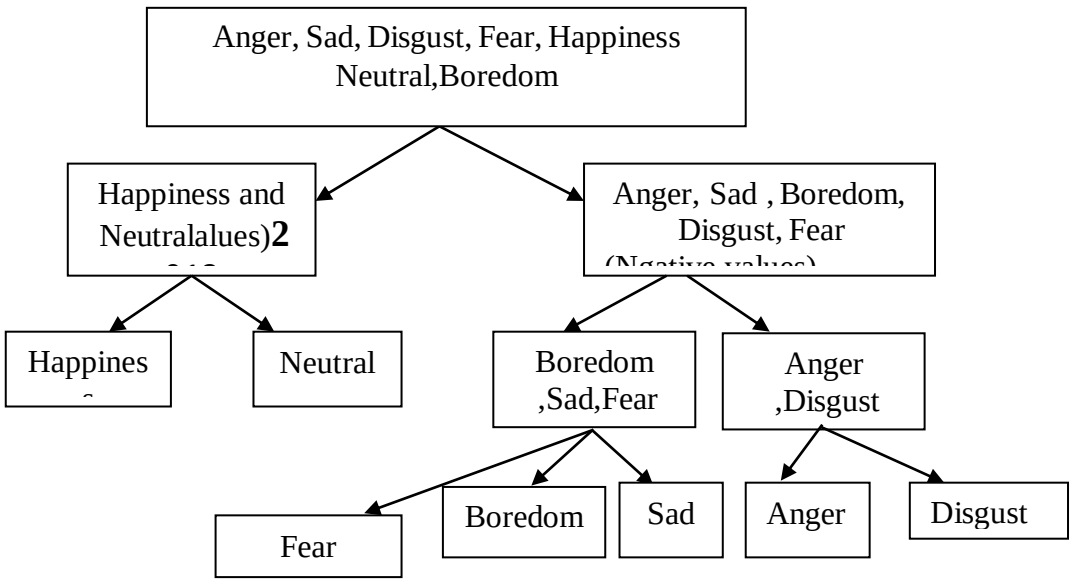


Fig4.1 Classification schema for speech emotion recognition

4.4 SPEECH EMOTION RECOGNITION

Speech emotion recognition is basically performed through pure sound processing without linguistic information. In terms of acoustics, speech processing techniques offer extremely valuable information derived mainly from prosodic and spectral features. Sometimes the process is assisted by Automatic Speech Recognition (ASR) systems, which contribute to classification using linguistic information. However, the use of ASR is limited due to fact that most of the experiments in the field have been assessed using databases of non-spontaneous and predefined speech and thus, there is no need for speech recognition. After sound processing and feature acquisition, it is quite common to follow a feature selection in search for the "golden set" of sound features. Finally, such a plethora of classification algorithms has been evaluated for speech emotion recognition, that attempting their comparison in this paper is, unfortunately, an impractical task. This is also due to the fact that there is a lack of uniformity in the way these methods are evaluated (different test sets, feature vectors and evaluation frameworks) and, therefore, it is inappropriate to make direct comparisons or explicitly declare which methods demonstrate the highest performance. In the next sections, a brief classification of papers that follow the basic processing pipeline (as highlighted in Fig. 1) are surveyed and categorized according to their major methodology for feature processing(with or without linguistic information), as well as their classification schema for emotion recognition.

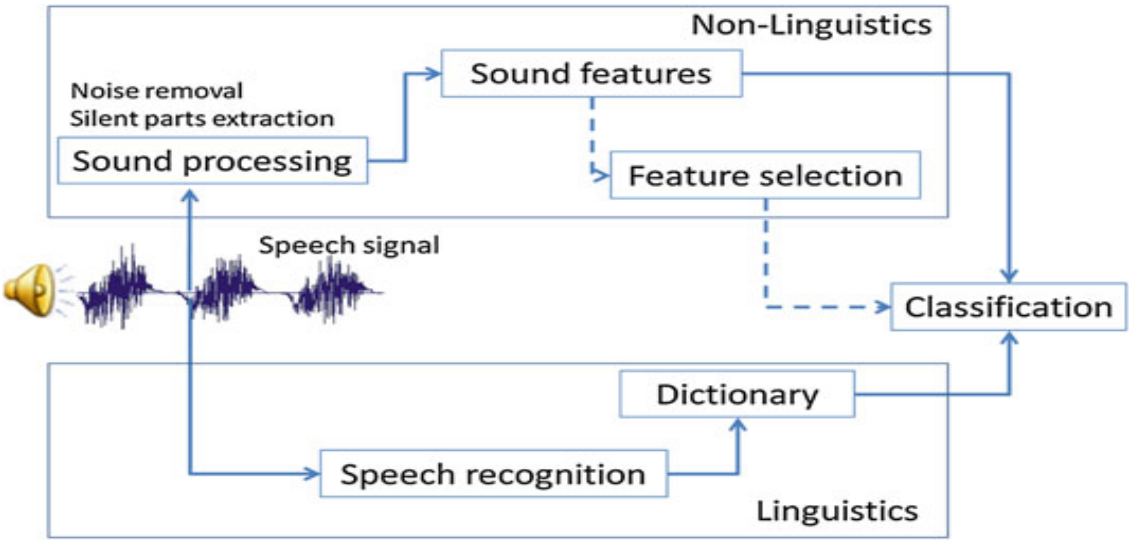


Fig4.2 Speech emotion recognition pipeline

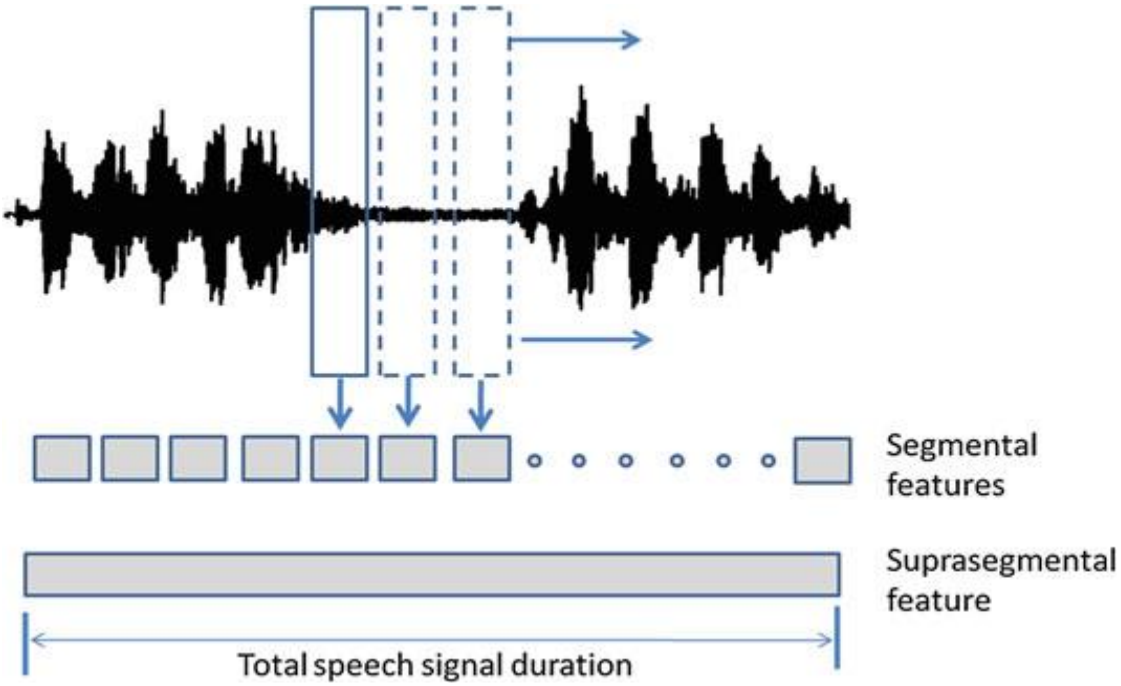


Fig4.3 Segmental and Suprasegmental features in a speech signal

4.5 MACHINE LEARNING ALGORITHMS USED IN SPEECH EMOTION RECOGNITION

There are various types of classification methods employed for using in the task of emotion recognition from speech. Hidden Markov models (HMM), support vector machines (SVM),

Gaussian mixture models and artificial neural networks are so far the most popular classifiers (E. Ayadi et al., 2011). Although Hidden Markov Model has many design issues such as determining the optical number of states, the type of observations and optimal number of observation symbols (E. Ayadi et al., 2011) performance of HMMs for speech emotion recognition task is satisfactory. In (Nwe et al., 2003), empirical assessments of many state numbers are tested. It is claimed that a four state HMM gives the optimal performance to determine the number of states. HMM is generally used to model the time dependency of the system. In (Nwe et al., 2003), it is claimed that frame size of 16 ms and overlapping duration 9 ms gives the best results. The best average rates were 78.5% and 75.5% for Burmese and Mandarin databases, respectively. This compares favorably with respect to human accuracy rate, which was 65.8%. Gaussian mixture models represent features with joint probability density functions GMMs are more appropriate for speech emotion recognition where only global features such as mean and variance of the fundamental frequency are used (E. Ayadi et al., 2011). Since GMM is based on the assumption that all vectors are independent, it cannot model temporal patterns. Determining the number of mixture model is an important problem with GMMs. In (Schuller et al., 2004), 74.83% average classification accuracy for speaker dependent approach and 89.12% for speaker-dependent approach is obtained using a sixteen component GMMs. This corresponds to a performance that is comparable to HMMs. Support vector machine is a supervised linear classifier. Determining a way to choose the kernel size is the problem of the SVM. Since there is not any certain way to choose proper kernel size, it is not guaranteed to segment the transformed features correctly (E. Ayadi et al., 2011). SVM classifier was employed in (Schuller et al., 2004). For speaker independent approach, classification accuracy was 76.12. Speaker dependent result was 92.95%. Another popular classifier is the artificial neural network (ANN). ANNs are known to be ef- fective in non-linear mappings (E. Ayadi et al., 2011). In (Petrushin, 2000), a two layer neural network is used. A classification result for normal state is 55-65%, for happiness is 60-70%, for anger is 60-80% and for fear is 25-50%. Among the other classifiers, ANN provides the worst performance.

4.6 DATABASE OF EMOTIONAL SPEECH

This database comprises 6 basic emotions (anger, boredom, disgust, anxiety, happiness and sadness) as well as neutral speech. Ten professional native German actors (5 female and 5 male) simulated these emotions, producing 10 utterances (5 short and 5 longer sentences), which could be used in everyday communication and are interpretable in all applied emotions. The recorded

speech material of about 800 sentences was evaluated with respect naturalness and recognisability in a forced-choice automated listening test by 20-30 judges. After selection, the database contained a sum of 494 sentences (286 voiced by women and 208 by men). The sentences were not equally distributed among the a variety of emotional states: 64 happy; 79 bored; 78 neutral; 53 sad; 127 angry, 55 frightened; 38 disgusted. As well as being grouped for the classification of the 7 different emotional states (7EMOTIONS) the sentences in the database were also grouped in such a way as to make a distinction between the following groups of states: Activation: (anger, disgust, fear, happiness) - (boredom, sadness) - (neutral); Emotion: (anger, boredom, disgust, fear, happiness, sadness) - (neutral); Evaluation: (anger, boredom, disgust, anxiety, sadness) - (happiness) - (neutral).

4.7 Features selection:-

Whereas it would appear, intuitively, that a large number of features would improve the discrimination capabilities of a classification system, in reality various studies have shown that this is not always true. By reducing the size of the classification vector, the system is provided with a more compact and more easily interpretable set of data, the performance of the learning algorithm is improved and the speed of the system increased . The main feature selection methods are divided into wrapper methods and filter methods. Wrapper methods establish the set of components by interacting with the classification algorithm: they are more accurate but require more computing time. Filter methods are independent as they do not require interaction with the classification algorithm and are thus faster. When there is a large amount of training data, filter models are a good choice on account of their computational efficiency and their independence of the learning algorithm. One method used to eliminate redundant and insignificant components is to identify components that are closely correlated

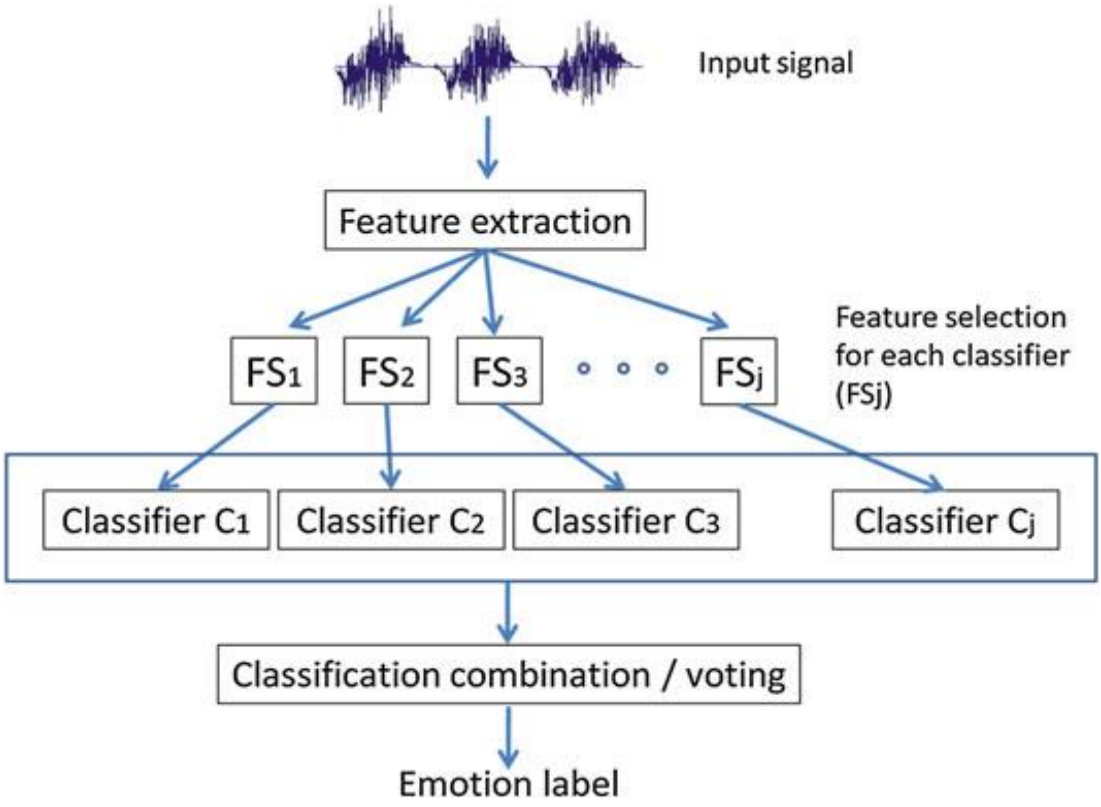


Fig 4.4 A typical architecture for a combined classifier or a voting scheme

with a class but not with each other. Analysis can be in the form of forward selection, starting with an empty list and at each step inserting a new attribute until the increase in performance drops below a pre-established threshold, or by backward elimination, starting from a vector containing all the components and eliminating the worst step by step. There are also more complicated search methods, including the best first method, which keeps a list of all the subsets of components evaluated, ordered according to performance measures in such a way that a previous configuration can be revisited. Genetic Algorithms are a search method based on the principle of natural selection. In this paper, of the feature selection techniques provided by WEKA, we used CFS Subset Eval. This algorithm uses as a feature assessor Correlation-based Feature Selection, which tries to recognize and discard mechanism that are closely correlated with one another. To determine the best subset we used a best-first search strategy and a stratified 10 cross validation procedure. We thus had 10 different sets of selected parameters. A parameter x may be selected for classification in all 10 subsets, while another parameter y may be selected for classification in only 9 out of the 10 subsets, a further features z may be selected in only 1 of the 10 subsets, and so on. Considering the parameters selected for the a variety of subsets, we created an combined of parameters "1-10", which will contain all the parameters selected for classification in at least one of the subsets, the aggregate "2-10" which contains all

the parameters selected for at least 2 of the subsets, and so on up to the aggregate "10" which contains only parameters selected for classification in all of the 10 subsets considered.

4.8 FEATURES DISCRETIZATION

In some cases the continuous domain of the components of the feature vector selected is unsuitable for certain classification algorithms. In order to fully develop the performance of these algorithms it may be suitable to discretize the continuous domain of the various components. There are unverified discretization methods such as equal-frequency binning, and supervised methods, such as the algorithms proposed by Fayyad and Irani and Kononenko. In this paper both techniques were used to identify the improvement in performance that can be obtained when classifying emotional states.

4.9 FEATURES NORMALIZATION

Starting from the original data, new attributes can be obtained so as to present the data in a form that is more appropriate for the learning scheme used. One technique frequently used for this reason is data normalization. The most commonly used data normalization techniques are those which try to ensure that all the components fall within a predefined range. In this work we used the min-max and z score techniques.

4.10 Emotional Training Set

Survey the creative writing, it becomes evident that emotion detection in speech is mostly assessed using digital sources that are datasets rather than databases. Datasets are small scale collections of material created to focus on a specific research and most importantly they are not widely available. Collections that are available to the community tend to full fill requirements related to validity and generalization and, therefore, the term "database" is the most appropriate for them. Generally, it is extremely difficult to produce a database representing the natural speech of a man or a woman in completely natural conversation. Many examples of humans talking exist, but very few of them illustrate speech in a natural environment. In the latter case, some databases use corpora (i.e. large collections) of spontaneous speech, usually consisting of clips from live television, radio programs or call centers, with natural speech recorded in real-world situations. On the other hand, such databases are not distributed easily, since their assessment and processing could raise serious ethical or copyright issues.

Thus, in most cases, speech databases/datasets use acted speech, since the easiest way to collect emotional speech is to have actors simulate it. However, some questions are raised related to the naturalness of the outcome. There are many reasons to suspect that there are significant differences between acted and spontaneous speech. Actors often simply read the utterances or the passages, failing to whole heartedly participate in their role. This could easily lead to the recoding of inaccurate characteristics in the speech signals. Moreover, actors may not capture the original context-related real-world emotions or exaggerate in their acting, making emotion recognition in acted speech easier than in spontaneous. It should be noted that there are numerous smaller experimental datasets, which are not publicly available. As mentioned above, capturing a faithful, detailed record of human emotion, as it appears in real life, is an incredibly challenging task. speech. The assembly of databases (or datasets) has not traditionally been considered a high-profile or intellectually challenging area. Focus is explicitly placed in good quality recording and large samples that usually contain high arousal emotions (e.g. anger, sadness), while real human emotions are left relatively off-focus.

4.11 Emotion recognition algorithm

4.11.1 Support Vector Machine (SVM)

The SVM is a high dimensional vector supervised learning method that is based on emotion assumptions. It predicts that the presence (or absence) of a specified feature of a class is not related to the presence (or absence) of all other features. It is very simple to program and execute it, its parameters are simple to assume, even on very large databases learning or training is very fast and effective and its accuracy is relatively better in comparison to the other techniques. The emotion recognition process along with training and testing phases.

In developed speech emotion recognition algorithm, features are extracted using a state-of-art computational auditory model. Extracted features are classified using a generated binary tree as given in the figure. In each branch, audio samples are classified into two segments. Segmentation is implemented using principal component analysis. In segmentation algorithm, audio samples first projected, then their distance to test sample is measured using Euclidean distance. Used computational auditory model generated total of 286 modulation filtered signals. From each modulation filtered signal, 2 features are extracted which are signal's mean and signal's standard deviation. When two features are fused, total length of the feature vector becomes 572. At first, extracted feature vector size is reduced using component analysis. SVM is a binary classification algorithm which separates the data finding a hyperplane that maximizes

margins. In this section, given binary tree is applied with SVM in. Binary tree given in figure 4.1 is process by six different SVM's. Each SVM is trained separately. SVM kernels were set as linear kernel. First branch is the classification of excited and non-excited emotions. Table results are derived with leave-one-out method on Berlin Speech Emotion Test. In table 4.8 segmentation results of excited, non-excited emotions is provided. Classification with SVM has a higher success rate with 97.16%. Second branch is segmentation of neutral and sad boredom emotions. In table 4.9, success rate of neutral, sad-boredom is given. Third branch is the segmentation of sad and boredom emotions. Rates are given in table 4.10. Segmentation of excited emotions is processed in two levels. First segmentation is happy-anger and fear-disgust. Results are given in table 4.11. Success rate of fear-disgust is given in table 4.13. Final branch is the segmentation of happy and anger emotions. Classification rate is given in table 4.12 which is equal to 87%. Compared with other algorithms, segmentation .These results have shown that, SVM provides higher classification accuracy. In addition to that features extracted from auditory model are discriminant to segment happy and anger than extracted short time spectral features.SVM are also called as maximum margin classifiers. Firstly, the SVM theory is used to solve binary classification problems , rendering SVM ideal for the case under consideration, hence we apply a psychologically-inspired binary cascade classification schema for identifying or to make a classifier . In this we use two different kernels.

Let **g***i* be the *i*th training vector.

1. Gaussian radial basis function kernel:

TSVM (**g**i , **g**j) = ex_((−_**v**i − **v**j)^2)/σ2

where σ is a scaling factor; and

2. Linear (Homogeneous):

TSVM (**v**i , **v**j) = **v** Si **v**j

For various *S* values, for male subjects, female subjects, and both genders.

SVM with Gaussian radial basis function kernel are tested for various σ values with $\sigma \in (0, 10]$, The best performance is obtained for $\sigma = 1$. For the case of SVM with Gaussian radial basis function kernel the two genders exhibit the same pattern: emotion recognition accuracy reaches at a fast rate its maximum for $\sigma = 1$,

Classification	**Happiness**	**Neutral**	**Boredom**	**Sadness**	**Anger**
Happiness	99.7	0	0	0	1.7
Neutral	2.9	90.5	1.5	0	0
Boredom	0	9.5	91.0	11.4	0

Sadness	0	0	6.0	88.6	0
Anger	4.5	0	0	0	90.1

Table2. Confusion matrix (%) for the SVM with Gaussian radial basis function kernel (σ =1)

Classification	**Happiness**	**Neutral**	**Boredom**	**Sadness**	**Anger**
Happiness	92.6	6	0	0	1.6
Neutral	0	98.9	1.3	0	0
Boredom	0	1.1	88.5	11.5	0
Sadness	0	0	8.9	88.5	0.8
Anger	1.6	0	0	0	87.7

Table 3. Confusion matrix (%) for the linear SVM

Speaker-independent emotion recognition accuracy of SVM with Gaussian radial basis function kernel for various σ values, for male subjects, female subjects, and both genders whereas it decreases strictly at a slower rate for greater σ values. The confusion matrix for σ = 1 is exhibited in Table 2 and the related accuracy equals 92.4%. Male emotion recognition accuracy is consistently greater than female emotion recognition, with exception of extreme low and high σ values. The lower bound accuracy presented by SVM with Gaussian radial basis function kernel is 50.7% and can be attributed to poor parametrization. Linear SVM has the advantage of no need for parametrization . The corresponding confusion matrix is sketched in Table 3. It achieves an emotion recognition accuracy equal to 95.5%.

Chapter 5

Simulation Results and Discussion

5.1 Emotion recognition based on speech analysis:

The introduction to the theory contains a review of emotion inventories used in various studies of emotion recognition as well as the speech corpora applied, methods of speech parameterization, and the most commonly employed classification algorithms. In the current study the EMO-DB speech corpus and three selected classifiers, the *k*-Nearest Neighbour (*k*-NN), the Artificial Neural Network (ANN) and Support Vector Machines (SVMs), were used in experiments. SVMs turned out to provide the best classification accuracy of 75.44% in the speaker dependent mode, that is, when speech samples from the same speaker were included in the training corpus. Various speaker dependent and speaker independent configurations were analyzed and compared. Emotion recognition in speaker dependent conditions usually yielded higher accuracy results than a similar but speaker independent configuration. The improvement was especially well observed if the base recognition ratio of a given speaker was low. Happiness and anger, as well as boredom and neutrality, proved to be the pairs of emotions most often confused.

5.2 Results of experiments:

Results for all three classifiers were evaluated based on the mean classification accuracy, both for the whole corpus and for each speaker independently. The best result in the SIF configuration was 68.7% and was achieved with the SVM classifier. The other two classifiers yielded worse results: 63.22% and 56% for the k-NN and ANN, respectively.

With regard to the SI-i and SD-i configurations, the results will be described separately for each of the classifiers. Confusion matrices for the classes being recognized will be presented and analyzed too.

$\log_2(\gamma)$	Parameter	8	32	128	512
−7	Avg. error	62.19	31.30	28.45	28.95
−7	Std. dev.	1.25	1.47	1.28	1.40
−7	SV share	72.33	50.68	33.62	24.21
−7	Train error	22.89	15.24	9.46	6.01
−5	Avg. error	31.54	28.55	28.99	29.12
−5	Std. dev.	1.54	1.38	1.40	1.37
−5	SV share	50.82	33.70	24.30	19.20
−5	Train error	15.28	9.46	5.99	3.96
−3	Avg. error	27.95	**28.51**	28.94	30.20
−3	Std. dev.	1.38	**1.43**	1.23	1.26
−3	SV share	34.04	**24.55**	19.41	17.19
−3	Train error	9.51	**5.90**	3.72	2.36
−1	Avg. error	28.51	30.08	32.64	35.24
−1	Std. dev.	1.27	1.25	1.34	1.60
−1	SV share	25.56	20.59	18.82	18.27
−1	Train error	5.73	3.31	1.67	0.69

Table4. Classification evaluation for various values of γ (in rows) and C (in columns) for the SVM classifier (values in percentages). Results for the configuration selected as optimal are printed in bold.

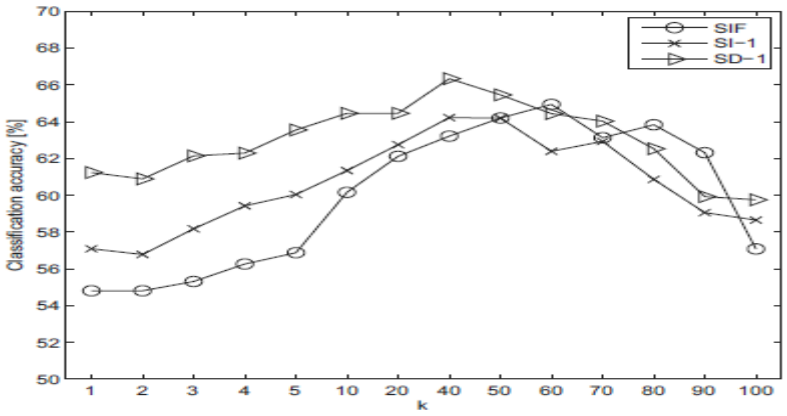


Fig 5.1 Emotion classification error for various *k* for the *k*-NN classifier, for speaker dependent and speaker independent configurations.

The following section the results will be summarized, discussed, and compared with other studies.

5.3 k-NN algorithm

The total results for the *k*-NN classifier are shown in Table 5. It can be observed that for $i < 5$ the classification accuracy for the SD-*I* configurations is 1–2 percentage points (p.p.) higher than

for the speaker independent configuration, while the performances for i = 5and i = 6are almost equal (within the confidence level of the test). The impact of adding speaker samples to the training set was highly dependent on the evaluated speaker. The general trend was that speakers who were classified with high results in SI- did not improve their results in the SD-i configuration, or even showed a slight drop. On the other hand, speakers with low results in SI-i (e.g., speakers #11 and #16) performed much better when their samples were present in the training set. The gain for SD-i reached over 9 p.p. in some cases. it can be seen that the average classification performance for each speaker has only a small variation. The highest differences are for i = 4, 5, 6, where the variation of average classification accuracy is caused by.

i	1	2	3	4	5	6
SI-i	64.41	64.33	63.23	65.71	71.62	64.35
SD-i	65.63	67.64	64.51	66.59	71.97	64.81

Table5. Classification results (in percentages) for various numbers of speaker samples (i) in the training set for the k-NN classifier.

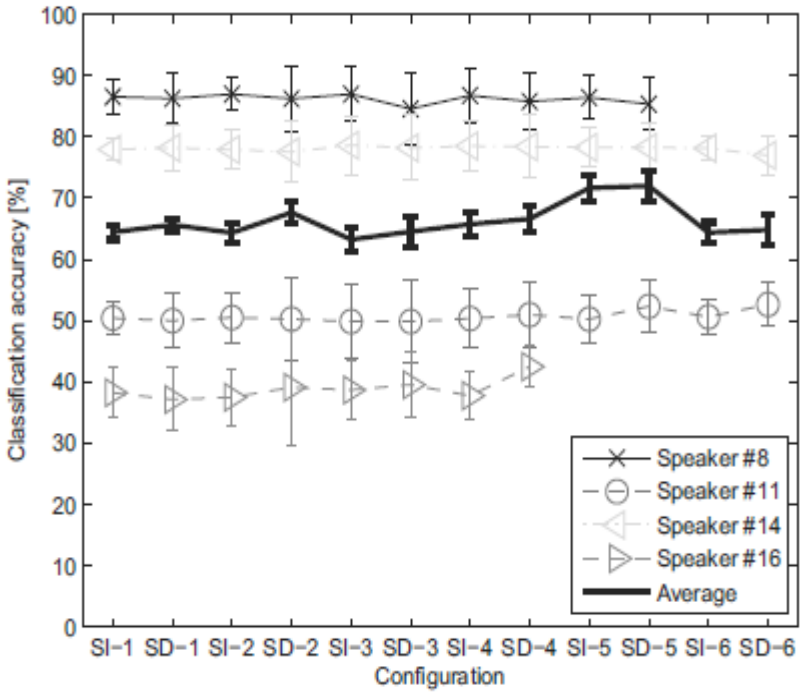


Fig 5.2 Emotion classification error for various configurations for the k-NN classifier

The elimination of speakers from the test set. The increase for i = 5 is caused by the elimination of the poorly classified speaker #16. The decrease for i = 6, on the other hand, is caused by the elimination of speaker #8, whose classification error was very low. The confusion matrix shown in Table 5 presents a high rate of misclassification between happiness and anger, equal to about

50%, for both SI and SD configurations. We believe that the reason for this is that both emotions are characterized by high arousal and are therefore difficult to distinguish. Further investigation of the results for each speaker in SI-1 revealed that five out of nine speakers did not have even one correctly classified sample of happiness.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**74.80**	0.00	0.00	2.53	13.69	9.30
Ang	1.51	**94.56**	1.69	0.00	0.00	2.24
Hap	8.08	52.24	**19.42**	0.00	5.67	14.59
Sad	7.53	0.00	0.00	**84.83**	3.87	3.78
Bor	28.14	0.73	0.99	6.48	**57.70**	5.96
Fea	21.10	17.27	3.43	0.00	2.50	**55.70**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**72.53**	0.01	0.00	2.89	15.63	8.94
Ang	2.02	**94.07**	1.67	0.00	0.00	2.25
Hap	7.36	49.44	**25.19**	0.00	4.05	13.97
Sad	4.87	0.00	0.00	**87.59**	3.81	3.74
Bor	26.24	0.97	0.62	6.53	**60.48**	5.16
Fea	22.90	14.68	3.84	0.02	2.82	**55.74**

Table6. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for *k*-NN (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

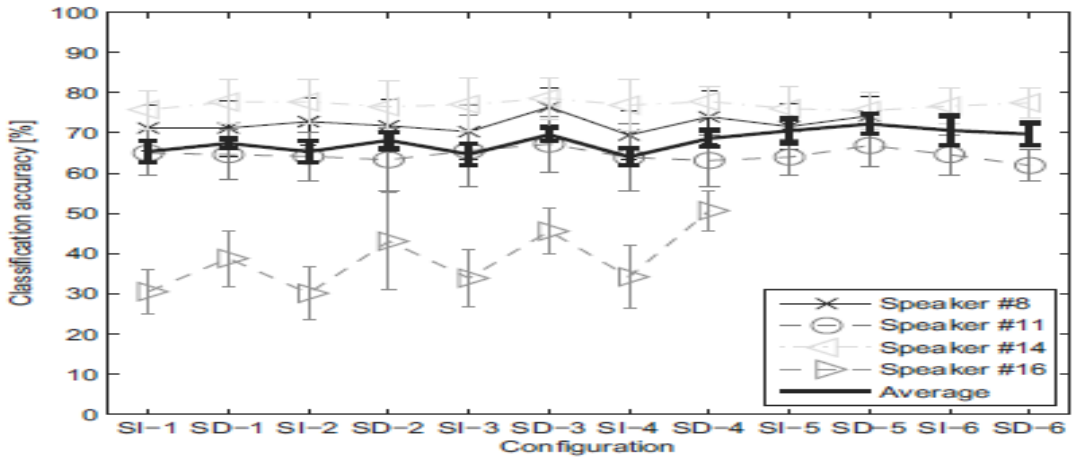


Fig 5.3 Emotion classification error for various configurations for the ANN classifier.

The results of the other four showed that 20% of answers were correct. In contrast, the situation of classifying anger as happiness was very rare. It is supposed that in the *k*-NN classifier anger was represented by a set of vectors (potential "neighbours") not accompanied by representatives of happiness, whilst happiness was often accompanied by representatives of anger. It is noticeable, however, that the recognition of happiness increased from 19.42% to 25.19% when

switching to the speaker dependent configuration, which probably caused an increase in the "distinct" representatives of happiness.

5.4 SVM algorithm

As shown in Table 7, for $i = 1, \ldots, 4$ the recognition accuracy of the SVM classifier variates at the level of 69% for SI-*i* and 72% for SD-*i*. These results are significantly better than those for the *k*-NN and ANN. Similarly to the previous classifiers, the speaker dependent configuration improves the recognition accuracy by ca. 3 p.p. For $i = 5, 6$ the increase is only minor. Figure shows the results of emotion recognition for the most characteristic speakers. The highest gain was again observed for speaker #16, for whom it reached 21 p.p. in the case of $i = 4$. Speaker #15 also yielded a remarkable improvement. For the other speakers the SD configuration did not improve the results much.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**54.17**	0.00	3.93	4.05	22.26	15.60
Ang	0.36	**78.45**	17.38	0.00	0.00	3.81
Hap	5.12	38.69	**46.07**	0.00	4.17	5.95
Sad	5.48	0.00	0.00	**84.17**	3.81	6.55
Bor	18.21	0.83	2.98	5.24	**65.83**	6.90
Fea	10.12	14.29	12.26	1.07	2.86	**59.40**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**59.76**	0.00	3.57	4.05	17.62	15.00
Ang	0.24	**80.48**	16.90	0.00	0.00	2.38
Hap	3.81	35.24	**50.95**	0.00	4.48	4.52
Sad	3.10	0.00	0.00	**88.81**	4.29	3.81
Bor	18.10	0.71	1.90	4.76	**68.57**	5.95
Fea	10.71	10.00	8.33	0.48	1.67	**68.81**

Table7. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for the ANN (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

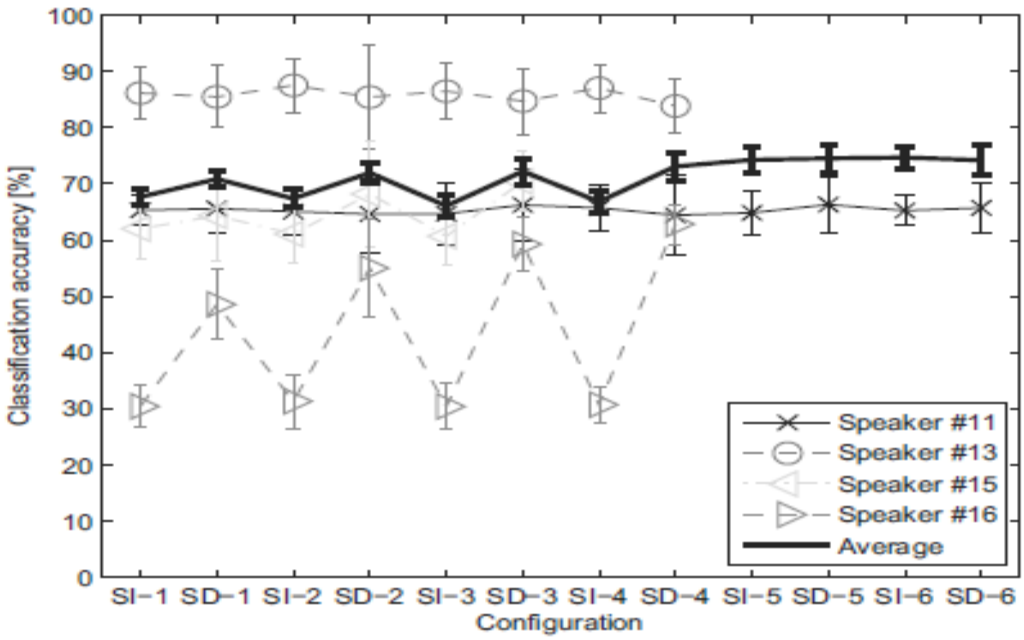


Fig 5.4: Emotion classification error for various configurations for the SVM classifier.

Table shows confusion matrices for the SVM in configurations SI-1 and SD-1. The high recognition rate of sadness compared to the other emotions is noticeable. This is the most well recognized emotion and the one that is least frequently misclassified as other emotions. The recognition of happiness improved further. Fear and boredom are sometimes confused with neutral emotion, in both configurations. For example, in the case of speaker #3, whose neutral emotion was recognized with a 100% success rate, 61% of samples of boredom were classified as neutral. On the other hand, for speaker #10, whose boredom was recognized with a 78% correct-classification rate, 51% of samples of neutral emotion were recognized as boredom.

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**63.17**	2.03	4.65	3.40	17.99	8.75
Ang	1.48	**83.34**	13.63	0.00	0.00	1.54
Hap	5.06	32.09	**52.47**	0.00	4.04	6.34
Sad	2.97	0.00	0.00	**87.15**	4.83	5.06
Bor	17.27	0.64	4.65	7.18	**65.32**	4.94
Fea	10.99	12.44	7.12	1.66	2.76	**65.03**

	Neu	Ang	Hap	Sad	Bor	Fea
Neu	**66.28**	0.71	4.30	3.85	17.91	6.96
Ang	2.06	**82.53**	14.56	0.00	0.00	0.85
Hap	3.14	31.44	**57.09**	0.00	3.40	4.93
Sad	2.51	0.00	0.00	**92.49**	2.24	2.76
Bor	16.55	0.81	5.19	7.88	**65.28**	4.29
Fea	11.98	9.31	7.02	1.30	3.17	**67.22**

Table8. Confusion matrices for the SI-1 (upper) and SD-1 (lower) configurations for the SVM (in percentages of recognitions). The diagonals show the percentages of correctly recognized emotions.

5.5 DISCUSSION OF RESULTS

All the tested classification methods yielded classification accuracies between 64% and 75%. It is worth remembering that in the case of six classes the choice level is 1/6 = 16.67%, so the accuracy results are far above this level. They are also higher than the estimated level of human performance in speaker independent conditions (60%).

With regard to the comparison between speaker dependent and speaker independent conditions, shows that in almost all configurations the speaker dependent configuration improved recognition; however, this increase was minor for the *k*-NN classifier. The higher improvement was observed for speakers who were originally poorly recognized (e.g., speaker #16). The best recognition results were achieved by the SVM, followed by the ANN and *k*-NN classifiers.

The *k*-NN classifier had the worst happiness recognition, i.e., 0% for five out of nine speakers. Speaker dependent recognition had a small influence on the recognition rate mostly due to the specificity of the *k*-NN algorithm: for complex problems it requires a high value of *k* (the number of neighbours, in this case *k* = 40). The addition of one to six samples to the training set

had only a slight influence on the decision border, compared with the number of 40. Only speakers #15 and #16 yielded better results in SD-*i*, which influenced the overall result.

The SVM classifier had the best overall recognition rate of 75.44% and the best performance in the SD-5 configuration, with a classification error below 25%. Emotion recognition was at a different level for each speaker, which places this classifier between the distributions achieved by the *k*-NN and ANN.

A somewhat strange behaviour of the tested classifiers for $i > 4$ was caused by deficiencies of the corpus used: EMO-DB unfortunately did not contain enough samples to obtain a representative training and testing set of speakers and their emotional recordings. Therefore the results for $i < 5$ should be treated as more reliable. All classifiers showed the presence of pairs of emotions which were often confused, for example, anger and happiness, boredom and neutrality .We believe that this was caused by high class infiltration, that is, there was no dimension that could distinguish between these emotions.

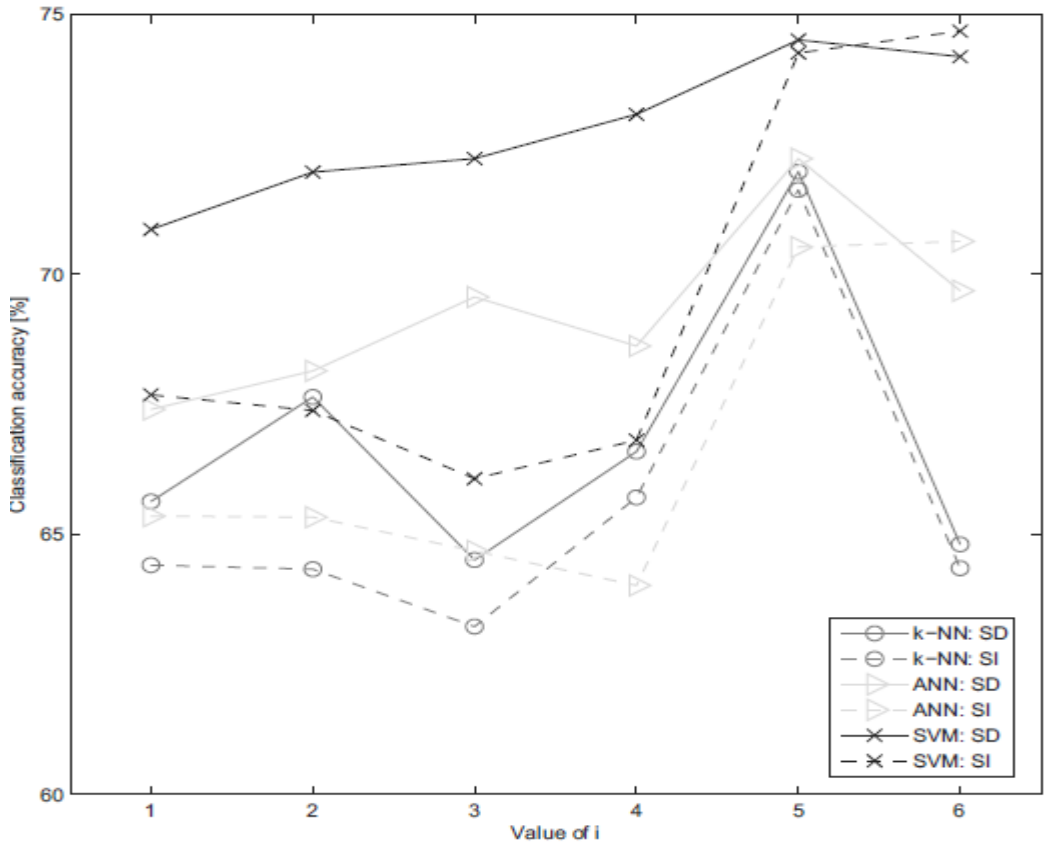


Fig 5.5 Comparison of speaker independent and speaker dependent configurations for all three classifiers tested.

SVM consisted of many one-versus-one classifiers, each distinguishing between a pair of emotions. For the happiness–anger pair the training error was the highest. Both these emotions show high arousal and are therefore often confused. The differ as for valence (positive vs.

negative); however, this feature is much more difficult to be captured using speech signal parameters. We think that replacing some of the currently used speech parameters with novel ones (e.g., TEO based) could possibly improve.

5.6 Emotion Classification:

In this project we recognise the mood of a user through their voice on the basis of their mood and classify into classifier.

5.6.1 Anger emotion :

Anger requires high energy to be expressed. Definition meaning of the anger is simple extreme displeasure. In case of anger, aggression increases in which control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, joy and sadness. The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions. Besides a faster speech rate is observed in angry speeches.

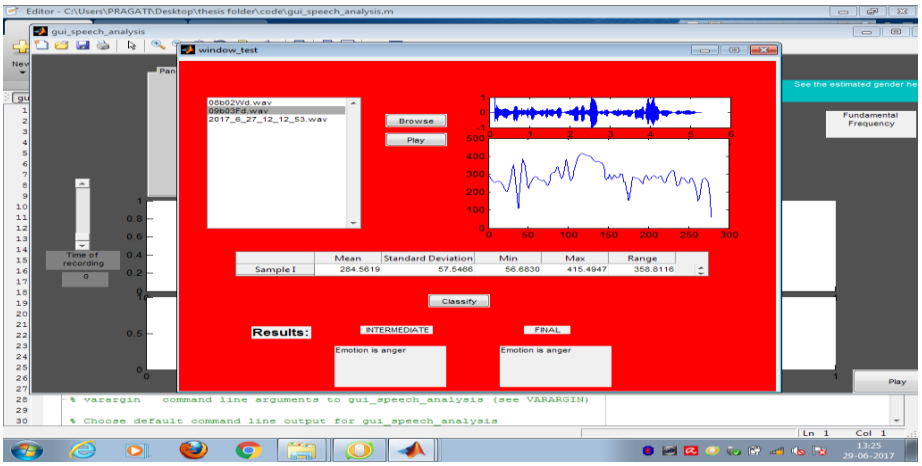


Fig 5.6 Anger Emotion

The widest observed pitch range and highest observed rate of pitch change are other findings about the emotion label anger when compared with other emotions. Besides a faster speech rate is observed in angry speeches.

5.6.2 Fear emotion:

In emotional dimension, fear has similar features to anger. High pitch level and raised intensity level are correlated with fear. It is stated that fear has a wide pitch range. Highest speech rate is observed in fear speeches. The pitch contour trend separates fear from joy. Although the pitch contour of fear resembles the sadness having an almost downwards slope, emotion of joy have a rising slope.

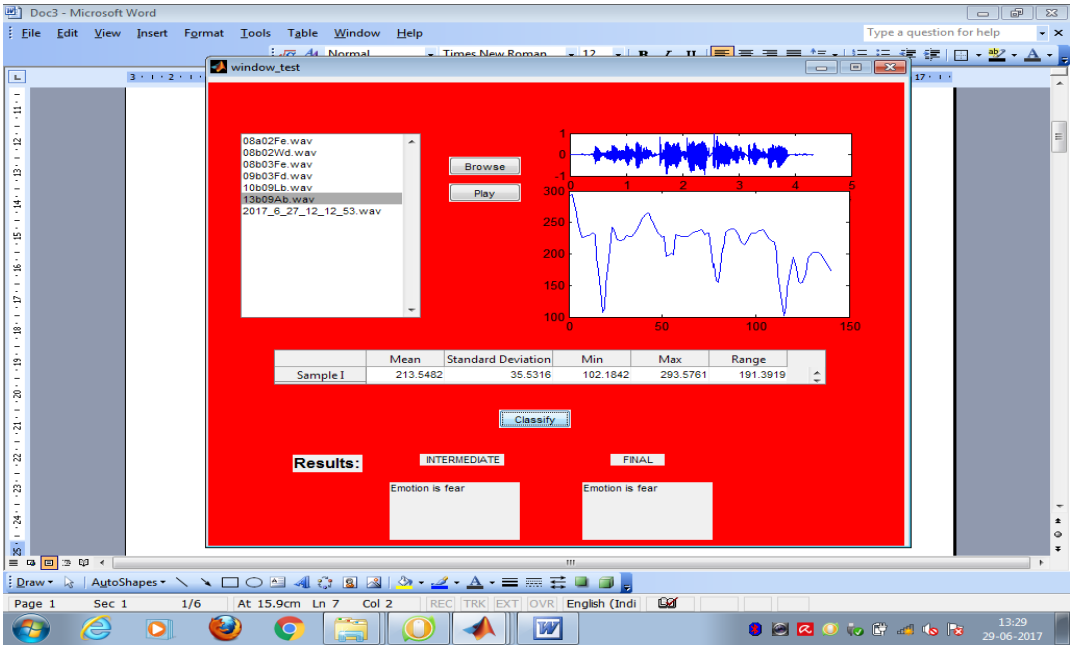


Fig 5.7 Fear Emotion

5.6.3 Disgust emotion:

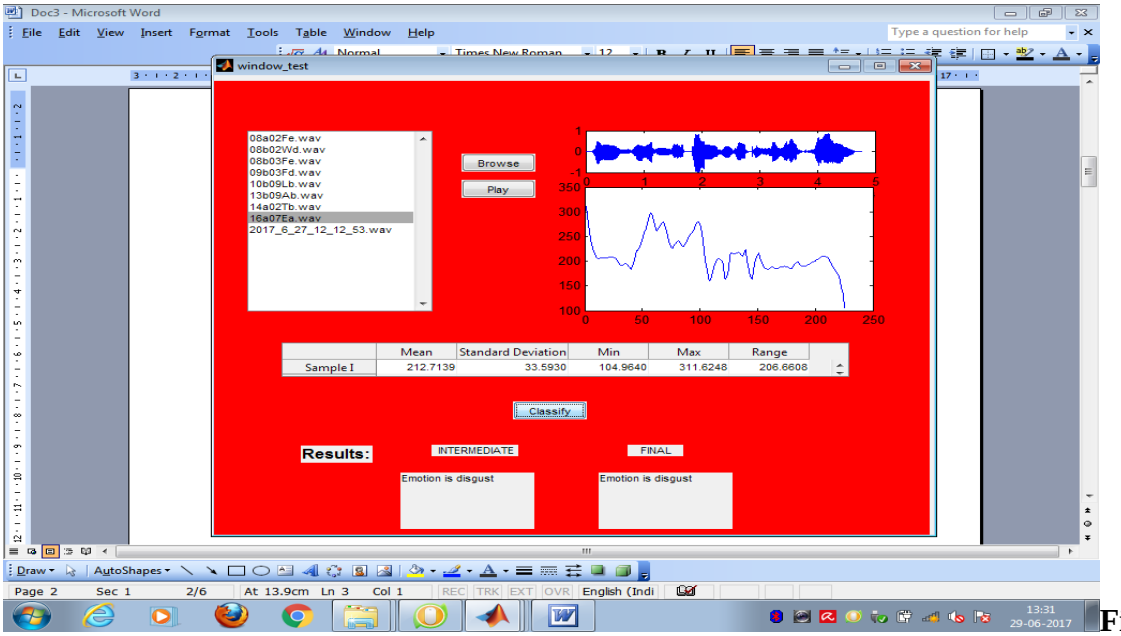


Fig 5.8 Disgust Emotion

In, low mean pitch level, a low intensity level, and a slower speech rate is observed when disgust is compared with the neutral state. Disgust is stated the lowest observed speech rate and increased pause length.

5.6.4 Happiness emotion:

Happiness exhibit a pattern with a high activation energy, and positive valence. Strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range and variance increases. In, it is stated that fundamental and formant frequencies increases in case of smile. Moreover, amplitude and duration also increase for some speakers.

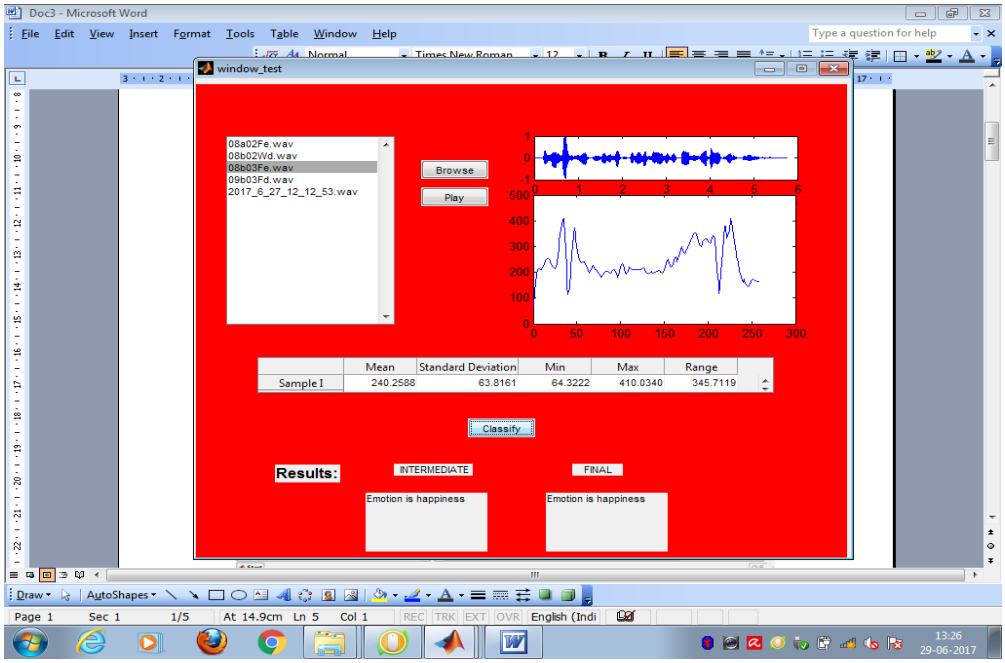


Fig 5.9 Happiness Emotion

5.6.5 SADNESS EMOTION:

In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo. Speech rate of a sad person is lower than the neutral one.

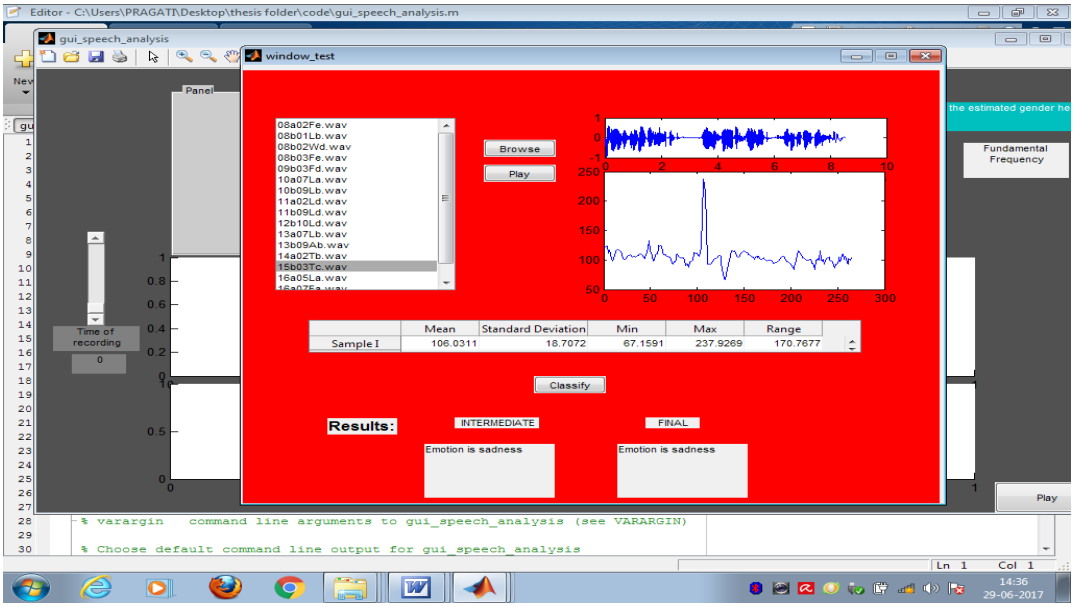


Fig5.10- Sadness Emotion

5.6.6 BOREDOM EMOTION: Boredom is a negative emotion with negative valence and low activation level same as sad. A lowered mean pitch and a narrow pitch range with a slow speech rate are defined as the properties of a bored expression.

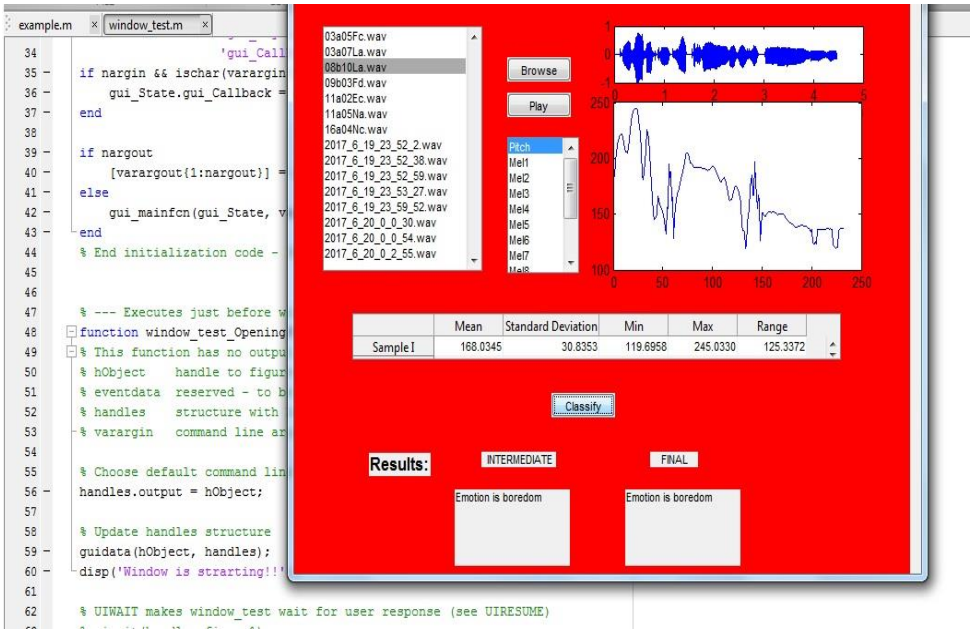
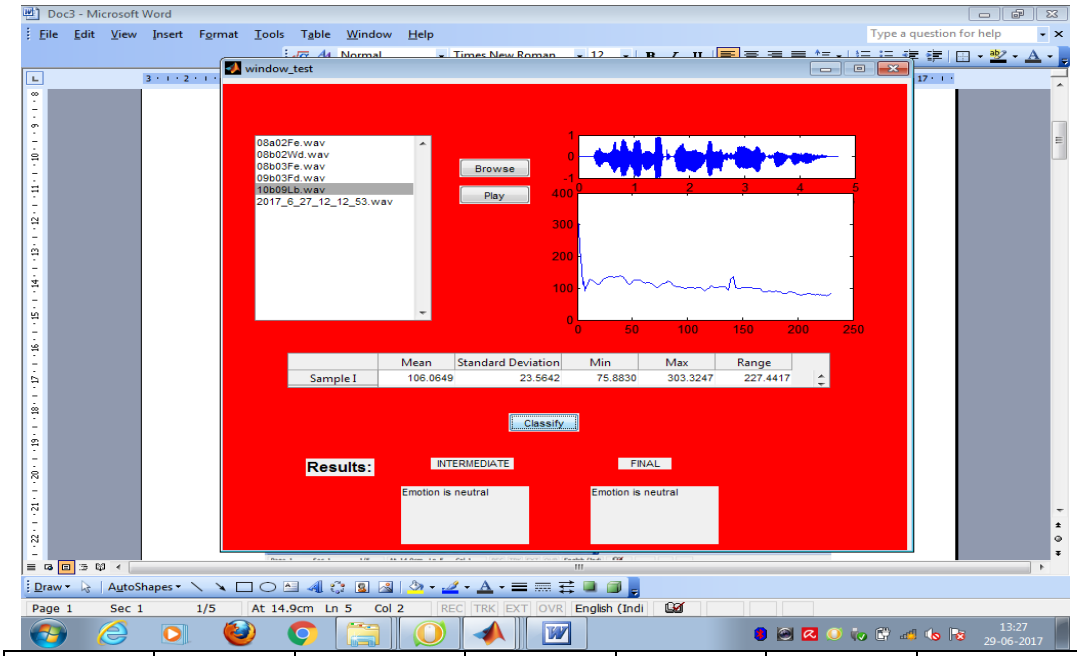


Fig 5.11 Boredom emotion

Neutral emotion:

Fig 5.12 Neutral Emotion



Emotion classification of speech signal

Sample	Mean	Standard deviation	Min	Max	Range	Emotion	Accuracy
08a02fe	204.5379	44.9043	77.7242	291.9683	293.5781	Happiness	Yes
08b01l	187.54	42.8591	78.7224	290.567	215.447	Happine	Yes

b	97			7	8	ss	
02b09a b	213.54 82	35.5316	102.428 2	293.578 1	191.391 9	Fear	Yes
100a51 d	106.47 24	18.9341	66.2311	165.972 0	165.972 0	Disgust	No
16a07e a	212.71 39	33.5930	14.9649	311.624 8	206.660 8	Disgust	Yes
03b01l b	113.59 12	35.5358	66.5358	470.588 2	403.779	Boredam	Yes
13a027 a	132.64 33	31.4061	80.8352	214.849 4	134.014 2	sadness	Yes
11a05N a	109.05 24	14.5833	52.4239	179.400 7	126.916 8	Neutral	Yes
03a04w c	227.92 00	48.6458	77.5875	300.113 3	222.525 9	Anger	Yes

Table-8 Emotion Classification

On the basis of table 8 we can easily see that the recognition ratio is 95% on the basis of the data set and on the basis of classifier. Hence, recognition accuracy is more as comparison to other algorithm.

CHAPTER-6

CONCULSION AND FUTURE SCOPE

6.1 Conclusions:

As technology evolves, interest in human like machines increases. Technological devices are spreading and user satisfaction increases importance. A natural interface which responds according to user needs has become possible with affective computing. The key issue of affective computing is emotions. Any research which is related with detection, recognition or generating an emotion is affective computing. User satisfaction or un-satisfaction could be detected with any emotion recognition system. Besides detection of user satisfaction, such systems could be used to detect anger or frustration. In such cases, user could be restrained like driving a car. In emotion detection tasks, speech or face emotion detections are the most popular ones. Easy access to face or speech data made them very popular. Speech carries a rich set of data. In human to human communication, via speech information is conveyed. Acoustic part of speech carries important info about emotions. In this work, human speech emotion process is tried to be simulated. More importantly, to test human speech emotion performance using only acoustic part, listening test is constituted with German speeches on listener who do not know any German. Therefore, extracted results provide a comparison between subjective and automatic speech emotion recognition task. Computational auditory model is used to generate a feature set. To convert auditory model output to feature set, simple transformation methods such as mean and standard deviation is used. Extracted features are classified into 7 discrete emotions using classification algorithms. MFCC are used for the feature extraction. Algorithm with the SVM's overall performance is tested. Finally results for different combination of the features and on different databases are compared and we get SVM recognition accuracy is more than other algorithm. Accuracy obtained from SVM is 95% basis of data set.

6.2 Future Scope:

In future work, performance of the generated algorithm could be improved. In feature extraction part, extracted features from auditory model may be enhanced. Instead of using mean and standard deviation, more complex methods could be used to extract features from auditory model output. Besides, modulated signals are not the only output generated by auditory model. Human auditory system transmit to the brain, phase information of the first three auditory filterbank output. Results have shown that when leave speech sample out method is implemented, highest accuracy rates are obtained for all three databases when compared with speaker dependent and independent cases. In leave one speech sample out

method, all speakers are included in the training part. This shows that, there is hyperplane which can classify all seven emotions. SVM selects the hyperplane from many choices which maximizes the margins. Since in speaker independent case, number of training samples is low, SVM selects the hyperplane accordingly. On the other hand, when leave one speech sample out method is implemented, generated hyperplane also segments the training samples in speaker independent case. To overcome this issue, and to generalize algorithm into speaker or language independence, a normalization in features could be searched. Besides, generated algorithm could be tested with noisy data, which fits to the real life data. In that case, algorithm could be extended to real life case. Since SVM is a binary classifier, binary decision tree s generated. Yet generated binary tree may not fit to the all languages. Therefore, instead of a binary classifier, multi class classifiers may increase success rate and could work properly for many languages. Besides multi class classifiers, using ensemble learning many different models could be fused. Due to very less knowledge about this field there are very few researches going on in the area of speech processing. But a large amount of work can be done by processing the spectral features effectively to recognize. Higher accuracy can be obtained using the combination of more features. Increasing the sigma value from the default value one, substantial results may be obtained.

REFERENCE :

1. Murray, I. R., & Arnott, J. L. (1993)," Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion" Journal of the Acoustical Society of America, 93, 1097–1108.
2. Ververidis, D., & Koropoulos, C. (2006), "Emotional speech recognition: Resources, features, and methods" Speech Communication, 48, 1162–1181.
3. Burkhardt, F., & Sendlmeier, W. F. (2000), "Verification of acoustical correlates of emotional speech using formant-synthesis" ISCA Workshop on Speech and Emotion, 4, 151–156.
4. Vidrascu, L., & Devillers, L. (2005), "Detection of real-life emotions in call centers".
5. Engberg, I. S., & Hansen, A. V. (1996), "Documentation of the Danish emotional speech database(des)". Aalborg University, Denmark.
6. Hanjalic, A. (2006), "Extracting moods from pictures and sounds towards truly personalized tv" IEEE Signal Processing Magazine, 23, 90–100.
7. Haq, S., & Jackson, P. J. (2012), "SVM tutorial classification, regression and ranking (G. Rozenberg, T. Back, & J. N. Kok, Eds.)" Springer.
8. Iliou, T., & Anagnostopoulos, C. N. (2010), " Classification on speech emotion recognition - a comparative study" International Journal On Advances in Life Sciences, 2, 18–28.
9. Ayadi, E. et al; (2011), "Survey on speech emotion recognition features, classification schemes and databases" Pattern Recognition, 44, 572–587.
10. Grimm et al; (2008), "The vera am mittag german audio visual emotional speech database" IEEE International Conference on Multimedia.
11. Casale et al; (2008), "Speech emotion classification using machine learning algorithm" IEEE International Conference on Semantic Computing, 158–165.
12. Tawari, A., & Trivedi, M. M. (2010), "Speech emotion analysis: Exploring the role of the content" IEEE Transactions on Multimedia, 12, 502–509.
13 T. D., & Narayanan, S. S. (2010), "Acoustic feature analysis in speech emotion primitives estimation" Interspeech .
14. Automatic recognition of speech emotion using long-term spectro-temporal features. Digital Signal Processing, 1–6.
15. C J., & Tao, J. (1997), "Getting started with susas: a speech under simulated and actual stress database". Euro Speech, 4, 1743–1746.
16. Kamran Soltani and Raja Noor Ainon (2002), "Speech emotion detection based on neural networks" In 9th International Symposium on Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.8
17. Thapanee Seehapoch & Sartra wongthanavasu (2013), "Speech Emotion Recognition using Support Vector Machine".
18. E. Dellandrea & Dou (2005) "Features extraction and selection for emotional speech classification". IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp.411- 416, Sept 2005.
19. T.-L. Pao et al; (2006), "Mandarin emotional speech recognition based on SVM and NN", Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, pp. 1096-1100, September 2006.
20. Lin Y & Wei G, "Speech emotion recognition based on HMM and SVM". Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol.8, pp. 4898-4901. Agu 2005.
21.Jana Tuckova and Martin Sramka(2010) "Emotional speech analysis using Artificial Neural Network" Proceedings of the International Multi conference on Computer Science and Information Technology, 141-147, 2010
22. H. Fletcher et al; (2001), "Speech and Hearing in Communication". The Bell Telephone Laboratories Series, D. Van Nostrand Company, Inc.

23. B. Schuller et al; (2004), "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", IEEE international conference on acoustics, speech, and signal processing, vol.1, pp. I-577-80, 2004.
24. Sirko Molau et al; (2011), "Computing mel-frequency cepstral coefficients on the power spectrum". IEEE Transaction, 2011.
25.Mandar Gilke et al; (2012), "MFCC based vocal emotion recognition using ANN". International Conference on Electronics Engineering and Informatics, 150-154, 2012.
26. Huang et al; (2005), "SVM based recognition of chinese vowels" Artificial Intelligence 3802, 812-819.
27. M. G. & Alpaydin (2011), "Multiple kernel learning algorithms". Journal of Machine Learning Research, July, 12, 2211-2268.
28. Kulkarn et al; (2011), "Comparison between SVM & Other Classifiers for SER".International Journal of Research and Technology, January, 2(1), 1-6.
29. Koolagudi et al; (2010), "Real Life Emotion Classification using VOP and Pitch Based Spectral Features". India: Jadavpur University.
30. Liqin Fu et al ;(2008) ,"Speaker Independent Emotion Recognition Based on SVM/HMMs Fusion System" IEEE International Conference on Audio, Language and Image Processing(ICALIP), pages 61-65, 7-9 July 2008
31. Peipei Shen et al; (2011), " Automatic Speech Emotion Recognition Using Support Vector Machine" IEEE International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT) volume2 , Page(s) : 621 - 625 , 12-14 Aug. 2011.
32. Akalpita Das et al; (2014), " A brief study on speech emotion recognition" , International Journal of Scientific & Engineering Research(IJSER), Volume 5, Issue 1,pg-339-343, January-2014.
33. Vinay et al; (2014), "Gender Specific Emotion Recognition Through Speech Signals", IEEE International Conference on Signal Processing and Integrated Networks (SPIN), 2014 , Page(s):727 – 733, 20-21 Feb. 2014.
34. Norhaslinda Kamaruddin et al; (2014), "Speech emotion identification analysis based on different spectral feature extraction methods", IEEE Information and Communication Technology for The Muslim World, 2014 The 5th International Conference, Pages:1-5, 2014.
35 A. D. Dileep & C. Chandra Sekhar, "GMM Based Intermediate Matching Kernel for Classification of Varying Length Patterns of Long Duration Speech Using Support Vector Machines", IEEE Transactions on Neural Networks and Learning Systems, Volume: 25, Issue: 8,Pages: 1421 - 1432, 2014.
36. S.Lalitha et al; (2014) "Speech Emotion Recognition" IEEE International Conference on Advances in Electronics, Computers and Communications (ICAECC), Page(s): 1-4, 2014.
37. S.Sravan Kumar & T.RangaBabu (2015), "Emotion and Gender Recognition of Speech Signals Using SVM", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 4, Issue 3, pg.- 128-137 May 2015.
38. Zhou y et al; (2009), "Speech Emotion Recognition using Both Spectral and Prosodic Features", IEEE, 23(5), 545-549, 2009.
40. Rabiner L. R. and Juang (2005), "Fundamentals of Speech Recognition", Pearson Education Press, Singapore, 2nd edition, 2005.
41. Albornoz E. M et al ; (2009), "Recognition of Emotions in Speech". Proceedings of 17th European Signal Processing Conference, 2009.
42. F. Yu et al; (2001) , "Emotion detection from speech to enrich multimedia content", in Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp.550-557, Beijing, China, October 2001.

43. Tsuyoshi Moriyama et al; (2009), "A synthesis method of emotional speech using subspace constraints in prosody". Journal of Information Processing Society of Japan, 50(3):1181–1191, 2009. (in Japanese).
44. Aishah Abdul Razak et al; (2005), "Comparison Between Fuzzy and NN Method for Speech Emotion Recognition", Proc. of the Third International Conference on Information Technology and Applications, pp.297 – 302, July. 2005.
45. Sondhi et al; (1968) "New Methods of pitch extraction", IEEE Trans. ASSP, 16(2): pp.262-266, 1968.
46. Joseph W. Picone, "Signal modeling techniques in speech recognition", Proc. of the IEEE, Vol. 81, No. 9, pp.1215-1245, Sep. 1993.
47. Johnson et al; (2006), "Generalized Perceptual Features for Vocalization Analysis Across Multiple Species", ICASSP.2006 Proceedings, 2006.
48. Xia Mao et al; (2007), "Mandarin speech emotion recognition based on a hybrid of HMM/ANN", INTERNATIONAL JOURNAL of COMPUTERS Volume 1, pp.321-324, 2007.
49. Milan Sigmund, "Voice Recognition By Computer", Tectum Verlag publication, pp.20-22.
50. Stankovic et al; (2011), "Improvement of Thai speech emotion recognition by using face feature analysis", Proceedings of the Nineteenth IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2011), Chiang Mai, Thailand, December 7-9, pp. 87, 2011.