

**GENDER DEPENDENT WORD-LEVEL EMOTION
DETECTION USING GLOBAL SPECTRAL SPEECH
FEATURES**

HARIS SIDDIQUE

**MASTER OF SCIENCE
UNIVERSITI UTARA MALAYSIA
2015**

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Malaysia

Abstrak

Dalam kajian ini , ciri-ciri spektrum global dipetik daripada perkataan dan ayat yang dikaji untuk pengecaman pertuturan emosi. MFCC (Mel Kekerapan Cepstral Pekali) telah digunakan sebagai maklumat spektrum untuk tujuan pengecaman. Ciri spektrum global mewakili statistik kasar seperti purata MFCC digunakan. Kajian ini juga mengkaji perkataan di kedudukan yang berbeza (di awal, pertengahan dan akhir) secara berasingan dalam ayat. Ciri pengekstrakan tahap perkataan digunakan untuk menganalisa prestasi pengecaman emosi berdasarkan perkataan di kedudukan yang berbeza. Sempadan perkataan dikenalpasti secara manual. Model berdasarkan jantungina atau model bebas jantungina juga dikaji untuk menganalisa kesan jantungina ke atas prestasi pengecaman emosi. Berlin Emo - DB (Pangkalan Data emosi) telah digunakan sebagai set data ucapan beremosi. Prestasi pengklasifikasi-pengklasifikasi yang berbeza juga dikaji. NN (Rangkaian Neural), KNN (K - Jiran Terdekat) dan LDA (Analisa Diskriminasi Linear) adalah pengklasifikasi yang digunakan. Emosi kemarahan dan neutral juga dikaji. Keputusan menunjukkan bahawa, dengan menggunakan semua 13 pekali MFCC memberikan hasil yang lebih baik daripada pengelasan gabungan lain pekali MFCC untuk emosi yang dinyatakan. Perkataan-perkataan di kedudukan permulaan dan berakhir menandakan posisi emosi lebih baik daripada kandungan emosi di kedudukan pertengahan. Prestasi model berdasarkan jantungina adalah lebih baik daripada jantungina model bebas jantungina. Selain itu, wanita adalah lebih baik daripada lelaki dari segi mempamerkan emosi. Secara amnya, prestasi NN adalah paling teruk daripada KNN dan LDA dari segi klasifikasi emosi marah dan neutral. Prestasi LDA adalah lebih baik daripada KNN sebanyak hampir 15% dengan menggunakan model bebas jantungina dan hampir 25% menggunakan model berdasarkan jantungina.

Kata Kunci: Koefisyen Kekerapan Mel Cepstral, pengekstrakan ciri, pengecaman ucapan beremosi, korpus simulasi ucapan beremosi, model klasifikasi

Abstract

In this study, global spectral features extracted from word and sentence levels are studied for speech emotion recognition. MFCC (Mel Frequency Cepstral Coefficient) were used as spectral information for recognition purpose. Global spectral features representing gross statistics such as mean of MFCC are used. This study also examine words at different positions (initial, middle and end) separately in a sentence. Word-level feature extraction is used to analyze emotion recognition performance of words at different positions. Word boundaries are manually identified. Gender dependent and independent models are also studied to analyze the gender impact on emotion recognition performance. Berlin's Emo-DB (Emotional Database) was used for emotional speech dataset. Performance of different classifiers also been studied. NN (Neural Network), KNN (K-Nearest Neighbor) and LDA (Linear Discriminant Analysis) are included in the classifiers. Anger and neutral emotions were also studied. Results showed that, using all 13 MFCC coefficients provide better classification results than other combinations of MFCC coefficients for the mentioned emotions. Words at initial and ending positions provide more emotion, specific information than words at middle position. Gender dependent models are more efficient than gender independent models. Moreover, female are more efficient than male model and female exhibit emotions better than the male. General, NN performs the worst compared to KNN and LDA in classifying anger and neutral. LDA performs better than KNN almost 15% for gender independent model and almost 25% for gender dependent.

Keywords: Mel Frequency Cepstral coefficients, Feature extraction, emotional speech recognition, simulated emotional speech corpus, classification models

Table of Contents

Permission to Use	i
Abstrak	ii
Abstract	iii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
CHAPTER ONE INTRODUCTION	1
1.1. Background.....	1
1.1.1. Dimensions of Emotions.....	2
1.2. Problem Statement	3
1.3. Research Questions	6
1.4. Research Objectives.....	6
1.5. Motivation	6
1.6. Contribution	7
1.7. Scope	7
CHAPTER TWO LITERATURE REVIEW	9
2.1. Introduction.....	9
2.2. Databases	10
2.2.1. Acted Emotional Speech Database	11
2.2.2. Elicited Emotional Speech Database	12
2.2.3. Natural Emotional Speech Database	13

2.3. Speech Features	16
2.3.1. Excitation source feature	19
2.3.2. Vocal tract features	21
2.3.3. Prosodic features	24
2.3.4. Works using combined features	28
2.4. Feature extraction	29
2.4.1. Frame Blocking	32
2.4.2. Windowing	32
2.4.3. Fast Fourier Transform (FFT).....	32
2.4.4. Mel-Frequency Wrapping	33
2.4.5. Cepstrum	33
2.5. Classifier	34
2.6. Linear and nonlinear classifier	37
2.7. Gender based models	38
2.8. Word boundary detection	39
2.9. Summary	41
CHAPTER THREE METHODOLOGY	43
3.1. Introduction.....	43
3.2. Framework	43
3.3. Speech Corpus	44
3.4. Pre-processing	47
3.5. Feature Extraction	48

3.6. Classifier	50
3.7. Evaluation of model	51
3.8. Summary	52
CHAPTER FOUR ANALYSIS OF RESULTS AND DISCUSSION	53
4.1. Introduction.....	53
4.2. Speech Corpus	53
4.3. Feature Extraction	54
4.4. Classifier	59
4.5. Gender independent and gender dependent models	60
4.5.1. Emotion recognition using utterance level global MFCC	61
4.5.2. Emotion recognition using word-level global MFCC	68
4.6. Summary	74
CHAPTER FIVE CONCLUSION	76
5.1. Introduction.....	76
5.2. MFCC Coefficients	76
5.3. Word level	77
5.4. Gender dependent and independent	77
5.5. Future work	77
References	78

List of Tables

Table 2.1: Emotional Speech Databases.....	15
Table 2.2: Excitations features used for different speech functions.....	21
Table 2.3: Literature on emotion recognition using spectral features.....	24
Table 2.4: Literature on emotion recognition using Prosodic features.....	26
Table 2.5: Literature on emotion recognition using combined features (source, system or prosodic).....	28
Table 2.6: Use Of First 13 MFC Coefficients In Previous Studies	31
Table 2.7: Literature on some common classifiers used for Emotion Recognition.....	35
Table 3.1: Speaker Information.....	45
Table 3.2: Code of text used in database.....	45
Table 3.3: Code of Emotions.....	46
Table 3.4: Example of file codes and values.....	48
Table 4.1: Accuracy of different classifiers in percentage	60
Table 4.2: Emotion recognition performance at utterance level	61
Table 4.3: Emotion recognition performance at utterance level while using 6, 9, 10 and 5, 8, 10 MFCC	64
Table 4.4: Emotion recognition performance at utterance level while using 5, 6, 8, 9, 10 and 11 MFCC	65
Table 4.5: Emotion recognition performance at word-level	69
Table 4.6: Performance of different MFCC combinations of initial words for general model in percentage	71

List of Figures

Figure 2.1: MFCC General Framework.....	32
Figure 3.1: Framework	43
Figure 3.2: Detecting Word Boundary	49
Figure 4.1: Difference of MFCC values for all words	54
Figure 4.2: Difference of MFCC values for all words while neglecting first 4 MFCC	55
Figure 4.3: Difference of MFCC values for all words while using 5,6,8,9,10 and 11 MFCC.....	56
Figure 4.4: Difference of MFCC values for all words while using 6,9 and 10 MFCC	57
Figure 4.5: Difference of Variance for all mean values of all words	58
Figure 4.6: Difference of MFCC values for all words for female	58
Figure 4.7: Difference of MFCC values for all words for male	59
Figure 4.8: Performance Analysis of classifiers for different models	60
Figure 4.9: Utterance level performance of the models using KNN and LDA	62
Figure 4.10: Difference of emotions using mean of 5,8 and 10 MFCC	63
Figure 4.11: Difference of emotions using mean of 6, 9 and 11 MFCC	63
Figure 4.12: Utterance level performance of the models using KNN and LDA	65
Figure 4.13: Utterance level performance of the models using KNN and LDA	66
Figure 4.14: Performance of Study A and B	67
Figure 4.15: Performance of Study A and B	68
Figure 4.16: Emotion recognition performance of initial words	70

Figure 4.17: Emotion recognition performance of Middle words	70
Figure 4.18: Emotion recognition performance of Ending words	71
Figure 4.19: Performance of different MFCC combinations of initial words for general model.....	71
Figure 4.20: Performance of Study A and B	72
Figure 4.21: Performance of Study A and B	73

CHAPTER ONE

INTRODUCTION

1.1 Background

Speech is the fastest medium of presenting messages in face-to-face communication. On the other hand, emotions are also another medium of communication, for example, a smaller set of gestures can describe the individual's emotional state to others. One debatable topic on it is that a smile or laugh is treated as a signal of happiness in all civilizations. Whereas, crying is treated every bit a sign of sadness or heartbreak, their assessments can change from culture to culture (Lewis et al., 2008).

Most human has emotional activities and we often produce emotions in our free time. Usually we take novels, films, music or other plays as a root of our amusement. These all portray real emotions, but about unreal events, and beside we know that these are fictions or unreal, we make an emotional attachment with them. Reason is the means we used to interact with or understand with everyday real world, we practice the same method with these informants, that's why a particular music makes us happy and other character of music makes us sad (Lewis et al., 2008).

According to Ling He (2010) emotions are psychological and physiological states that take in actual and liberated responses. Emotions comprise person's state of mind and the way a person interacts with others as well as with the environment. Sometimes emotion is related to 'mood'. Normally emotion is short-timed physiological and psychological state that last from a few minutes to a few hours, whereas mood is long-timed emotional state that can last from hours to weeks (Ling He, 2010).

Perceiving of emotions through speech acoustics is now a hot topic to research. In that location is substantial evidence now that can show specific acoustic features that are related to the emotions of the speaker system. More pointed question in this area is which acoustic features are associated with which discrete emotional state of the person, because reliable association with different emotional state is still unconvincing (Lewis et al., 2008). One cannot say that this particular feature is for this emotional state.

1.1.1 Dimensions of emotion

There are many difficulties while detecting and classifying the emotions into different categories. (Hanjalic, 2005; Ling He, 2010) mentioned three dimensions that can be utilized to classify emotions into different classes. These dimensions include Valence, Arousal and Control. Valence is described as different emotional states ranging from positive to negative. This can be related as types of emotions. Arousal describes the energy of emotion. Arousal also can be related as the intensity of the emotion. Control dimension is applied to present emotions that have similar valence and arousal. This dimension ranges from ‘no control’ to ‘full control’. E.g is differentiating between grief and rage emotions.

To communicate in a social environment, emotion perception is important. Researchers have instituted that for human-to-human interaction, emotional expertise can play a vital role as a component of intelligence (Ramakrishnan & Emary, 2011). Although human-to-computer interaction has been studied for years and it is different from human-to-human interaction, human-to-computer interaction does follow the basics of human-to-human interaction (Wu et al., 2009b). Another study mentions that only 10% of human life are unemotional (Ramakrishnan & Emary, 2011). Due to new application growths with

respect to human-computer interaction, there is an advanced development in emotional cues investigation (Ramakrishnan & Emary, 2011). From all these studies, there is a need to make human-computer interaction more sense-and-feel rather than point-and-click.

To make human-computer interaction more real, careful considerations should be made on how emotions are encoded and communicated through speech parameters. Although there is no exact answer to the question of what the ‘correct’ emotion is for a given speech sample, it is possible to predict some emotions from prosodic information of the given speech. Recent advances in speech processing, computer and human’s positions are redefined about each other (Wu et al., 2009b).

1.2 Problem Statement

Speech signals provide information between speakers and provide information about emotions, feelings, attitude, personalities, mental state and stress level of the speakers (Wu et al., 2009b; John et al., 2012). If speakers will not able to understand or detect each other’s emotions, human-to-human communication will not be as effective. Human-to-machine communication suffers with the same case that is used in most of Human-computer interaction applications because of the inability of the machine to understand or generate emotions (Ramakrishnan & Emary, 2011).

In recent years, many researchers worked on emotion detection using speech for better Human-computer interaction (Rao et al., 2010; Chauhan et al., 2010; Staroniewicz, 2011; Vicsi & Sztahó, 2011; Koolagudi et al., 2011; Han et al., 2012; Rao et al., 2013; Sethu et al., 2013). Selection of features that are extracted from speech and at which level they are extracted is very crucial (Ramakrishnan & Emary, 2011). Most researchers used speech

features (spectral or prosodic) at frame-level (Iliou & Anagnostopoulos, 2009b; Chauhan et al., 2010; Khanna & Kumar, 2011; Staroniewicz, 2011) for emotion recognition task. Some researchers have used prosodic information of the speech at different class-levels (sentence, word or syllable) (Koolagudi et al., 2011; Zhang, 2012; Rao et al., 2013). While very few studies have been reported on the use of spectral features at class-level (words, vowels and consonants) (Bitouk et al., 2010; Koolagudi et al., 2011). Koolagudi et al. (2011) stated that most recognition systems used frame-level feature extraction but Bitouk et al. (2010) stated that class-level (words, vowels and consonants) feature extraction provide higher classification rate than the frame - level.

The most commonly used spectral features for emotion recognition are Mel-Frequency Cepstral Coefficients (MFCC) (Han et al., 2012; Koolagudi & Rao, 2012; Khanna & Kumar, 2011; Mao et al., 2009). Ramakrishnan and Emary (2011) stated that MFCC is “new standard” for recognizing emotions. No written report, if whatever, has used MFCC features at word level. One cannot neglect the importance of this spectral feature in speech processing task at different class-levels.

In speech emotion classification task, data sets are important need. Male and female have different speech qualities from each other and because data sets used for emotion detection systems are speech files, difference of gender affects the system accuracy. As shown in some studies (Rong et al., 2009; Rao et al., 2013), gender dependent models have better accuracy in opposite of gender independent. From the results mentioned in given references, it can be seen clearly that male and female models have better results than of general (male and female mix) model. Moreover, according to the results

mentioned in Bitouk et al. (2010) spectral features of vowel and consonant levels have better performance than traditional frame-level.

Koolagudi et al. (2011) and Rao et al. (2013) conducted studies of emotion recognition at word level. Both used prosodic features with Support Vector Machine (SVM) classifier. Whereas Koolagudi et al. (2011) detected word boundaries manually and Rao et al. (2013) detected word boundaries automatically. Rao et al. (2013) reported low classification rate for anger and neutral, whereas, Koolagudi et al. (2011) reported slightly better classification rate. Ayadi et al. (2011) stated that the average classification rate of SVM for emotional speech recognition systems is 75% to 81%. Whereas classification rates given in Koolagudi et al. (2011) and Rao et al. (2013) were lower than average i.e 63.5% and 47.5% respectively that can't be used for Human Computer Interaction (HCI) applications. Reason of such classification rates may be used of prosodic features at word level as Zhou et al. (2009), Kuchibhotla et al. (2014) and Koolagudi and Rao (2012) stated that prosodic features are not as effective as spectral features for emotion recognition systems. Use of the classifier can also be another reason of such classification rates as Ramakrishnan and Emary (2011) stated that short-term features, that were used in Koolagudi et al. (2011) and Rao et al. (2013), requires the dynamic classifier like Hidden Markov Models (HMM).

This study tends to obtain better classification rate in the field of Speech Emotion Recognition (SER), it proposes the use of Gender dependent emotion recognition model in which spectral features will be extracted at word-level and different classifier will be studied. Due to lack of research at word-level among other class-levels, as discussed above, word-level feature extraction is proposed in this study.

1.3 Research Questions

1. How to detect word boundaries in a spoken sentence?
2. What speech features would represent emotions?
3. How to classify emotions based on extracted features?

1.4 Research Objectives

This study provides better emotional speech recognition model by using the spectral speech features at word level. The effect of different genders on emotion classification is also studied. To achieve this main objective, three objectives are considered for this particular study.

1. To develop improved emotional speech recognition model that uses the concept of word level feature extraction.
2. To identify the best combination of MFCC features to be extracted for recognizing specific emotions based on at word level.
3. To study the behavior of the linear and nonlinear classifiers for the classification of emotions.

1.5 Motivation

Most researchers used prosodic information at a different class - level for emotion recognition as described in section 1.2 (Koolagudi et al., 2011; Zhang, 2012; Rao et al., 2013). According to Koolagudi and Rao (2012), spectral features perform better as compared to the source (excitation) and prosodic features in emotion recognition task. It also mentioned that of high arousal emotions like anger, fear, happiness, etc., prosodic

features perform better in the recognition process. After combining these two features, there is a notable improvement in performance. This study is also motivated by the findings mentioned by Koolagudi et al., (2011) that very less work has been done on features extraction at different class-level. Moreover, Han et al., (2012) and Wang et al. (2011) stated that Neural Network (NN) is less time consuming and have better recognition performance in speech emotion recognition; K-Nearest Neighbor (KNN) is the simplest classifier that uses the idea of similar observation belongs to similar classes (Kuchibhotla et al., 2014; Pao et al., 2008) and Linear Discriminant Analysis (LDA) can be used to reduce the dimensionality of classification process (You et al., 2007).

1.6 Contribution

This study will contribute in the use of spectral features at word level for emotion recognition task while using NN, KNN and LDA as classifiers. This study will provide gender effect on emotional speech recognition. Moreover, performance of male and female can also be found in this study. Different word positions for more efficient emotion recognition system will be mentioned in this study. Moreover, efficient MFCC coefficient combinations will be mentioned in this study.

1.7 Scope

This study focused on the use of spectral speech features extracted at word-level for emotion recognition. As very little work has been done at word level as mentioned in previous sections for emotion recognition, word-level will be used to extract speech features for this particular study. Only two emotions, Anger and Neutral are considered for this study. This study uses secondary data for processing, as it is difficult to collect

voiced data for different emotions. Berlin Emotional Database (Emo-DB) will be used for this study. As voiced data in the database are in sentence form, word boundaries will be identified manually for further processing.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Speech is most common and natural medium to communicate with other humans. It not only contains the information about the message, but also about emotions, speaker and language. Most of the speech systems are efficient for studio recorded, acted emotional speech, but are not efficient for real life's emotional speech. Here the classification rate of the system in the testing phase is different from in the training phase (Khanna & Kumar, 2011; Vicsi & Sztahó, 2011; Ramakrishnan & Emary, 2011; Anagnostopoulos & Iliou, 2010). One cause of this can be difficulty in exhibiting and classifying emotions existing in speech. Emotions make speech real and natural.

A human can understand the message through the emotion expressed in the speech and other nonlinguistic cues like gesture, posture in case of video and punctuation in the case of written text. This study restricts itself to the emotions in the speech. The basic goals of every speech processing system should be how to understand the emotions that are present in the speech and how to classify correct emotion for a small interval of the speech that best describe the message.

This chapter gives some literature review on Speech Emotion Recognition with respect to emotional databases, which are used to develop recognition systems, features extracted from speech that describe the emotions and classification methods for recognizing the emotions. Human computer interaction in emotional speech has given a boost to this area of research.

2.2 Databases

Currently there is no benchmark for the data sets used for speech emotion recognition systems (Koolagudi & Rao, 2012). There is no such database, which can be widely used as the research standard. Most of researchers use prerecorded, simulated databases for their research (Henríquez et al., 2011; Staroniewicz, 2011; Han et al., 2012). This lead to a great difference in the availability of the databases in terms of number of speakers, number of emotions, type of the data (acted, natural), setup for the recording, the purpose of the research (recognition of the speaker or emotion) and the language used (German, Polish or Chinese according to above mentioned references). Ververidis and Kotropoulos (2006), conducted a great deal of research for summarizing 64 emotional speech databases which have 29 different emotions in different language (Ververidis & Kotropoulos, 2006). Research conducted by Ververidis and Kotropoulos (2006) comprises all the databases that exists at that time.

Databases used for speech emotion recognition can be separated into three types based on data collected in them (Ververidis & Kotropoulos, 2006).

1. Acted emotional speech database (Simulated)
2. Elicited emotional speech database
3. Natural emotional speech database

2.2.1 Acted emotional speech database (Simulated)

In this type of database, data are collected from proficient artists. Artists are asked to express the emotions in different types by reading some sentences. Through this approach, one can collect a vast range of emotions easily. LDC (Linguistic Data Consortium) speech corpus (Bitouk et al.,2010), Emo-DB (Emotional Database) (Burkhardt et al.,2005), are some examples of such databases.

LDC comprises 7 professional actors (4 male and 3 female) which simulate 15 different emotions. Each actor simulates 10 utterances. This database is in English language. This database is not free but is available on a commercial basis with some fee that is why most of researchers do not prefer this database (Bitouk et al., 2010; Sethu et al., 2013).

Emo-DB is simulated emotional speech database that is recorded in German Language. It comprises 10 professional actors, 5 male and 5 female. Male actor's age ranges 25 to 32 and female actor's age ranges 21 to 35. Seven different emotions sampled in this database includes anger, boredom, disgust, fear, happiness, sadness and neutral. Each actor is asked to simulate 10 different daily life pre-selected sentences in each emotion. There is a total of 800 utterances in this database that comprises 10 actors for 7 emotions in 10 different sentences each, plus few second versions of few emotions. Sennheiser MKH 40 P48 microphone was used for recording with 16kHz sampling frequency (Rao et al., 2013). After listening test by 20 humans few utterances were deleted on the basis of quality and nature. Utterances with 80% discrimination ratio and over 60% naturally are kept in the corpus. After test, 535 utterances were selected for 7 emotions in the corpus (Zhang, 2012). Number of human evaluators i.e "20" is discussed in Bitouk et al., (2010) for conducting listening tests. This database is free for public use.

Most researchers used Emo-DB for their research (Gaurav, 2008; Bitouk et al., 2010; Henríquez Rodri'guez et al., 2012; Zhang, 2012; Koolagudi & Rao, 2012; Henríquez et al., 2011; Rao et al., 2013). In Gaurav (2008), 120 and 78 (198 total) utterances from Emo-DB were chosen for anger and neutral emotions respectively. For anger, 40 utterances (33.33%) were used as training set whereas remaining 80 (66.67%) utterances were used for testing the models. In the case of neutral, 40 (51.28%) utterances were used for training and remaining 38 (48.72%) utterances were used for testing. In Rao et al. (2013), datasets are divided based on the number of speakers. Speech data of 8 speakers are used for training and remaining 2 speaker's speech data is used as a testing. This type of model is called speaker independent model because training and testing performed by different speakers.

Some advantages of such databases are they can commonly be used as mentioned above; their result can be compared easily as emotions are labeled in the databases. A wide range of languages is available for such type of databases (Ververidis & Kotropoulos, 2006). The disadvantage of such databases is that data in such databases are periodic in nature, which is not accurate in actual or real conditions. Such database comprises recited speech not spoken speech.

2.2.2 Elicited emotional speech database

Elicited emotional speech databases are collected by creating fake situations for the speaker without the speaker's knowledge. The speaker is involved in an emotional conversation with another subject (human or machine). The other subject generates different situations for the speaker and through the conversation; different emotions are stimulated without speaker's knowledge. Such databases are natural as compared to

simulated ones, but sometimes speakers cannot be as expressive as before, if they identify that they are actually being recorded.

One example of such database can be seen in Wu and Liang (2011). A total of eight members, in two groups selected for data collection. Group A with 6 members and group B with 2 members. Two computer systems were made; one was a dialog system that used to collect data regarding daily life of group A members by asking questions about day-to-day life for a period of one month. Another system was a computer game, which was used to collect data of group B. Emotional feedback was stored in the system while members were busy in the game. 2033 utterances were collected in laboratory atmosphere. Sampling rate was 16KHz which comprises neutral, happy, sad and angry emotions in both groups.

Such database gives almost natural emotional speech data but contains artificial information. Not all emotions may present in the database. If speaker gets to know that they are being recorded then the records will be non-natural.

2.2.3 Natural emotional speech database

A natural conversation in some situations is used as the data in this type of database. Such database may or may not contain all emotions. Sometimes it is difficult to recognize these emotions clearly. Such data may be recorded from a call center (Vicsi & Sztahó, 2011), helicopter or airplane cockpits (Hansen et al., 1997), conversation between doctor and patient and so on. The availability of vast range of emotions is difficult sometimes in such databases. In addition, there is a legal issue to build such databases such as copyright or privacy. Such databases are completely natural and expressive that is useful

for real world applications. Not all the emotions are present. Different utterances may overlap each other. It may contain irrelevant data.

Database design and emotion collection heavily depend on the purpose for which it is going to be used. Emotional speech corpus of one single speaker can be used for emotional speech synthesis but for the recognition of emotion, corpus of multi-speaker expressing different emotions will be needed. According to Douglas-Cowie et al. (2003) some general issues regarding speech corpus should be considered. These issues address the need of corpus based on language, method to collect the data and number of emotions present in the corpus.

- Database scope should be decided properly in-terms of number of participants taking part in recording and in-terms of emotions that to be recorded in the database.
- Type of the databases should be decided properly also that is natural, acted or elicited. This can help developers to classify database in-terms of its applications and quality.
- Labeling of the emotions is a big task to perform in the database development. Developers should take utmost care while labeling the emotions. An acceptable methodology would be, getting this task using multiple experts and should choose a decision that is given by the majority of the experts.
- Database size plays an important role in speech emotion recognition. Size define the reliability, scalability and generalize the ability of the developed system.

Table 2.1 describes about the set of emotional speech databases that are found during literature review. The table shows that very few corpuses are found for the German, Chinese and Polish language. English language dominated the corpuses in the table.

Table 2.1

Emotional Speech Databases.

#	Corpus	Language	Access	Size	Source	Emotions
1	EmoDB Burkhardt et al. (2005)	German	Public and free	800 utterances (10 actors X 7 emotions X 10 utterances + some second versions)	Professional Actors	Anger, joy, sadness, fear, disgust, boredom, neutral
2	LDC Emotional Prosody Speech and Transcripts Bitouk et al. (2010)	English	Commercially available	7actors X 15 emotions X 10 utterances	Professional actors	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
3	SUSAS Hansen et al. (1997)	English	Public with license fee	16,000 utterances, 32 actors (13 females + 19 males)	Speech under simulated and actual stress (Natural and acted)	Anger, Natural, Loud, Lombard
4	Simulated Speech Database Wu et al.(2011)	Chinese	Own Research purpose	2,033 utterances, 8 actors, two groups	Computer Simulations (Elicited)	neutral, happy, angry, and sad
5	Acted emotional speech database Piotr. (2011)	Polish	Own Research purpose	2118 utterances, (13 actors X 10 utterances X 7 emotions X several repetitions)	Acted	Anger, Sadness, Happiness, Fear, Disgust, Surprise And Neutral

The main purpose of the emotional speech recognition is to automatically identify the emotional state of the speaker. Whereas data used in this field as a sample is very crucial.

As mentioned before, there are three types of databases used for this purpose (Acted or Simulated, elicited and natural), each has its own advantages and disadvantages.

According to Ververidis and Kotropoulos (2006), there are copyright issues in collecting data for speech corpus as many TV channels or radio broadcasters have issues to share their data. One such example can be seen in (Hansen et al., 1997), SUSAS was created to represent actual stress and emotional speech data. Cockpit recordings are used to represent stressed situations on the other hand, the actors were asked to say pre-defined words while riding on a roller coaster for the emotional speech database. On the other hand, in a studio atmosphere selected emotions can be easily collected by asking professional actors to record. Another reason is; natural emotions are difficult to classify as opposite to simulated ones (Anagnostopoulos & Iliou, 2010). Moreover, researchers studying emotions in speech find it hard to collect natural data for underlined emotions and recorded data samples are useful for the emotional speech recognition task (Ramakrishnan & Emary, 2011). Berlin emotional speech database (Emo-DB) is internationally known and used for emotional speech processing (Koolagudi & Rao, 2012; Rao et al., 2013).

2.3 Speech Features

Feature selection for emotion speech recognition is a big task to perform. This decision is very crucial of the underlying emotions. As its importance described by Ramakrishnan and Emary (2011), extracted features should efficiently characterize emotions. Chosen features should represent the intended information for the different emotions. Different features provide different information in overlapped ways for different emotions. Proper set of feature vectors affects the emotion classification.

In machine learning, features are measurable attributes of samples. Extracting discriminating features has been always essential for any pattern recognition application, including emotion recognition. As research on Speech Emotion Recognition (SER) moves towards the exploitation of an increasingly complex feature space, specific features used by different works vary significantly. Nevertheless, from the speech production point of view, source (excitation) and system (spectral) features are almost different. From the feature extraction point of view also, source, system, and prosodic features are different. System features represent the response of vocal tract system derived from the autocorrelation analysis, which mostly represents the first and second order correlations among the speech samples. On the other hand, excitation source features are represented by a sequence of linear prediction (LP) residual samples, representing higher order correlations. Prosodic features represent duration, intonation and intensity patterns present in the speech signal (Koolagudi & Rao, 2012).

Therefore, in view of feature extraction, speech production and perception aspects source, system and prosodic features are non-overlapping. In this regard, the emotion specific information carried by these features may also be non-overlapping in nature (Koolagudi & Rao, 2012). Based on this intuition, in this research, we have explored the contribution of emotion specific information from the system (Spectral) features at word level. There are two most widely used feature groups: Prosodic and Spectral. They are introduced here with main focus on the feature extraction technique that will be used in this research.

Speech features that are used for the information extraction from speech signals and the area of the analysis that will be used to extract features from are very important. Many researchers have followed the common method to extract features by dividing the signal

into a number of small intervals (Khanna & Kumar, 2011; Staroniewicz, 2011; Koolagudi & Rao, 2012). These intervals are known as ‘frames’. Feature vectors that are extracted from these frames independently are called local features. Some researchers used statistical features of the whole utterance of the speech signal (Henríquez et al., 2011; Han et al., 2012; HenríquezRodri’guez et al., 2012). These features are called global features. Global features are extracted as the whole statistical values of the all local features extracted from one utterance.

Researchers have different opinions on the selection of feature type for the emotion speech recognition (Rao et al., 2013). Moreover, Rao et al. (2013) stated that many researchers prefer global over local features. Classification time for executing global features is also very low as compared to local features. One more advantage of global features has is number of features are less than local ones. However, from the results shown in Iliou and Anagnostopoulos (2009), we can see that global features are only efficient in distinguishing the high arousal emotions e.g happy, anger, and fear with low arousal emotions e.g sad, neutral. Researchers have argued that global features fail to classify the emotions in same arousal e.g anger with happy or sad with neutral (Rao et al., 2013). One more disadvantage of the global features is that, because features are chosen on statistical basis, temporary information present in the speech signals is lost completely.

Local features are efficient in training a large number of vectors. Some classifiers such as Hidden Markov Model (HMM) and Support Vector Machine (SVM) may not be as efficient for global features as for local features. Local features may lead to higher efficiency for such classifiers, as these classifiers require parameters in large number.

Different researchers have used different features for their research. These features can be divided into three main types (Koolagudi et al., 2011). These types are as follows.

- I. Excitation source features (source features).
- II. Vocal tract features (system/spectral features).
- III. Prosodic features.

2.3.1 Excitation source features

Features that are derived from the excitation source from the signal are known as source features. Excitation signal is derived from a speech by extracting vocal tract (VT) characteristics from the speech. First, speech is filtered using Linear Prediction Coefficients (LPCs). This LPC provides VT information from speech. Then, VT is separated from the speech by inverse filter process. The resulting signal is called Linear Prediction Residual (LP residual). The importance of excitation source features is first mentioned by Makhoul (1975). LP residual mostly contains the information about source features. Other excitation source features that are extracted from the signal are the features of glottal pulse, glottis phases (open or close) and so on.

The use of the LP residual and glottal volume velocity (GVV) features for the recent research can be seen in Kodukula (2009). Many researchers have neglected the excitation source features for the speech systems. One reason for this can be popularity of spectral and prosodic features. Moreover according to Bajpai and Yegnanarayana (2008), LP residual comprises high order of connections between its samples. These samples may be

captured till some extent by using excitation strength, GVV, glottal pulse shapes and phases as discussed above and so on.

Studies regarding excitation source features of speech tell that sources features have almost all information for all formats of speech systems such as message identification, speaker identification, language specific information and emotion identification information. Some of the literature is given in Table 2.2.

Table 2.2 indicates that excitation source features are as important as spectral and prosodic features. Many researchers for emotion speech recognition did not explore source features as they have done with spectral and prosodic features. Source features may have important information for emotions in speech. Very little work has been done on excitation source features (Iliev et al., 2010; Chauhan et al., 2010; Koolagudi et al., 2010). According to Yegnanarayana et al. (2009) and Bapineedu et al. (2009) excitation source features does provide all information for speaker identification, message identification and emotion classification but according to Kodukula (2009), the performance of speech emotion recognition systems, that uses excitation source features, depends on the accuracy of LP analysis in the VT response, quality of speech signals and order of LP analysis. Moreover, the time varying nature of VT, invariably and existing methods to extract VT information adds ambiguity of LP information (Kodukula, 2009). On the other hand, use of glottal features like glottal pulses and glottal phases requires extra hardware for recognition systems thus increases the complexity of the system.

Table 2.2

Excitations Features Used For Different Speech Functions.

#	Excitation Features	Purpose
1	LP residual Rao et al. (2007a)	Used for detecting the excitations from speech.
2	LP residual Yegnanarayana et al. (2009)	Used for determining the delay between segments of different speakers in multi-speaker environment.
3	LP residual Bapineedu et al. (2009)	Used for characterizing the loudness, lombard effect, speaking rate, and laughter segments.

2.3.2 Vocal tract features

As discussed earlier features are extracted from speech in the form of frames. Normally spectral system features are extracted using 20-30 ms frame window. Mostly these features are extracted from frequency domain of speech signal. A short time spectrum can be obtained by Fourier transform of the speech signal. These spectrum features normally include formants, spectral energy, slope and the bandwidth of formants. Whereas cepstral features of the signal is obtained via taking a fourier transform on log magnitude spectrum mentioned in the book “Fundamentals of speech recognition” (Rabiner & Juang, 1993). MFCC (Mel frequency cepstral coefficients), LPCC (Linear prediction cepstral coefficients), PLPC (Perceptual Linear Prediction Coefficients) and formant features are features derived commonly from cepstral domain. These features are also known as spectral or system features and represents the VT information. Many researchers for various speech-processing tasks like speech recognition, speaker recognition, emotion recognition from speech, stress recognition from speech and so on have used spectral features.

In Mubarak et al. (2005), researchers used MFCC to differentiate speech and non-speech information from speech. They have witnessed that phonetic (speech) information is present in lower order MFCC features, whereas higher order MFCC features presents non-speech or music information.

The importance of lower order MFCC features as described by Mubarak et al. (2005) used in Neiberg et al. (2006). Low frequency ranging 20Hz to 300 Hz is proposed to model pitch. Neiberg et al. (2006) used MFCC, MFCC-low (pitch model) and plain pitch as feature sets for classifying emotions from two different speech corpuses. English and Swedish emotional speech databases are used for the speech input. Neiberg et al. (2006) also indicate that MFCC-low has better performance than plain pitch features for emotion recognition.

Different spectral features like MFCC, LPCC, LFPC (log frequency power coefficients) are used as feature set to distinguish anger with neutral emotion (Pao et al., 2007). By increasing popularity of MFCC and LPCC, Kamaruddin and Wahab (2009) aimed to test LFPC in comparison with MFCC and LPCC. Kamaruddin and Wahab have found that LFPC perform better than MFCC and LPCC.

A great deal of research conducted in Bitouk et al. (2010), the authors used MFCC for three classes namely stressed vowels, unstressed vowels and consonants for emotion recognition based on independent speaker. So far, researchers were using frames to extract features from speech signals but in this study, Bitouk et al. used class-level spectral features that are described in the begging. The result shows that class-level features outperform the prosodic and utterance-level spectral features. Moreover, the

combination of these class-level and prosodic features gives more efficient results than before. Bitouk et al. (2010) stated that consonant class contains more emotion related information than stressed and unstressed vowel features.

Researchers are treating spectral features as strong correlated to the rate of change in articulator movement and to varying shapes of the VT (Khanna and Kumar, 2011). According to Bitouk et al. (2010), the average classification rate of emotion recognition systems of spectral features is inversely proportional to the length of utterance whereas Koolagudi and Rao(2012) stated that MFCCs are the only series of real numbers that does not provide any meaning to the real world, but it can be used to develop efficient emotion recognition systems and, spectral features outperform source and prosodic features in classification process. Due to the common usage of MFCC, Ramakrishnan and Emary (2011) stated that MFCC are “new standard” for recognizing emotions. Table 2.3 mentioned some work on the system features for emotion classification. It is clearly observed that most of the researchers use frame-level signal processing approach, in which whole speech signal is processed frame by frame. Considering 20ms frame size and 10ms of the gap between frames, information is only retrieved from frames whereas one cannot neglect the importance of information present in the gap. One study has emphasized the importance of information present in the whole signal (Bitouk et al., 2010). Bitouk et al. (2010) have shown that information in vowels and consonants is also more important to be extracted for emotion recognition.

Table 2.3

Literature on Emotion Recognition Using System Features.

#	System Features	Purpose
1	MFCC features Mubarak et al. (2005)	MFCC's higher order and lower order information is discriminated. It's been shown that higher order MFCC contains music information and lower order MFCC contains speech information.
2	MFCC and MFCC-low features Neiberg et al. (2006)	MFCC-low and plain pitch feature's efficiency is given here and mentioned that MFCC-low is more efficient for emotion recognition than plain pitch features.
3	MFCCs, LPCCs and LFPCs features Pao et al. (2007)	Classification of 2 emotions in Mandarin language. Anger and neutral emotions are classified.
4	LFPC features Kamaruddin and Wahab (2009)	It's been shown that LFPC performs better than other system features like MFCC and LPCC.
5	MFCC features from consonant, stressed and unstressed vowels (class-level MFCCs) Bitouk et al. (2010)	Opposed to many researchers, authors used class-level speech features to denote emotion information rather than frames.
6	MFCC features Khanna and Kumar (2011)	MFCC features are used for classify 6 emotions. Experiment conducted on Danish emotional database.
7	MFCC features Staroniewicz (2011a)	EER (Equal error rate) is been calculated for speaker basic emotion. In total, 7 emotions were classified.

2.3.3 Prosodic Features

Prosodic reflects many speech features that tell about the speaker or speech itself e.g. it can tell about a form of utterance whether it is a statement, a question or a command. One study has given the importance of these prosodic features as they provide meaningful information for emotion classification for the speech signal (Rao et al., 2010). In this context, Rong et al. (2009) have conducted a survey which tells about the different prosodic features many researchers had used for emotion speech recognition. It can be seen clearly that pitch, intensity and energy are features which most researchers used for their research (Rao et al., 2010; Rong et al., 2009; Rao et al., 2013). Sometimes these features are called acoustic features. Acoustic is a field of science that deals with waves

that include vibration, sound, ultrasound and infrasound. Intonation is also a very important feature for emotion that deals with the change in spoken pitch, still it is pitch feature but describing the pitch difference. Tao and Kang (2005) and Rao et al. (2010) also mentioned about the acoustic features in their study. F0, duration and intensity have given importance.

From mentioned studies, it can be seen that pitch, energy and duration are highly related with emotions. Researchers have mentioned their importance in their research (Lee & Narayanan, 2005; Schroder & Cowie, 2006; Rao et al., 2010). Moreover, Ververidis and Kotropoulos(2006) used a different pitch related short-time features with energy, formant and their bandwidths and speaking rate for emotion analysis. In Lugger and Yang (2007), the authors used 8 different prosodic and voice quality features to classify 6 different emotions. Berlin database was used for emotional speech. In this research, researchers used speaker independent emotional information for Bayesian classifiers. One research conducted on Mandarin language (Wang et al., 2008) to classify 6 different emotions using pitch, energy and duration prosodic features. A total of 88% accuracy of emotion recognition is reported. The authors used genetic algorithm and SVM (Support Vector Machine) classifiers for classification. The Large number of prosodic feature vectors in Iliou and Anagnostopoulos (2009b) are used for independent speaker recognition. Features include pitch, duration and energy are extracted from the speech signals in 35 different dimensions. The seven emotions were classified using neural networks classifier. Experiments were conducted on Berlin emotional speech corpus. The result shows that 51% of accuracy is obtained from this research.

Mentioned studies and some older literature (Nwe et al., 2003; Schroder & Cowie, 2006; Koolagudi et al., 2009; Ververidis et al., 2004; Iida et al., 2003) shows that most of the studies on emotion speech recognition are carried out on the utterance level using global static prosodic features. Some researchers used local prosodic features for emotion speech analysis (Rao et al., 2010). In some studies, emotion analysis is carried out using prosodic features at/with a sentence, word and syllable levels (Rao et al., 2007b; Wu & Liang, 2011; Rao et al., 2013). It is important to study the behavior of global and static information of prosodic features (Rao et al., 2013). Also it is important to study this information at the segment and semantic level of the word, syllable or sentence. This importance is highlighted by Koolagudi et al. (2011), Wu and Liang (2011) and Rao et al. (2013), that using prosodic features with a word, sentence or syllable level and in combination with the semantic meaning of the spoken sentence can lead to more efficient and accurate emotion recognition.

Table 2.4:

Literature on Emotion Recognition Using Prosodic Features.

#	Prosodic Features	Purpose
1	pitch, energy and duration Lee and Narayanan (2005); Schroder and Cowie(2006)	Relation of prosodic features with emotions is emphasize
2	short-time features related with pitch, energy, formant and their bandwidths and speaking rate Ververidis and Kotropoulos(2006)	Analysis of emotions
3	prosodic and voice quality features Lugger and Yang (2007)	speaker independent emotional information used to classify 6 different emotions
4	pitch, energy and duration prosodic features Wang et al. (2008)	6 different emotions are classified with 88% accuracy. Genetic algorithm and SVM (Support Vector Machine) classifiers are used for classification.
5	pitch, duration and energy in 35 different dimensions Iliou and Anagnostopoulos(2009b)	Independent speaker recognition. 51% of accuracy is obtained.

6	Static features of prosody with respect to time (duration contour, pitch contour and energy contour) Rao et al. (2010)	Static prosodic features are used for indian content.
7	Pitch, formant and spectrum related features Wu and Liang (2011)	Prosodic information is used with semantic labels to classify emotions.
8	Energy and pitch features are extracted from utterance, word and syllable level Koolagudi et al. (2011)	Prosodic features are extracted from segmental level to classify emotions. Syllable boundaries are separated using vowels.
9	Pitch, duration and energy features are extracted. Rao et al. (2013)	Local and global prosodic features are used in combination for emotion analysis. Features are analyzed region-wise in the spoken sentence.

Some researchers in this context have emphasized on the importance of region of analysis to perform recognition task (Koolagudi et al., 2011; Rao et al., 2013). They have found that the last region of the spoken sentence contains more emotion specific information than other areas of the sentence. Only few studies have been reported in the context that uses both global and local prosodic features with syllable (word or sentence) (Rao et al., 2013). It is mentioned that global prosodic features represent gross static and local prosodic features represents the finer variation in the prosody (Rao et al., 2013; Koolagudi et al., 2011; Wu & Liang, 2011). According to Koolagudi and Rao (2012), prosodic features do provide emotion specific evidence for emotion recognition. While according to Rao et al. (2013), prosodic features like intensity, duration and intonation makes the speech natural. These features are associated with syllables, words and sentences. Prosody structures the flow of speech and represents the perceptual properties of the speech that can be used in various speech tasks. But Zhou et al. (2009) stated that prosodic information alone is not suitable for emotional speech recognition task and prosodic features does not or slightly improve the system performance if using with spectral features.

2.3.4 Works using combined features

Most of the researchers are using different features in combination for more accurate and efficient emotion speech recognition (Wu & Liang, 2011; Khanna & Kumar, 2011; Koolagudi & Rao, 2012; Henríquez Rodri'guez et al., 2012; Kishore & Satish, 2013; Rao et al., 2013; Sethu et al., 2013). In previous sections that discussed about source, spectral and prosodic features are used separately by the researchers. Some work (Bozkurt et al. 2009; Wu et al. 2009; Zhou et al. 2009; Iliev et al. 2010; Han et al. 2012; Koolagudi and Rao 2012) have used combination of these features for their studies. Suitable combination of these features may help to improve system performance. Many studies reported better performance with combined features as compared to with individual ones.

In Zhou et al. (2009), it is suggested for combining articulatory features with spectral features to classify emotions in Mandarin language. To recognize 7 discrete emotions in Berlin emotional speech corpus, spectro-temporal features for long-term are used in Wu et al. (2009). It is reported that these features have better performance over short-term prosodic and spectral features. 88.6% of accuracy was reported using these features. Some of other important work in this regards can be seen in Table 2.5.

Table 2.5:

Literature on Emotion Recognition Using Combined Features (Source, System Or Prosodic)

#	Features	Purpose
1	Spectral features include MFCC and LSF (Line spectral Frequency) and their derivatives. Mean, first derivatives of pitch and intensity are include in Prosodic features. Bozkurt et al. (2009)	Experiments on INTERSPEECH 2009 emotion challenge corpus are conducted.
2	long-term spectro-temporal features. Wu et al. (2009)	long-term spectro-temporal features are compared with short-term spectral and some prosodic features.

3	articulatory and spectral features. Zhou et al. (2009)	Authors investigated the combination of articulatory and spectral features.
4	Glottal and MFCC features. Iliev et al. (2010)	Optimum path forest classifies is used to classify emotions.
5	first three formants, their bandwidths, harmonic to noise ratio, spectral energy distribution, voice to unvoiced energy ratio, and glottal flow are chosen as quality voice features whereas prosodic features includes energy, duration, pitch, F0 and their statistics and derivative statistics. Han et al. (2012)	The mismatch between training and test Conditions is studied
6	Glottal features, LPCC features and pitch, duration and energy features are chosen for the experiment. Koolagudi and Rao (2012)	Emotion recognition systems are developed using mentioned features independently and in combination and system's performances are measured.

2.4 Feature Extraction

Previous studies on emotion recognition via speech use global prosodic features and spectral features that are typically engaged in speech recognition. The most commonly used spectral features for emotion recognition are Mel-Frequency Cepstral Coefficients (MFCC) (Han et al., 2012; Koolagudi & Rao, 2012; Khanna & Kumar, 2011; Mao et al., 2009; Rong et al., 2009; Kishore & Satish, 2013). Due to its popularity in speech processing tasks, MFCC are becoming “new standard” (Ramakrishnan & Emary, 2011). As in emotion recognition, MFCCs are extracted using a 25 ms Hamming window at intervals of 10ms. The majority of spectral methods for emotion recognition make use of either frame-level or utterance-level features as mentioned before. Frame-level approaches model how emotion is encoded in speech using features sampled at small intervals (typically 10–20 ms) and classify utterances by combining predictions from all of the frames. On the other hand, utterance-level methods rely on computing statistical functions of spectral features over the entire utterance.

The ability to perceive emotions from speech depends equally upon speaker to express and upon listener to perceive. Human ears process the speech components in a nonlinear manner (Koolagudi et al., 2011b; Rao & Koolagudi, 2012). Therefore, a nonlinear mel scale filter bank is used to perceive human emotions. These mel scale filter banks have low frequency as compared to original speech signal. Lower frequency components of speech comprise more emotion specific information (Rao & Koolagudi, 2012; Mubarak et al., 2005). MFCC provides these low frequency components better that are used in nonlinear auditory observation of the speech. The MFCCs are robust speech features that are used in most speech processing tasks (Rao & Koolagudi, 2012). The main purpose of the MFCC is to mimic the human ear behavior through cepstral representation (Kishore & Satish, 2013). According to Rong et al. (2009) MFCC are only real numbers that does not have physical property and are not understandable to real world but these MFCC are very useful for machine learning and can provide better emotion recognition.

Different number of Mel Frequency Cepstral (MFC) coefficients has been used for emotion recognition. According to Kishore and Satish (2013) researchers have used first 9 to first 13 MFC coefficients for emotion recognition. It is due to the properties of cosine form which compress the signal energy in first few coefficients (Kishore & Satish, 2013). Moreover, Rao and Koolagudi (2012) stated that emotion recognition systems created using higher number MFC coefficients (first 13 MFCC) performs better as compared to less number of MFC coefficients (first 8 MFCC). In Mubarak et al. (2005), 8 to 28 MFC coefficients were evaluated for error rate in speech processing. It is found that error rate increases as the number of MFC coefficient increases. Lowest error rates were observed for 11 to 13 MFC coefficients for speech.

When an utterance is given for feature extraction, the total speech utterance is divided into a number of portions/frames and speech features are generated for each one of them. MFCC coefficients are extracted for each portion. The total number of portions depends on the length of a speech utterance. Longer the utterance will be, the higher number of portions it will be divided into.

This study has used first 13 MFCC coefficients as the spectral information. Table 2.6 mentioned some literature on the use of 13 MFCC in the field of emotion recognition. A general framework for MFCC feature extraction is given in Figure 2.1.

Table 2.6:

Use Of First 13 MFCC Coefficients In Previous Studies.

#	Emotions	Results	Findings
1	Anger, disgust, fear, happy, neutral, sad, sarcastic and surprise. (Koolagudi et al., 2011b)	Male (Average: 77.37%) Female (Average: 80.75%)	<ul style="list-style-type: none"> Human ears process the speech components in a nonlinear manner. MFCC also represents human speech in nonlinear manner. 13 MFCC performs better than prosodic features.
2	Anger, Disgust, Fear, Happiness, Neutral, Sad, Sarcasm and Surprise. (Rao & Koolagudi, 2012)	Average: 63.38% 8 MFCC = 55% 13 MFCC = 63% 21 MFCC = 65%	<ul style="list-style-type: none"> First 8, 13 and 21 MFCC are used in the study. MFCC is a lower order representation of speech signal. MFCC contains more emotion specific information. Higher number of MFCC performs better than lower number of MFCC but after some point, it does not improve results. MFCC are robust features for speech processing tasks.
3	Anger, Disgust, Fear, Happy, Neutral and Sad. (Kishore & Satish, 2013)	Average 75 %	<ul style="list-style-type: none"> MFCC mimics the human ear behavior. First 13 MFCC coefficients contain most of the signal information. MFCC provides good representation of the local spectral properties.

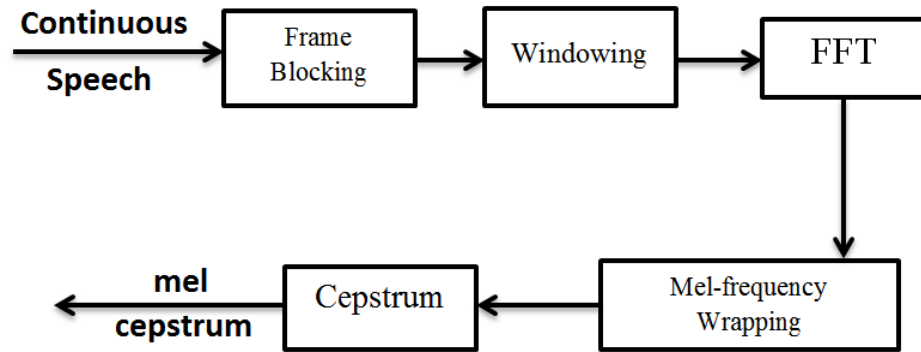


Figure 2.1: MFCC General Framework (Khanna & Kumar, 2011)

2.4.1 Frame Blocking

This step involves the blocking of speech signal into frames. The input signal is blocked into N samples, with adjacent frames being separated by M ($M < N$). The first frame consisting of first N samples while the second frame begins M sample after the first frame, and overlaps it by $N - M$ samples and this process continue (Khanna & Kumar, 2011).

2.4.2 Windowing

This concept is used to minimize the signal distortion by using the window to taper the signal to zero at the beginning and end of each frame (Khanna & Kumar, 2011).

2.4.3 Fast Fourier Transform (FFT)

The next step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain (Khanna & Kumar, 2011).

2.4.4 Mel-Frequency Wrapping

In this step, input frequency is converted in mel frequency. Conversion is done by the given formula. the pitch of a 1 kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the equation 2.1 to compute the mels for a given frequency f in Hz (Khanna & Kumar, 2011).

$$f_{mel} = 2595 * \log 10(1 + \frac{f_{Hz}}{700}) \quad (2.1)$$

The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval.

2.4.5 Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT) (Khanna & Kumar, 2011).The MFCCs may be calculated using the equation 2.2.

$$C_n = \sum_{k=1}^K (\log S_k) \cos[n(k - 0.5) \frac{\pi}{K}] \quad (2.2)$$

Here $n=1, 2, 3, \dots, K$. K denotes the number of mel cepstrum coefficients. $K=13$ is chosen for this study.

2.5 Classifiers

Literature shows that there are many pattern recognizer or classifiers has used for speech systems like speaker recognition, speech recognition, and emotion classification and so on (Mao et al., 2009; Rao et al., 2013; Staroniewicz, 2011; Han et al., 2012; Zhou et al., 2009; Scherer et al., 2008; Anagnostopoulos & Iliou, 2010; Wu & Liang, 2011; Koolagudi et al., 2011; Kishore & Satish, 2013). Many researchers have not mentioned the importance and justification for which they have chosen particular classifier only few have mentioned this importance (Henríquez et al., 2011). Many researchers have used classifier based on past researches. Sometimes classifier selected based on the experiments conducted on different classifiers (Iliou & Anagnostopoulos, 2009; Han et al., 2012). In Scherer et al. (2008), it is mentioned that some of the most common features that are used in emotion recognition have either linear values or non-linear values. Classifier selection mainly dependent on the type of data it is going to be classified.

Data linearity or non-linearity is depends on the feature sets on which classification is done. As described in section 3, emotions recognition is done based on features. These features sets are called data characteristics. Classifier selection depends on these characteristics. Linear characteristics are used to develop a linear classifier. Whereas, non-linear characteristics are used to develop a non-linear classifier. Table 2.7 shows some common classifiers used in the literature.

Table 2.7

Literature On Some Common Classifiers Used For Emotion Recognition

#	Classifier	Features	Ref.
1	HMM (Hidden Markov models) (non-linear)	Prosodic	Mao et al., 2009; Sethu et al., 2013
		Spectral	Mao et al., 2009; Zhiyan and Jian, 2013
2	SVM (Support vector machines) (non-linear)	Prosodic	Rao et al., 2013; Wu and Liang, 2011; Staroniewicz, 2011; Koolagudi and Rao, 2012
		Spectral	Staroniewicz, 2011; Vicsi and Sztahó, 2011; Staroniewicz, 2009
3	ANN (Artificial neural networks) (non-linear)	Prosodic	Wu and Liang, 2011; Staroniewicz, 2011; Han et al., 2012; Zhang, 2012; Mao et al., 2009; Iliou and Anagnostopoulos, 2009; Zhou et al., 2009; Anagnostopoulos and Iliou, 2010
		Spectral	Staroniewicz, 2011; Scherer et al., 2008; Mao et al., 2009; Iliou and Anagnostopoulos, 2009; Shi and Song, 2010; Anagnostopoulos and Iliou, 2010
4	GMM (Gaussian mixture models) (non-linear)	Prosodic	Wu and Liang, 2011; Yong-Wan et al., 2009; Koolagudi et al., 2011
		Spectral	Staroniewicz, 2011a; Koolagudi and Rao, 2012; Kishore and Satish, 2013
5	KNN (K – nearest neighbor) (non-linear)	Prosodic	Kuchibhotla et al. (2014); Staroniewicz (2011); Wang and Guan (2004)
		Spectral	Kuchibhotla et al. (2014); Staroniewicz (2011); Pao et al. (2008); Wang and Guan (2004)
6	LDA (Linear Discriminant Analysis) (Linear)	Prosodic	Kuchibhotla et al. (2014); Staroniewicz (2011); Wang and Guan (2004)
		Spectral	Kuchibhotla et al. (2014); Staroniewicz (2011); Wang and Guan (2004)

As discussed earlier, classifier selection is based on the characteristics of the features. If data nature known, then it is easy to select classifier. As in emotion recognition system, researcher deals with speech, whose data is unknown to the researcher in the beginning or is rarely known. So researchers prefer non-linear classifier for classification as it can classify linear and non-linear data better than linear classifier.

The design structure of NN has two output modes. One called as binary coding and second as one-to-one. In binary coding, the system has less number of output neurons for

larger number of corresponding outputs. According to Han et al. (2012), binary coding has a less classification rate than one-to-one. In one-to-one mode, same number of output neurons used for corresponding number of outputs. For example, 5 neurons should be used for 5 outputs in one-to-one. Han et al. (2012) also stated that the number of neurons in hidden layer in multilayer architecture, should be appropriate for the architecture. If the number of neurons is too many, it will not be able to converge, if number is too small, recognition error will be large. The number of neurons in hidden layer should be expressed by the formula (Han et al., 2012).

$$N_{no} = \frac{(In_{no} \times Out_{no})}{2}$$

Here N_{no} represents the number of neutrons at hidden layer; In_{no} represents the number of neutrons at input layer and Out_{no} represents the number of neutrons at output layer.

According to Kuchibhotla et al. (2014) and Wang and Guan (2004), KNN is non-parametric method for classification. It classifies the unknown and unlabeled data by examining its k nearest neighbors of known and labeled class. Pao et al. (2008) stated that KNN is simple and powerful classification method that classifies the data based on the idea that similar observations should belong to similar class. KNN uses training samples directly and represent each sample in d -dimensional space (Kuchibhotla et al., 2014). Where d is the number of speech features. KNN tries to find the nearest neighbor to the unlabeled data from the training sample based on the predefined distance measure. According to Kuchibhotla et al. (2014) and Pao et al. (2008), Euclidean distance mostly

used in KNN. Once the training samples list of nearest neighbor is obtained, testing samples are classified based on the majority of nearest neighbors.

LDA is a statistical method of classification that uses the variance of the classes. It classifies the data by minimizing the covariance within a class and maximizing the covariance between classes (Kuchibhotla et al., 2014). Classification is done on the basis of given speech features as input.

2.6 Linear and Non-Linear Classifiers

In general, pattern recognizers used for speech emotion classification can be categorized into two broad types namely linear classifiers and non-linear classifiers. A linear classifier performs the classification by making a classification decision based on the value of a linear combination of the object characteristics. These characteristics are also known as feature values and are typically presented to the classifier in the form of an array called as a feature vector.

The non-linear weighted combination of object characteristics is used to develop non-linear classifiers. During implementation, proper selection of a kernel function makes the classifier either linear, or non-linear (Gaussian, polynomial, hyperbolic, etc.). In addition, each kernel function may take one or more parameters that would need to be set. Determining an optimal kernel function and parameter set for a given classification problem is not really a solved problem. There are only useful heuristics to reach satisfying performance. While adopting the classifiers to the specific problem, one should be aware of the facts that, non-linear classifiers have a higher risk of over-fitting, since

they have more dimensions of freedom. On the other hand a linear classifier has less degree of freedom to fit the data points (Koolagudi & Rao, 2012b).

In this study, linear and nonlinear classifiers have been studied. Linear Discriminant Analysis (LDA) has been selected as a linear classifier and Neural Network (NN) and K-Nearest Neighbor (KNN) have been selected as nonlinear classifiers. NN is less time consuming and have better recognition performance in speech emotion recognition (Han et al., 2012; Wang et al., 2011); KNN is the simplest classifier that uses the idea of similar observation belongs to similar classes (Kuchibhotla et al., 2014; Pao et al., 2008) and LDA can be used to reduce the dimensionality of classification process (You et al., 2007).

2.7 Gender-Based models

From literature, it can be seen that some studies have tried to build gender-based models for their studies (Rong et al., 2009; Rao et al., 2013). Whereas one study has mentioned some conditions that can lead a system to a poor performance, in which speakers are also one big factor (Han et al., 2012). In Rong et al. (2009), it is found that gender-dependent data sets perform better than gender-independent. Two different models were created by the researchers to test the result of male, female and common models. Moreover, two sets of features, 84-set and 16-set feature vectors are used to evaluate the models differently. From the results it is clear that the model's result gathered from gender-dependent data sets are more accurate than result gathered from gender-independent model. Model performance of gender-dependent, increased by 4.34% on average while dealing with different sets of features (3.99% for 84-set features and 4.69% for 16-set features). Moreover, even better results are given by male model with improvement by 8.28% and

9.25% for 84-set and 16-set features respectively. In the same research, it is mentioned that the females are more emotional than male and on the basis of results gathered from 84 and 16 feature sets, one can say that female and male speakers express their emotions differently. Moreover, from the results we can say that male can express natural emotions better than females; on the other hand, females are better in expressing acted emotions than males while using a selected number of features.

In another study conducted by Rao et al. (2013), it is mentioned that female's emotions are more easy to recognize than male. According to results presented by the Rao et al. (2013), female model performs better than male model. Average performance for both models was 45% for male and 51% for female. Female model performs average 6% better than the male model in recognizing the emotions. Moreover, it is shown in the study that the female has a high pitch for both anger and neutral emotions than male. On the other hand, males have more energy in their tone than females for both anger and neutral emotions. Model using pitch features for classification may suffer from the overlapped condition of the pitch for anger and neutral emotions. It is shown in the results that mean for a male pitch for anger emotion is 195 Hz and female pitch for neutral emotion is 267 Hz. One can mismatch these emotions for male and female while dealing with the same model. Same is the case with energy features, these features, in individual, may give inappropriate results in gender-based models.

2.8 Word boundary detection

Word boundary detection is the process of detecting word boundary in an utterance. In the study conducted by Koolagudi et al. (2011), prosodic speech features have been studied for emotion recognition performance at utterance, word and syllable level. Word

boundaries were selected manually. Energy and pitch parameters were considered for the study. Altogether, eight emotions (anger, disgust, fear, happy, neutral, sad, sarcastic and surprise) were considered for the research. SVM classifier has been used for classification. Simulated emotion speech corpus IITKGP-SESC (*Indian Institute of Technology-Simulated Emotion Speech Corpus*) was used by the Koolagudi et al. (2011). For word level, the average of 46.5%, 30% and 31.5% accuracy was achieved for beginning, middle and ending words respectively for anger and neutral emotions.

Rao et al. (2013) also conducted study on word level for gender independent and gender dependent models. Local and global prosodic features were used for the research. IITKGP-SESC and Emo-DB were used for speech utterances. Classification was performed using SVM. Word boundaries were identified automatically using vowel onset points (VOPs). VOP detection was done using the excitation source, modulation spectrum and spectral peaks evidences. Such method is known as a combined method for VOP detection. Hilbert envelop (HE) of the LP residual was used as excitation source evidence, variation in the spectrum of speech signal used as a modulation spectrum and the sequence of the sum of ten largest peaks of spectra of speech frames was used as spectral peak evidence in Rao et al. (2013). All these three features were combined to enhance the VOP information. Peak in the positive region of the combined VOP evidence was taken as VOP location. Although 95% of the combined VOP result of proper detection was given, 31.5%, 46.5% and 51.5% average accuracy was noted for anger and neutral emotions by using global prosodic features at initial, middle and final word position respectively.

Results of Rao et al. (2013) and Koolagudi et al. (2011) conflict each other. Rao et al. (2013) have better results for middle and ending words for anger than beginning words whereas Koolagudi et al. (2011) has better results for beginning words for anger than remaining words section. Rao et al. (2013) suggest that words in final position are more emotional specific than others whereas Koolagudi et al. (2011) suggest that for anger, initial words are more informative and for neutral, last words are more informative.

In case of automatic word boundary detection, performance of the boundary detection model is an important factor in emotion classification model. If word boundaries are poorly detected, features extracted from those words will also be not as efficient as they should be. One example can be seen in Agarwal et al. (2010), it is mentioned in the study that for neutral emotions, the average accuracy of proposed word boundary detection algorithm is 88% and for anger, it is 79%. If one uses such methodology for emotion detection, there result may not be as efficient as it should be regarding features and classifiers. For example, if 100, 100 utterances will be taken for both training and testing phases, word boundary detection will give only 88 and 79 utterances for neutral and anger emotions respectively. Means while doing testing, the boundary detection algorithm will not detect a number of utterances efficiently. Hence emotion detection model will give less accurate results than manually detected process.

2.9 Summary

In this chapter, introduction of emotional speech recognition systems was given. Common databases that are used for emotion recognition systems were discussed. Types of speech features and common speech features in each type were also mentioned. Commonly used classification techniques were also mentioned in this chapter. It is found

that mostly researcher uses classifiers on the basis of previous studies. Concept of data linearity and non-linearity was also mentioned that can be useful in classifier selection. The impact of different genders on speech emotion classification was given to differentiate the importance of gender in emotional speech recognition systems. The feature extraction level was also highlighted in this chapter to study the efficient level of feature extraction for more accurate emotion recognition systems. Whereas this study only focused on gender dependent emotional speech recognition system that uses the word level for speech feature extraction and NN, KNN and LDA as classifiers to study the behavior of data linearity and non-linearity on classifier. This study uses MFCC spectral feature as speech features, MFCC architecture is also provided in this chapter that will be used in the next chapter in feature extraction phase. For a less complicated emotion recognition model, manual detection of word boundaries will be done.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter presents the proposed framework for emotion detection using speech. It describes the speech corpus used in this study, pre-processing method that is sufficient for the proposed model, features that will be extracted from speech using chosen corpus and classification technique that is used to create class models for selected emotions.

3.2 Framework

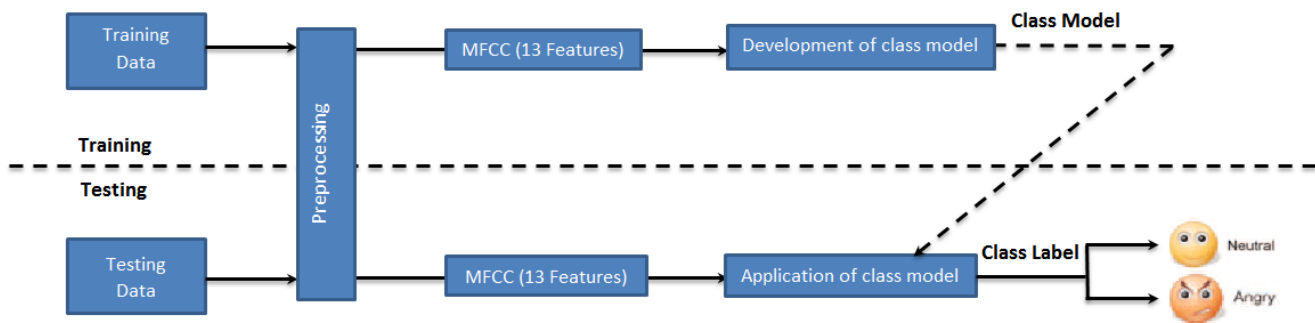


Figure 3.1: Framework of the study

This study have proposed gender-dependent word-level emotion detection model that uses spectral speech feature for emotional speech recognition as shown in Figure 3.1. As described in previous chapters, very little work has been done on spectral features on word level. Study uses Emo-DB emotional speech database as datasets. Further sections discuss about the features and database used in this study.

3.3 Speech corpus

The speech corpus used for this study is Emo-DB (Burkhardt et al., 2005). In emotion recognition process, main purpose of the model is to identifying the emotional state of the speaker automatically. Emo-DB is acted emotional speech database that is used by many researchers as mentioned in section 2.2.1. According to Rao et al. (2013), Berlin emotional speech database (Emo-DB) is internationally known and used for emotional speech processing.

According to literature, there are three different kinds of speech databases (see section 2.2) for emotion recognition task. Emo-DB is simulated emotional speech database that is recorded in German Language. There are a total of 800 utterances in this database but during quality and naturality analysis (Bitouk et al., 2010), only 535 are chosen for the database. Details on Emo-DB are mentioned in chapter 2.

As this study only focused on two emotions anger and neutral, total of 206 utterances are chosen for these two emotions in which anger comprises 127 utterances and neutral comprises 79 utterances. Two models will be created by using same framework for both male and female gender using same framework. Male model will be trained and tested by using male training and testing data respectively for selected emotions, whereas female model will be trained and tested by using female training and testing data respectively for selected emotions. In the end both results will be merged for overall model performance analysis. For this purpose, data should be divided according to model type.

For anger, there are 60 utterances for male and 67 utterances for female. For neutral, there are 39 utterances for male and 40 utterances for female. For anger 85 (~67% of

total) utterances (40 for male, 45 for female) will be chosen for training the model and 42 (33% of total) utterances (20 for male, 22 for female) will be chosen for testing the model. For neutral 53 (67% of total) utterances (26 for male, 27 for female) will be chosen for training whereas 26 (~33% of total) utterances (13 for male, 13 for female) will be chosen for testing the model.

Table 3.1 describes about the actors participated in the recording of the database. Participants' code, gender and age are given in the table. Table 3.2 describes about the text code used in the database and about text codes, their respective German text and their English translations.

Table 3.1

Speaker Information

Code	Gender	Age
03	Male	31
08	Female	34
09	Female	21
10	Male	32
11	Male	26
12	Male	30
13	Female	32
14	Female	35
15	Male	25
16	Female	31

Table 3.2

Code Of Text Used In Database.

Text	German Text	English Translation
Code		
a01	Der Lappenliegt auf demEisschrank.	The tablecloth is lying on the fridge.

a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es so weit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

Table 3.3 describes about the emotions included in database. Emotions code, their respective emotions in German and in English is given in the table. Letter ‘N’ is used for neutral emotion in both English and German version.

Table 3.3

Code Of Emotions

Letter for English	Emotion (English)	Letter for German	Emotion (German)
A	Anger	W	Ärger (Wut)
B	Boredom	L	Langeweile
D	Disgust	E	Ekel
F	Fear/Anxiety	A	Angst
H	Happiness	F	Freude
S	Sadness	T	Trauer
N		Neutral	

3.4 Pre-Processing

A system may lead to a bad performance if its training and testing environments are different. The conditions that can lead to this mismatch between training and testing environments may be divided into three classes: Differences of speakers, Changes of recording channels and Effect of noisy environment (Han et al., 2012). This study uses pre-recorded voiced data that is recorded in studio noise free environment using same channel for recording. Noisy environment and difference in channel may not affect this model. Difference of speaker is the factor that can decrease the performance of this model.

This study aims to build different models for male and female for emotions classification as based upon mentioned findings in second chapter. Two different models will be made while following same framework for different genders. Both models will be trained and tested differently.

As importance of automatic and manual word boundary detection is discussed in section 2.8, manual word boundary detection is used for this study. Words will be separated using “Audacity v 2.0.5” tool. This tool is used for audio editing and recording. Every utterance will be divided into number of words in it. Each word will then form a new file where file codes will be same. A new digit will be added in the last of file that will denote about the word position in the sentence. One example can be seen in Table 3.4.

Table 3.4

Example of File Codes and Values

File Code	Text	Value
03a01Nc	Der Lappenliegt auf demEisschrank.	Normal Sentence
03a01Nc1	Der	1 st word
03a01Nc2	Lappen	2 nd word
03a01Nc3	liegt	3 rd word
03a01Nc4	auf	4 th word
03a01Nc5	dem	5 th word
03a01Nc6	Eisschrank	6 th word

In this upper given table, “03a01Nc” is an original file name used in the berlin database Emo-DB. First two parameters ‘03’ describes about speaker (see table 3.1). Second parameter ‘a01’ tells about the sentence (see table 3.2). Third parameter ‘Nc’ describes about the file emotion in which first letter ‘N’ is for emotion (see table 3.3) and second letter ‘c’ is for version as some file in database have more than one version as described in second chapter.

3.5 Feature Extraction

Word-level feature extraction is done in this study. Word boundaries have been detected manually using Audacity Software. As database is in German language, sentences are translated into English word by word using Bing translation. Once words are correctly detected for particular sentence, they are separated using Audacity. Figure 3.2 illustrate

the process of separating the word from the sentence. Figure 3.2 (A) represents the full sentence file whereas selection area denotes one word. After exporting the selected area, separate file has been created that represents one word, Figure 3.2 (B) denotes one separate file of whole single word that is created from Figure 3.2 (A).

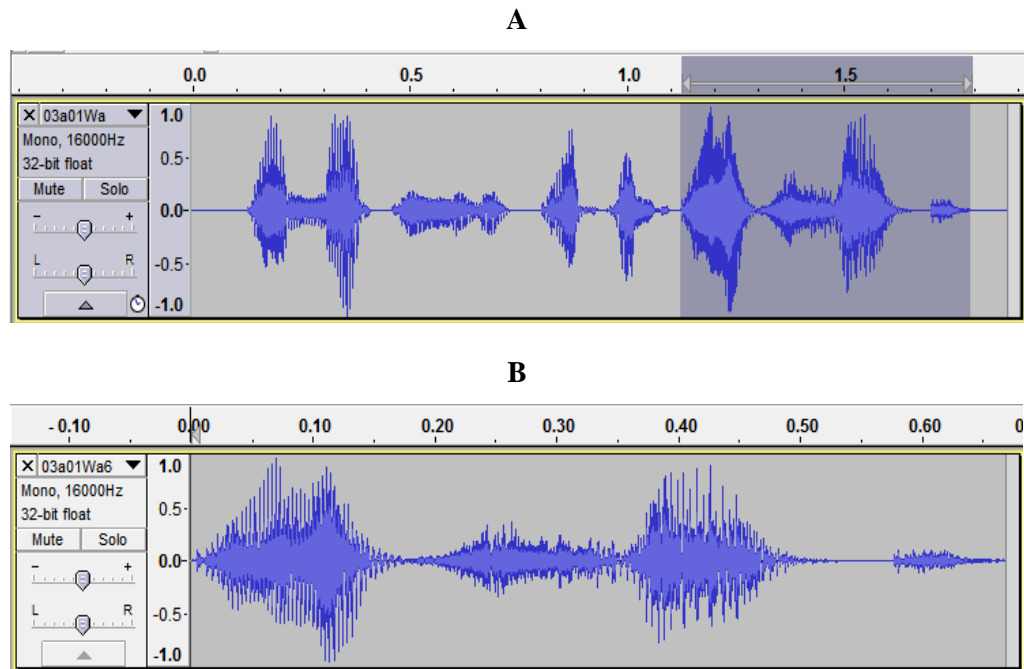


Figure 3.2. Detecting Word Boundary for ‘03a01Wa’

This study focused on the use of spectral (MFCC) features at word level. Importance of spectral features (local and global) is given in second chapter in detail. $K=13$ is selected for the formula mentioned in section 2.4.5 that gives 13 MFCCs for each speech file. 13-MFCC features are used as spectral information and mean, standard deviation and variance of 13 MFCC are extracted as global features. MFCC architecture mentioned in section 2.4 implemented in the Matlab, their mean values extracted using Mat lab coding, standard deviation and variance calculated manually.

3.6 Classifier

According to Rong et al. (2009), it doesn't matter which classification model you are using for classification, the main purpose is to analyze the extracted features and to find out valuable patterns that can predict the speech data instance to a certain emotional state accurately. Ling He (2010) stated that the type of classifier used to generate class models and make the classification decision does not appear to have as high impact on the classification accuracy as the type of features.

This study uses NN, KNN and LDA for classification. According to Han et al. (2012), NN has better emotion recognition speed and less calculation load than others, it is suitable for computer chips with lower computing capability, KNN is the simplest classifier (Kuchibhotla et al., 2014; Pao et al., 2008) and LDA can be used to reduce the dimensionality of classification process (You et al., 2007). Detailed discussion of NN, KNN and LDA is given in section 2.5. Performance of the model totally depends on selected features for the model.

There are two phases in the model. First is training phase, second is testing phase. In training phase classifiers are trained for classification purpose. Class models are created for each emotion in this phase. Speech features are evaluated and used to train the model. While in testing phase, created class model are used to classify the given speech into particular emotion using selected extracted features. System efficiency is measured in this phase. Efficiency is measured in terms of accurately classified utterance divided by total number of utterances given for classification.

In section 3.3, number of utterances used for training and testing are mentioned. This proposed model will use 70% of utterances for training purposes while remaining 30% utterances will be used for testing purpose. Number of utterances for both male and female are also given in section 3.3.

Two models created by using same framework for both male and female gender using same method. Male model trained and tested by using male training and testing data respectively, whereas female model trained and tested by using female training and testing data respectively.

To compare the proposed models, a general model following the same framework but with combined data of male and female created. Same 70% utterances will be used for training and remaining 30% utterances will be used for testing the model as same percentage will be used to train and test the gender models. Male and female models will be compared one by one with general model on the basis of classification. Although any percentage cannot be given at this stage of study but according to results presented in Rao et al. (2013), this study expects female model to perform better than other two models. Moreover, position of words that will give high recognition performance in the sentence, will also be evaluated. This study also expects models to perform different for both different emotions (neutral and anger). It expects female model to perform even better for neutral emotion.

3.7 Evaluation of the model

Classifiers that are used to generate the class models and are used for classification purposes does not affect the emotional speech recognition process (Rong et al., 2009;

Ling He, 2010). The main purpose of the classifier is to analyze the given speech features. For the evaluation of this study, KNN, LDA and NN are used as classifiers. Classification rates are selected as the performance metrics. The classification rates mentioned in Rao et al. (2013) and Koolagudi et al. (2011) are taken as the tool to evaluate this study, as the main objective of this study is to provide better performance for emotion recognition. Results of gender dependent and independent models for word level of this study are compared with each other as the results for different genders at word level are not mentioned in Rao et al. (2013) and Koolagudi et al. (2011). The effect of word positions (initial, middle and Final) on emotion recognition is also evaluated in this study. Classification rates of each position are compared with other positions to observe the more accurate word position for emotion recognition. Classification rates of each word position are also compared with the respective classification rates of Rao et al. (2013) and Koolagudi et al. (2011).

3.8 Summary

This chapter describes about the methodology applied to fulfill the objective of this study. Framework for this study is given in this chapter that will be used to create gender dependent and independent models. Speech corpus that will be used as input to the models is also discussed in this chapter. Process of word boundary detection and feature extraction are discussed in this chapter. Classifiers and datasets used for training and testing purposes are given in section 3.6 of this chapter. In the end, method to evaluate models is also given.

CHAPTER FOUR

ANALYSIS OF RESULTS AND DISCUSSION

4.1 Introduction

This chapter discusses about the Speech corpus used for this particular study, word boundaries that are detected manually, speech features are extracted from words and classification results that are calculated by different classifiers. Classification result shows that NN (Neural Networks) is not as sufficient for particular study as KNN (K-Nearest Neighbor) and LDA (Linear Discriminant Analysis). In the end, results for proposed models are given. This chapter also discusses the results obtained from this study and compares them with results present in previous studies. The comparison of previous studies that were done on either word-level or gender dependent and independent or both is also given in this chapter. Results show that accuracy that is reported in this study is better than previous studies. Use of only two emotions may be a reason for higher results of this study.

4.2 Speech Corpus

Emo-DB (Emotional Database) was selected as speech files for this particular study. Detail introduction of Emo-DB is given in section 2.2 and data construction in database is provided in section 3.3. Number of speech files used for training and testing purposes also mentioned in section 3.3. Process of words separation from whole sentence is given in section 3.4.

4.3 Feature Extraction

13 MFCC features have been extracted for this study from word-level. Mat lab codes were used to extract local and global MFCC features from sound files which were individual words themselves. Only mean, standard deviation and variance of 13 MFCC are extracted as global features.

Figure 4.1 illustrates the difference of 13 MFCC for anger and neutral emotion. In the figure 4.1, x-axis represents the MFCCs and y-axis represents the values of each coefficient. Each MFCC is analyzed on the basis of its mean values. Figure 4.1 uses the mean values of all words for both male and female gender. From the figure, it can be seen clearly that 1st MFCC has very low values than others. Study finds the same performance of MFCC for all words values independently.

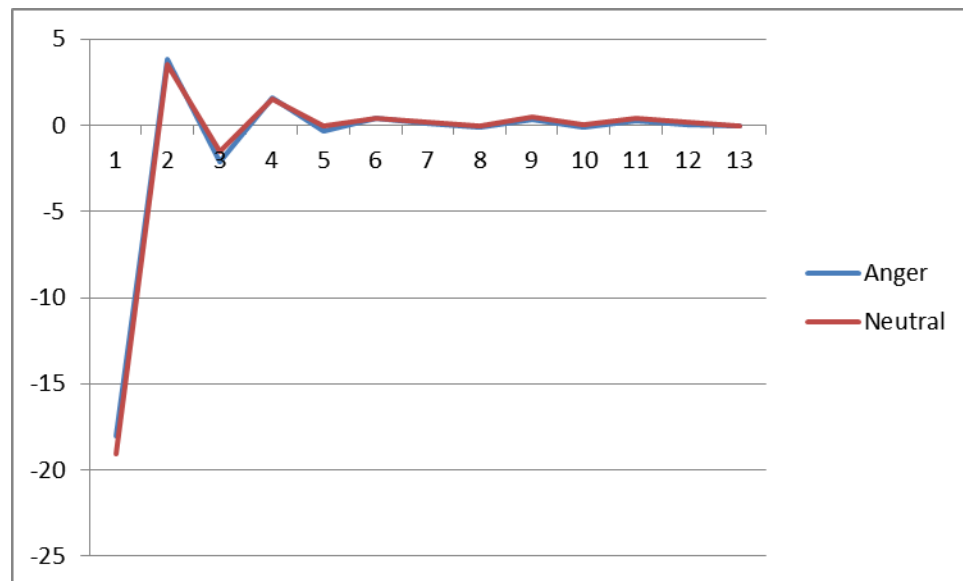


Figure 4.1: Difference of MFCC values for all words

Figure 4.1 does not provide good difference of anger and neutral emotions for human observations, but by ignoring first 4 MFCC, a more differentiable image can be seen as given in figure 4.2. Although figure 4.1 does not provide good results for humans but results show that best performance is obtained using 13 MFCC.

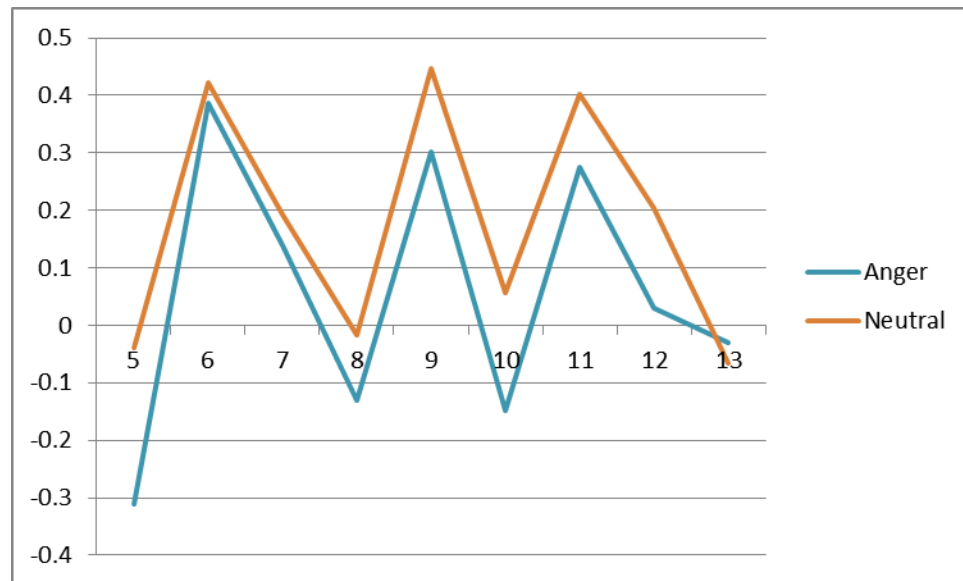


Figure 4.2: Difference of MFCC values for all words while neglecting first 4 MFCC

Gaurav (2008) has proposed some MFCCs that were used in his study. Maximum of 65% accuracy has been reported in his study while using proposed MFCCs with other spectral and prosodic features using KNN classifier. It is found that mentioned MFCCs are not as effective as all 13 MFCCs for this study. When set of MFCCs mentioned by Gaurav (2008) experimented in this study, almost same accuracy has been achieved i.e 65.59% using KNN. Gaurav (2008) has also proposed other features but those are not applicable for this particular study due to the scope of this study. Total of six emotions were classified by Gaurav (2008). Figure 4.4 illustrates the result for MFCCs mentioned by Gaurav (2008).

Other combination of MFCCs is also studied in this study. MFCCs are selected based on simple random selection (Koolagudi & Rao, 2012), to examine the efficient combination of MFCCs. Obtained results show that combination of 5th, 6th, 8th, 9th, 10th and 11th MFCCs is more efficient than other combinations. These results are discussed in detail in section 4.5.1 and can be seen in table 4.4. Figure 4.3 illustrates the difference of emotions while using mentioned combination.

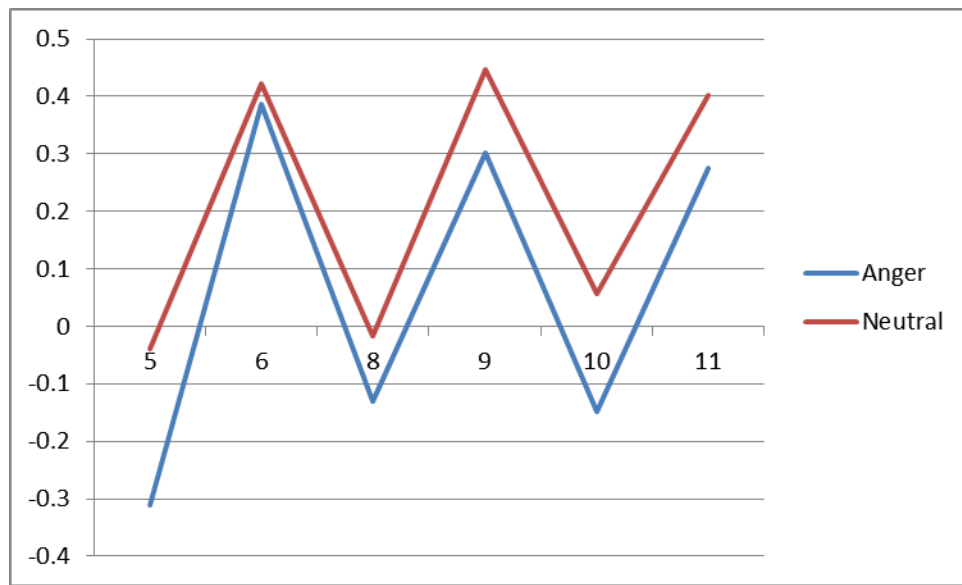


Figure 4.3: Difference of MFCC values for all words while using 5th, 6th, 8th, 9th, 10th and 11th MFCC

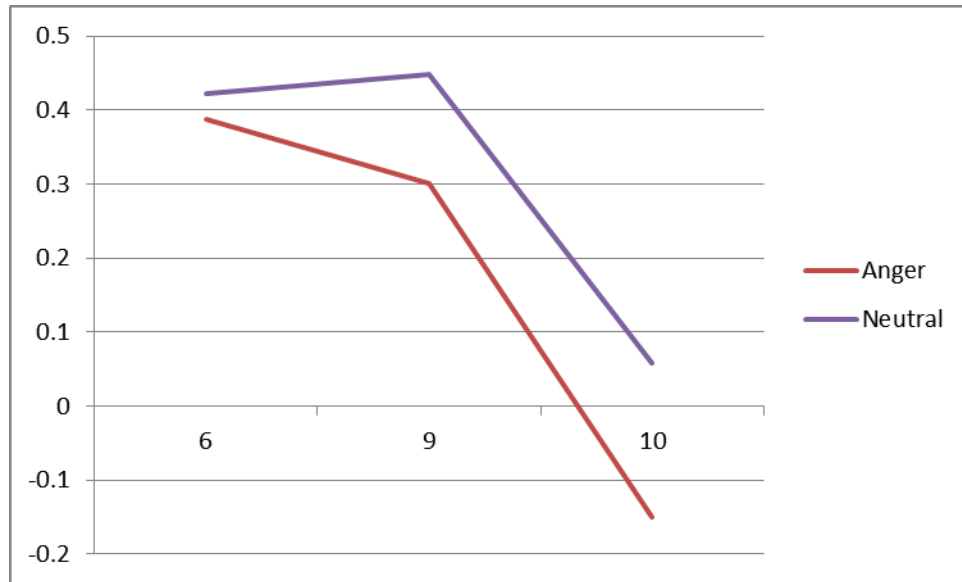


Figure 4.4: Difference of MFCC values for all words while using 6th, 9th and 10th MFCC

Figure 4.5 illustrates the difference of variance of all mean values for all words. It can be seen that middle words for anger exhibit more emotion specific information for anger, but classification result shows that initial and ending parts of utterances are more emotion specific than middle. These results are discussed in section 4.5.2 in this chapter in detail.

As this study focused on genders, male and female MFCC values are also examined. Both seem to provide same results while dealing with 13 MFCC. Figures 4.6 and 4.7 illustrate 13 MFCC for female and male respectively.

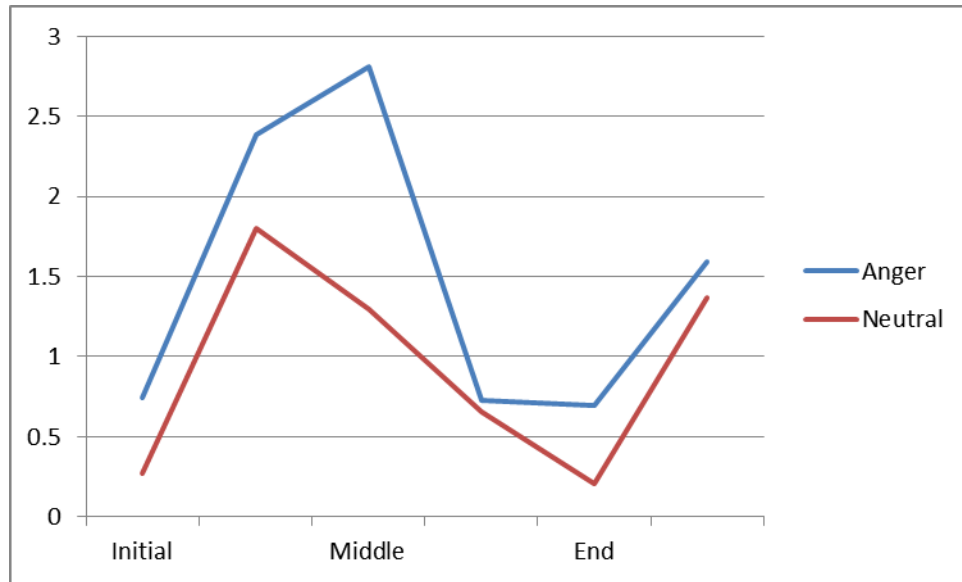


Figure 4.5: Difference of Variance for all mean values of all words

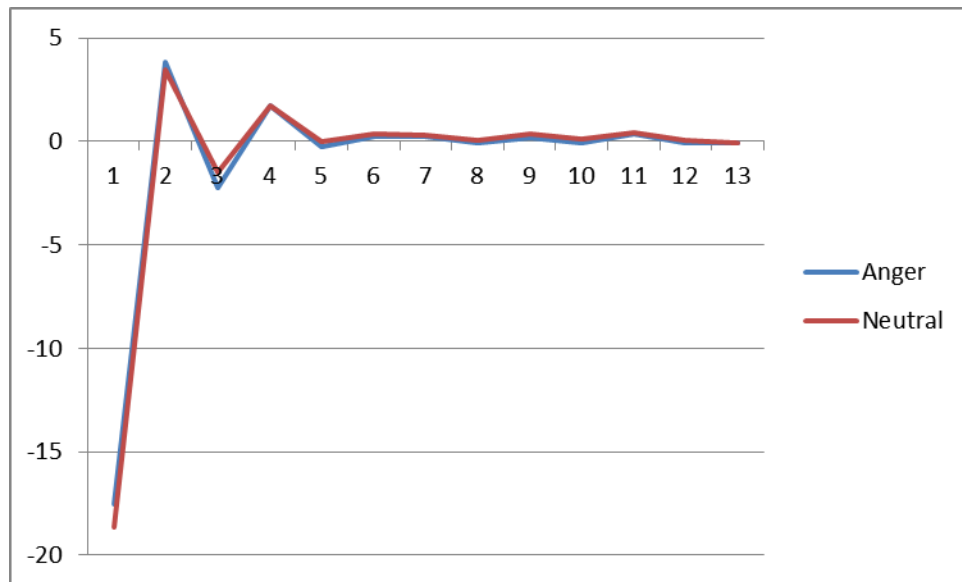


Figure 4.6: Difference of MFCC values for all words for female

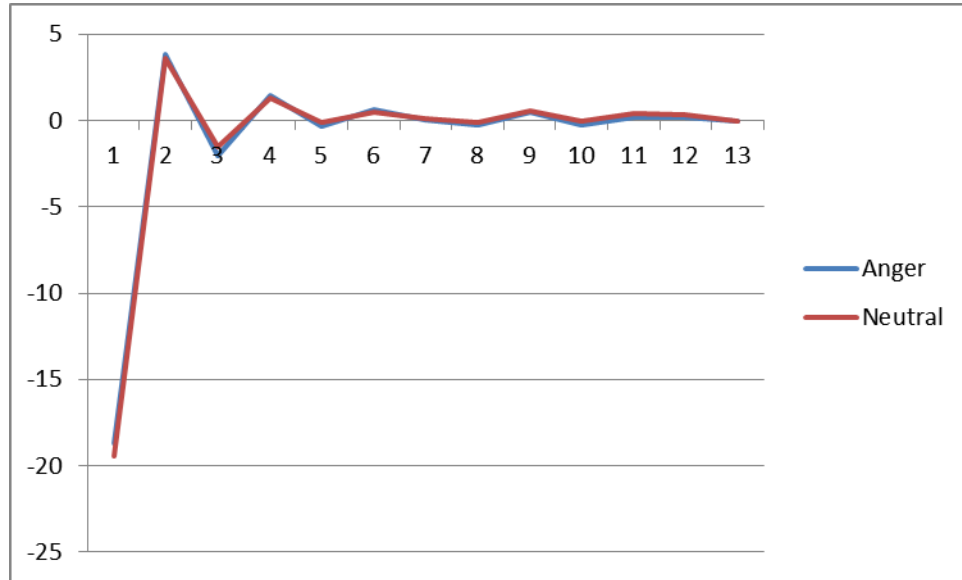


Figure 4.7: Difference of MFCC values for all words for male

4.4 Classifier

NN, KNN and LDA classifiers are used for classification in this study. Performance of the model totally depends on selected features for the model however when data is given to NN to classify, results were very low. Highest accuracy was achieved by male model i.e 22.44 %. The reason of such results is because NN requires large amount of data to process and provided data is not sufficient for NN to train. Two other classifiers KNN and LDA are tested based on their fewer requirements for training and testing. 70% data is used for training and 30% data is used for testing purpose. For NN, 0.1 learning rate was used, 0.9 was set as momentum factor and mean square error set as 0.02%. For KNN this study has made a few experiments using different values of k in KNN and it is found that $k=2$ give better recognition rate. Results of classifiers while using 13 MFCC at the utterance level are given in table 4.1. From the results it can be seen that LDA dominates other two classifiers in terms of classification rate. Figure 4.8 illustrates the performance

of classifiers for proposed models. Due to low efficiency of NN; KNN and LDA were used as classifiers for further testing.

Table 4.1

Accuracy of Different Classifiers In Percentage.

Classifier	General	Male	Female
KNN	69.53	60.91	63.56
LDA	84.67	85.79	86.83
NN	21.91	22.44	20.12

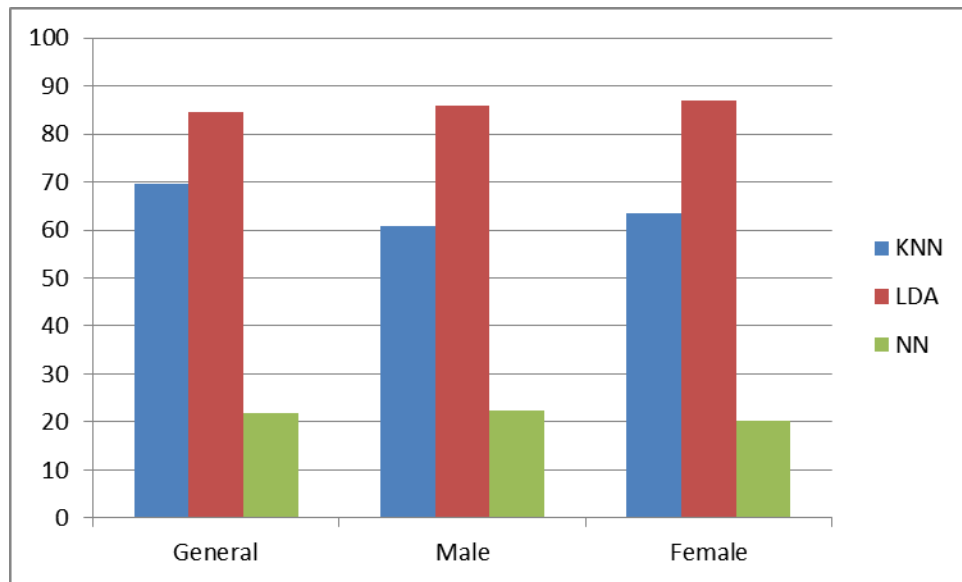


Figure 4.8: Performance Analysis of classifiers for different models

4.5 Gender independent and gender dependent models

In this study, different models are separately developed for different genders using global MFCC features. Total of three models are developed for gender independent, male and female separately, whereas KNN and LDA are used as classifiers for word level models. NN is used in utterance level model but due to low efficiency (see table 4.1); it is ignored while studying the words position in the utterance (see section 4.5.2). Each model is

trained and tested using its respective datasets as mentioned in chapter 3. This study considered only two emotions i.e anger and neutral of Emo-DB for studying the role of male and female in recognition performance at different word levels. Moreover, performance is also calculated for whole utterance while using global MFCC features at word-level.

4.5.1 Emotion recognition using utterance level global MFCC

Mean of MFCCs are used as features vectors for all three classification models. 13 mean values of 13 MFCC are obtained for each word in the utterance for each actor and for each emotion. Emotion recognition performance of different models is given in the table 4.2. Highest recognition performance is obtained from general model for KNN classifier i.e 69.53%. Male and female models have given accuracy of 60.91% and 63.56% respectively. While for LDA, female model has given highest accuracy rate of 86.83%, whereas male and general models have provided accuracy of 85.79% and 84.67% respectively. Figure 4.9 illustrates this difference of accuracy between different models more accurately for KNN and LDA.

Table 4.2

Emotion Recognition Performance At Utterance Level

Classifier	Gender	Utterance
	Dependent/Independent	Level (%)
KNN	General(Independent)	69.53
	Male	60.91
	Female	63.56
LDA	General	84.67
	Male	85.79
	Female	86.83

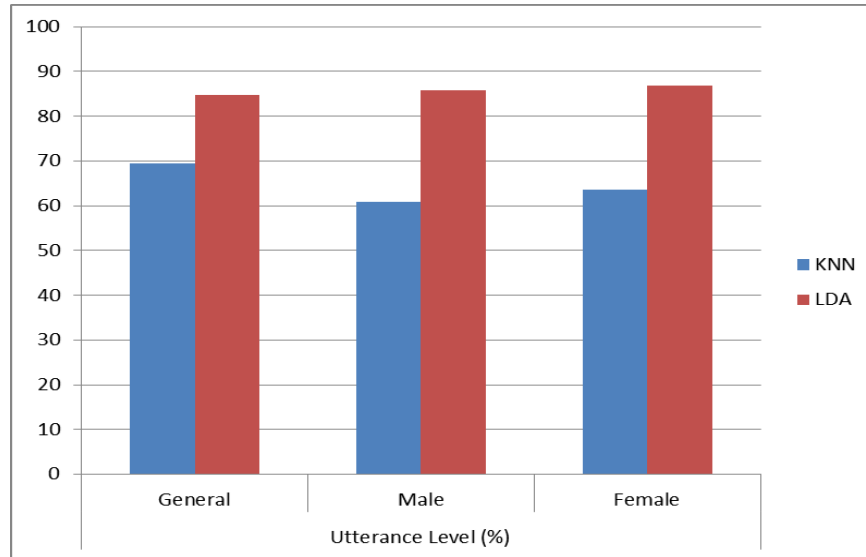


Figure 4.9: Utterance level performance of the models using KNN and LDA

It can be seen from results that KNN has low performance as compared to LDA. One reason of it can be data linearity and classifier. It can be seen from above mentioned figures that while using some specific MFCC combinations, as shown in figures 4.10 and 4.11, data can be linearly separable whereas, KNN is non-linear classifier and LDA is linear classifier. Results show that linear data (at input layer of NN) is not appropriate for non-linear classifier (NN) (see figure 4.10 and 4.11). Figures 4.4 and 4.10 represent two types of data (linear or non-linear), from the results mentioned in table 4.3, it can be seen that non-linear classifier (KNN) performs better with non-linear data (figure 4.4). Whereas linear data (figure 4.10) decreases the performance of KNN by 6.4% (gender independent model) to 14.4% (gender dependent model). Hence data type affects the performance rate of emotion recognition system. Moreover, from the results of all models, obtained from LDA, it can be said that female are more emotional than male.

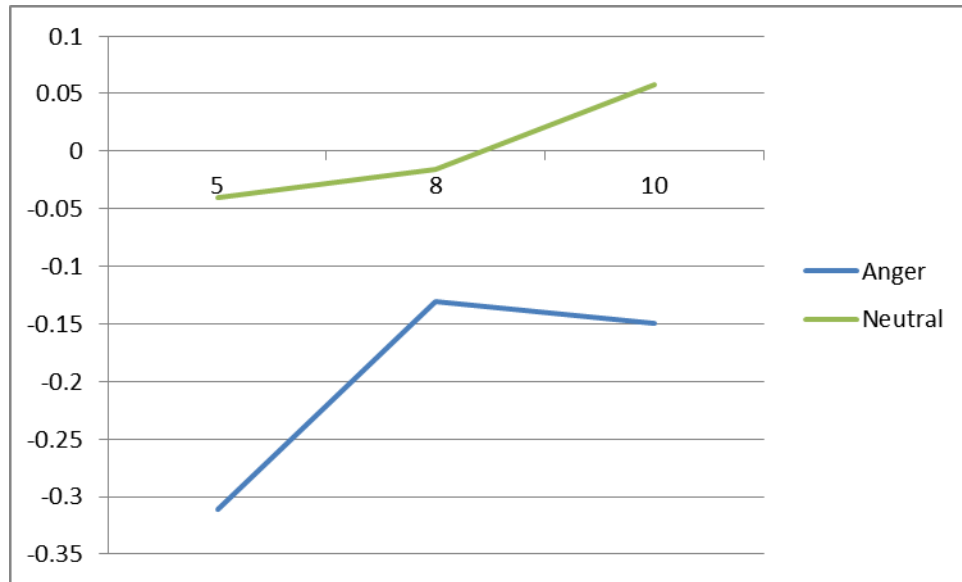


Figure 4.10: Difference of emotions using mean of 5th, 8th and 10th MFCC

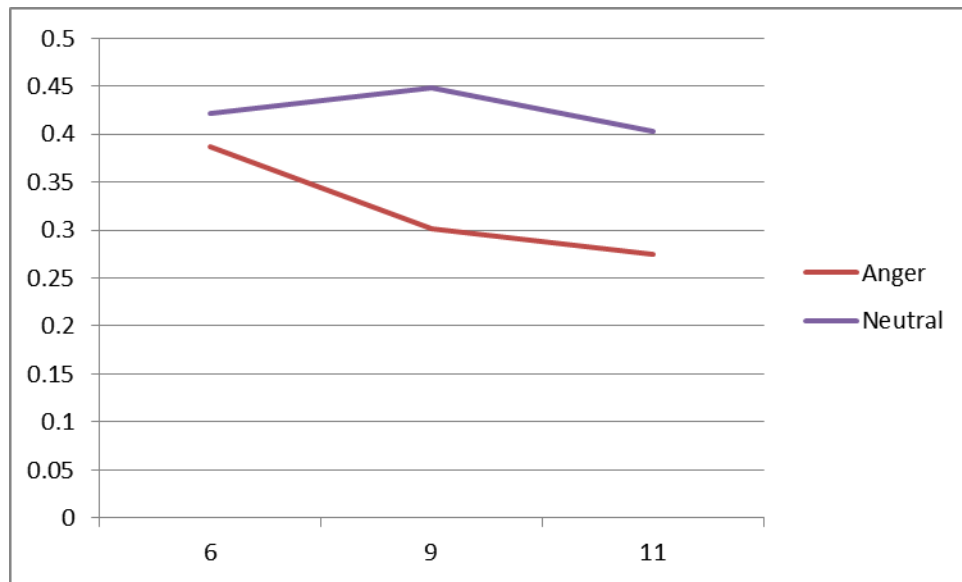


Figure 4.11: Difference of emotions using mean of 6th, 9th and 11th MFCC

While using the global features of MFCCs that were proposed by Gaurav (2008), i.e 6th, 9th and 10th, can also be seen in figure 4.4 and results are given in table 4.3, highest accuracy is obtained from female model i.e 75.41% while using KNN. As seen in figure

4.4, this combination of coefficients is not linearly separable, KNN provides maximum performance. This result also verifies the statement that non-linear data is appropriate for non-linear classifier and linear data is appropriate for linear classifier. Whereas general and male models provide accuracy of 65.59 % and 52.86 % respectively for KNN. On the other hand, LDA provides 68.43%,68.40% and 70.15% for general, male and female models respectively. This result also proves that females are more emotional than males and while using 13 MFCC, model provides better results for particular study. Figure 4.12 shows graphic form of performance of given results.

Table 4.3

Emotion Recognition Performance at Utterance Level While Using 6th, 9th, 10th And 5th, 8th, 10th MFCC

Classifier	Gender Dependent/Independent	Utterance Level (%)	
		6, 9, 10	5, 8, 10
KNN	General(Independent)	65.59	59.2
	Male	52.86	49.5
	Female	75.41	61
LDA	General	68.43	68.7
	Male	68.40	63.2
	Female	70.15	66

While using the global values of MFCCs, shown in figure 4.3, results are mentioned in table 4.4, it can be seen that male model have highest accuracy i.e 75.74% for KNN classifier while female and general models have 69.78% and 65.74% respectively. LDA again gives slightly better results than KNN, providing 75.50% highest accuracy for female model. Male and general models have accuracy of 73.04% and 69.05%

respectively. Figure 4.13 provides graphical representation of the results mentioned in table 4.4.

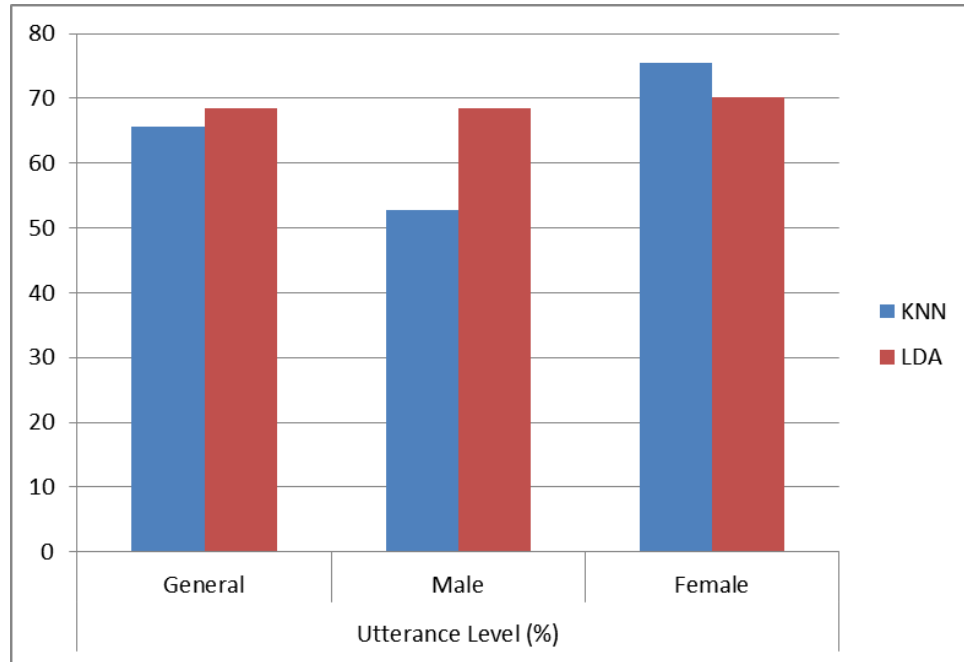


Figure 4.12: Utterance level performance of the models using KNN and LDA

Table 4.4:

Emotion Recognition Performance At Utterance Level While Using 5th, 6th, 8th, 9th, 10th And 11th MFCC

Classifier	Gender	Utterance
	Dependent/Independent	Level (%)
KNN	General(Independent)	65.74
	Male	75.74
	Female	69.78
LDA	General	69.05
	Male	73.04
	Female	75.50

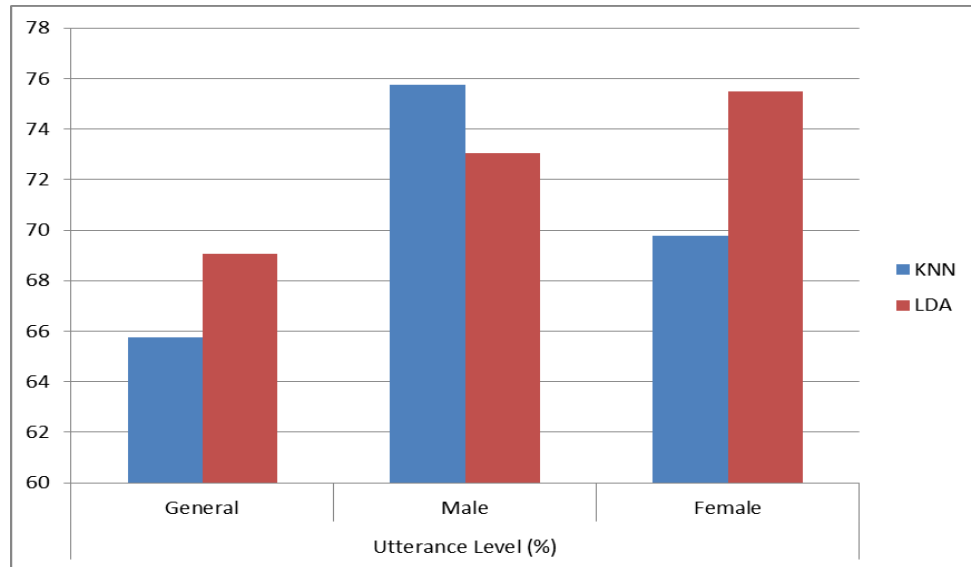


Figure 4.13: Utterance level performance of the models using KNN and LDA

In the study conducted by Koolagudi et al. (2011), prosodic speech features have been studied for emotion recognition performance at utterance, word and syllable level. Word boundaries were selected manually. Energy and pitch parameters were considered for the study. Altogether, eight emotions (anger, disgust, fear, happy, neutral, sad, sarcastic and surprise) were considered for the research. SVM classifier has been used for classification. Simulated emotion speech corpus IITKGP-SESC (*Indian Institute of Technology-Simulated Emotion Speech Corpus*) was used by the Koolagudi et al. (2011). For the entire utterance, average of 63.5% accuracy has been reported for anger and neutral emotions. As this study is concern, total of 77.1% accuracy has been achieved while considering the results of general models at utterance level for both KNN and LDA classifiers. Accuracy of conducted study also improves further for gender based models. Figure 4.14 provides the comparison of this study and study by Koolagudi et al. (2011).In

the figure 4.14, A denotes the study conducted by Koolagudi et al. (2011) and B refers this particular study.

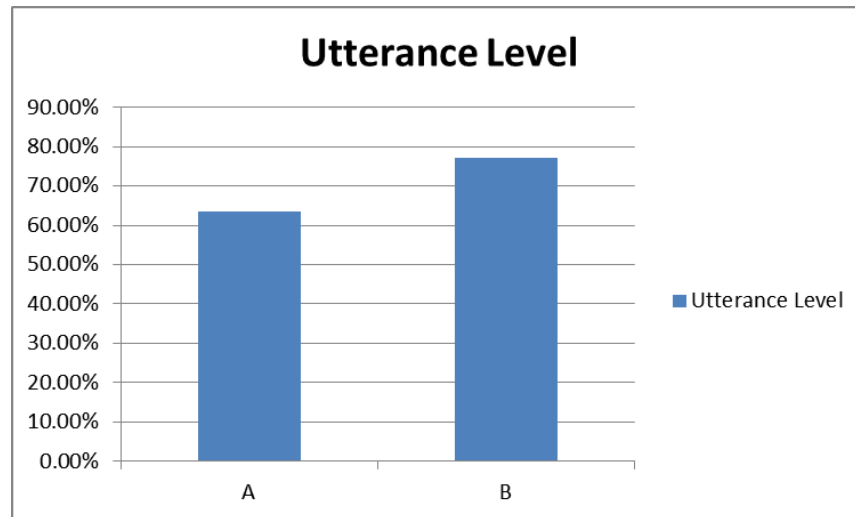


Figure 4.14: Performance of Study A and B

Rao et al. (2013) also conducted study on utterance level for gender independent and gender dependent models. Local and global prosodic features were used for the research. IITKGP-SESC and Emo-DB were used for speech utterances. Classification was performed using SVM. It is mentioned that global and local prosodic features can be used to classify anger and neutral emotions. For anger and neutral, average performance of 47.5% achieved by using global prosodic features extracted over entire utterance. While this particular study reported average of 77.1% accuracy while using global spectral feature at entire utterance. Performance of the study by Rao et al. (2013) improves further to 60.5% if used global and local prosodic features at utterance level. Figure 4.15 illustrates the difference of accuracies of Rao et al. (2013) (denoted as A) and this study (denoted as B) for general models at utterance levels.

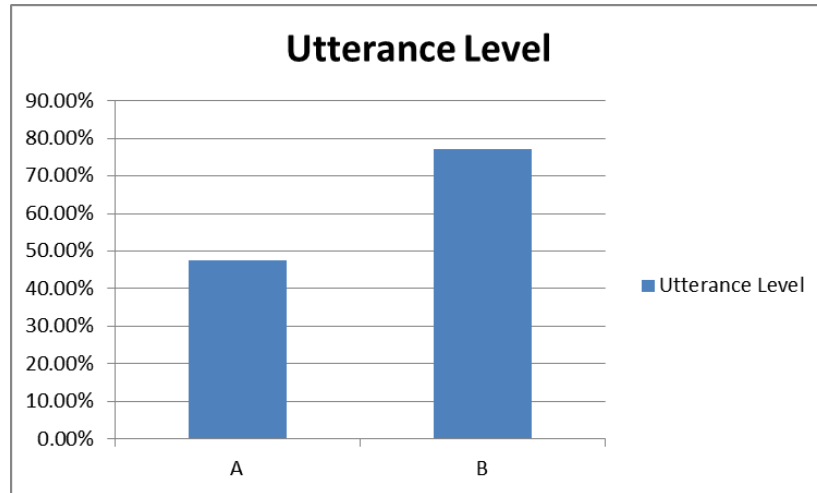


Figure 4.15: Performance of Study A and B

4.5.2 Emotion recognition using word-level global MFCC

To study the relevance of emotion recognition system at word-level, three different models are developed by using words at initial, middle and end positions for gender dependent and gender independent data. Emotion recognition performance of these models is given in table 4.5. It can be seen from the table that initial and ending positions of the sentence exhibits more emotion specific information than middle words. Accuracy of initial and ending position models is always high for gender dependent and independent data regardless of classifier. Mean of all 13 MFCC are used as feature vectors same as utterance level classification. Total average of all three models is given by calculating average of accuracies of the models.

Highest accuracy is obtained from female model for initial words i.e 97.77% for LDA classifier. Overall accuracy of female model for all words is 93.15% for LDA. The overall accuracy of general and male models are 84.38% and 87.91% respectively. For KNN classifier, highest overall accuracy is obtained from again female model i.e 88.2%.

From these findings, it can be said that females are more emotional than male and have more emotion specific information in their speech. Moreover, general models have less accuracy than gender dependent models for both classifiers. Another interesting pattern is that, for gender dependent models, ending words have high accuracy than others words for KNN, while for LDA, initial words have highest accuracy. Figures 4.16, 4.17 and 4.18 illustrate the performances of initial, middle and ending words models respectively.

Table 4.5: Emotion Recognition Performance At Word-Level

Classifier	Gender Dependent/Independent	Word Level (%)			
		Initial	Middle	End	Total Avg
KNN	General(Independent)	76.57	54.54	73.56	68.22
	Male	92.60	74.51	93.39	86.83
	Female	93.89	76.40	94.32	88.20
LDA	General	93.93	74.20	85.03	84.39
	Male	96.89	74.97	91.87	87.91
	Female	97.77	89.51	92.18	93.15

Emotion recognition performance was also calculated for different MFCC combinations like 5, 6, 8, 9, 10, and 11 and for 6, 9 and 10 MFCC. Results of these combinations were comparable at utterance level, but for word-level, their results are very low as compared to all 13 MFCC. Results are calculated for these combinations only for initial words for general model and compared with the results of all 13 MFCC. Table 4.6 shows the results obtained from experiments. It shows that around 20% less accuracy is given by 6, 9 and 10 MFCC for both classifiers. Such difference in performance is not negligible; therefore, results for remaining models are not calculated. Same is the case with other combination of MFCC, around 13% to 18% less accuracy is obtained for KNN and LDA respectively. Figure 4.19 illustrates this difference.

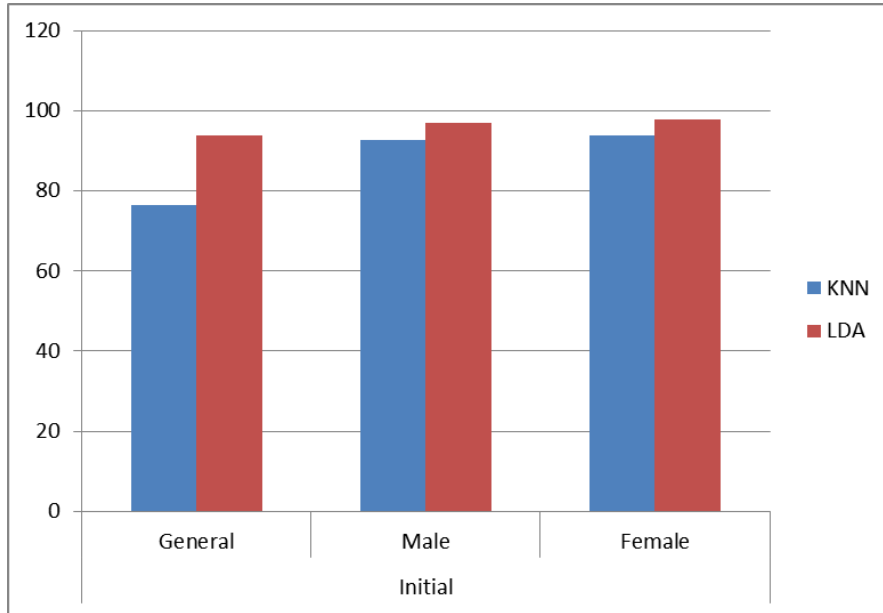


Figure 4.16: Emotion recognition performance of initial words

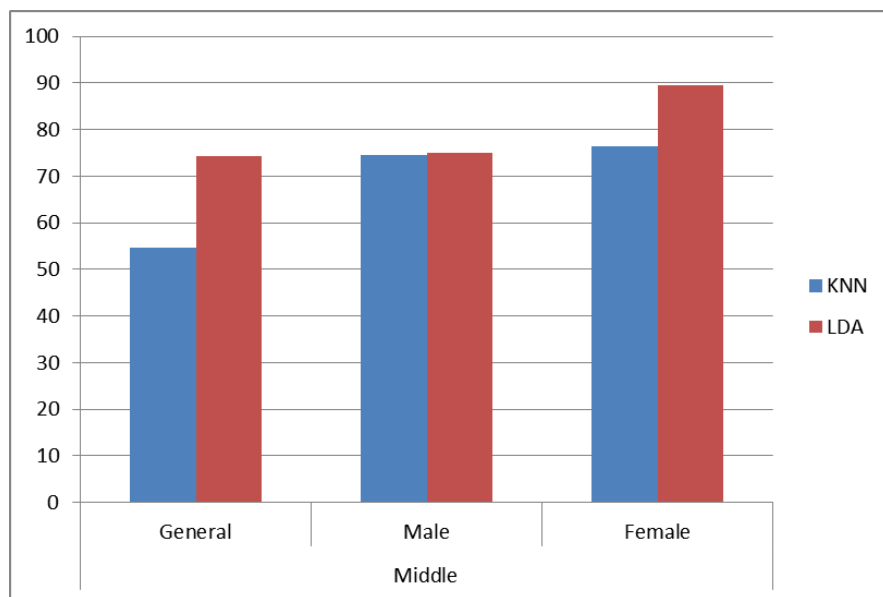


Figure 4.17: Emotion recognition performance of Middle words

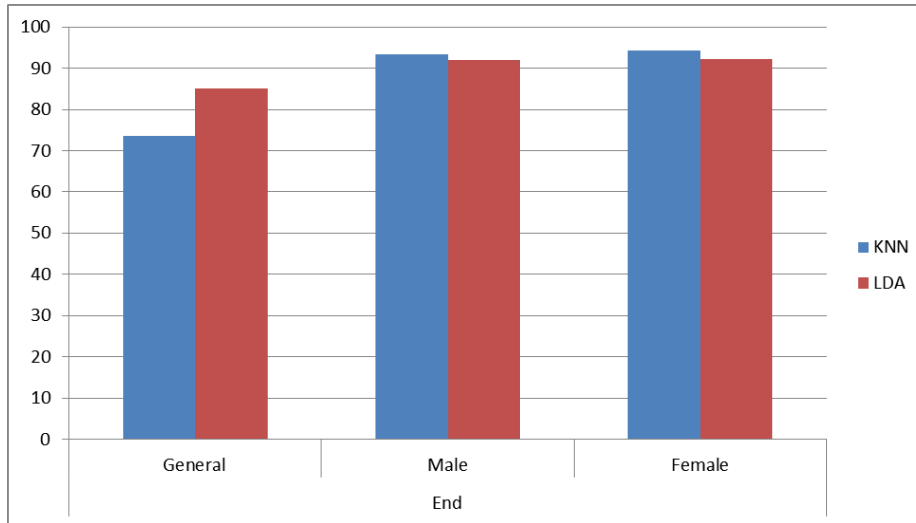


Figure 4.18: Emotion recognition performance of Ending words

Table 4.6:

Performance Of Different MFCC Combinations Of Initial Words For General Model In Percentages

Classifier	All 13	6, 9, 10	5, 6, 8, 9, 10, 11
KNN	76.57	52.02	63.33
LDA	93.93	71.54	74.90

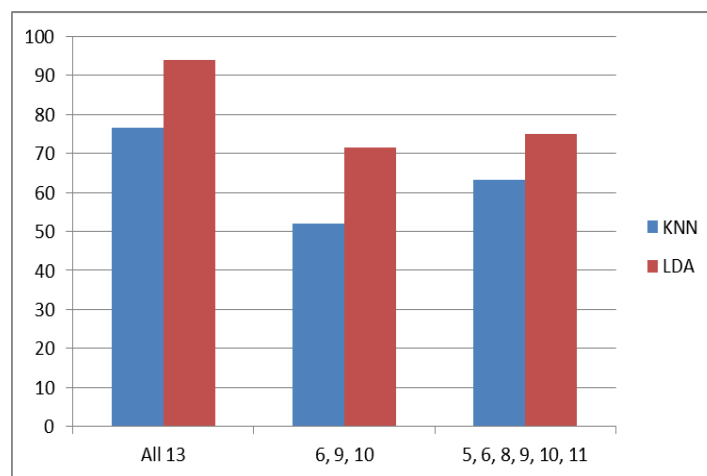


Figure 4.19: Performance of different MFCC combinations of initial words for general model

In the study conducted by Koolagudi et al. (2011), as mentioned above in section 4.5.1, for word level, average of 46.5%, 30% and 31.5% accuracy was achieved for beginning, middle and ending words respectively for anger and neutral emotions. As this study is concern, average accuracy of 85.25%, 64.47% and 79.29% for KNN and LDA has been reported for initial, middle and ending words respectively for general models. Figure 4.20 illustrates the difference of two studies at word level. Koolagudi et al. (2011) also mentioned that initial and ending positions of the words contain more emotional specific information for anger and neutral emotions, this particular study also obtained the same results. In the figure4.20, A denotes the study conducted by Koolagudi et al. (2011) and B refers this particular study.

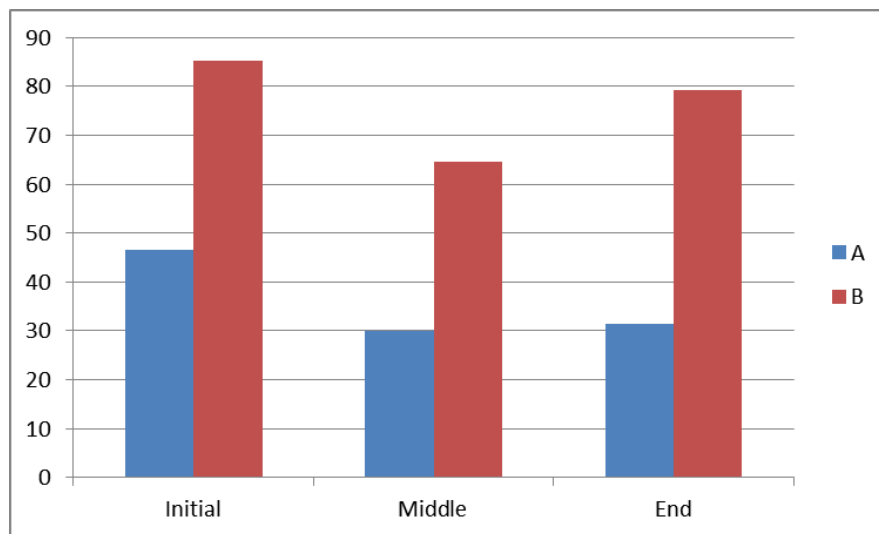


Figure 4.20: Performance of Study A and B

Rao et al. (2013) also conducted study on word level for gender independent and gender dependent models, as mentioned in section 4.5.1. Average of 45% and 51% accuracy were achieved for male and female models respectively for anger, disgust, happiness,

fear, neutral, sadness, sarcastic and surprise emotions. 31.5%, 46.5% and 51.5% average accuracy noted for anger and neutral emotions by using global prosodic features at initial, middle and final word position respectively (Rao et al., 2013). While this study reported the average of 87.68%, 68.48% and 87% for initial, middle and ending words respectively using global spectral features for KNN of general and gender based models. These results further improve for LDA classifier. For general models, 76.57%, 54.54% and 73.56% have been noted for KNN in this study. Rao et al. (2013) has mentioned the accuracy of 62% of combine local and prosodic features for anger and neutral at all positions of the word. Whereas this study has reported 68.23% and 84.38% for KNN and LDA respectively at all positions of the word of general models. 76.3% can be seen as average of both classifiers. Rao et al. (2013) also mentioned that words in the final position of the sentence provide more emotion specific information than other regions. Figure 4.21 illustrates the difference of accuracies of Rao et al. (2013) (denoted as A) and this study (denoted as B) for general models at different levels.

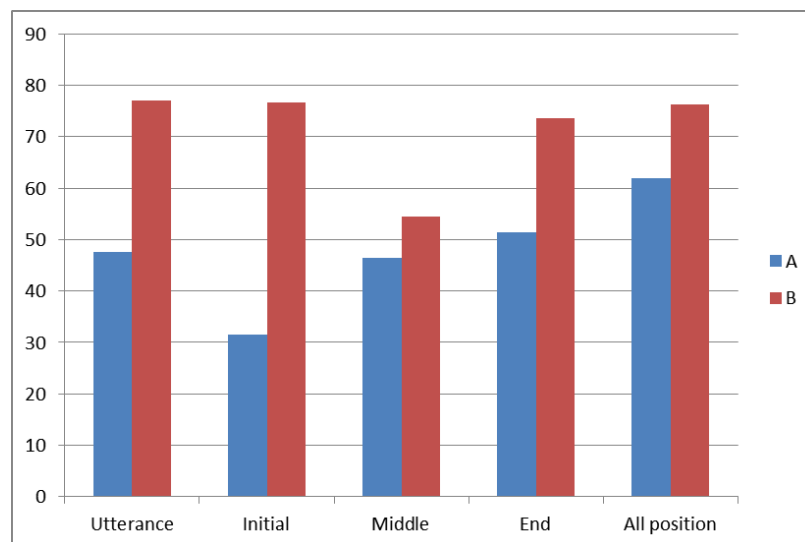


Figure 4.21: Performance of Study A and B

4.6 Summary

1. Only two emotion, anger and neutral have been considered for this study while using Emo-DB for speech utterances.
2. Results show that, emotion recognition performance of models using all 13 MFCCs is better than performance of other combinations of the MFCCs.
3. NN classifier has been proposed for this study but results show that NN is not effective for selected amount of features and emotions.
4. KNN and LDA classifiers have been evaluated with selected amount of features and emotions and results show that these classifiers perform better than NN.
5. Performance of gender dependent and independent models is also studied. It is observed from the results that, gender dependent models are more efficient and female models are more accurate than male models.
6. Word positions are also evaluated in this study. It is shown in the results that initial and ending positions of the words are more emotion specific.
7. Performance of MFCCs proposed by Gaurav (2008) is measured with the performance of MFCCs used in this study and it is found that MFCCs used in this study are more efficient.
8. Results of gender dependent models are more accurate than results presented in Bitouk et al. (2010) and Rao et al. (2013).
9. Results obtained at utterance and word level in this study are higher than results presented in Koolagudi et al. (2011) and Rao et al. (2013) for anger and neutral emotions. Whereas this study uses global spectral features and, Koolagudi et al.

(2011) and Rao et al. (2013) have used prosodic features and global prosodic features respectively.

10. Performance of spectral features and prosodic features at word level can be compared in this study and in Koolagudi et al. (2011) and Rao et al. (2013) to see importance of speech features at word level.
11. Koolagudi et al. (2011) and this study have mentioned that, words in initial and final/ending positions provide more emotion specific information, while Rao et al. (2013) mentioned that words at final position are more efficient.
12. Performance of words at all positions is also calculated in this study and in Rao et al. (2013). It is observed that, results of this study are higher than results in Rao et al. (2013) while considering anger and neutral.

CHAPTER FIVE

CONCLUSION

5.1 Introduction

This Chapter provides a conclusion of this study in which spectral features analysis of emotional speech utterances have been performed at different levels of utterance for emotion recognition task. Two emotions, anger and neutral of Emo-DB are studied. Altogether, three classifiers NN, KNN and LDA are used to develop models. MFCC is used as spectral information. Global spectral features are derived by computing statistical parameters like mean of spectral features extracted from sentence and words for developing the model. Word boundaries are manually identified.

5.2 MFCCs

In this study, different combinations of MFCCs are studied. Most of combinations of MFCC are selected randomly but those combination that are presented in the previous studies, are also studied. It is found that previously presented combinations are not as much efficient as all 13 MFCC. Other tested combination also does not provide better results than 13 MFCC for both utterance and word level for this particular study. Utterance level model also studied to compare the results of word level and utterance level of this study and also with previous studies. It is found that this study is more efficient with the studies of Koolagudi et al. (2011) and Rao et al. (2013) for utterance level feature extraction.

5.3 Word Level

In this study, the contribution of word positions for emotion recognition is also studied. Three different models have been developed for emotion recognition using global spectral features extracted from initial, middle and ending positions of the utterance for gender dependent and independent datasets. From the results obtained from word level spectral analysis, it is observed that, initial and ending positions of the words provide more emotion specific information than words at middle position as mentioned in previous studies. Moreover, results show that feature extraction for emotional speech recognition at word level is more efficient than at utterance level.

5.4 Gender dependent and independent

Performance of gender dependent and independent models is also observed in this study. Three different models have been developed for emotion recognition for general (gender independent), male and female. Results obtained from gender dependent and independent models, show that gender dependent models are more efficient than gender independent model. Female model provides better accuracy in terms of classification rate than other models. Hence it can be said that females are more emotional than males.

5.5 Future Work

This study focused on only two emotions, anger and neutral. Global spectral speech features are used for classification. In future, more emotions may be included in this study. Performance of local spectral, other spectral features, excitation source and prosodic features can be studied with the combination of these global spectral features. Other classifiers such as, GMM, SVM and HMM can be studied to evaluate performance of emotion recognition systems.

References

- Agarwal, A., Jain, A., Prakash, N., & Agrawal, S. S. (2010). Word boundary detection in continuous speech based on suprasegmental features for hindi language. *2010 2nd International Conference on Signal Processing Systems*, V2–591–V2–594. doi:10.1109/ICSPS.2010.5555691
- Anagnostopoulos, C.-N., & Iliou, T. (2010). Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research. In M. Wallace, I. Anagnostopoulos, P. Mylonas & M. Bieliková (ed.), *Semantics in Adaptive and Personalized Services*, Vol. 279 (pp. 127-143). Springer. ISBN: 978-3-642-11683-4.
- Bajpai, A., & Yegnanarayana, B. (2008). Combining evidence from subsegmental and segmental features for audio clip classification. *TENCON 2008 - 2008 IEEE Region 10 Conference*, 1–5. doi:10.1109/TENCON.2008.4766692
- Bapineedu, G., Avinash, B., Gangashetty, S. V., & Yegnanarayana, B. (2009). Analysis of lombard speech using excitation source information. In *INTERSPEECH-09*, Brighton, UK, 6–10 September 2009 (pp. 1091–1094).
- Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, 52, 613-625.
- Bozkurt, E., Erzin, E., Erdem, Ç. E., & Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals. *INTERSPEECH* (p./pp. 324-327), : ISCA.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *INTERSPEECH* (p./pp. 1517-1520), : ISCA.
- Chauhan, A., Koolagudi, S. G., Kafley, S., & Rao, K. S. (2010). Emotion recognition using lp residual. In *IEEE TechSym 2010*, West Bengal, India, April 2010. IITKharagpur: IEEE.
- Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2), 33–60. doi:10.1016/S0167-6393(02)00070-5
- Gaurav, M. (2008). Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech. In A. Das & S. Bangalore (eds.), *SLT* (p./pp. 313-316), : IEEE. ISBN: 978-1-4244-3472-5

- Han, Z., Lun, S., & Wang, J. (2012). A Study on Speech Emotion Recognition Based on CCBC and Neural Network. *2012 International Conference on Computer Science and Electronics Engineering*, 144–147. doi:10.1109/ICCSEE.2012.128
- Hanjalic, a. (2005). Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1), 143–154. doi:10.1109/TMM.2004.840618
- Hansen, J. H. L., Bou-Ghazale, S. E., Sarikaya, R., & Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database.. In G. Kokkinakis, N. Fakotakis & E. Dermatas (eds.), *EUROSPEECH*, : ISCA.
- Hansen, J. H. L., Sangwan, A., & Kim, W. (2012). Speech Under Stress and Lombard Effect - Impact and Solutions for Forensic Speaker Recognition. *Forensic Speaker Recognition*. (A. Neustein & H. A. Patil, Eds.). doi:10.1007/978-1-4614-0263-3
- Henríquez Rodríguez, P., Alonso Hernández, J. B., FerrerBallester, M. a., Travieso González, C. M., & Orozco-Aroyave, J. R. (2012). Global Selection of Features for Nonlinear Dynamics Characterization of Emotional Speech. *Cognitive Computation*, 5(4), 517–525. doi:10.1007/s12559-012-9157-0
- Henríquez, P., Alonso, J. B., Ferrer, M. A., Travieso, C. M., & Orozco-Aroyave, J. R. (2011). Application of Nonlinear Dynamics Characterization to Emotional Speech.. In C. M. Travieso-González & J. B. A. Hernández (eds.), *NOLISP* (p./pp. 127-136), : Springer. ISBN: 978-3-642-25019-4
- Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, 40(1-2), 161–187. doi:10.1016/S0167-6393(02)00081-X
- Iliev, A. I., Scordilis, M. S., Papa, J. P., & Falcão, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language*, 24(3), 445–460. doi:10.1016/j.csl.2009.02.005
- Iliou, T., & Anagnostopoulos, C.-N. (2009). Comparison of Different Classifiers for Emotion Recognition. *2009 13th Panhellenic Conference on Informatics*, 102–106. doi:10.1109/PCI.2009.7
- Iliou, T., & Anagnostopoulos, C.-N. (2009b). Statistical Evaluation of Speech Features for Emotion Recognition. *2009 Fourth International Conference on Digital Telecommunications*, 121–126. doi:10.1109/ICDT.2009.30
- Kamaruddin, N., & Wahab, A. (2009). Features extraction for speech emotion. *Journal of Computational Methods in Science and Engineering*, 9(9), 1–12.

- Khanna, P., & Kumar, M. S. (2011). Application of Vector Quantization in Emotion Recognition from Human Speech.. In S. Dua, S. Sahni & D. P. Goyal (eds.), *ICISTM* (p./pp. 118-125), : Springer. ISBN: 978-3-642-19422-1
- Kishore, K. V. K., & Satish, P. K. (2013). Emotion Recognition in Speech Using MFCC and Wavelet Features. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International* , vol., no., pp.842,847
- Kodukula, R. SRI. (2009). Significance of excitation source information for speech analysis. Unpublished doctoral dissertation, Dept. of Computer Science and Engineering, IIT Madras, Chennai, India.
- Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, 15(2), 265–289. doi:10.1007/s10772-012-9139-3
- Koolagudi, S. G., & Rao, K. S. (2012b). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2), 99-117.
- Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011). Speech Emotion Recognition Using Segmental Level Prosodic Analysis. *2011 International Conference on Devices and Communications (ICDeCom)*, 1–5. doi:10.1109/ICDECOM.2011.5738536
- Koolagudi, S. G., Maity, S., Vuppala, A. K., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech Database for Emotion Analysis. In S. Ranka, S. Aluru, R. Buyya, Y.-C. Chung, S. Dua, A. Grama, S. K. S. Gupta, R. Kumar & V. V. Phoha (eds.), *IC3* (p./pp. 485-492), : Springer. ISBN: 978-3-642-03546-3
- Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. *International Conference on Signal Processing and Communications (SPCOM)*, 1–5. doi:10.1109/SPCOM.2010.5560541
- Koolagudi, S. G., Reddy, R., Yadav, J., & Rao, K. S. (2011b). IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In *Devices and Communications (ICDeCom), 2011 International Conference on* (pp. 1-5). IEEE.
- Kuchibhotla, S., Yalamanchili, B. S., Vankayalapati, H. D., & Anne, K. R. (2014, January). Speech Emotion Recognition Using Regularized Discriminant Analysis. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013* (pp. 363-369). Springer International Publishing.
- Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs.. *IEEE Transactions on Speech and Audio Processing*, 13, 293-303.

- Lewis, M., Haviland-Jones, J. M., & Barrett, L. F., (2008). *Handbook of Emotions*. Third Edition. The Guilford Press ISBN: 1609180445
- Ling He. (2010). Stress and Emotion Recognition in Natural Speech in the Work and Family Environments. Unpublished doctoral dissertation, School of Electrical and Computer Engineering, RMIT University.
- Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *ICASSP*, Honolulu, Hawaii, USA, May 2007 (pp. IV17–IV20). New York: IEEE Press.
- Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 561–580.
- Mao, X., Chen, L., & Fu, L. (2009). Multi-level Speech Emotion Recognition Based on HMM and ANN. *2009 WRI World Congress on Computer Science and Information Engineering*, 225–229. doi:10.1109/CSIE.2009.113
- Mubarak, O. M., Ambikairajah, E., & Epps, J. (2005). Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. In *8th international symposium on signal processing and its applications*, Sydney, Australia, Aug. 2005.
- MY, S. A. (2014). An Improved Feature Extraction Method for Malay Vowel Recognition based on Spectrum Delta.
- Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion Recognition in Spontaneous Speech Using GMMs Department of Speech, Music and Hearing, KTH, Stockholm, Sweden Classifiers, 809–812.
- Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4), 603–623. doi:10.1016/S0167-6393(03)00099-2
- Pao, T.-L., Chen, Y.-T., Yeh, J.-H., Cheng, Y.-M., & Chien, C. S. (2007). Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech. In A. Paiva, R. Prada & R. W. Picard (eds.), *ACII* (p./pp. 741-742), : Springer. ISBN: 978-3-540-74888-5
- Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. (2005). Combining acoustic features for improved emotion recognition in mandarin speech. In *Affective Computing and Intelligent Interaction* (pp. 279-285). Springer Berlin Heidelberg.
- Pao, T. L., Liao, W. Y., & Chen, Y. T. (2008). A Weighted Discrete KNN Method for Mandarin Speech and Emotion Recognition. *Speech Recognition Technologies and Applications*, 550-552.

- Rabiner, L. R., Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall. ISBN: 978-0-13-015157-5
- Ramakrishnan, S., & Emary, I. M. M. (2011). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3), 1467–1478. doi:10.1007/s11235-011-9624-z
- Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, 16(2), 143–160. doi:10.1007/s10772-012-9172-2
- Rao, K. S., Prasanna, S. R. M., & Sagar, T. V. (2007b). Emotion recognition using multilevel prosodic information. In *Workshop on image and signal processing (WISP-2007)*, Guwahati, India, Dec. 2007. Guwahati: IIT Guwahati.
- Rao, K. S., Prasanna, S. R. M., Yegnanarayana, B., & Member, S. (2007a). Determination of Instants of Significant Excitation in Speech Using Hilbert Envelope and Group Delay Function, 14(10), 762–765.
- Rao, K. S., Reddy, R., Maity, S., & Koolagudi, S. G. (2010) Characterization of emotions using the dynamics of prosodic features. School of Information Technology Indian Institute of Technology Kharagpur.
- Rao, K. S., & Koolagudi, S. G. (2012). *Emotion Recognition Using Speech Features*. Springer.
- Rojas, R. (1996). *Neural networks: a systematic introduction*. Springer.
- Rong, J., Li, G., & Chen, Y.-P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*, 45(3), 315–328. doi:10.1016/j.ipm.2008.09.003
- Scherer, S., Schwenker, F., & Palm, G. (2008). Emotion Recognition from Speech Using Multi-Classifer Systems and RBF-Ensembles.. In B. Prasad & S. R. M. Prasanna (ed.), *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, Vol. 83 (pp. 49-70) . Springer . ISBN: 978-3-540-75397-1.
- Schroder, M., & Cowie, R. (2006). Issues in emotion-oriented computing toward a shared understanding. In *Workshop on emotion and computing (HUMAINE)*.
- Sethu, V., Ambikairajah, E., & Epps, J. (2013). On the use of speech parameter contours for emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1), 19. doi:10.1186/1687-4722-2013-19

- Shi, Y., & Song, W. (2010). Speech emotion recognition based on data mining technology. *2010 Sixth International Conference on Natural Computation, (Icnc)*, 615–619. doi:10.1109/ICNC.2010.5583142
- Staroniewicz, P. (2009). Recognition of Emotional State in Polish Speech - Comparison between Human and Automatic Efficiency, 33–40.
- Staroniewicz, P. (2011). Automatic Recognition of Emotional State in Polish Speech.. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller & G. Scarpetta (eds.), *COST 2102 Training School* (p./pp. 347-353), : Springer. ISBN: 978-3-642-18183-2
- Staroniewicz, P. (2011a). Influence of Speakers' Emotional States on Voice Recognition Scores.. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud & A. Nijholt (eds.), *COST 2102 Conference* (p./pp. 223-228), : Springer. ISBN: 978-3-642-25774-2
- Tao, J., & Kang, Y. (2005). Features Importance Analysis for Emotional Speech Classification.. In J. Tao, T. Tan & R. W. Picard (eds.), *ACII* (p./pp. 449-457), : Springer. ISBN: 3-540-29621-2
- Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods.. *Speech Communication*, 48, 1162-1181.
- Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *ICASSP* (pp. I593–I596). New York: IEEE Press.
- Vicsi, K., & Sztahó, D. (2011). Problems of the Automatic Emotion Recognitions in Spontaneous Speech; An Example for the Recognition in a Dispatcher Center.. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller & G. Scarpetta (eds.), *COST 2102 Training School* (p./pp. 331-339), : Springer. ISBN: 978-3-642-18183-2
- Wang, J., Han, Z., & Lung, S. (2011, October). Speech emotion recognition system based on genetic algorithm and neural network. *International Conference on Image Analysis and Signal Processing (IASP)*, (pp. 578-582).
- Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and Optimal Classification of Speech Emotion Recognition. *2008 Fourth International Conference on Natural Computation*, 407–411. doi:10.1109/ICNC.2008.713
- Wang, Y., & Guan, L. (2004). An investigation of speech-based human emotion recognition. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on* (pp. 15-18). IEEE.
- Wu, C., Member, S., & Liang, W. (2011). Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels, 2(1), 10–21.

- Wu, C., Yeh, J., & Chuang, Z. (2009b). Emotion perception and recognition from speech. *Affective Information Processing*. Retrieved from http://link.springer.com/chapter/10.1007/978-1-84800-306-4_6
- Wu, S., Falk, T.H., & Chan, W. Y., (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. *Digital Signal Processing, 2009 16th International Conference on* , vol., no., pp.1,6, 5-7 July.
- Yegnanarayana, B., Member, S., Swamy, R. K., & Murty, K. S. R. (2009). Determining Mixing Parameters From Multispeaker Data Using Speech-Specific Information, *17(6)*, 1196–1207.
- Yong-Wan, R., Kim, D.-J., Lee, W.-S., & Hong, K.-S. (2009). Novel acoustic features for speech emotion recognition. *Science in China Series E: Technological Sciences*, *52(7)*, 1838–1848. doi:10.1007/s11431-009-0204-3
- You, M., Chen, C., Bu, J., Liu, J., Tao, J. (2007). Emotion recognition from speech signals combining PCA and LDA. *17(6)*, 1196–1207.
- Zhang, H. (2012). Emotional Speech Recognition Based on Syllable Distribution Feature Extraction, In *Advances in Intelligent and Soft Computing (AISC) Volume 122*, 2012, pp 415-420.
- Zhao, X., Zhang, S., & Lei, B. (2013). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*. doi:10.1007/s00521-013-1377-z
- Zhiyan, H., & Jian, W. (2013). Speech emotion recognition based on wavelet transform and improved HMM. *2013 25th Chinese Control and Decision Conference (CCDC)*, 3156–3159. doi:10.1109/CCDC.2013.6561489
- Zhou, Y., Sun, Y., Yang, L., & Yan, Y. (2009). Applying Articulatory Features to Speech Emotion Recognition. *2009 International Conference on Research Challenges in Computer Science*, 73–76. doi:10.1109/ICRCCS.2009.26