AN APPROACH FOR FEMALE SPEECH EMOTION RECOGNITION IN THE INDIAN CONTEXT

A thesis submitted by

AGNES JACOB

for the award of the degree of

DOCTOR OF PHILOSOPHY

(Faculty of Engineering)

Under the guidance of

Dr. P. MYTHILI

DIVISION OF ELECTRONICS, SCHOOL OF ENGINEERING
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
KOCHI - 682 022
INDIA

December 2014

AN APPROACH FOR FEMALE SPEECH EMOTION RECOGNITION IN THE INDIAN CONTEXT

Ph. D. Thesis in the field of Speech Processing

Author

Agnes Jacob
Research scholar
Division of Electronics Engineering
School of Engineering
Cochin University of Science and Technology
Kochi - 682 022, Kerala, India
E-mail: agneswills3@gmail.com

Research Advisor

Dr. P. Mythili
Associate Professor
Division of Electronics Engineering
School of Engineering
Cochin University of Science and Technology
Kochi - 682 022, Kerala, India
E-mail: mythili@cusat.ac.in

December 2014

Dedicated to…….

The Almighty,

&

My parents

Certificate

This is to certify that the thesis entitled ***"An approach for Female Speech Emotion Recognition in the Indian context"*** *is a bonafide record of research work carried out by Mrs. Agnes Jacob under my supervision and guidance in the Division of Electronics Engineering, School of Engineering, Cochin University of Science and Technology, Kochi. No part of this thesis has been presented for the award of any other degree from any other university.*

Kochi
6th December 2014

Dr. P. Mythili
(Supervising Guide)
Associate Professor
Division of Electronics Engineering
Cochin University of Science and Technology
Kochi, Kerala, India.

Declaration

I hereby declare that the work presented in this thesis entitled, "An approach for Female Speech Emotion Recognition in the Indian context" is based on the original research carried out by me, under the supervision and guidance of Dr P. Mythili, Associate Professor, Division of Electronics, School of Engineering, Cochin University of Science and Technology, Kochi - 22. This work did not form part of any thesis submitted for the award of any degree, diploma, or other similar title or recognition from this or any other institution.

Kochi - 682 022 *Agnes Jacob*
6th December 2014

Acknowledgements

I express my gratitude to my research Guide, Dr. P. Mythili, Associate Professor, Cochin University of Science and Technology, for motivating me to undertake this interdisciplinary work by her professional, personal guidance and constant support at all stages of this research.

I heartily thank Dr. R. Gopikakumari, for her constructive comments in all the Ph. D. reviews and several fruitful discussions.

I thank the Head of the Department and other Faculty of the Division of Electronics for their valuable suggestions.

I sincerely thank Dr. Sumam Mary Idicula, Head of the Computer Science Department, for her valuable suggestions in improving the presentation of this thesis.

I am indebted to A.J Paul, Baby Paul, Shanavas K.T, Rema, Anjith, Philip Cherian and Biju for their kind cooperation at various stages of this research. I take this opportunity to specially thank all other people in the department for their goodwill. I would like to thank the Principal and office staff for all other support.

I thank all the people who patiently rendered their voice for the recordings and those others who helped me with the recordings. I remember with gratitude all others who have contributed to this research by their ideas and inspiring words.

I am truly grateful to my parents for their help, encouragement and prayers that enabled me to complete this work. I thank my mother for her unfailing trust in my abilities. Thanks to Wills, Elizabeth, Sridevi and all other near and dear ones for their firm support.

.

Agnes Jacob

Abstract

Speech emotion recognition (SER) has an increasingly significant role in the interactions among human beings, as well as between human beings and computers. Emotions are an inherent part of even rational decision making. The correct recognition of the emotional content of an utterance assumes the same level of significance as the proper understanding of the semantic content and is an essential element of professional success.

Prevalent speech emotion recognition methods generally use a large number of features and considerable signal processing effort. On the other hand, this work presents an approach for SER using minimal features extracted from appropriate, sociolinguistically designed and developed emotional speech databases. Whereas most of the reported works in SER are based on acted speech with its exaggerated display of emotions, this work focuses on elicited emotional speech in which emotions are induced. Since female speech is more expressive of emotions, this research investigates the recognition of emotions from the speech of educated, urban females in the age group of 24 to 42 years. The context of this research is set by SER for English - a global language, and for two Indian languages namely, Hindi the national language and Malayalam the native language of Kerala.

The investigations are done using prosodic features (intensity, pitch, duration / speech rate), their variations (jitter and shimmer), and spectral features (first four formants and their bandwidths), extracted from the emotional speech databases developed exclusively for this research. This approach makes use of multiple classifiers for SER, as well as for verifying the consistency of the obtained SER rates. The KMeans, Fuzzy C - Means (FCM), K - Nearest neighbor (KNN), Naive Bayes (NB) and Artificial Neural networks (ANN) are used for the recognition of neutral and the six basic emotions comprising happiness, surprise, anger, sadness, fear and disgust. Happiness and surprise are of positive valence, whereas anger, sadness, fear and disgust are of negative valence. One of the objectives of this research is to investigate the valence dependency of SER rates.

After introducing the basic concepts of SER and surveying the relevant literature, this thesis discusses the results of statistical analysis and SER in English, Hindi and Malayalam at the suprasegmental level, using three popular prosodic features. Similar investigations are carried out at the segmental level too in English, for each feature, in order to compare the performance of SER between these two levels. The segmental utterances chosen for analysis in English were five widely used vowels which have independent existence and meaning. Similar segmental utterances in Hindi or Malayalam were not chosen for SER since these do not possess independent existence from the semantic or emotional perspective. English being a syllable based language; SER based on syllable speech rates is also analyzed. The individual contributions of each prosodic feature, as well as their combined role in SER are assessed for each language. Incidences of universality in the vocal expression of emotions across English, Hindi and Malayalam are identified. The KMeans, NB, KNN and the ANN classifiers have been used for the prosodic feature based classification of emotions. Additionally, classification by the FCM method also is investigated for segmental and suprasegmental utterance in English. The classification results indicate improved SER for statistically well discriminated feature values. The final results are validated with new emotional speech samples and the results of human SER. Since there are no available results for such prosodic feature based SER in Indian English, Hindi as well as Malayalam, the obtained results are compared with those available in literature, for the prosodic features in other languages.

The thesis next addresses SER in English, Hindi and Malayalam using micro perturbations in pitch, called jitter, as well as very small variations in intensity, called shimmer. Jitter and shimmer, are proposed as features for SER, since it is difficult to bring about minute variations in intensity and pitch, artificially, without actually experiencing the emotions. Therefore, more than certain other observable prosodic features, which can be acted, jitter and shimmer are expected to reflect true emotions only. The investigations are carried out separately, at the segmental level and suprasegmental level, based on jitter, shimmer and their combination (at the suprasegmental level). Subsequently, universality in

emotion recognition across the three languages is assessed for each emotion. Performance comparisons with other features and with SER in other languages demonstrate the effectiveness of jitter and shimmer in speech emotion recognition.

Finally this research investigates the use of a minimum number of formants and bandwidths in an efficient, yet simple approach to classify neutral and six basic emotions in English, Hindi and Malayalam. For each language, the best vocal tract features - formants and bandwidths are identified by the KMeans, KNN and NB classification of individual features. This is followed by the ANN classification using the best features. In English, the formant based classification accuracies for each emotion, are compared between the segmental and suprasegmental levels. The effect of reduction in the number of emotion classes, on the emotion classification accuracy, and the universality in vocal expressions of emotions across the three languages, are also investigated. This chapter further reports the results of the classification of vowels on the basis of first four formants. Quantitative information regarding the discrimination of vowel utterances based on formants, and the identification of emotions suitable for utterance discrimination has not been reported so far. Lastly, this thesis presents insightful modeling of the SER in Malayalam by using Decision trees and Logistic regression, based on formants and bandwidths.

This thesis concludes by assessing its main contributions from the perspective of the proposed objectives and gives suggestions for future work.

Contents

List of Figures

List of Tables

LIST OF SYMBOLS AND ABBREVIATIONS

Symbols	
B1	First Bandwidth
B2	Second Bandwidth
B3	Third Bandwidth
B4	Fourth Bandwidth
F0	Fundamental frequency
F1	First Formant frequency
F2	Second Formant frequency
F3	Third Formant frequency
F4	Fourth Formant frequency
H0	Null Hypothesis
H1	Research Hypothesis
Abbreviations	
ACR	Absolute Category Rating
AFVC	Advanced Feature Vector Classification
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AP	Acoustic Prosodic
ATIS	Air Travel Information System
CD	Compact Disc
dB	Decibels
3DEC	Data-Driven Dimensional Emotion Classification
DES	Danish Emotional Speech
EI	Emotional Intelligence
EMA	Electro Magnetic Articulograph
EMODB	Berlin Emotional database
EP	Emotion Profile
EQ	Emotional Quotient
FCM	Fuzzy C Means

GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IQ	Intelligence Quotient
IT	Information Technology
kNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LPC	Linear Predictive Coding
LSTM	Long Short Term Memory
MFCC	Mel-Frequency Cepstrum Coefficient
MLP	Multi Layer Perceptron
MOS	Mean Opinion Score
MSE	Mean Squared Error
NB	Naive Bayes
P	Significance Value
RSS	Ratio of Spectral flatness measure to the Spectral center
SD	Standard Deviation
SEM	Standard Error of Mean
SEMAINE	Sustained Emotionally colored Machine - human Interaction using Nonverbal Expression
SER	Speech Emotion Recognition
SNK	Student Neumann Keul
SR	Speech Recognition
SUSAS	Speech Under Simulated and Actual Stress
SVM	Support Vector Machines
TIMIT	Speech database recorded at Texas Instruments and transcribed at the Massachusetts Institute of Technology
VAM	Vera Am Mittag

Chapter 1

INTRODUCTION

Contents

This introductory chapter of the thesis begins with an overview of the concepts of speech, emotion, significance of speech emotion recognition and its implications in the Indian context. The subsequent sections present the motivation, problem statement, research objectives and a brief introduction to the features used in these investigations in SER. The specific contributions of this thesis that distinguish it from previous work in this field are mentioned. The chapter concludes with an outline of the structure of this thesis.

1.1. Overview

From time immemorial, communication among human beings has been inherently multimodal - visual and aural being the primary modes. While the visual mode is the most effective in capturing information, speech remains the preferred and most convenient means of conveying information. The role of oral communications has been enhanced by the shift from machine-centric to human-centric, human-computer interfaces which has become the need of the day.

Speech as the medium of communication, conveys information on several layers; most important being the linguistic layer and the paralinguistic layer. The linguistic layer carries the semantic information in the text of the

utterance. The paralinguistic layer of communication is non-linguistic, non-verbal and tells the listener about the speaker's current affective, attitudinal or emotional state. Paralinguistic features include variations in pitch, intensity and spectral properties that have no linguistic functions and are therefore irrelevant to word identity [1]. This research in SER therefore, focuses on the paralinguistic layer of speech.

Etymologically the word emotion is a composite formed from two Latin words; 'ex' or out, outward and 'motio' or movement, action, gesture. This classical formation refers to the immediate nature of emotion as experienced by humans and attributed in some cultures and ways of thinking, to all living organisms. An emotion is defined as a psychological state or process that functions in the management of goals. It is typically elicited by evaluating an event as relevant to a goal. It is of positive valence when the goal is advanced, and is of negative valence when the goal is impeded. The core of an emotion is readiness to act in a certain way, with the prioritization of some goals and plans, rather than others [2]. Thus an emotion involves physiological arousal, expressive behavior and conscious experience.

Emotions can be primary or secondary. Whereas primary emotions are innate, secondary emotions refer to feelings attached to objects, events, and situations, through learning. Emotion came to have its contemporary meaning only in the late nineteenth century [3]. Even though emotion received little research attention earlier, it gained acceptance in the last half a century probably due to Ekman's influential, cross cultural studies of human facial expressions, which implied an innate biological basis for emotional experience [4].

This research has been conducted on neutral and the basic set of emotions comprising happiness, surprise, anger, sadness, fear and disgust.

Whereas happiness and surprise are of positive valence, anger, sadness, fear and disgust are of negative valence.

SER ultimately aims at the automatic detection of emotions in speech signals by analyzing the vocal behavior as a marker of affect (e.g. emotions, moods); focusing on the nonverbal aspect of speech [5]. The acoustic feature based SER assumes that the emotional arousal of the speaker is accompanied by spontaneous physiological changes that affect respiration, phonation, and articulation. These in turn produce emotion specific patterns of acoustic parameters [6]. Besides SER, presently research is being done on recognition of emotions from facial expressions, as well as based on multimodal databases comprising both audio and video information.

The increasing significance of research in speech and emotions has varied perspectives. Emotions are an essential part of our existence, acting as sensitive catalysts and assist in the development and regulation of interpersonal relationships [7]. Over the past few decades numerous studies have been done on speech, either for its synthesis or for various analysis purposes. These include deducing the age, sex, height and weight of the speaker, as well as for evaluating his / her state of health with respect to certain specific diseases. Even though there have been many studies done separately on emotions as well as voice, these have been concerned with other factors such as speaker sobriety, emotions and credibility of the content conveyed. Certain other investigations were concerning emotion and memory, forensic profiling - which deals with the critical task of identification of voices recorded by phone tapping, and the detection of deception from voice [8], to list a few popular studies. Recently, SER has become significant in call center applications, in order to improve customer service.

Current research on the neural circuitry of emotion suggests that emotion makes up an essential part of human decision-making. Therefore the need to face a changing and unpredictable world makes emotions necessary for any intelligent system (natural or artificial) with multiple motives and limited capacities and resources [9].

Recent research is concerned with the effects of emotions and moods. A mood has similar basis to an emotion, but lasts longer (hours to weeks). Emotions are often associated with brief (lasting few seconds to minutes) expressions of face and voice along with perturbation of the autonomic nervous system. Such manifestations often go unnoticed by the person who experiences the emotion. Moods are less specific, less intense and less likely to be triggered by a particular stimulus or event [10]. Whereas an emotion tends to change the course of action, a mood tends to resist disruption. The experience of emotions is important since, personality traits mostly with an emotional basis, last for several years or even a lifetime [11].

Emotional Intelligence is increasingly being promoted as necessary for successful teamwork [12]. Emotional analysis of one's speech serves to enhance the self-awareness, thereby improving the personal power and functioning. It helps one to analyze and modify one's own speaking style, or adopt a benchmarked speaking style. There is a pressing need for such awareness in the prevalent cross-cultural work environments, especially in the Information Technology (IT) sectors. Such emotion consciousness leads to improved inter-personal communication skills and enhances the Emotional Quotient (EQ) of the employees; which will in turn, influence their career graph favorably.

Results of SER based on various features can be made use of, in the synthesis of emotional speech as well.

1.2. SER in the Indian Context

This research focuses on emotional expressions in three languages, English, Hindi and Malayalam, the choice of which was made, based on the following facts.

Though popular emotional speech databases in English are available on the Internet and have been the basis for most of the researches in this field, these are inappropriate for SER in the Indian context due to their foreign accent. Besides, there were no readily available, public speech databases for Hindi, the national language and for Malayalam, the native language of Kerala. There are no existing reports of similar analysis in the Indian context, across these three languages.

Indian English is one of the main regional standards of English. Within this regional variety, a number of highly differentiated local dialects are found. None of the publicly available, popular emotional speech databases like the Danish Emotional Speech (DES) corpus, German Aibo emotion corpus and the Berlin database of Emotional speech (EMO-DB) include these versions [13].

English is basically a stress-timed, syllable based language. Articulatory gestures associated with the opening and closing of the jaws and lips are synchronised to the syllable. Meaningful words can be formed using one or more syllables. Hence in this research, SER based on syllable rates too are investigated for English. A part of this investigation in English focuses on SER at the segmental level using the vowels a, e, i, o, u. These sound segments possess an independent, stand alone nature unlike vowels in Hindi and Malayalam. Phonetically these include diphthongs too. Whereas a pure vowel remains constant and does not glide, diphthongs are sounds which glide from one vowel to another.

- The English speech database developed in this work, is based on the standardised global English spread by the media and the internet, and collected from Indian speakers.

Hindi is the fifth most spoken language in the world with about 188 million native speakers and is written in the Devanagiri script. Hindi is the national language of India and is spoken by about 43% of the population of India [14]. It is an Indo Aryan language, with Sanskrit as the major supplier of Hindi words. There are more than twenty dialects used for speaking Hindi. Hindi is mostly phonetic in nature i.e. there is one to one correspondence between written symbols and the spoken utterances, unlike in English. Hindi is syllabic in nature: each syllable contains a vowel as nucleus, followed or surrounded by consonants.

- The Hindi database used in these investigations was collected from South Indian females, speaking Hindi as their second language.

Malayalam is the official language of Kerala, the most literate state in India.. It is spoken by thirty five million people, mostly in Kerala and Lakshadweep . It belongs to the south Dravidian family of languages and is not syllable-based. There are fifty six alphabets in Malayalam. Though there are vowels and consonants, unlike in English, salient acoustic features are not solely attributed to vowels. Malayalam as spoken now has been influenced by Sanskrit and Tamil, apart from other regional, social and religious factors. Some of the most significant among the various dialects are the Thiruvananthapuram, mid-Travancore, Malappuram and North Malabar dialects.

- The Malayalam database used in this work has been based on the purest form as adopted by the educated, urban class, among the mid-Travancore Malayalees. The context of this research in SER is thus set by the choice of these three languages.

1.3. Feature Set Used

This research proposes to investigate SER using prosodic features, their variations and certain spectral features, the choice of which was made considering the following facts:

The characteristics of speech that have to do with individual speech sounds are referred to as segmental, while those that pertain to speech phenomena involving consecutive phones are referred to as suprasegmental and relates to the prosody. Prosodic components of speech are the constantly present, observable characteristics of speech, proven to be the most important in discriminating emotions, according to human perception [13]. Some of the most important prosodic features are the pitch statistics, loudness or intensity, speech rate and voice quality [15], [1].

Pitch has been defined as the percept of the fundamental frequency in the vibration of the vocal chords in voiced speech, and as the primary acoustic cue to intonation and stress in speech [16]. The pitch signal, also known as the glottal waveform, has information about emotion, since it depends on the tension of the vocal folds and the sub glottal air pressure which in turn, varies with emotion of the speaker [17]. Intensity or loudness has been correlated to activation - an important dimension of emotion. High intensity values imply high, and low intensity values imply low activation. Speech rate has been related to vocal effort, which again depends on emotions.

Voice quality features like jitter and shimmer are quantified as cycle to cycle variations of fundamental frequency and speech waveform amplitude respectively. Thus jitter and shimmer represent the variations in prosodic features. Measurements of jitter and shimmer commonly form part of a comprehensive voice examination and have been used in certain clinical studies to detect voice pathologies [18], [19].

Besides prosodic features, certain spectral features are also popular for SER. Some of the important spectral features are formants, which are vocal tract resonances that modify energy from the sound source. The shape of the vocal tract is modified by the emotional states, since the intake of air changes with emotions, especially those with strong arousal. Therefore acoustic variations due to emotion are expected to be reflected in the formants as well as their respective bandwidths.

1.4. Motivation

During the recent years, a principal focus of research in speech has been on the manner in which emotions are displayed in vocal interactions. SER has gained an increasingly significant role in human-computer interactions. In computerized speech interface systems, where in commands and data are conveyed through speech, the correct recognition of emotions leads to better machine response to human emotions and mental states. Likewise, a proper understanding of the dynamics of emotions by human beings, contributes to their EQ which in turn is linked to their social success in a positive manner. Emotion as the subject of scientific research has multiple dimensions: behavioural, physiological, subjective, and cognitive.

Hence this research is interdisciplinary; has applications in diverse fields, and is therefore challenging. The motivation to analyze the emotional speech of females was based on the observation that female speech is more expressive of emotions.

The present education as well as job scenario in Kerala (where these investigations were carried out), calls for increased multilingual capacity of people in English, Hindi and Malayalam. This is due to the considerable influx of (Hindi speaking) people from other non-Malayalam speaking states. Hence it was considered relevant to investigate the SER in English, Hindi and

Malayalam, for a better understanding of the dynamics of emotions in these three languages.

Since the prosodic feature set of speech comprises observable characteristics such as the intensity, speech rate and pitch, it was decided to first investigate SER with those features. In addition to these, features representing small variations in pitch (jitter) and amplitude (shimmer) were investigated, anticipating these to reflect emotions in a more genuine manner than the basic prosodic features. Further, it was envisaged to compare the performance of time domain and frequency domain features for SER in these three languages. This motivated investigations on the role of spectral features comprising the first four formants (F1, F2, F3 and F4) and their bandwidths (B1, B2, B3 and B4), in the vocal expression of emotions.

Whereas acted speech is known to give exaggerated values for the acoustic correlates, spontaneous or natural speech in various emotional colors is difficult to collect in a short span / warding off variations caused by nature (such as illness or aging). Besides, there can be serious ethical issues involved in the collection of spontaneous emotional speech samples. Hence it was decided to investigate elicited speech, wherein emotions are induced. However the emotional quality of such elicited speech was verified by appropriate perception tests.

1.5. Problem Statement

The main research question is to identify minimal inputs for female speech emotion recognition in English, Hindi and Malayalam. Tackling this challenging problem raises a number of important sub research questions like:

i. Can neutral and the basic set of emotions be recognized with minimal time domain and / or spectral features for English, Hindi and Malayalam?

ii. How can a minimal feature set be selected for each of the three languages?

iii. Are the SER rates in any language, valence dependent?

iv. Is there any similarity (universality) in the vocal expression (feature values) of emotions and SER across English, Hindi and Malayalam? And if so, to what extent?

v. For English, with its stand-alone nature of vowel utterances, can SER be achieved at the segmental level itself using vowels?

vi. Is there any scope for segmental level Speech Recognition (SR) using spectral features?

vii. How can this SER problem be modeled intuitively?

1.6. Objectives and Scope

Recognizing emotions from female speech in the Indian context, using minimal inputs is a challenging pattern recognition problem with direct applicability in the field of computer interactions and speech analysis. The preliminary steps to this end were the design and development of emotionally rich speech databases in English, Hindi and Malayalam in female voice, for SER in the Indian context. The other main objectives of this research are identified as follows:

i. To study elicited emotions as opposed to earlier, popular studies based on acted emotions or spontaneous emotions.

ii. To use a less complex approach for emotional speech analysis, but include various linguistic, psychological and social aspects in the design of the speech database, so as to achieve better SER accuracies.

iii. To statistically analyze the acoustic correlates (feature values) of emotional speech in the six basic emotions and the neutral for each feature (mean pitch / speaking rate / duration / intensity/ jitter / shimmer / any of the first four formants or their bandwidths), and for each language.

iv. To assess the individual contribution of each of the above mentioned features to SER in the three languages, and to identify the best features for SER in each language.

v. To evaluate the emotion valence dependency of the SER rates for each feature and in each language.

vi. To investigate universality in manifestation of any specific emotion, across English, Hindi and Malayalam, mainly in terms of the values of the above listed features, and to note the universal characteristics for each emotion class.

vii. To analyze and compare the feature values of emotions, and the SER at both the segmental and suprasegmental levels for English.

viii. To quantitatively evaluate the accuracy of SER at both segmental and suprasegmental levels using the K-Means, FCM (for prosodic feature based SER in English), KNN, NB and the ANN classifiers.

ix. To investigate the feasibility of emotional speech recognition at the segmental level in English using vowel formants, and to identify the most favorable and unfavorable emotions for the same.

x. To model the SER problem appropriately, taking into account the emotional content and valence of emotions.

1.7. Contributions of the Thesis

The major contributions of this thesis are summarized as follows:

i. This research has been designed and carried out in the Indian context. Female speech emotion recognition has been investigated from an interdisciplinary perspective using a minimal feature set, comprising prosodic features or their variations or the spectral features. These databases comprise semantically appropriate, non neutral content for each emotion, as opposed to most other works using only neutral content [1], [13]. The emotionally rich, speech databases developed exclusively for this research fully account for the fact that the production and perception of emotions can vary with the mother tongue of the speaker / user. The investigations were carried out on databases that were manually segmented taking care to preserve phonetically accurate boundaries for the utterances.

ii. The approach used here is both straight forward and reliable as the obtained results are comparable with those of statistical analysis, human SER rates, and has been validated with new samples (for prosodic feature based SER).

iii. Classification accuracies of 95.6%, 97.14% and 86.76% respectively were obtained for SER in English, Hindi and Malayalam with spectral features comprising the first four formants and their bandwidths. Three classifiers namely, the K-Means, KNN and the NB were used as the base classifiers to identify the best spectral feature subset for effective SER. This was followed by the final ANN classification based on the optimal feature set.

iv. Universality in the manifestation of each emotion, across the three languages was evaluated.

v. Several of the important results obtained, showed higher SER rates at the segmental level, compared to the suprasegmental level. This indicates potential savings in time and effort. Whereas the accepted stand is that prosody exists at the suprasegmental level, results from these investigations on SER with prosodic features at the segmental level in English, proved otherwise.

vi. The SER rates in each language were found to be independent of the valence of the emotions in almost all cases.

vii. The results obtained through this work were based on elicited emotional speech of females, whereas most of the available results were obtained based on acted speech.

viii. Statistically discriminated feature values gave improved SER rates as well SR rates.

ix. Wherever applicable, the characteristics of various voice features and the classification accuracies reported in the literature were compared with the results of this investigation in emotional English speech in the Indian context.

x. Formant based segmental SR was investigated and the emotions most favourable for speech recognition at the segmental level were identified.

xi. This SER problem for Malayalam was modeled at different levels using decision trees and logistic regression.

1.8. Outline of the Thesis

This thesis is organized in seven chapters, as follows. Each chapter including the present discusses the various aspects of this research work. Chapters 1 to 3 present the introduction, findings from literature survey, and the methodology, respectively. Chapters 4 to 6 present the results obtained with various features along with the inferences and interpretation of the obtained results. The SER rates are compared with previously reported results, obtained by other approaches and also for other languages. Comparisons of SER rates for the three languages illustrate the extent of universality in the vocal expressions of emotions. The conclusions drawn from these investigations are presented in Chapter 7.

Chapter 1 presents the introduction and overview of this research work. It also gives the motivations, problem statement, objectives, research contributions, and the thesis outline.

Chapter 2 surveys the literature presenting the relevant previous works in this area along with the necessary background for this work.

Chapter 3 discusses the methodology adopted for this investigation. The basis for the selection of languages, emotions, subjects (speakers), features and classification methods are clearly explained. The important steps in the design and development of the speech databases and validation of the emotional content of the databases are presented. Besides, this chapter acquaints the reader with the fundamental principles of the various techniques used in this approach of SER, such as the feature extraction, statistical analysis followed by classification using multiple classifiers namely, the K-Means, FCM, KNN, NB and the ANN classifiers. Finally, the salient steps to model SER using decision trees and logistic regression are discussed.

Chapter 4 presents the results and discussions of SER based on the prosodic features of intensity, pitch and duration / speech rate. The performance of the automatic SER is quantitatively assessed by the SER rates of the various classifiers. The pitch contours of the various emotions are analyzed, grouped manually, and the salient features identified across emotions and languages can be used for rule based classification of the test emotional pitch contours.

Chapter 5 presents the relevant results obtained by the use of jitter and shimmer, indicating substantial improvement in SER rates over that obtained with the basic prosodic feature set. Improved SER rates were obtained for emotions not recognized well with the basic prosodic feature set as given in Chapter 4.

Chapter 6 discusses SER based on the eight vocal tract spectral features comprising the first four formants along with their bandwidths for all three languages. The obtained results show that spectral features are efficient for SER in all three languages. Results of modeling of SER in Malayalam by decision trees and logistic regression are presented. Further, the emotions most favorable for speech recognition at the segmental level are also identified.

Chapter 7 consolidates the results of this investigation and highlights the main findings of this research. It summarizes the research by evaluating its contributions, mainly from the perspective of the stated objectives. Finally, the chapter concludes the thesis proposing the directions for further research to enhance the performance of SER systems.

Chapter 2

LITERATURE SURVEY

Contents

This chapter presents a brief, yet comprehensive survey of the literature on speech emotion recognition covering the methods and techniques available till date. The background of the research in emotions, underlying theories and challenges of the research in SER, popular public emotional speech databases, features and classifiers used in SER, are presented. Glimpses of the state of the art research highlight the salient contributions of several prominent researchers in this domain. The literature related to this topic had been collected from various available resources including libraries and the internet.

2.1. Introduction

The initial sections of this Chapter present the necessary conceptual background for SER, along with its key elements and highlight the significance of this research. The various aspects of SER reveal the complexity of research in this area. The subsequent sections present relevant previous works, thereby identifying the scope for this research.

Historical Background of Research in emotions: The initial researches in emotions can be traced to have originated from a psychological perspective [20]. The research in emotions started in the 1800's with Darwin's observations of emotions in human beings and animals. It has been dominated by four main

theoretical traditions, as detailed by Cornelius [21]. As per the Darwinian perspective, initially articulated in the late 19th century by Charles Darwin, emotions evolved via natural selection and therefore have cross-culturally universal counterparts. Ekman's work on the facial expressions of basic emotions is a representative of the Darwinian tradition [22]. The Jamesian perspective of William James in the 1800's holds that emotional experience is largely due to the experience of bodily changes that could be visceral, postural or facially expressive. The cognitive perspective propounded that thought and in particular, cognitive appraisals of the environment form the underlying causal explanations for emotional processes.[23]. The social constructivist perspective of Averill holds that emotions are cultural constructions that serve particular social and individual ends [24]. It emphasizes the importance of culture and context in understanding emotional interactions in society and focuses on constructing knowledge based on this understanding [25, 26].

- This thesis is based on the social constructivist view, as it focuses on SER at the specific social levels of educated, urban, females in the Indian context.

2.2. The Basics of SER

Research in human SER has an extensive theoretical background owing to the strong interplay of various factors. In 2001, Cowie et al. [1] identified two interacting channels of human communication: the implicit channel and the explicit channel. The implicit channel tells people "how to take" what is transmitted through the explicit channel. Human communication, as manifested through a combination of verbal and nonverbal channels, is significantly modulated by various linguistic, emotional, and idiosyncratic aspects. Whereas the linguistic aspect defines the verbal content of what is expressed, the idiosyncratic aspects are dependent on culture and social environment [27].

Voice has been recommended as a promising signal in affective computing applications as it is low-cost, nonintrusive, and has fast time resolution [28].

Emotion has been defined to begin with a stimulus and encompasses feelings, psychological as well as physiological changes, impulses to action, and specific goal oriented behavior [29]. Basic emotions are more primitive and universally recognized than the others [17]. The basic emotions belong to a psychologically irreducible set and are also known as the archetypal emotions. The list of basic emotions first proposed by Ekman for facial expressions comprised anger, fear, sadness, disgust, happiness and surprise. Non-basic emotions are called ‘‘higher-level’’ emotions and are rarely represented in emotional databases. In 2014, Jack et al. [30] has reported experimental results regarding classic facial expressions of sixty western, white Caucasians indicating that basic emotion communication through facial expressions comprised fewer than six categories. These basic facial expressions when perceptually segmented over time resulted in only four emotion categories, namely, happiness, surprise / fear, anger / disgust and sadness.

Significance of emotion: Emotions are an essential part of our existence. Emotional distress impels people to seek help, and the repair of emotional disorders is the primary concern of psychotherapy [29]. Emotional Intelligence (EI) is an indispensable facet of human intelligence for successful interpersonal, social interactions [31]. The concept of emotional intelligence, pioneered by Daniel Goleman holds that, self awareness which includes emotional awareness is crucial for personal success. Swati Patra [32] has evaluated that whereas Intelligence Quotient (IQ) accounts for only about 20% of a person’s success in life, the balance 80% can be attributed to EI. EI refers to the ability to monitor one’s own and others’ emotions, to discriminate among them, and to use such information to guide one’s thoughts and actions [33]. Thus emotionally

intelligent persons can use their superior awareness and insight to deal more successfully with the everyday challenges in life.

- Thus there is scope for investigating emotions from the perspective of behaviour and the research in SER can be challenging.

2.3. The Complex Nature of SER

The complexity of speech emotion research can be understood from the perspective of its interdisciplinary nature, varied taxonomy of emotions, linguistic issues, and above all, its dependence on human beings rather than inanimate objects. Added to these, are the influences of varied culture and communication styles.

Interdisciplinary nature: The main sources of emotion research have been traced to psychology and linguistics, with some input from biology [1]. Subsequent research in SER has been based on various acoustic features, mostly from a speech signal processing perspective.

- Hence the need for interdisciplinary research in this area has been repeatedly emphasized [34], [28].

Emotion taxonomy: A primary problem in emotion classification is the lack of an appropriate theoretical model of emotions. A lack of consensus for a definition of the word "emotion" has been reported, despite considerable agreement about the activation, functions, and regulation of emotions [35], [36].

Use of multidimensional description of Emotions: Emotion psychology research has shown that as an alternative to categories, a multidimensional description of emotions provides a greater level of generality and allows for describing the intensity of emotions. Multi-dimensional descriptions of emotions are necessary for describing inter speaker as well as intra speaker emotion expression variability [37]. The popular three dimensional emotion space concept

describes emotion in terms of three basic entities or primitives namely, valence, activation and dominance. Valence is decided by the type of emotion and can be either positive (pleasant) or negative (unpleasant). It is positive for happiness and surprise and negative for anger, sadness, fear and disgust. Activation refers to the arousal and manifests the strength or intensity of the emotional experience [38]. The dominance dimension distinguishes between emotions overlapped in the two dimensional space, such as fear and anger, which have the same valence and activation levels. Whereas anger is a dominant emotion, fear is a submissive emotion. The emotion primitive concept has been applied to label the spontaneous emotions in the excerpts from the "Vera Am Mittag," (VAM) German TV talk show, on a continuous valued scale [39]. Recommendations for four dimensions comprising valence, potency (indicating the degree of control), arousal, and predictability have been put forward. [40, 41].

- One of the objectives of this research has been to assess the valence dependency of SER rates for each of the seven emotions in English, Hindi and Malayalam, based on the various acoustic features (individually and in combination).

Expression and experience of emotions: There are multiple reasons for the complexity in expressing an emotion or attitude. Firstly, an emotion may be expressed involuntarily or voluntarily. Secondly, an attitude that is expressed could be an attitude towards the listener, towards what is being said, or towards some external situation [15]. Above all, the internal experience of emotion is highly personal and often confusing, particularly because several emotions may be experienced simultaneously.

Human beings as objects of study: Whereas in the physical, general, cognitive realm, objects are studied, in the emotional realm, people are the objects of study. This makes the study on SER more complex [33].

Assimilation of speech sounds: Assimilation causes a phoneme to be realized by a different allophone (phoneme variant) and is a matter of concern in the analysis of emotional speech. In natural connected speech, sounds belonging to one word can cause changes in sounds belonging to neighboring words. Roach [15] has pointed out that assimilation varies in extent, according to speaking rate and style. It is common in rapid casual speech than in slow, careful speech and affects consonants more than vowels**.**

- Therefore, in this research such effects of assimilation were anticipated to influence vocal expressions of sadness (which is typically slow) and happiness (which is typically fast) in different ways. This motivated the investigations on stand - alone vowel utterances (such as a, i, o, u), which paved way for this research in SER at the segmental level too.

Speech perception: A study of Spanish - Catalan bilingual subjects, classifying seven simulated vowel stimuli, demonstrates that even early and extensive exposure to a second language is not sufficient to attain the ultimate phonological competence of native speakers. Thus it is well attested that speech is perceived through the filter of one's native language. It was further pointed out that comparing immigrant and native speakers would create very serious problems [42].

- This explains why it was imperative to develop a separate English speech database in the Indian context itself for speech analysis, as done in this research, inspite of other publicly available English databases.

Gender and communication style: The findings of several researchers in the area of gender and communication indicate that men and women differ greatly in the way they communicate and interact with others [43]. Gender is viewed as the learned behavior that a culture associates with being male or female, whereas sex refers to the biological category. Communication style is the way people perceive themselves interacting and communicating with others.

- The observation of such gender based differences in communication styles motivated this study to be specifically confined to the emotional speech of women rather than being generalized across genders**.**

Role of culture and universality in SER: When speakers from different parts of the country or of different ethnic or class backgrounds talk to each other, it is likely that their words may not be understood in the true sense [44]. Ekman [4] attributed apparent differences in human emotional expressions from area to area, to the fact that different cultures taught different display rules. Yrizarry et al. [45] investigated cultural influences on emotion perception and concluded that different cultures colour the perception of universal emotions depending on their particular dynamics, psychological goals, or characteristics. According to Matsumoto [46], collectivistic cultures, emphasize values such as conformity, obedience, and in-group harmony, at least as ideologies. Thus different guidelines for the regulation of expressive behaviour are produced and exist because the meanings of social relationships differ from one culture to the next.

- Aforesaid facts once again substantiate the need and significance of separate, language wise emotional speech analysis in the varied Indian cultural context for English, Hindi and Malayalam, as undertaken in this research.

2.4. Emotional Speech Databases

Literature on SER reveals that the quality of the speech database is of paramount importance. This section first lists the three main types of emotional speech databases used till date, along with the salient features of each. Examples of the most popular public emotional speech databases are also given.

Emotional speech databases are essential not only for psychological studies, but also for automatic emotion recognition, as standard recognition methods are statistical and need to be learnt by examples. Well designed emotional speech databases provide good opportunities for developing affective recognizers or affective synthesizers [47]. SER has been based on three classes of speech databases namely, simulated / acted or completely spontaneous / natural emotions or elicited / induced. Simulated or acted speech is expressed in a professionally deliberated manner, and the advantages and drawbacks of using acted speech for SER have been extensively discussed in literature. Banziger and Scherer [48] studied the role of intonation in emotional expressions based on an acted speech corpus in German. Ververidis and Kouropoulos [17] reported acted speech as the most reliable for SER, since professional speakers deliver emotionally colored speech of high arousal. Sixty four available speech data collections were reviewed, of which, some of the most popular databases are the following:

The public, Berlin emotional database, EMO-DB consists of acted speech of five male and five female speakers in German. The public, DES database consists of emotional utterances of two male and two female actors and was constructed as a part of Voice Attitudes and Emotions in Speech Synthesis project. Both of these public databases have been widely used by other researchers. The Electromagnetic Articulograph (EMA) database in

American English comprises acted speech of l680 sentences, uttered by two females and one male speaker in four different emotions.

Yacoub et al. [49] has used an acted speech database obtained from the Linguistic Data Consortium (LDC). Two popular databases provided by LDC are the TIMIT and the Air Travel Information System (ATIS). TIMIT, a corpus containing broadband recordings of eight major dialects of American English is named so, as the speech was recorded at Texas Instruments, Incorporated (TI) and transcribed at MIT. Acted speech has been used for the validation of results and recommended for proof of a concept, rather than for the construction of a real-life application for the industry [50]. Detailed analysis of the results of emotion recognition studies carried out [47-52] leads to the conclusion that high recognition rates were characteristic of acted speech.

Natural speech is simply spontaneous speech, where all emotions are real [17]. The ATIS speech database was the first in a series of recordings of natural speech. The natural speech recordings were collected over an extended period, in order to overcome Labov's well-known Observer's Paradox, wherein the presence of an observer or a recording device influences the speech of the observed person [53]. Another problematic issue is that hidden recordings, which promise the acquisition of truly spontaneous behavior of people in their usual environment, are privacy intruding and unethical [31]. Even though natural emotions cannot be as easily classified as simulated ones, several experimentalists have focused on the natural expressions of the basic emotions.

Natural speech is difficult to record especially in the preparation of large databases and for a large variety of emotions. The effects of noise from the surroundings also have to be taken care of [54]. In 2011, Weninger et al. [55] recommended any research on emotion recognition from speech in realistic conditions to account for the Lombard effect. As per the Lombard effect,

increasing interferences such as background noise encountered in spontaneous speech induce people to speak louder which may be wrongly interpreted by the listener, as angry speech. Hansen et al. [56] constructed the natural speech database, Speech under Simulated and Actual Stress (SUSAS). Another popular, annotated, natural emotional database consists of 12 hours of audio-visual recordings of the German TV talk show VAM [39]. This corpus contains spontaneous, very emotional speech recorded from unscripted, authentic discussions between the guests of the talk show. Emotions occurring in everyday human computer interactions are spontaneous, with high interspaced or intra-speaker variations [13]. Results of all such studies show that the complexity of the emotional speech recognition increases with the naturalness**.**

Elicited speech is that in which the emotions are induced [17]. Johnstone et al. [57] used computer games to elicit emotional speech in order to study the extent to which emotional changes reflect factors other than arousal, such as valence. It was pointed out that the small number of studies that have attempted real emotion induction have been predominantly the bipolar inductions, such as high - low stress.

- Thus, survey on emotional speech research has revealed that elicited emotional speech databases have been sparingly used, when compared to the acted or spontaneous type. Identification of such a gap motivated this research to be based on elicited type of emotional speech even though it was very difficult to collect appropriate speech samples.

Multimodal databases: Multimodal databases contain information of more than one modality, such as both image and speech. In 2012, McKeown [58] developed a high-quality; emotional, multimodal database as part of a project called Sustained Emotionally colored Machine - human Interaction using Nonverbal Expression (SEMAINE). The various problems concerned

with the establishment of a readily accessible, benchmark audiovisual database of human affect expressions have been discussed [31].

Languages of databases: Languages have different accents leading to pronunciation differences. The same language is pronounced differently by people from different geographical places, social classes, ages and different educational backgrounds. Dialect refers to a variety of a language which is different from others not only in pronunciation but also in vocabulary, grammar and word order [15]. A study of thirty two emotional speech databases has been presented in [59]. Further to this, an up-to-date record of the available emotional speech data collections providing information on the number of emotional states, the language, the number of speakers, and the kind of speech is available [17].

- This research focuses on vocal expressions of emotions of educated, urban females in three languages namely, English, Hindi and Malayalam. Pronunciation differences due to different dialects of the same language are outside the purview of this research. Comprehensive reviews of emotional speech databases in various languages have revealed English as the most popular language for SER.

The most common emotions in the reviewed databases, listed in the order of decreasing frequency were, anger sadness, happiness, fear disgust, joy and surprise. The lack of appropriate publicly available annotated databases is one of the major barriers to research advances in emotional information processing [39]. Whereas standard databases like TIMIT and ATIS are available for recognition of European languages such as English, the major hurdle in speech research for Hindi or any other Indian language is the deficiency in resources like speech and text corpora. Language specific effects prevent the use of databases from non-Indian languages for research in Indian

languages [60]. Due to all these reasons, researchers in Indian languages have to first design and develop their own speech corpus and database as was done in this research.

The Emotional Speech Corpus: A corpus is a collection of naturally - occurring samples from which a database can be constructed [61]. The nature of the text material that is used to build the database is very important when designing the corpora for emotional speech analysis.

- Consequently in this research, investigations were conducted on speech corpus comprising emotionally rich, non neutral semantic content, except for the neutral emotion.

2.5. SER- State of the Art

This section first highlights the popular acoustic features and classifiers reportedly used in previous works in SER. This is followed by brief introduction of many other classifiers and features used to achieve maximum SER rates.

Popular acoustic features: Pitch, formants, short-term energy, Mel Frequency Cepstral Coefficients (MFCCs) and the Teager energy operator based features have been identified as important in SER.

Prosody: The vocal parameters that have been best researched by psychological studies in relation to emotion, and which are also intuitively the most important ones are the voice quality and prosody comprising pitch, intensity, and speaking rate [13], [14], [31], [47], [49] and [62]. Intonation is a part of supra-segmental phonology that includes aspects related to pitch, length and loudness [48]. Analysis of the mean fundamental frequency (F0) over the utterance, mean energy, syllable based speech rate etc. showed that these acoustic

cues are good at recognizing emotions [63]. The typical acoustic features strongly involved in an emotional speech signal include the following [64]:

i. Level, range and contour shape of the fundamental frequency F0.

ii. Level of vocal energy, which is perceived as intensity of voice, and the distribution of the energy in the frequency spectrum, which affects the voice quality.

iii. Formants, which affect articulation.

iv. Speech rate.

Universality with respect to pitch contours was concluded when native speakers of different languages produced the same pitch contours for the same emotions (on different or same linguistic contents) [48].

Intensity is a prosodic feature of speech with relatively simple structure and has been cited to be useful in recognizing certain languages and certain emotional states. Intensity has been taken as one of the several feature vectors in many SER experiments [17], [50], [55], [65- 68].

Speech rate has been reported to be affected by physiological arousal associated with emotions [1]. Karlsson et al. [69] experimented with the segmental duration of vowels collected from six Stockholm speakers, under neutral and stressed conditions, for the implementation of a novel speaker verification system. The use of speech rate in terms of syllabic units, and syllabic unit segmentation, using the maxima and the minima of energy contour has been recommended [17]. Variations in speaking styles by the manipulations of vowel durations in certain adjectives and adverbs have been reported [70]. Based on observations of prosodic features being carried by syllables, investigations were carried out with syllable speech rates in English [71], [72].

- This research therefore proposes to investigate the ability of the aforesaid prosodic features namely pitch, intensity, and duration / speech rate to discriminate emotions across English, Hindi and Malayalam as well as within each of these languages.

Jitter and Shimmer: The local jitter has been proposed as the most common measure of voice quality and is usually expressed as a percentage. The evaluation of different methods to estimate the amount of jitter present in speech signals is presented in [18]. Praat was found to give good results for the jitter measurement of sustained vowels [73]. The earliest demonstration of possible applications of jitter in SER showed a decrease in the recognition of joy and fear, due to the smoothening of the pitch contours [48]. In 2007, Li et al. [74] used jitter and shimmer for detection of speaking style, as well as the arousal level in both human speech and animal vocalization. Experiments conducted in speaker verification, with the Switchboard-I conversational speech database, showed jitter and shimmer as excellent complementary features of spectral and prosodic parameters. Besides, values of jitter and shimmer above a certain threshold could be related to pathological voices [75].

- Therefore this research proposes to use both jitter and shimmer for SER.

Formants and Bandwidths: Ververidis and Kotropoulos [17] identified formants as an important set of features for SER, as they describe the shape of the vocal tract during emotional speech production. In 2007, Alvarez [47] used the prosodic features, jitter, shimmer, first three mean formants, and their corresponding bandwidths for the classification of emotions. Maximum accuracies of 64.9% and 68.7% respectively were reported for male speech in Basque and female speech in Spanish.

- This research uses the first four formant frequencies and their respective bandwidths to analyse their scope for discriminating various emotions. Further the formant based classification of vowel sounds is also investigated.

Popular Classifiers: Appropriate emotional speech classification techniques such as ANNs, Hidden Markov Model (HMM), KNN, and Support Vector Machine (SVMs) have been reviewed by Ververidis and Kotropoulos [17]. ANNs, HMMs, Maximum-likelihood Bayes classifier, Gaussian Mixture density Models (GMM) and fuzzy membership indexing, are the widely utilized classification techniques for the vocal expression of emotions [31], [52]. ANN-based classifiers are used for emotion classification due to their ability to find non-linear boundaries separating the emotional states. Static classifiers like SVMs, ANNs, and decision trees are recommended for SER based on global statistics features whereas, HMMs require strong assumptions about the statistical characteristics of the input and are recommended as a dynamic modeling technique for short-term features [14]. HMMs have almost exclusively been applied to acted data and often outperform static modeling techniques with feature types, such as MFCCs.

In 2003, Nwe et al. [76] proposed the use of short time Log Frequency Power Coefficients to represent the speech signals. An average accuracy of 78% was reported with the HMM classification of six basic emotions in Burmese and Mandarin. Yacoub et al. [49] applied neural network and SVM classifiers to 37 prosodic features of an acted English speech database from the LDC. Highest accuracy of 94% was obtained in recognizing hot anger versus neutral speech.

Ververidis and Kouropoulos [77] implemented the Sequential Forward Selection (SFS) technique to eighty-seven statistical features comprising pitch,

spectrum and energy etc. extracted from the DES database, comprising five emotions. A classification rate of 54% was achieved with a Bayes classifier using the five best features among the statistics of energy, pitch, and formants. When considering gender information, classification rates of 61.1% and 57.1% were obtained for male and female subjects respectively, using ten features, by a Bayes classifier with gaussian probability density functions. The best result in their work was obtained by a GMM for male samples, at 66% classification rate [78].

In 2005, Schuller et al. [79] applied evolutionary programming on prosodic, voice quality and articulatory feature contours of two public databases, namely the Berlin Emotional Speech Database (EMO-DB), and the DES Corpus. SER system performances based on single feature groups too were reported. The highest SER rates obtained were 27.46% and 48.16% on the basis of duration and intensity respectively. The highest SER rates for intonation and formants were 62.09% and 63.73% respectively for the EMO-DB database. Genetic generation and selection of features along with SVM classifier resulted in 87.7% SER rate.

In 2005, Lee and Narayanan [51] proposed the recognition of domain specific negative and non-negative emotions from speech signals in commercial automatic call center dialog system, using language and discourse information in conjunction with twenty one different acoustic correlates of emotion in speech signals. These features included utterance - level statistics corresponding to F0, energy, duration, and the first and second formant frequencies. Maximum classification accuracy of 80.9% was reported with ten features.

In 2005, Athanaselis et al. [80] reported speech recognition rate of 70% under anger for English and experimentally concluded that emotions have major effects on recognition rate. It was observed that emotions are perceived not only from the voice prosody and voice quality, but also from the verbal content of the spoken utterance, and the situational context.

In 2006, Ramamohan [81] proposed a sinusoidal model for the classification of stressed emotional speech in Telugu and English. Frequency, amplitude and phase features of the sinusoidal model of speech were analyzed and classified with a vector-quantization classifier and a HMM classifier. An average success rate of 91.7% for happiness, anger and neutral in Telugu and 91.4% for English was obtained with frequency features.

In 2006, Navas et al. [82] classified neutral and the six basic emotions, of a Basque speech database of neutral and non neutral semantic content. The maximum SER rate obtained was 98.4% with a speech data base of neutral semantic content using GMM and MFCC. Subjective evaluation and automatic recognition tests indicated that neutral content results in exaggerated expressions of emotions and are hence undesirable for SER.

In 2007, Schuller et al. [54] used the acted EMO-DB, the DES and the spontaneous Aibo Emotion Corpus to illustrate the difficulty of speaker-independent recognition of spontaneous emotions compared to acted data. Random Forest classifiers with 100 trees per forest gave maximum accuracies of 72.5% and 57.1% on the EMO-DB and DES databases respectively.

In 2007, Busso and Narayanan [27] investigated the influence of articulation and emotions on the interrelation between facial gestures and speech for neutral, sadness, happiness, and anger. A multi-linear regression framework was used to estimate facial features from acoustic speech parameters. The results showed that facial and acoustic features are strongly interrelated, with correlation higher than 0.8.

In 2007, Mingyu You [83] proposed Enhanced Lipschitz Embedding for SER in Chinese. Based on geodesic distance estimation, 64-dimensional acoustic features related to prosody and formants were embedded into a six-dimensional space in which speech data with the same emotional state were generally

clustered around one plane. The data was classified into six emotional states (neutral, anger, fear, happiness, sadness and surprise) by a trained linear Support Vector Machine (SVM) system. The highest classification rate obtained was nearly 78%.

Neiberg in 2008, used Linear Discriminant Analysis (LDA) and a GMM based classifier on MFCCs to evaluate the automatic detection of negative emotions in speech. Both classifiers were tested on an extensive corpus from Swedish voice controlled telephone services and the results indicated the detection of anger with reasonable accuracy (average recall 83%) in natural speech. The GMM method performed better than the LDA [84].

In 2008, Wang and Guan [85] proposed a novel multi classifier scheme for recognition of emotions from prosodic, MFCC, and formant frequency features of audiovisual signals collected from speakers of English, Mandarin, Urdu, Punjabi, Persian, and Italian, which gave an overall recognition rate of 82.14%.

In 2009, Bozkurt et al. [86] investigated prosody related, spectral and HMM-based features for SER with model GMM based classifiers. The spectral features used were MFCC, line spectral frequency (LSF) features and their derivatives, whereas the prosody-related features consisted of mean normalized values of pitch, first derivative of pitch and intensity. The highest accuracy obtained was 65.25 % and 46.7 %, for the class-2 SER problem and for the class-5 SER problem respectively, for spontaneous speech.

In 2009 itself, Park et al. [87] proposed an Advanced Feature Vector Classification (AFVC) for SER using service robots interacting with diverse users who are in various emotional states. The method used short emotional samples from the Emotional Prosody Speech and Transcripts of the LDC, without any intrinsic emotional content and classified discriminative feature

vectors comprising pitch, log energy, zero crossing rate and 12 dimensional MFCCs. 96.9% recognition rate was obtained for class-2 (negative vs. non-negative) emotion recognition. Classification accuracies of 76.5% and 59.4% were obtained for class 3 and class 5 SER problems respectively, with the combination of GMM and AFVC.

Zhongzhe Xiao et al. [64] proposed harmonic and Zipf based features along with a multi-stage classification scheme driven by a dimensional emotion model for the classification of emotional speech. Zipf features characterize the inner structure of signals, particularly rhythmic and prosodic aspects of vocal expressions, with a better discriminative power in the valence dimension. The classification rate was 71.52% on the Berlin data set of six emotions, to which a gender classification was first applied. An SER rate of 81% was obtained on the DES dataset with five emotion states.

Eyeben et al. [88] in 2009 introduced a novel open-source emotion recognition engine, open EAR which integrates audio recording and audio file reading, state-of-the-art paralinguistic feature extraction and pluggable classification modules. They made use of a KNN classifier, a Bayes classifier, and a module for Support-Vector classification of six databases including the EMO-DB, for benchmarking. The highest recall rate was reported for EMO-DB, at an unweighted average class-wise recall rate of 88.8%.

In 2009, Busso et al. [89] presented a novel framework based on Kullback - Leibler Divergence and logistic regression models to identify, quantify, and rank the emotionally salient aspects of the F0 contour in English, German and Spanish. Analyzing the pitch statistics at the utterance level was found to be more accurate and robust than analyzing the pitch statistics for shorter speech regions (e.g., voiced segments). The recognition accuracy of the system with a two step approach was over 77% based solely on pitch features.

Kim et al. [90] proposed a novel speaker-independent feature namely, the Ratio of the Spectral flatness measure to the Spectral center (RSS), for emotion recognition. Gender and emotion were hierarchically classified by using RSS, pitch, energy, and the coefficients. A maximum recognition rate of 62.9% was achieved in the speaker-independent mode, for a class-4 SER (joy, neutral, anger, and sadness) problem on an acted Korean emotional speech database.

In 2010, Gharavian et al. [91] evaluated the role of happiness, anger, neutral and interrogative emotional states in the performance of continuous speech recognition as well as SER based on pitch. Results indicated more than 68% deterioration in speech recognition rates when compared to neutral speech. The SER rate reported for neutral was 61.43%. Gharavian and Sheikhan [92] investigated with various forms and combinations of the first three formants and GMM classifiers. Average recognition rate of 69% was obtained for happy, angry, interrogative and neutral emotional states.

In 2010 itself, Stefan Steidl et al. [93] investigated the influence of specific emotions in accuracy of word recognition in the spontaneous speech of children. Sammon transformation was applied in the MFCC space to visualize neutral, anger, emphatic and motheresse emotional speech from 51 children of the German FAU Aibo Emotion Corpus. The best recognition accuracy was obtained for anger and emphatic at 85.1% and 84.4% respectively.

In 2011, Yoon and Park [94] proposed a two step, hierarchical emotion recognition system for heterogeneous speech database of male and female speakers. SVM classification of neutral and anger gave a maximum accuracy of 97.3% based on a feature set comprising pitch, energy and 12^{th} order MFCC. Sundberg et al. [40] analyzed stimuli from the Geneva Multimodal Emotional Portrayal database, collected from 10 professional French-speaking actors. Multiple discriminant analysis gave a maximum SER rate of 87.1% based on 12

parameters. In 2011, Ramakrishnan and El Mary [50] presented an overview of 10 interesting applications illustrating the importance of SER in modern human-computer interfaces. HMM and SVM classifiers were used to discriminate emotions based on pitch, formants and MFCC features. The maximum classification accuracies obtained were 79.7% and 77.6%, based on the EMO-DB (seven class) and DES (five class) databases respectively. The extent to which emotions could be discriminated on the basis of valence and arousal was also investigated.

Caponetti et al. [95] in 2011 introduced a long short-term memory (LSTM) recurrent neural network capable of recognizing long-range dependencies between successive temporal patterns in emotion recognition. SER rates of 71.5% and 75.2% respectively were obtained from the LSTM network on the two different feature sets namely MFCC and the Lyon Cochlear model. In 2011, Mower et al. [96] proposed an emotion classification paradigm, based on Emotion Profiles (EPs) wherein the emotional content of naturalistic human expressions was interpreted by providing multiple probabilistic class labels, rather than a single hard label. Four-way binary classification of prosodic and spectral features by a SVM classifier showed that EPs discriminate between types of highly ambiguous utterances for happiness, neutral, anger and sadness with 68.2% accuracy.

Classification of non prototypical emotions in reverberated and noisy Speech in German was conducted by Weninger et al. [55], using supervised non-negative matrix factorization. Relevant spectral information was extracted from a signal by reducing the spectrogram to a single column to which emotion classification could be applied, giving a maximum accuracy of 65.8%.

An approach to speech emotion recognition based on multiple classifiers using Acoustic Prosodic information (AP) and Semantic Labels is presented by

Chung and Wei [97]. For AP-based recognition, acoustic and prosodic features including spectrum, formant, and pitch - related features were extracted from the Chinese database. Three types of models, GMMs, SVMs, and Multi-Layer Perceptrons (MLPs), are adopted as the base-level classifiers. A combination of acoustic prosodic information and semantic labels could achieve average SER rate of 83.55% for happiness, neutral, anger and sadness.

Three different methodologies namely, short-term statistics, spectral moments, and autoregressive models for integrating subsequent feature values in SER were investigated by Ntalampiras and Fakotakis [98], in 2012. Additionally, a group of parameters based on wavelet decomposition was used and all the sets were fused on the feature and log-likelihood levels. The HMM classification of features from the EMO-DB database resulted in overall accuracy of 92.2%.

Koolagudi et al. [99] has presented recent literature on SER including issues related to emotional speech corpora, different types of speech features, models and classifiers used for SER. Thirty two representative speech databases have been reviewed from the point of view of their language, number of speakers, number of emotions, and purpose of collection.

In 2012, Yun and Yoo [100] proposed speech emotion classification using a discriminant function based on GMMs. The GMM parameter set was estimated by margin scaling with a loss function to reduce the risk of predicting emotions with high loss. MFCCs, log energy, pitch, zero-crossing rate and the corresponding delta and acceleration coefficients of the EMO-DB, SUSAS, DES and VAM databases were used. The highest accuracy obtained was 87.8% for the acted EMO-DB database. In the same year, Feraru [101] followed an interdisciplinary theme for the classification of happiness, sadness and fury in Romanian, at an average accuracy of 83% with the KNN classifier on F0, F1 and F2 values.

In 2012 itself, Hassan and Damper [102] implemented a hierarchical classification technique called Data-Driven Dimensional Emotion Classification (3DEC) based on Non-metric Multi-Dimensional Scaling. Four ways to extend binary SVMs to multiclass SER were investigated using 6552 features per speech sample, extracted from three databases of acted emotional speech (DES, EMO-DB and Serbian database) and a database of spontaneous speech (FAU AIBO Emotion Corpus). In the speaker independent mode, the proposed method gave highest SER rate of 80.1% for the Serbian database.

In 2013, Attabi and Dumouchel [103] proposed an anchor models system, in which an emotion class was characterized by its measure of similarity relative to other emotion classes. GMMs were used as front- end systems to generate feature vectors used to train complex back - end systems such as SVMs or a MLP to improve the classification performance on the AIBO Emotion Corpus.

Scherer K.R. [6] has given special emphasis to the Brunswikian lens model and reviewed the conceptualization of the speaker's emotional state, listener's attribution, advantages and disadvantages of various research paradigms.

With respect to speech recognition at the segmental level, the perception of vowels in isolation, without any of the co articulation effects of neighbouring phones is known to be based on their steady state spectra, usually interpreted in terms of the location of the first three formants, F1 - F3. Virtually all vowels were suggested to be identified based on F1 - F2 alone. Rear vowels having low value for F2 and concentration of energy in the frequency range below 1 KHz are perceived with only one formant in the F1 - F2 region. Front vowels have much separation between F1 and F2 and require two resonances [16]. Results of investigations on formant and pitch based discrimination of Japanese vowels by Hiroya and Takako [104] indicated that, for ordinary buzz-excited vowels, the two lowest formants are the most important.

Validation of emotional content in the speech corpus: The validation of the expressive content of an acted oral corpus for speech synthesis has been presented by Iriondo [105] in 2007. The objective validation was conducted by means of automatic emotion identification techniques using statistical features obtained from the prosodic parameters of speech. The subjective validation was done by means of perceptual listening tests performed with a subset of utterances.

- This research makes use of both subjective and objective validation.

  Vogt et al. [13] has reported that a reduction in the number of emotion types to be identified, and a greater resemblance between experimental data and the validation test data yields better validation results. Validation studies on automated analyzers of human SER, address commonly the question of, whether the interpretations reached automatically are equal to those given by human observers judging the same stimulus material [31]. Vogt has further observed that real-time emotion recognition is mostly attempted only in prototypical applications due to various unresolved difficulties.

2.6. Chapter Summary

The Literature survey has enabled the proper identification of the challenges in SER as well as the important elements of SER. This chapter has exposed a wide variety of techniques for SER. An overview of state of the art SER systems along with the different classifiers used has been presented here. As per the literature on SER, the most popular classifiers have been based on the Bayesian techniques, ANN or SVM. The public databases give an insight into the widely used databases as well as the popular acoustic features. The most commonly investigated languages for SER are English and German. There are very few reports of SER in Indian languages and none regarding investigations

on the universality among Indian languages. Besides, there are no available results comparing SER at the segmental and suprasegmental levels in Indian English. It can be noted that SER which started from a psychological perspective and with acted databases, has matured well over time.

It can be concluded that very little work has been done on elicited emotional speech when compared to the acted speech, due to the difficulty in preparing the elicited speech database. The specific roles of culture and gender, in speech emotion recognition problems have paved way for this research in SER of urban, educated females in English, Hindi and Malayalam. Further, this chapter has highlighted the interdisciplinary nature of speech emotion recognition.

Chapter 3

METHODOLOGY

Contents

This chapter describes the various steps involved in this interdisciplinary, experimental work on the classification of female, elicited, emotional speech in English, Hindi and Malayalam. Characteristic elements of the adopted approach are the choice of languages, speakers, and utterances, development of the emotional speech database, features, statistical analysis and the classification techniques along with the essential validation of the database. Segmental utterances in English and suprasegmental utterances in all three languages have been analysed based on certain prosodic features, their variations, as well as vocal tract, spectral features. The feature values were analysed statistically, prior to classification by the KMeans, NB, KNN and the ANN classifiers. Salient aspects of modeling SER using decision trees and logistic regression are included.

3.1. Introduction

The outline of this chapter is as follows. To begin with, this section provides a framework of the various implementation steps for SER. This is followed by a more detailed description of each procedure along with the principal challenges faced at each stage. The three principal elements of an emotion recognition system have been identified as signal extraction, feature calculation and classification [13].

The Implementation Framework: The approach of this experimental research is based on low complexity SER for English, Hindi and Malayalam, emphasizing the selection of appropriate speech units and a minimum set of features, rather than the application of advanced techniques or the adoption of new features. Figure 3.1 gives the schematic of the implementation framework for SER.

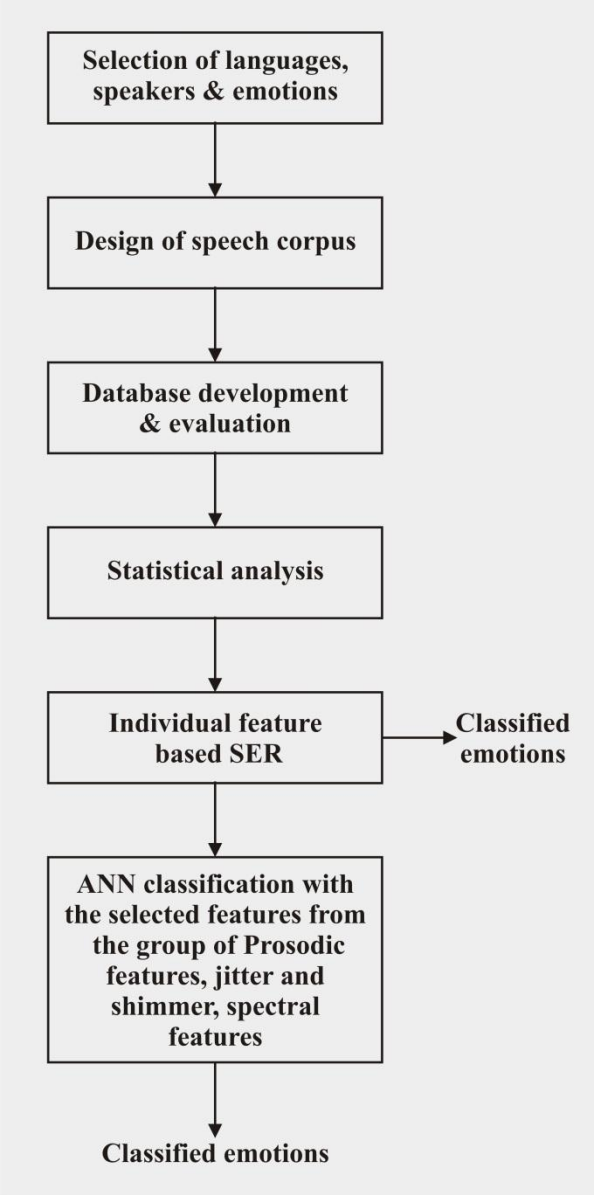


Figure 3.1 Schematic of the speech emotion classification

Emotional speech units of varied lengths are analyzed for English, in order to identify the minimal inputs for maximum SER rates. Investigations on SER at the segmental level are possible and significant for English, due to the stand alone nature of the chosen vowel sounds (a, e, i, o, u). Apart from the choice of the right speech units, other important issues in the database preparation are the choice of emotions, choice of subjects and choice of languages. After recording the suitable database, the wave files are segmented manually, labeled and stored. The next step is the perceptual analysis of the database to check its validity. This is followed by feature extraction. A well balanced set of time domain and frequency domain features comprising certain prosodic features along with their variations and four formants along with their respective bandwidths are investigated. Statistical analyses followed by multiple classifications have been used to assess the SER efficiency of each individual feature, for each emotion, and in each language.

This section has outlined a general framework for the analysis of vocal expressions of emotion. The implementation of various layers of this framework are summarised in the subsequent sections.

3.2. The Work Design

This section first presents the premises of this research. From the literature survey and everyday experience, it is found that certain aspects particularly significant in this experimental study are:

i. Social context of the specific utterance.

ii. Social position of the speaker.

iii. Geographical origin of the speaker.

iv. Gender of the speaker.

3.2.1. The Research Purview

This study has been carried out on utterances with known semantic content. Only the attitude of the speaker had to be decided for the classification of emotions in speech. Since a message / mood / attitude is conveyed through words (semantic part), tone, pitch, etc., this work involves linguistics and voice analysis along with speech analysis. Even though this work is gender dependent, it is not speaker dependent. However the obtained results could be applied only to typical voices, since atypical voices are avoided in the analysis. Further, this research uses a cross sectional approach in which speech of people of different ages are studied over a short period; the prime advantage being, it could be conducted more quickly and reliably. Whereas, in longitudinal research, the emotional speech is analyzed for the same group of people for a longer time span, over which voice qualities would probably change. Detection of the semantic content of the utterance has been kept outside the purview of this research.

3.2.2. Sampling

Sampling has been used in this study of the emotional speech in three languages since the population of interest (emotional speakers) is spread over a large geographical area. A sample is a portion of the population. Statistical analysis of the observations from the sample (speech signal) helps to make inferences about the population (entire set of values) [106]. Further, sampling has the well known advantages of greater accuracy, reduced cost, greater speed and convenience. This study made use of non probability sampling, in which samples are picked up based on judgment or convenience and the chance of selection of any particular unit is unknown. In judgment sampling, the opinion or judgment of some experts form the basis for sample selection. In

convenience sampling, the sample units are chosen primarily based on the convenience of the investigator.

3.2.3. Attributes of the Speech Databases

This subsection provides information on the salient attributes of the speech databases which include the choice of emotions, languages, speakers and utterances.

Emotions: These investigations focused on the neutral and six basic emotions proposed by Ekman [107], namely happiness, surprise anger, sadness, fear and disgust. The first two are positive valence emotions, whereas the last four are negative valence emotions. Neutral was included in the study so that its parameter values for various voice features could be taken as the reference. These seven emotional states were considered as adequate to cover the range of emotions one usually encounters in everyday life. Further, these analyses have been based on elicited emotions as opposed to most earlier popular studies based on acted emotions or spontaneous emotions. Elicited emotional speech is that in which emotions are induced.

Languages: The three languages were chosen for SER from the perspectives listed as follows, and the emotional speech databases were constituted accordingly.

i. English, as spoken by non native English speakers, specifically Indians.

ii. Hindi, the national language of India.

iii. Malayalam, the native language of Kerala, where this study was carried out.

Since there are were no public databases for these seven emotions in the chosen Indian languages and for Indian English, exclusive emotional speech databases were developed for English, Hindi and Malayalam. As the research

on emotional speech relies on the richness and appropriateness of the databases, the databases were designed taking into account various gender, social and linguistic aspects as suggested by Giri [43] and Jones [108]. Since spontaneous emotions are very difficult to record and acted emotions have exaggerated expressions, a database of elicited emotions was developed in each language. The databases consisted of short, but often used utterances in these languages.

Speakers: It was decided to investigate female speech as it is emotionally more expressive than male speech. The ten female subjects (speakers) selected were educated, urban, Indian, non-professional speakers of English, Hindi and Malayalam; in the age group of 28 - 42 years. The speakers who participated in this study are professionally teachers, medical doctors, pharmacists, and office managers. Only educated females with typical voice were selected. Speakers with atypical (hoarse) voices were not chosen for the recordings. Moreover, only people who were healthy and conversant in the three languages were chosen as speakers. Their speech was ensured to be free from the influence of their native language accent. The speech samples were collected at suitable times so as to avoid the possible physiological / hormonal changes. This is because hormonal changes in women could result in hoarseness in their voices. Speakers had volunteered to spare time and effort for this study. They were informed of the purpose of the recordings to the extent that it would be analyzed from the perspective of emotional content. The exact objectives of the study were however not revealed to them. Due to their professional and other commitments, it was not easy to arrange subjects willing to be available for the long recording sessions.

3.3. Design of the Speech Corpus

The primary need for emotion researchers in Indian languages is to first develop their own emotional speech corpus and database as pointed out in Section 2.4 of the literature survey. The database was designed to be rich in the context of the seven emotions investigated. The speech corpus has been designed based on certain principles of psychology that indicate great variability in the male and female vocal interaction styles. Bern S.L [109] has reported the maintenance and transmission of gender linked characteristics. Whereas the male conversation is motivated to establish dominance, the female conversation serves to foster connection. Such gender differences in interaction styles tend to be greater when the interaction is brief and among strangers and therefore had to be accounted for, in the database design.

Pragmatics, which relates to the rules for participation in conversations, social conventions for communicating, sequencing sentences and responding appropriately to the others was incorporated in the design of these emotional speech databases by the choice of words appropriate to each emotion type.

Gumperz and Herasimchuk [110] have pointed out that it is impossible to interpret situated meanings of a word apart from the total context of what has been said before and what is said afterwards. Thus the interpretation of an utterance is not constant but context dependant. Hence it is rather complicated and also less rewarding to analyze excerpts from long utterances. Also long utterances may contain more than one emotion. Considering all these facts, the utterances for this study were chosen to be scripted, short and independent. The semantic content was mostly chosen to match various emotions.

Mostly utterances with non neutral or specific emotional content are used, as per the recommendation of Navas [82]. Since nearly thirty five percent of the social meaning of a conversation is considered to be carried by the words used, it

was decided to use scripted dialogues, so as to have better control over the affective states. The speech corpus was designed to accommodate features of language used by women. These include disclaimers, qualifiers, minimal responses, hedges and questions. Examples of utterances analyzed are greetings, addressing, enquiries, apology, and request for permission, affirmation etc. Another significant feature incorporated in the database is the use of fillers. Indian speakers of English, especially ladies tend to use fillers liberally in their conversations. Examples of these are words like 'actually', 'really' and 'eh'.

Choice of the Utterance for Analyses: The choice of utterances determines the complexity of the segmentation process, especially if the speech data is recorded in chunks (to be segmented later), rather than as isolated utterances. Roach defines an utterance as the continuous piece of speech beginning and ending with a clear pause [15]. The smallest independent (as in the case of vowels), meaningful utterance is the phoneme. Phonemes are an abstract set of units forming the basis of speech. The shortest piece of speech is the single syllable and a minimum syllable is a single vowel in isolation. Vowels including diphthongs are voiced and are the phonemes with the greatest intensity, ranging in duration from 50 to 400 milliseconds in normal speech. Vowels are distinguished primarily by the location of the first three formant frequencies. Utterances can range from a single phoneme to multiple words. A significant difference in natural connected speech with respect to isolated utterances is, the way that sounds belonging to one word can cause changes in sounds belonging to neighboring words (due to the co-articulation effects in continuous speech).

Therefore in English, SER was investigated at both the segmental and suprasegmental levels.

3.3.1. Segmental Utterances in English

In English, the proposed research includes a comparison of SER rates at the segmental level, with those at the suprasegmental level, in an attempt to identify minimum inputs for effective SER. Hence, additional analysis was done on minimal utterances comprising vowel sounds and diphthongs for all the emotions considered. As opposed to consonants these vowels have been selected owing to their stand alone nature as independent utterances in the form of articles, fillers, pronouns and exclamations. Moreover, these are minimal length, salient speech units, capable of carrying the emotive load. Since relevant literature rules prosody to exist at the suprasegmental level only, the choice of segmental utterances for SER, especially on the basis of prosody, is challenging. Hence the the results on segmental SER are significant. The segmental utterances used in this investigation are given in Appendix A.

3.3.2. Suprasegmental Utterances

Syllables, single words, phrases as well as short utterances have been used in these investigations. Utterances beyond six words have not been used in this research due to the possibility of either the mix up or the dilution of emotions; both of which are undesirable. Table 3.1 presents certain frequently used sample utterances which have been selected to constitute the speech corpus. The Hindi and Malayalam databases were also designed along similar lines, mostly consisting of translation of content from the English speech corpus. Even though using different texts for each emotion demands considerable effort, this research made use of sufficient speech corpora of matching emotional content. Sample utterances from the Hindi and Malayalam databases are presented in Appendix B.

Table 3.1: Sample utterances for each emotion

Emotions	Sample utterances
Happy	All are welcome.
Surprise	Oh, what a surprise!
Neutral	Come on Monday.
Anger	You are not welcome.
Sad	It was an accident.
Fear	I am so scared.
Disgust	The food is stale!

At the same time, in order to facilitate better comparison of the characteristics of different emotions, it was decided to additionally record utterances of same semantic content for all the emotions. Table 3.2 presents sample utterances (of the same content) common to all emotions, for English. Similar texts for Malayalam and Hindi are given in Appendix B.

Table 3.2: Sample texts from the English database common to all emotions

Number	Utterances
1	Please come.
2	Will you?
3	Is it?
4	Okay
5	Yes
6	Thank you
7	Are you busy?
8	My God

Thus SER in English, Hindi and Malayalam were based on short single worded utterances as well as multi-worded utterances of up to six words. Analyses were done on such suprasegmental utterances in order to incorporate the effects of assimilation in emotional speech analysis and therefore arrive at results applicable to real life situations.

3.4. Database Development

A well-developed emotional speech database is the most essential prerequisite for the successful analysis and classification of emotions. This section deals with the capturing of emotions, recording, segmentation, labelling, storing and finally, perceptual evaluation of the speech database through subjective listening tests.

3.4.1. Method of Capturing Emotions

Cowie et al. [1] lists the following main methods to elicit emotional speech material, most of which (all, except the fourth) have been used in this work:

i. Context-free simulations,

ii. Reading material / scripts with appropriate emotional content,

iii. Prompting,

iv. Computer broadcasts and

v. Dialogues.

Context free simulations refer to attempts to generate emotional behavior in a vacuum such as short utterances intended to convey a specific emotion. Such context free simulations are controlled and balanced, but may not be natural sounding and is likely to be structurally biased. Reading material with appropriate emotional content is a step less artificial, mainly because well chosen passages can induce genuine emotion. Prompts are better at inducing certain emotional states such as sadness, anger, disgust and amusement rather than emotions such as love and serenity. Strong emotive prompts may be highly charged stories, extracts from film or television, or pieces of music and tend to produce monologues. Computer games have been increasingly used to elicit

induced emotions with automatic data recording and questionnaires; but are more suitable for visual databases. Broadcasts are in the form of unscripted discussions on radio. Dialogues represent interactive situations which are an important source for emotional speech because emotion has a strong communicative function, and interactive contexts tend to encourage its expression. Small groups of people who know each other well enough can talk freely or are set to talk on a subject that is likely to evoke emotional reactions.

To ensure repeatability and reliability of the various experiments, four biological repeats and ten experimental repeats respectively were used. It was very difficult to elicit emotions. Therefore, prompts and dialogues were also used to ensure adequate emotional content in the scripted utterances. Most often, as cited in the previous works which preferred to use acted emotions, the manifestation of elicited emotions was found to be weak.

Fear was the most difficult to elicit in the speech of these subjects and was often found to reach the level of panic. Further, since vocal expressions contains shades of different emotions rather than any single basic emotion, the task of eliciting basic emotions became all the more challenging.

3.4.2. Recording the Elicited Emotional Speech Database

Recording suitable input material for emotional speech analysis is not trivial and obtaining realistic recordings of any kind is a difficult problem. Following the method of Chateau et al. [111], speakers were asked to record the utterances in their own styles, according to their own interpretation of each of the seven emotions. All recordings were done in quiet noise free environment and in different sets on to a computer hard disk, using the Realtek audio system, with noise suppression. A high quality microphone was used for the recordings and the whole session of a speaker was recorded on the same day.

Recordings of different speakers were done in separate sessions to prevent influencing each other's style of speaking. They rehearsed and listened to their own sample recordings and made necessary changes to their utterances for the final recordings taken. Since emotions had to be elicited, several trials were required in order to obtain four good samples of each utterance, unlike acted emotions collected from professionals. A minimum of four recordings were therefore taken for each sample utterance, in all three languages and for the seven emotions considered. The WavePad and the WaveEdit softwares were used for editing these speech files. Their performance was additionally verified with the WaveSurfer software which can also be used for sound visualization and manipulation. The four samples of each utterance were taken at a stretch so as not to incorporate any possible speaking style changes over time (also to avoid changes in voice quality due to a possible sudden cold of the speaker). Noise suppression was applied and the sound files were saved as radio quality, 16 kHz mono recordings, of 16 bits, following the recommendation of Karlsson et al. [112]. Both reliability as well as repeatability of the emotional speech data were ensured in this work by multiple recordings of utterances by the same speaker (biological repeats) as well as repetition of the same set of utterances by multiple speakers (experimental repeats). In order to avoid the presence of multiple emotions over the utterance length, the recorded sentences were kept short enough.

The database was recorded adhering to the directions of Campbell [53] for the collection of superfluous data over a long period and controlling the recording environment. This served to overcome Labov's well-known Observer's Paradox (introduced in Section 2.4). According to the recommendation of Pantic and Rothkrantz [31], audio recordings of several individuals were included in order to avoid effects of the unique speech properties of a particular individual. Even though the number of such

experimental repeats (different speakers) is important, care was taken not to include atypical voices in the speech database. The Lombard effect and the Labov paradox could be avoided using sufficient biological repeats. Above all, the speakers were informed beforehand about the objective of the research, and their consent was obtained in conformity with standard research ethics.

The most challenging part in the recordings was the vocalisation of emotions for very small utterances comprising each of the five English vowels (a, e,i, o,u). Each emotion in full had to be conveyed by these utterances of very small duration.

3.4.3. Segmentation of the Speech Database

Speech segmentation is the determination of the beginning and ending boundaries of acoustic units which are chosen as the elements of emotion analysis. It is an important sub problem of speech analysis. There are manual, automatic and semi automatic methods of segmentation.

The use of a human transcriber in manual segmentation ensures that the segment boundaries and labels (at least at the narrow phonetic level) are perceptually valid. Manual segmentation is extremely costly in terms of time and effort and as a rule of thumb, the effort for segmentation and labeling increases dramatically and is inversely proportional to the size of the labeled units.

Automatic segmentation refers to the process whereby segment boundaries are assigned automatically by a program, using zero crossing ratio, power spectral density, and formant trajectories. Pitch contours and intensity contours too have been used [62]. The output boundaries may not be entirely accurate, especially if the training data was sparse, and also at voiceless transitions. The presence of noise causes considerable degradation in automatic segmentation methods, which are thus one of the most important reasons for the reduced emotion recognition rates [54]. The amount of inaccuracy that is

acceptable will depend on the intended use of the database, and its overall size. Automatic methods do not achieve the same performance as that of a human segmenter and labeller. Semi-automatic segmentation refers to the process whereby the automatic segmentation is followed by manual checking and editing of the segment boundaries. Due to the reasons cited, manual segmentation has been used in this work. After segmentation, these wave files were labeled appropriately and stored in compact discs (CDs).

3.4.4. Evaluation of the Speech Database

One of the prime challenges in the research in emotional speech is the difficulty in creation of an authentic emotional speech database. Speech emotional quality has been defined as the quality of speech samples in terms of the emotional content that describe the listener's global impression as elicited by the auditions [111]. Hence listener-centred, perceptual listening tests are conducted for subjective evaluation of the sound files, to ensure the proper emotional quality of the recordings. Such subjective tests are obviously time-consuming and expensive.

The Mean Opinion Score (MOS) method, specified in the International Telecommunication Union standard ITU-T P.800, has been used in this research for subjective evaluation of the perceptual speech quality. The MOS scale was originally meant to describe speech clarity or intelligibility. MOS is probably the most popular and simplest, subjective method to evaluate the speech quality in general. MOS is a five level scale from bad (1) to excellent (5) and it is also known as Absolute Category Rating (ACR). The listener's task is simply to evaluate the speech sample with the scale described in Table 3.3.

Table 3.3: The five point MOS scale

Score	MOS Rating(ACR)
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

The database of nearly 1400 segmental files as well as nearly 2250 suprasegmental files in each language were subjected to perceptual listening tests. An emotional speech sample with MOS rating greater than or equal to 4 was considered good. The recorded database was evaluated by ten listeners without any hearing difficulties. All the listeners were in the age group 15 - 50 years, educated and without any hearing pathologies. Following the observation of Roach [15] that headphones improve the sound quality even on a cheap machine, the listeners listened over headphones and indicated the perceived emotion from a list of six emotions apart from the neutral. While complex factors like intelligibility and naturalness are the major criteria in any speech evaluation scheme, here the emphasis was on the speech emotional quality which is basically a subjective, perceptual factor. Table 3.4 consists of a summary of the listening tests results for each desired emotion. Only those sound files rated as good, or above it were considered to be recognized properly and included for further analyses.

Table 3.4: Percentage of emotions recognized correctly by human listeners

Language	Percentage of emotions perceived correctly emotions						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
English (segmental)	93	92	91	95	94	93	91
English (suprasegmental)	90	92	90	96	94	93	95
Hindi	96	94	92	91	96	92	91
Malayalam	98	97	96	98	98	93	97

Thus the emotional content of the speech databases were validated for the purpose of further analyses.

3.5. Acoustic Feature Extraction

The choice of features to be used for the speech emotion recognition is a critical research issue due to its dependency on multiple factors. The most crucial factors being the choice of research domain (time versus frequency), the speaker category, size of the analysis unit, number of emotion classes to be discriminated, algorithmic complexity, popularity of features and the expected baseline recognition. This section presents the features used in this approach of SER.

The open source speech processing software Praat has been used for feature extraction. The software Praat is a research publication and productivity tool for phoneticians. It is a comprehensive speech analysis, synthesis and manipulation package which can perform general numerical and statistical analysis.

The main aim of this research was to identify minimal feature sets for SER. Information gathered from literature survey, led to the focus on SER based on three prosodic features, their variations and certain important spectral features. Accordingly, the mean pitch, average intensity, mean duration / speech rate, local jitter, shimmer, first four formants and their bandwidths were extracted using the Praat software. Further details on the extraction of these features are as follows.

Mean Intensity: The vowel intensities were measured in decibels. For the averaging method in decibels (dB), the mean intensity between the times t_1 and t_2 over the utterance of interest, is defined as

$$Mean\ intensity = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} |\, x(t) \,|^2 \, dt \qquad (3.1)$$

where, $x(t)$ is the signal.

The intensity at every time point is a weighted average over many neighbouring time points. The weighting is performed by a Gaussian window that has a duration that is determined by the minimum pitch setting.

Duration / Speech rate: In this investigation, duration was measured for segmental utterances. The durations of the segmental utterances were measured from the Praat intensity contour and included its entire continuous, non zero values. Syllable speech rates and word speech rates were calculated for suprasegmental utterances. Speech rate in syllables per minute and words per minute were found out manually by the method of Jones [108], after finding the duration of the utterance. The word rates were calculated along the lines of syllable speech rates. Though it is time consuming and tedious, this manual method of speech rate calculation is perceptually and linguistically correct due to the implicit accounting for elisons and assimilations that are common in continuous utterances. On the other hand, any algorithm for the automatic segmentation and speech rate calculation makes use of some sort of rule based approximations.

Pitch: Praat's pitch tracking algorithm which is used to extract pitch values, performs acoustic periodicity detection on the basis of an accurate autocorrelation method [113]. The autocorrelation pitch detector is one of the most robust and reliable of pitch detectors since the autocorrelation computation is made directly on the waveform and is largely phase insensitive [114]. It is more accurate than other methods and gives F0 (Fundamental frequency) with an accuracy of 10^{-6}. Therefore, the pitch was calculated on the basis of a short time autocorrelation function [115]. Candidates for F0 of a continuous signal $x(t)$ at a time t_{mid} are found from the local maxima of the autocorrelation of a short segment of the sound centred on t_{mid}. As per the algorithm, for any signal x(t), a piece with duration T (the window length), centred around t_{mid} is taken from

which, the mean μ_x is subtracted and the result is multiplied by a window function $w(t)$, so that we get the windowed signal as,

$$a(t)=[x(t_{mid} - ½T + t) - \mu_x]\, w(t) \quad (3.2)$$

The window function $w(t)$ is symmetric around t=½ T and zero everywhere outside the time interval [0, T]. The sine-squared or Hanning window was chosen, which is given by,

$$w(t) = ½ - ½\, Cos(2\pi t/T) \quad (3.3)$$

The Praat's pitch detection method estimates a signal's short-term autocorrelation function on the basis of a windowed signal, by dividing the autocorrelation function of the windowed signal by the autocorrelation function of the window.

$$r_x(\tau) \approx r_a(\tau) / r_w(\tau) \quad (3.4)$$

where $a(t)$ represents the windowed signal and $r_x(\tau)$ is the short term autocorrelation function of the signal. $r_a(\tau)$ and $r_w(\tau)$ are autocorrelations of the windowed signal and the window function respectively. Finally, the places and heights of the maxima of the continuous version of $r_x(\tau)$ are found out.

Study of pitch contours: The pitch contours present a visual representation of pitch, which helps one to easily understand the variations in the prime intonation parameter of the utterance. The pitch contours give a comprehensive picture of the entire pitch profile over the duration of an utterance rather than just a few instantaneous or representative values of pitch (such as the mean). Reports of investigations with pitch contours and their models have been presented [88], [116]. The pitch contours for various utterances under each of the seven emotion classes were studied so as to identify the salient characteristics of each emotion class, in each language, that can be used for the rule based classification of test pitch contours.

For this, the pitch contours were initially plotted using Praat and re-checked using Sigma Plot. These pitch contours were then visually examined for intra group similarities within any particular emotion class and for inter group similarities between the various emotion classes. The number of peaks in the contour, the skewness of the pitch contour, possible pitch breaks, the extent of flatness, the nature of the contour at the beginning and end of the utterance, are the main features based on which the contours were grouped. Based on the identified characteristics for each emotion group, a test contour could be assigned to that particular emotion class with which its observable features matched the best, visually.

Jitter and Shimmer Analyses: The most common jitter measurement is the local jitter which is usually expressed as a percentage. The feature extractions were performed using the Praat software which has been reported to work well on long sustained vowels. Local jitter was calculated as the average absolute difference between consecutive periods, divided by the average period. Local shimmer is the average absolute difference between amplitudes of the waveform for consecutive periods, divided by the average amplitude. The duration of a period is determined by looking for best matching wave shapes found out by a maximum of the cross-correlation function. The waveform-matching method used here, averages away much of the influence of additive noise. The jitter and shimmer values used in this research have obtained from the voice report provided by for the chosen wave files. This research also investigates SER based on jitter and shimmer of suprasegmental utterances too.

Jitter and shimmer are defined respectively as,

$$Jitter = \frac{\frac{1}{N-1}\sum_{i-1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i-1}^{N}T_i} \tag{3.5}$$

$$Shimmer = \frac{\frac{1}{N-1}\sum_{i-1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i-1}^{N}A_i} \tag{3.6}$$

Where, N is the number of frames in an utterance, T_i is the pitch period of the i^{th} frame, and A_i is the amplitude of the i^{th} frame. The jitter and shimmer values were separately extracted.

Formants and Bandwidths: Although there are a variety of automatic formant analysis approaches [117], the calculation of accurate formant features from the speech signal is a nontrivial problem. The formants were computed by solving the roots of the Linear Predictive Coding (LPC) polynomial and were extracted using Praat speech processing software. In the LP model of speech production, the vocal tract is modeled by a series of concatenated tubes of varying cross-sectional area. Pole angles of the vocal tract are associated with the resonant frequencies. The pole radius is related to the concentration of local energy and the bandwidth of spectral resonance of a formant. In order to obtain a good LP fit, often the order of the LP model is set higher than twice the number of expected formants; usually an LP model order of 13 or more is used. The formants carry much of the information used to distinguish phonemic identities [118].

The formant contours of a sound were observed as functions of time. The formant analysis parameters are set with the Formant menu so as to obtain any of the first four formants.

Formant based Discrimination of Emotions and Utterances: The ability to identify emotions at the segmental level is crucial as it saves time and effort. Besides, it often increases the credibility of the recognized emotion, since it is difficult to act out such minimal length utterances. In order to correctly classify emotions on the basis of formants of the vowels - a, e, i, o, u; the cross sensitivity of formants to the utterance type has to be minimized. This can be achieved by using only those formant values that are sensitive to emotion, without the influence of utterance differences. This in turn, requires using only formants of the same utterances, or formants of different utterances, but with statistically insignificant differences in values. On the other hand, the inter dependence between utterances and formants is put to good use in speech recognition at the segmental level. Besides, the dependence of formant based utterance discrimination on the emotional content of the utterance has not yet been reported. These facts motivated this investigation on the classification of the various English vowel formants, under each of the seven emotions so that the results of the same may be used in speech recognition as well as speech emotion recognition. The identification of emotional content that is most favourable for SR is important since it indicates SR can be made easier, compared to similar analyses with neutral emotions.

3.6. Statistical Analysis

Statistical analysis of the features, mainly by the analysis of variance (ANOVA), was done (using the Graph Pad - Instat software) prior to the classification. This is one of the features of this approach since in general, higher classification accuracies were obtained in this research, with the statistically discriminated feature values. The significance and application of ANOVA are given by Andrew [119] and Talar [120] respectively.

Significance of Gaussian (normal) distribution: The extracted feature values were checked for the adherence to the Gaussian (normal) distribution. The exact

shape of the normal distribution (the characteristic "bell curve") is defined by a function that has only two parameters: mean and standard deviation The assumption of a normal distribution serves to make inferences about the mean and other properties of the observations such as 68% of the values lie within one standard deviation (SD) from the mean, 95% of the values lie within 2 SDs of the mean. The SD quantifies scatter of the data and increase in sample size does not increase the scatter. Whereas SD is an important statistic whose value is unaffected by sample size, the mean or average changes with a single high / low value. SD is used to calculate the variance of the data.

$$Variance = (SD)^2 \quad (3.7)$$

The Standard Error of Mean, $SEM = \frac{SD}{\sqrt{N}}$ (3.8)

where *N* is the sample size.

Statistical significance: Tests for statistical significance serve as a common yardstick that can be easily understood and are used to estimate the probability that a relationship observed in the data, occurred only by chance. The steps in testing for statistical significance are as follows:

i. State the research hypothesis (H1).

ii. State the null hypothesis (H0).

iii. Select a probability of error level (alpha level), before the study.

iv. Select and compute the test for statistical significance. Find out P.

v. Interpret the results.

With any inferential statistical procedure, it is important to state the hypotheses of interest clearly before undertaking any statistical analyses of the data, these were deduced as follows, from the answers to relevant research questions: **The Research Hypothesis** states the expected relationship between

two variables (e.g. features and emotions, in this research), and is stated either in general terms or includes direction. These are stated for this research as follows:

- General: Feature values (for instance, duration) are influenced by the emotion class.
- Direction: Sadness has greater duration than anger.

The null hypothesis is usually a hypothesis of "no difference" and is defined before the start of the study. It is easier to find disconfirming evidence against the null hypothesis rather than to find confirming evidence for the research hypothesis.

Null Hypothesis for this research: There is no difference between the duration values of Group A (sadness) and group B (anger).

Type I and Type II errors: These are relevant in researches investigating the relationship between variables. The Type I error occurs when a relationship is assumed to exist, when in fact the evidence is that it does not. In a Type I error, the null hypothesis is rejected and the research hypothesis is accepted. The acceptable probability of committing a Type I error is called alpha (α).

The Type II error occurs when the researcher assumes that a relationship does not exist when in fact, the evidence is that it does. In a Type II error, the researcher accepts the null hypothesis and rejects the research hypothesis. The probability of committing a Type II error is called beta (β).

The significance level α therefore refers to a pre-chosen probability. The P value is used to indicate the probability of committing the type one error and is calculated after the study.

If the $P \leq \alpha$, H0 is rejected. It is then concluded that the difference in two sample means is high.

If $P > \alpha$, H0 is accepted [106].

The choice of the significance level is arbitrary. In this investigation, α was chosen as 0.05, which is the commonly chosen value, and indicates the willingness to accept 5% probability of assuming a relationship when it really does not exist. Table 3.4 gives the symbolic representation of P values along with its meaning in terms of asterisk rating. 'ns' means no significant difference.

Table 3.5 Symbolic representation of P values along with its meaning.

Symbol	Meaning
Ns	$P > 0.05$
*	$P \leq 0.05$
**	$P \leq 0.01$
***	$P \leq 0.001$

Along with one-way ANOVA, standard parametric test such as the Student Neumann Keul's (SNK) step wise, multiple comparison test chosen as post test, was performed using Graph Pad InStat version 3.05 for Windows, Graph Pad Software. The SNK method is used to identify sample means which are significantly different from each other and often used whenever a significant difference between three or more sample means has been revealed by ANOVA.

When comparing three or more groups (as most frequently used in this research), instead of performing a series of t tests, one-way ANOVA followed by post tests have been used, as it takes into account all the comparisons. The assumptions for this test are:

- Data are sampled from population with identical SDs. This assumption was tested through the Bartlett's test.

- Data is sampled from Gaussian distribution. This was tested using the method of Kolmogorov and Smirnov, to find how much the feature values vary from one another.

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.

In the case of segmental utterances, sample values that belong to normal distribution and those that significantly differ, between various emotion classes, resulted in higher classification accuracies.

3.7. Classification

Classification maps feature vectors onto emotion classes through learning by examples. Vogt [13] opines that the quality of a classifier can be determined by comparison to human raters in listening tests or to other classification algorithms. The former is more meaningful for practical purposes and shows the complexity of the particular task, even though it usually involves much effort to conduct such a study.

Nowadays, research in SER is focused on finding powerful combinations of classifiers that advance the classification efficiency in real life applications. Accordingly, this section mainly focuses on the ANN, K-means, FCM, NB and the KNN classifier used in this work to obtain the best SER accuracies. Except for the classification with spectral features, all the classifiers played a similar role in the classification of emotions based on individual features. In the classification of formants and bandwidths, for each language, the selection of features to the final minimal feature set was made by assessment of the recognition rate (RR) obtained using the three base classifiers with feature values in their raw or modified format.

K-means clustering: Nayak et al. [121] had applied the k-means clustering in the MATLAB analysis and classification of pathological conditions. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters; say k clusters, fixed apriori. The main idea is to define k centroids, one for each cluster. Clustering is done such that patterns in the same cluster are alike and patterns belonging to different clusters are different.

Fuzzy C-Means (FCM) clustering: FCM produces results close to that of the k-means clustering. FCM clustering is an iterative process. The cluster centres are moved to the right location within a data set by iteratively updating the cluster centers and membership grades, for each data point. The FCM has been limitedly used in this work for classification of feature values at the segmental and suprasegmental level of prosodic features to check its effectiveness in SER. This classifier gave very low SER rates for Hindi and Malayalam, which were therefore considered irrelevant.

K Nearest Neighbor (kNN) classifier: The input feature vector is classified based on the class represented by the majority of the k-nearest feature vectors obtained during the training process. Given an input feature vector, the algorithm finds k closest feature vectors representing different classes. The class represented by the majority of the k nearest feature vectors is assigned to the input vector. Its accuracy relies on the selection of an optimum number of neighbours and the most suitable distance measuring method. Euclidean distance has been chosen in this work.

The training procedure of the kNN classification algorithm requires only storage of feature vectors and class labels of training samples. This makes the kNN classifier one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. The notion of a distance measure is essential to the kNN approach. When a new sample data *x*

arrives, kNN finds the k neighbors nearest to the unlabeled data from the training space based on some distance measure [122]. Nearest can be taken to mean the smallest Euclidean distance $d_e(\boldsymbol{a}, \boldsymbol{b})$, in n-dimensional feature space, which is the usual distance between two points

$$\boldsymbol{a} = (a_1 \ldots . a_n), \text{ and } \boldsymbol{b} = (b_1 \ldots . b_n) \tag{3.9}$$

defined by,

$$d_e(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2} \tag{3.10}$$

Naïve Bayes Classifier: The NB classifier is a straightforward and widely used method for supervised learning [123]. It is one of the fastest learning algorithms, without any complicated iterative parameter estimation, which makes it particularly useful for very large datasets. It is based on Bayes' theorem with independence assumptions between predictors. In Bayesian classification learning, a new instance is assigned to the class with the maximum posterior probability. The probabilities of the class membership are calculated from the Bayes theorem. If the feature value is denoted by *x*, and a class of interest is *C*,

$P(x)$ is the probability distribution for feature x in the entire population,

$P(C)$ is the prior probability that a random sample is a member of class *C*, and is known before we know the feature value.

$P(x/C)$ is the conditional probability of obtaining a feature value *x*, given that the sample is from class *C*. Our goal is to estimate the probability that a sample belongs to class *C*, given that it has a feature value *x*, which is denoted by $P(C/x)$, on the basis of $P(x/C)$, $P(C)$ and $P(x)$.

The probability of the joint event that a sample comes from class C and has the feature value x is,

$$P(C)\,P(x/C) = P(x)\,P(C/x) \tag{3.11}$$

which yields the Bayes theorem

$$P(C/x) = \frac{P(C)\,P(x/C)}{P(x)} \tag{3.12}$$

$P(C/x)$ is called the posterior probability since it gives the probability of the class after we observe the value of the features.

ANN Classifier: Due to the superior recognition abilities of human beings, a common research trend is to simulate the recognition mechanisms of biological systems. ANNs were therefore designed to mimic the biological neural networks found in the human brain. ANN classifiers are used for emotion classification due to their ability to find non - linear boundaries separating the emotional states. Feed forward ANNs, in which the input feature values propagate through the network in a forward direction on a layer-by-layer basis, were reported to be used the most frequently [17]. A two layer feedforward, back propagation network with sigmoid output neurons and sufficient neurons in its hidden layer was used for classification. The hidden layer has different number of neurons for the different problems.

There are seven nodes in the output layer since seven emotions are classified. First, the classification problem is defined through the set of inputs and corresponding targets. In this supervised learning technique, the network is trained to classify the inputs according to the targets. The Mean Squared Error (MSE) is the average squared difference between the outputs and targets. The percent error (% E) indicates the fraction of samples which are misclassified. The total number of samples was randomly divided into three classes for training, validation and testing. The classification was repeated for different network sizes in order to arrive at the optimum size of the network. The SER rates from these four classifiers helped to validate the classification of emotions.

3.8. Validation

Performance validation enhances the quality and acceptability of the reported results and is therefore important. In the emotional speech based research, the first and foremost requirement is to ensure appropriate emotional content in the speech database, since it primarily determines the quality of the obtained results. This has been incorporated in this research as detailed in Section 3.4.4.

Validation of results: The emotion recognition rate has been found out, following the method used by Aggarwal and Dave [60] for a speech recognition problem, where randomly chosen words were tested and the language recognition rate (i.e. accuracy) was calculated. The emotion recognition rate has been defined along the same lines as:

$$Emotion\ Recognition\ rate = \frac{Successfully\ detected\ instances}{Number\ of\ \ instances\ in\ the\ test\ set} \quad (3.13)$$

3.9. Models for SER

Classification models are built using known data and known set of responses. Such models are used to obtain predicted responses to new data. This section presents the methodology for arriving at two intuitive models for the SER in Malayalam. The first model is based on a classification tree whereas, the second one employs logistic regression. Both models built here, used the first four formants along with their respective bandwidths as predictors (features).

Decision trees or classification trees predict responses to data by following the decisions in the tree from the root (beginning) node down to a leaf (end) node [124]. The root node has no incoming edges (branches); all others have exactly one incoming edge. Nodes with outgoing edges are called internal or test nodes. In a binary tree, each test node has two child nodes. Terminal nodes are the leaves where the category label is read. Thus the leaf node contains the response. Hence decision trees can be viewed as hierarchical

classifiers capable of multistage classification. Decision trees are non-parametric classifiers capable of rapid classification with good accuracy. If used as acoustic models, they can offer additional advantages over GMMs: they make no assumptions about the distribution of underlying data; they can use information from many different sources and they are computationally very simple [125,126].

Here, the decision tree based modeling was done using binary trees. The importance of each predictor was assessed algorithmically, so that only the important predictors were selected for modeling. The decision tree can be viewed textually or graphically. The decision tree used here has been based on the standard classification and regression tree (CART) algorithm. Each step in prediction involves checking the value of one predictor. The entire data is considered and binary splits on every predictor are examined. The split with the best optimization criterion is selected and imposed. This procedure is repeated recursively for the two child nodes. Splitting can stopped when the node is pure; containing only observations of one class. Pruning is the process of reducing a tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch. A sequence of sub trees is produced by pruning. Pruning can be done by specifying levels or some criterion such as the error. Trees are pruned based on an optimal pruning scheme that first prunes branches giving less improvement in error cost. The cost of each node is the classification error for this node multiplied by the probability for this node. Tree complexity is measured by the number of nodes, number of leaves, tree depth and number of attributes. The trained decision trees are used to predict the class labels for new data. The decision tree based modeling was done using MATLAB.

Regression investigates and models the relationship between a response (Y) and predictors ($x_1, \ldots, x_{k.}$). In many contexts, the response is binary; i.e., taking the values 0 or 1.

Let p denote the probability of a 1. This probability is related to the values of the explanatory variables $x_1, \ldots, x_k$. This cannot be written as $p = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ because the right-hand side is not constrained to lie in the interval [0, 1], which it must, if it is to represent a probability. One solution to this problem is to employ the logit link function, to obtain a valid equation given by,

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{3.14}$$

which leads to,

$$\frac{p}{1-p} = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k\} \tag{3.15}$$

and $$p = \frac{\exp\{\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k\}}{1+\exp\{\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k\}} \tag{3.16}$$

Thus the link function in logistic regression maps the interval (0, 1) onto the whole real line. This guarantees that the predicted probability of an event using the logistic regression model will produce a number between 0 and 1. The quantity ln (p/(1 − p) is referred to as the log odds. The procedure for estimating the coefficients $\beta_0, \beta_1 \ldots \beta_k$ using this relation and carrying out tests of significance on these values is known as logistic regression. Thus logistic regression is based on an ordinary regression relation between the logarithm of the odds in favor of the event occurring at a particular setting of the explanatory variables and values of the explanatory variables $x_1, \ldots, x_k$.

In this thesis, logistic regression; both binomial and nominal, have been performed using the popular statistical software Minitab [127]. A binomial logistic regression is often referred to simply as logistic regression. It predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables

(predictors) that can be either continuous or categorical [128]. It is the most common type of logistic regression and is referred to as binary logistic regression in Minitab. A model with one or more predictors is fit using an iterative reweighted least squares algorithm to obtain maximum likelihood estimates of the parameters. Further, the modeling technique has made use of Stepwise regression in order to identify a useful subset of predictors. The most significant variable is systematically added or the least significant variable is removed during each step of stepwise regression. For more than two categories of the dependent variable, multinomial logistic regression is used. Nominal variables are categorical variables that have three or more possible levels with no natural ordering.

In order to ensure the validity of obtained results, the samples are first checked to satisfy certain assumptions of binomial logistic regression such as: (i) the two categories of the dependent variable need to be mutually exclusive and exhaustive. (ii) One or more independent variables should be continuous or nominal. (iii) There should be independence of observations. If there is no independence of observations, there is likelihood to have repeated measures. (iv) There should be no multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated with each other. This leads to problems with understanding which variable contributes to the explanation of the dependent variable. This is assessed by the value of the variance inflation factor. Variance inflation factors close to 1 indicate that the predictors are not correlated (v) there should be no outliers.

If there are k distinct response values, Minitab estimates k-1 sets of estimated coefficients. These are the estimated differences in log odds or logits of levels of the response variable relative to the reference event. Each set contains a constant and coefficients for the covariates ($x_1, \ldots, x_{k)}$. Regression

generally uses the least squares method which derives the regression equation by minimizing the sum of the squared residuals. Regression results indicate the direction, size, and statistical significance of the relationship between a predictor and response. Coefficients represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. The sign of each coefficient indicates the direction of the relationship. Therefore, positive coefficients indicate that the event becomes more likely and negative coefficients indicate that the event becomes less likely. The p value for each coefficient tests the null hypothesis that the coefficient is equal to zero (no effect). Therefore, low p values suggest the predictor is a meaningful addition to the model.

It is desirable to choose a model (link function and predictors) that results in a good fit to the data. Goodness of fit statistics can be used to compare the fits of different models. Minitab provides three goodness-of-fit tests: Hosmer-Lemeshow, Pearson and Deviance of which the former two are used in this thesis report. The Hosmer-Lemeshow test assesses the model fit by comparing the observed and expected frequencies. The test groups the data by their estimated probabilities from lowest to highest (in ten groups by default) and performs a Chi-square test to determine if the observed and expected frequencies are significantly different. The Pearson goodness of fit test assesses the discrepancy between the current model and the full model. In both cases, if the p value for the goodness of fit test is lower than the chosen significance level, the predicted probabilities deviate from the observed probabilities in a way that the binomial distribution does not predict. Some of the main reasons for the deviation would be incorrect link function, omitted higher-order term for variables in the model or omitted predictors that are not in the model.

3.10. Chapter Summary

This chapter has sequentially detailed the various steps of this research. The methodology adopted is in accordance with the various research objectives. Some of the main features of this approach as revealed in this chapter are, the development and use of a speech database of elicited emotions, the analysis of emotional speech of females in three languages and at both the segmental as well as the suprasegmental levels in English. Statistical evaluation and ANOVA of feature values, and the selection of a minimal feature set for SER are certain other features of this approach. Development of perceptually validated, exclusive emotional speech databases of non neutral content, the use of manual segmentation and multiple classifiers are the remaining highlights of this approach for SER. The results are validated with feature values from new samples. The salient aspects of modeling SER for Malayalam, separately using decision trees and logistic regression have been presented.

Chapter 4

SPEECH EMOTION RECOGNITION BASED ON PROSODIC FEATURES

Contents

This chapter discusses the results of statistical analysis and classification of emotional speech in English, Hindi and Malayalam at the suprasegmental level, using three popular prosodic features namely, intensity, duration / speech rate and pitch. Similar investigations were carried out at the segmental level too in English, for each feature, in order to compare the performance of SER between these two levels. English being a syllable based language, SER based on syllable speech rates was also analyzed. The individual contributions of each feature, as well as their combined role in speech emotion recognition were assessed for each language. Incidences of universality in the expression of emotions across English, Hindi and Malayalam were identified. The Fuzzy C-Means, KMeans, Naïve Bayes, KNN and the ANN classifiers have been used for the prosodic feature based classification of emotions. Results indicate improved SER for statistically well discriminated feature values. The final results were validated with new emotional speech samples and the results of human SER. There are no available results for such prosodic feature based SER in Indian English, Hindi as well as Malayalam. Hence the obtained results are compared with those available in literature, for the prosodic features in other languages.

4.1. Introduction

The Section 2.5 of Chapter 2 has revealed prosody to be one of the most important acoustic correlates of emotion, according to human perception [14], [129]. The prosodic features within the purview this investigation are, intensity, speech rate or duration, and pitch. The principal aim of these experiments was to investigate the possibility of recognizing neutral and the six basic emotions in terms of the three prosodic features individually, as well as in combination, for each language. Other specific aims were to,

i. Investigate the existence of prosody at the segmental level.

ii. Compare the SER rates across the three languages, for the same feature set.

iii. Assess the valence dependency of SER rates.

iv. Compare the obtained SER rates with the ANOVA based discrimination of emotions.

Sections 4.2 to 4.4 discuss SER based on individual prosodic features namely, intensity, duration / speech rate and pitch of the utterance in English, Hindi and Malayalam. Section 4.5 presents the results of SER based on the complete prosodic feature set. Section 4.6 discusses pitch contour based SER. Section 4.7 gives the comparison with the state of the art. The chapter summary is given in Section 4.8.

4.2. Intensity based SER

This section discusses intensity based SER at the suprasegmental level in all three languages, and at the segmental level in English. The statistical discriminations of feature values for the various emotion pairs were evaluated using ANOVA.

4.2.1. Intensity Analysis of Segmental English utterances

Figure 4.1 gives the visual representation of a sample segmental speech sound 'I' uttered in happiness.

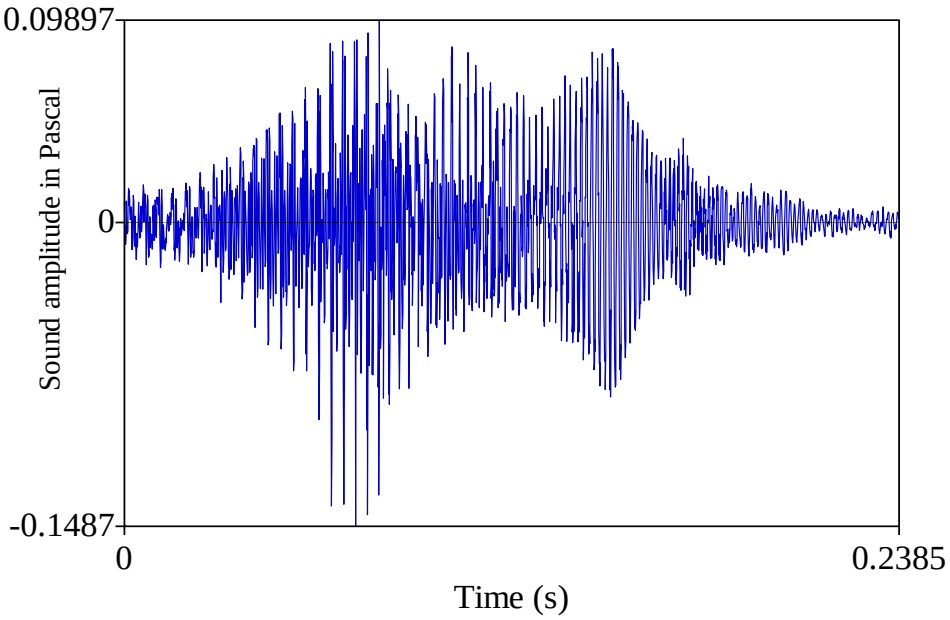


Figure 4.1: Sample sound file representation of "I"

The very short duration of vowel sounds made it difficult to portray the desired emotional content, added to the fact that only elicited emotions were used in this investigation.

Statistical analysis of the vowel intensities under the various classes of emotions was carried out prior to the classifications. The important statistical measures along with the results of ANOVA are given in Table 4.1. The extent to which the emotions could be differentiated was also identified in terms of the significance levels corresponding to $P < 0.001$, $P < 0.01$ and $P < 0.05$.

Table 4.1: Summary statistics of ANOVA of Intensities of Segmental English utterances

Statistical parameters of vowel intensities	Intensities in dB for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean	71.63	74.2	69.18	77.35	69.39	70.95	70.47
SD	6.08	6.96	5.96	5.01	5.13	7.02	5.9
SEM	0.6024	0.6892	0.5910	0.498	0.508	0.695	0.583
Minimum intensity	60.95	61.12	55.38	64.04	56.26	53.25	59.32
Maximum intensity	88.72	88.90	81.56	86.69	79.68	85.62	82.16
Intensity Range	27.77	27.78	26.18	22.65	23.54	32.37	22.84
Emotions discriminated with $P < 0.05$	surprise, neutral, anger, sad	all pairs	happy, surprise, anger, sad	all pairs*	happy, surprise, neutral, anger	surprise, anger	surprise, anger

*$P < 0.001$

Statistical analysis of the intensity of vowel sounds showed anger to be the best discriminated from other emotions. The mean intensities of all other emotions are different from that of anger, with a very high significance level ($P < 0.001$). Surprise too was discriminated from all other emotions. Fear and disgust were the least discriminated, statistically. Fear was characterized by the highest SD and SEM.

A more detailed, utterance wise examination of the mean intensities of the five segmental utterances under each of the seven emotions, was done in order to understand better, the influence of emotions on the intensity of any specific utterance. Table 4.2 provides information on the utterance specific, average intensities under each of the seven emotions, which is also illustrated in Figure 4.2.

Table 4.2: Utterance specific mean intensities at segmental level

Emotions	Mean Intensity of segmental utterances (in dB)					Range of Mean intensity (dB)
	A	e	i	o	u	
Happy	75.34	73.51	72.98	77.04	73.16	4.06
Surprise	80.16	76.39	74.52	80.88	80.18	6.36
Neutral	72.8	71.11	68.89	73.48	71.19	4.59
Anger	77.92	79.27	75.05	81.68	77.96	6.18
Sad	71.2	74.87	67.87	71.87	70.51	7
Fear	72.94	74.72	70.55	75.91	77.26	6.71
Disgust	70.04	75.13	69.03	73.13	73.49	6.1

The intensity range across the seven emotions was higher for a / o / u, compared to those of e / i. When choosing the average intensity values for SER, preference was given to those intensity values (utterances) with maximum intensity range across the seven emotions and minimum variation across the different utterances of the same emotion. This approach was feasible since these investigations were conducted on utterances of known semantic content. For instance, considering the case of happy emotion, conformity among the intensity values was noted within the utterance group of e, i u as well than within another group comprising a and o, as indicated in Figure 4.2. Identification of such suitable sub groups is important for SER, since the intensity values for different emotions are otherwise very close to each other.

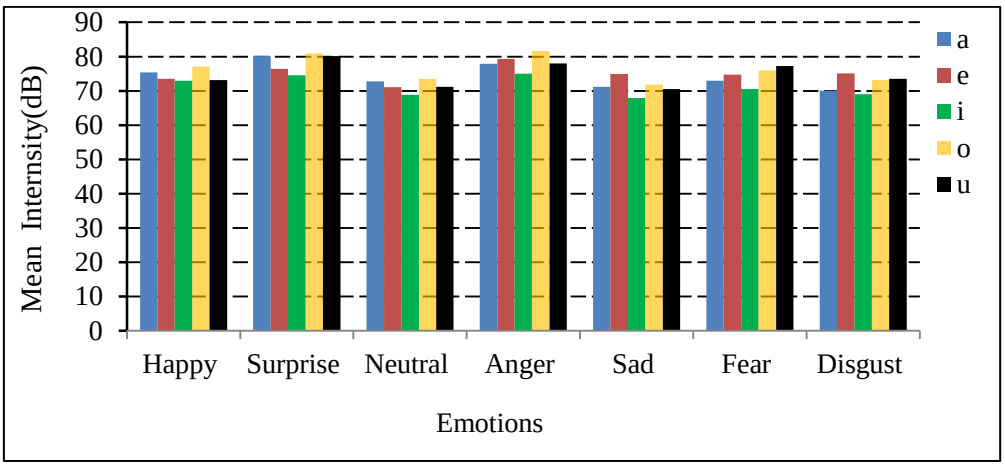


Figure 4.2: Vowel specific utterance intensities for seven emotions

SER based on Segmental intensity: The segmental intensity based SER rates obtained with each of the five classifiers are presented in Table 4.3. The SER cut off rate was fixed at 60%, arbitrarily.

Table 4.3: Consolidated Segmental Intensity based SER rates

Emotions	SER rates in percentage for the various classifiers				
	FCM	K Means	KNN	NB	ANN
Happy	10.26	30.3	30.8	64.1	43.3
Surprise	30.77	22.5	53	62.5	40
Neutral	30.77	11.8	23.1	33.3	10
Anger	38.46	22.5	40	82.5	58.3
Sad	30.77	30.3	40	51.3	50
Fear	7.7	21.5	17.5	50	33.3
Disgust	23.1	20.6	28.2	26.6	33.3
Average	24.55	22.79	33.23	52.9	38.31

Anger and surprise were the best classified emotions. Neutral, disgust and fear were poorly classified. The SER rates agreed with the results of ANOVA, as the emotions (anger and surprise) that were statistically discriminated the most, were the ones that were better classified, by each classifier. The NB classifier gave relatively higher classification accuracies (except for disgust), and was the only one which gave SER rates above 60%. SER rates above cutoff were obtained only for anger and both the positive valence emotions. The SER rates were valence independent. The results of intensity analysis and SER, based on suprasegmental English utterances are presented next.

4.2.2. Intensity analysis of Suprasegmental English utterances

The statistical analysis of English utterances up to five words in length, showed anger and sadness to be the best discriminated from all other emotions. Table 4.4 gives the summary statistics of ANOVA of suprasegmental utterance intensities, along with their emotion discriminations.

Table 4.4: Summary statistics of ANOVA for English utterance intensities

Statistical parameters of intensities	Intensities in dB for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean intensity	74.45	76.44	72.81	82.27	70.47	73.1	78.69
SD	2.005	2.68	1.71	1.96	1.42	2.51	0.55
SEM	0.4276	0.5714	0.3652	0.4170	0.3017	0.5348	0.1167
Minimum Intensity	71.2	72.37	67.6	78.68	67.23	68.73	77.14
Maximum Intensity	77.96	81.6	75.39	86.04	73.16	76.678.	79.52
Intensity Range	6.76	9.23	7.79	7.36	5.93	7.95	2.38
Emotions discriminated by ANOVA; $P < 0.001$	all except neutral, fear	all except disgust	all except happy, fear	all pairs	all pairs	all except happy, neutral.	all except surprise

Those emotions that were discriminated from the others were done so with a very high significance level ($P < 0.001$). Surprise and disgust could not be discriminated from each other. The mean intensity profiles across the seven emotions for segmental and suprasegmental utterances are presented in Figure 4.3. Similarity in intensity profiles was observed for all emotions, except for disgust, for which the mean intensity was much higher for suprasegmental utterances.

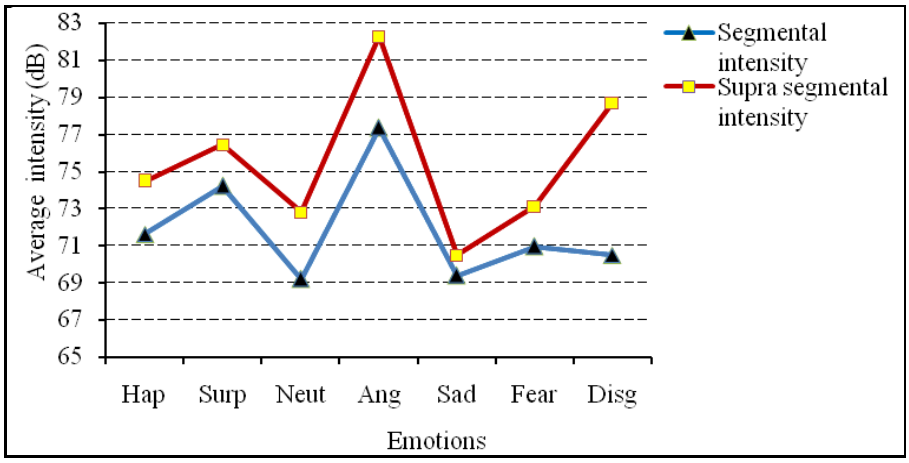


Figure 4.3: Intensity profiles at Segmental and Suprasegmental levels

This indicates that except for disgust, possible similarities could be expected in the SER at both levels.

SER based on Suprasegmental intensity: The intensity based SER rates of the four classifiers for suprasegmental English utterances are given in Table 4.5.

Table 4.5: Consolidated Suprasegmental Intensity based SER rates in English

Emotions	SER rates in percentage for various classifiers				
	FCM	K Means	KNN	NB	ANN
Happy	40.9	40.9	50	33.4	100
Surprise	13.63	30.4	38.5	66.7	40
Neutral	45.45	37.5	14.3	50	57.1
Anger	36.36	75	21.4	100	83.3
Sad	54.55	66.7	72.7	100	83.3
Fear	27.27	17.3	54.5	16.7	100
Disgust	95.45	94.4	44.4	100	100
Average	44.8	51.74	42.3	66.1	80.53

SER rates above 90% were obtained for disgust, with four different classifiers, indicating consistency in the recognition of disgust based on intensities of suprasegmental utterance for English. This further agrees with the ANOVA results given in Tables 4.1 and 4.4 of better statistical discrimination for disgust, at the suprasegmental level compared to the segmental level. However, neutral was poorly recognized by all five classifiers.

4.2.3. Comparison of intensity based SER at Segmental and Suprasegmental levels in English

The intensity based SER rates for English varied with the analysis level, classifier and emotion. Figure 4.4 shows the comparison of the best classification rates at the two levels across the five classifiers.

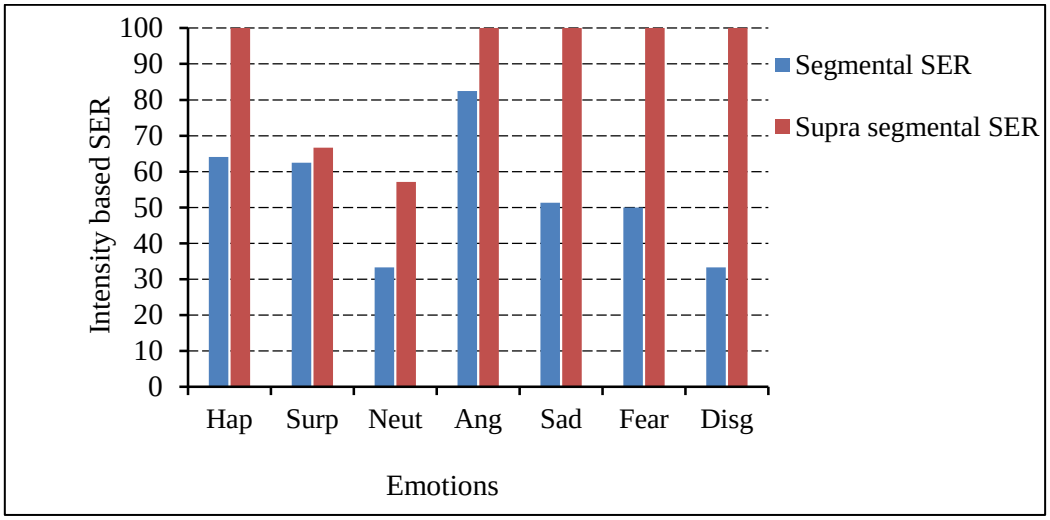


Figure 4.4: Comparison of the best emotion classification rates at the segmental and suprasegmental levels.

Based on Figure 4.4 and Tables 4.1, 4.3, 4.4 and 4.5, it can be concluded that, at the suprasegmental level,

- 100% SER rate was obtained for each emotion (other than surprise and neutral), which therefore benefitted from the increased size of the analysis unit.
- Anger and sadness showed consistently high recognition accuracies with at least three classifiers.
- Emotions that were the best discriminated had the highest SER rates; emotions with poor statistical discriminations mostly had poor recognition rates.
- With increase in analysis unit size, surprise showed only a marginal increase in the classification rate.

At the segmental level, anger was recognized the best, at 82.5% with the NB classifier in agreement with the results of ANOVA. The best classification

accuracies at the segmental and suprasegmental level were obtained with the NB and ANN classifiers. Further, the SER rates were valence independent.

4.2.4. Intensity analysis of Hindi utterances: The intensity analysis approach was the same as detailed for English. The results of ANOVA are presented in Table 4.6.

Table 4.6: Summary statistics of ANOVA of Suprasegmental Hindi utterance intensities

Statistical parameters of utterance intensities	Intensities in dB for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean intensity	75.79	80.56	78.47	78.74	73.55	74.99	77.33
SD	3.05	2.31	2.33	2.69	4.56	2.69	3.28
SEM	0.5988	0.4534	0.4650	0.5270	0.9720	0.5270	0.644
Minimum intensity	68.04	75.88	74.70	73.87	63.8	63.88	69.9
Maximum intensity	81.13	84.36	83.24	84.25	79.59	80.90	81.9
Range	13.09	8.48	8.54	10.38	15.79	17.02	12.08
SD	3.05	2.31	2.33	2.69	4.56	2.69	3.28
Emotions discriminated by ANOVA $P < 0.05$	all except fear, disgust	all pairs	all except anger, disgust	all except neutral, disgust	all except fear	all except sad, happy	surprise, sad, fear,

Surprise was the best discriminated from all other emotions on the basis of mean intensity of Hindi utterances. Disgust was the least discriminated emotion. The similar values of mean utterance intensities for sadness, fear and happiness indicate poor statistical discrimination among these emotions.

The statistically analyzed intensity values of the Hindi suprasegmental utterances were given to the various classifiers. The results of classification by k-means, kNN, Naive Bayes and the ANN classifiers are given in Table 4.7.

Table 4.7: Consolidated Intensity based SER for Hindi utterances

Emotions	SER rates in percentage for the various classifiers			
	K Means	**KNN**	**NB**	**ANN**
Happy	34.6	23.1	61.52	50
Surprise	38.5	38.5	62	60
Neutral	32	25	40	50
Anger	19.2	8	69.22	25
Sad	18.2	36.4	36.4	100
Fear	42.3	38.46	53.8	50
Disgust	11.5	15.4	23.1	28.6
Average	28.04	26.41	49.43	51.94

Following are the conclusions from intensity based SER for Hindi, based on Tables 4.6 and 4.7.

- Sadness was the best recognised (at 100%).
- The second highest SER rate across all other emotions and classifiers was only 69.2%. These low values for the intensity based SER rates could be attributed to the fact that all Hindi utterances in the database except for sadness were perceptually loud, irrespective of the emotion and content.
- Surprise showed relatively consistent and good recognition rate (above 60%) with two of the classifiers. This again agreed with the maximum statistical discrimination for surprise.
- Disgust was the least recognized, which agrees with the poor statistical discrimination results for disgust.
- Further, the SER rates were found to be valence independent.

4.2.5. Intensity analysis of Malayalam utterances

The Malayalam utterances analyzed were excised from the speech of ten native speakers. Important statistical details are as given in Table 4.8.

Table 4.8: Summary statistics of ANOVA of Suprasegmental utterance intensities for Malayalam

Statistical parameters of intensities	Intensities in dB for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean	76.72	78.71	71.22	78.38	71.15	74.03	70.31
SD	1.77	2.19	2.73	3.92	4.28	3.53	1.79
SEM	0.3695	0.4428	0.5697	0.8166	0.8920	0.7360	0.3738
Minimum intensity	73.66	74.55	66.35	71.02	62.9	65.69	67.04
Maximum intensity	80.6	82.38	76.05	87.15	79.02	78.70	72.96
Intensity range	6.94	7.83	9.7	16.13	16.12	13.01	5.92
Emotions discriminated by ANOVA $P < 0.05$	all pairs	all except anger	all except sad, disgust	all except surprise	all except neutral, disgust	all pairs	all except neutral, sad

Emotions were discriminated mostly at a low level of significance. Even so, happy and fear were the better discriminated emotions. Neutral, disgust and sadness were less discriminated. The highest value for mean intensity and maximum intensity was for surprise. The largest intensity range was for anger.

The intensity based SER rates for Malayalam utterances are given in Table 4.9.

Table 4.9: Consolidated Suprasegmental Intensity based SER rates for Malayalam

Emotions	SER rates in percentage for various classifiers			
	K Means	**KNN**	**NB**	**ANN**
Happy	42.9	50	90	62.5
Surprise	61.9	70	70	83.3
Neutral	33.3	30	20	100
Anger	14.3	50	20	66.7
Sad	28.6	30	40	100
Fear	28.6	50	50	50
Disgust	52.4	50	90	60
Average	41.5	47.14	54.29	74.64

Surprise showed the highest consistency in emotion recognition rates with above 60% SER by the four classifiers. This was followed by happiness

and disgust. The ANN classifier gave the highest average SER rate across the seven emotions. The statistical discrimination as well as SER rates based on utterance intensities in Malayalam, were found to be valence independent.

4.2.6. Comparison of Intensities and SER rates of English, Hindi and Malayalam utterances

This section presents the salient features of intensity based emotion analysis and classification of suprasegmental utterances in each language. The mean intensity profiles for the seven emotions in the three languages are given in Figure 4.5.

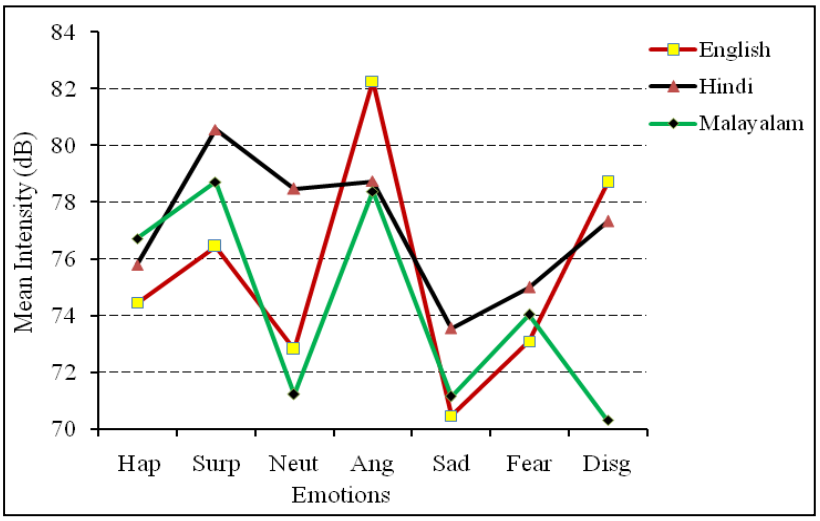
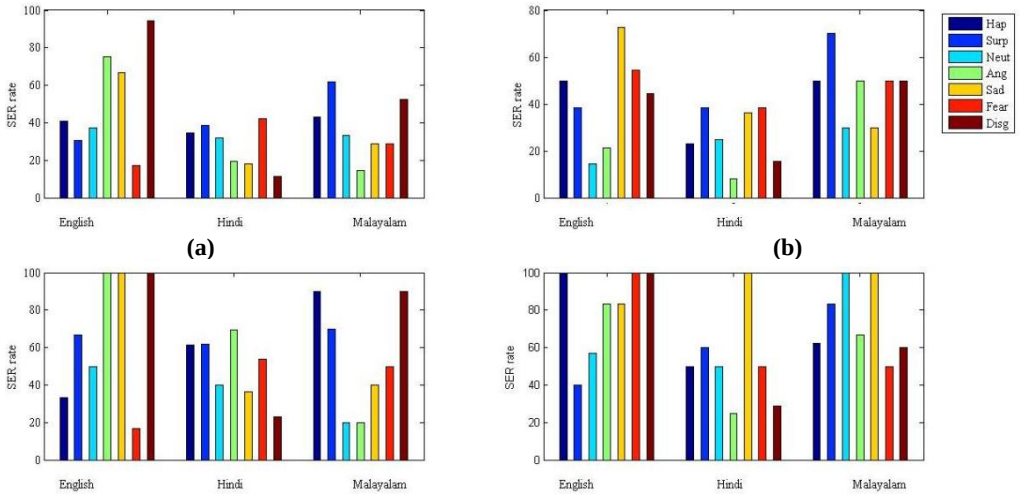


Figure 4.5: Emotion specific intensities of Suprasegmental utterances in English, Hindi and Malayalam

The intensity values in English, Hindi and Malayalam were the closest for fear. Irrespective of the language, sadness showed the least value for mean intensity. The highest of all average intensities was for angry utterance in English. The intensity of utterances under disgust was notably low for Malayalam whereas, it was higher for Hindi and English.

Figure 4.6 presents the consolidated picture of classifier performances for intensity based SER in English, Hindi and Malayalam for all four classifiers.



a) KMeans classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier

Figure 4.6: Comparison of Intensity based SER rates at suprasegmental level in English, Hindi and Malayalam.

The important inferences drawn from the SER in the three languages are as follows:

English

- 100% classification accuracy was obtained for each of the emotions other than surprise and neutral, by at least one classifier.
- The highest average emotion classification rate across the classifiers was for disgust, followed by sadness and anger.
- The ANOVA results in Table 4.4 had indicated good discrimination for anger and sadness.
- The overall best classification rates were obtained with the ANN classifier, followed by the NB classifier.

Hindi

- Highest average classification rate was obtained for surprise; the only emotion that was classified with more than 60% accuracy by more than one classifier.

- The least classification rates were for disgust, as indicated by ANOVA.
- 100% classification accuracy was obtained for sadness with the ANN classifier which also gave higher average SER than other classifiers.

Malayalam

- 100% SER rate was obtained for neutral and sadness with ANN classifier.
- Very high SER rates were obtained for all emotions other than fear and anger.
- ANN and the NB classifier gave the best classification rates.

At the suprasegmental level, the SER rates were found to be valence independent, for all three languages.

4.3. Duration / Speech rate based SER

This section focuses on the analyses of duration / speech rate based measures for SER in English, Hindi and Malayalam. Results of segmental, syllable based and word based investigations for SER are presented in Sections 4.3.1, 4.3.2 and 4.3.3 respectively. This is followed by discussions of SER at the suprasegmental level in Hindi and Malayalam.

4.3.1. Duration analysis of Segmental English utterances

The segmental duration analysis in English was done using the five vowel utterances a, e, i, o, u in the seven emotions. Statistical analysis was carried out on the average duration values using standard ANOVA. Table 4.10 presents the summary statistics of the vowel duration measurements in terms of the statistical parameters of mean, standard deviation, standard error of mean, maximum and minimum average durations.

Table 4.10: Summary statistics of ANOVA of Segmental durations

Statistical parameters of segmental duration	Segmental durations in seconds for the various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean duration	0.3781	0.4064	0.3280	0.3238	0.4209	0.3899	0.4733
SD	0.0584	0.0899	0.0565	0.0947	0.0709	0.073	0.0912
SEM	0.0107	0.0164	0.0103	0.0170	0.0129	0.0134	0.0167
Minimum duration	0.2201	0.2848	0.2330	0.1798	0.2678	0.2144	0.2830
Maximum duration	0.464	0.663	0.479	0.612	0.560	0.518	0.663
Range	0.244	0.378	0.2460	0.4322	0.2922	0.3036	0.377
Emotions discriminated ($P < 0.01$)	neutral, anger, disgust	neutral, anger, disgust	all except anger	all except neutral	all except happy	neutral, anger, disgust, sad	all pairs

Anger had the least mean duration, but the highest SD and SEM. The results of statistical discriminations indicated disgust as the most discriminated, followed by anger and neutral. Statistically, the least discriminated emotions on the basis of segmental durations were happiness, sadness and fear**.** The mean, minimum and maximum average durations are as plotted in Figure 4.7 and indicate the scope for emotion discrimination based on these three duration statistics.

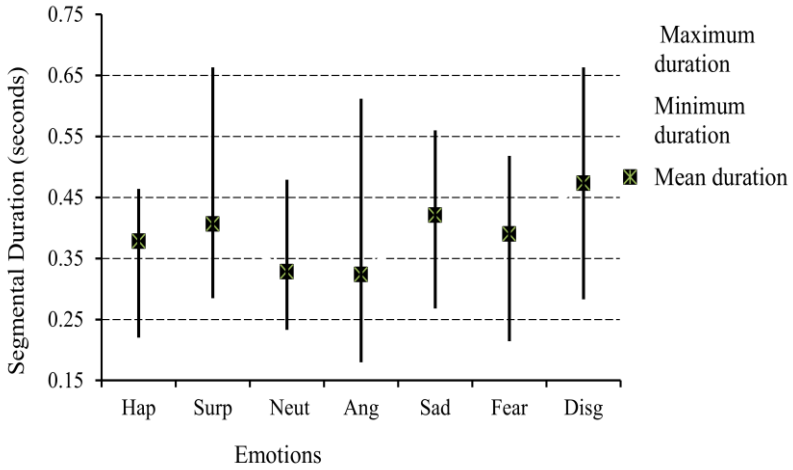


Figure 4.7: Minimum, mean and maximum segmental duration

For instance, even though the mean duration for neutral and anger are close, the maximum durations are relatively far apart and can thus be discriminated. The segmental mean durations were found to be independent of the valence of emotions, supporting the obtained results for the statistical discrimination of emotions, which were also valence independent.

Segmental duration based SER: The emotion classification rates for segmental durations obtained using the FCM, KMeans, Naïve Bayes, KNN and the ANN classifiers are presented in Table 4.11.

Table 4.11: Consolidated Segmental duration based SER rates

Emotions	SER rates in percentage for the various classifiers				
	FCM	**KMeans**	**KNN**	**NB**	**ANN**
Happy	52	43.3	71.4	28.6	40
Surprise	8	6.6	20	13.3	100
Neutral	40	50	28.6	64.3	75
Anger	20	26.7	33.3	26.7	50
Sad	20	13.3	40	53.3	40
Fear	28	43.3	6.7	33.3	25
Disgust	44	50	28.6	57.1	75
Average	30.29	33.31	32.66	39.51	57.86

With the cutoff in classification accuracy for good emotion recognition fixed arbitrarily at 60%, happiness, surprise, neutral and disgust were well recognized. The classification results of disgust, neutral, sad and fear, agreed with the statistical discrimination results of ANOVA given in Table 4.10. The duration based SER rates were valence independent. Comparatively better SER rates were obtained with the ANN classifier.

4.3.2. Syllable rate analysis of English utterances

This investigation was done on syllables that were linguistically identified from the utterances, for the sake of better classification accuracy. The results of the statistical analysis (ANOVA) of syllabic rates are presented in Table 4.12.

Table 4.12: Summary statistics of ANOVA of syllable based speech rates

Statistical parameters	Speech rate for various emotions (in Syllables per minute)						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean speech rate	284.8	191.61	262.94	334.6	227.2	276.8	223.38
SD	80.6	41.83	54.79	60.98	59.13	40.17	31.44
SEM	19	9.86	12.92	14.37	13.94	9.47	7.41
Minimum speech rate	180	118	202	248	113	199	159
Maximum speech rate	397	244	359	467	330	355	271
Range of Speech rate	217	126	157	219	217	156	12
Emotions discriminated ($P < 0.001$)	all except fear, neutral	all except sad, disgust	surprise anger	all pairs	happy, anger, fear	all except happy neutral,	happy, anger, fear

The results showed significant differences ($P < 0.001$) between various emotions as given in Table 4.12. This can be attributed to the fact that English is a syllable based language. Anger had the highest mean speaking rate and was the best discriminated from all other emotions. This was followed by happiness and surprise. Even though the mean syllable rates of sadness and disgust are close, these emotions could be discriminated on the basis of the minimum and maximum syllable speech rates.

SER based on Syllable rates: The classification of emotions based on the syllable rates was assessed in term of the performance of the four classifiers and are presented in Table 4.13.

Table 4.13. Consolidated SER rates in English based on Syllable rates

Emotions	SER rates in percentage for the various classifiers			
	K Means	**KNN**	**NB**	**ANN**
Happy	33.3	42.9	57.14	50
Surprise	38.9	33.3	44.4	75
Neutral	22.2	62.5	37.5	60
Anger	27.8	33.3	86.7	66.7
Sad	16.7	22.2	62.5	66.7
Fear	38.9	25	62.5	50
Disgust	44.4	11.1	66.7	50
Average	31.74	32.9	59.6	59.8

Anger was the best classified as it had the highest overall SER as well as a maximum SER of 86.7%. This is in agreement with the results of statistical discrimination by ANOVA as given in Table 4.12.

The overall SER performance of both the NB and the ANN classifiers were similar and both gave comparatively better SER rates, than the k-means and the KNN classifiers. The obtained SER rates based on the NB classifier were higher for the negative valence emotions compared to that for the positive valence emotions.

4.3.3. Word Speech Rate Analysis in English

ANOVA was performed on the tabulated word-speech rates to identify statistically different emotion classes. The summary statistics are presented in Table 4.14.

Table 4.14: Summary statistics of ANOVA of word rates in English

Statistical parameters	**Speech rate for different emotions in words per minute**						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean word rate	202.7	151.2	183.44	244.1	174.06	233.5	189.67
SD	46.3	20.8	31.4	42.9	38.5	31.25	28.24
SEM	10.91	4.91	7.41	10.12	9.08	7.37	6.66
Minimum word rate	140	118	142	200	107	168	125
Maximum word rate	291	182	233	334	226	275	237
Speech rate range	151	64	91	134	119	107	112
Emotions discriminated $P < 0.01$	all except sad, disgust	all except sad	happy, anger, fear	all except fear	anger, fear	all except neutral, anger	surprise, anger, fear

Word speech rates for anger differ significantly from that of all other emotions, other than fear. The word rates were valence independent. Certain important statistics of word speech rates are presented in Figure 4.8.

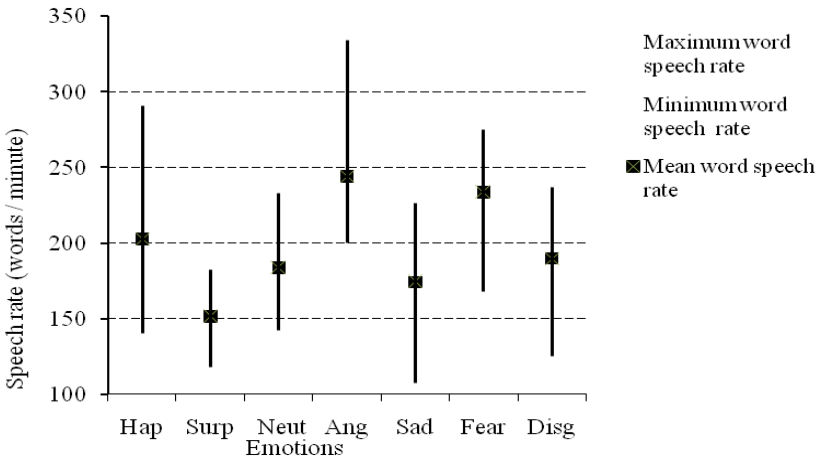


Figure 4.8: Minimum, mean and maximum word rates in English

Even though the mean word speech rates of anger and fear are confusingly close, their maximum speech rates are sufficiently far enough for discrimination. The results of word speech rate based emotion classification are given in Table 4.15.

Table 4.15: Consolidated SER rates in English based on Word rates

Emotions	SER rates in percentage for the various Classifiers			
	KMeans	KNN	NB	ANN
Happy	33.3	66.7	16.7	100
Surprise	27.8	42.9	85.7	75
Neutral	38.9	44.4	55.6	80
Anger	16.7	44.4	88.9	33.3
Sad	27.8	28.6	28.6	50
Fear	44.4	55.6	44.4	40
Disgust	50	11.1	66.7	66.7
Average	34.1	41.9	55.3	63.6

All emotions except sadness and fear could be classified with accuracies above 60%. The highest average SER rate was obtained with the ANN classifier, followed by the NB classifier.

Comparison of duration / speech rate based SERs in English: The performance summary of the duration / speech rate based emotion recognition at the three levels in English is given in Table 4.16, with 60% as the cut off in SER rate

Table 4.16: Comparison of SER rates for various analysis units

Emotions	Segmental Level	Suprasegmental Level	
		Syllable	Word
Happy	√	X	√
Surprise	√	√	√
Neutral	√	√	√
Anger	X	√	√
Sad	X	√	X
Fear	X	√	X
Disgust	√	√	√

Maximum number of emotions were recognized based on syllable rates. Even though positive valence emotions have been recognized based on duration and word rate, no strong valence dependency of SER rates was observed based on the obtained results.

4.3.4. Observations from Duration / speech rate analysis for English

The analysis of the statistical parameters obtained from duration, syllable rate and word rate measurements helped to identify the following, for the rule based classification of emotions.

- Happiness had the maximum standard deviation on the basis of both word and syllable speech rates. It had the largest range of average speech rate in words per minute as well as the smallest range of segmental duration. It was statistically well discriminated on the basis of speech rates.
- Surprise had the lowest average word rate and syllable rate along with lowest values of SD, SEM, minimum, maximum and range in word rate. It had the highest minimum and maximum values in average vowel duration. It was statistically well discriminated based on speech rates.
- Neutral had the least value for mean duration and its standard deviation, and was well discriminated based on vowel duration and word rate.
- Anger was characterized by the least vowel duration as well as the highest speech rates in terms of words and syllables. It was statistically well discriminated at all levels.

- Sadness had the least values of minimum word and syllable speech rates. Sadness could not be discriminated from surprise on the basis of speech rates as no significant difference was noted between their average values.
- Fear had mean values close to those of happiness at the vowel and syllable levels respectively.
- Disgust had the longest vowel duration and could be statistically distinguished from all other emotions with a very high level of significance ($P < 0.005$, for emotion pairs of disgust with fear / anger / happy). It had the least SEM and range for syllable speech rate.

4.3.5. Word Speech rate analysis of Hindi utterances

The results of the statistical analysis of Hindi speech rates are presented in Table 4.17.

Table 4.17: Summary statistics of ANOVA of Hindi word speech rates

Parameters	Speech rate in words / minute for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean word speech rate	195.57	164.24	206.56	227.05	187.47	160.13	148.08
SD	24.85	32.28	21.18	36.41	34.31	35.2	32.49
SEM	6.64	7.83	4.99	7.95	8.86	8.8	9.01
Maximum speech rate	250	218	248	300	263	230	194
Minimum speech rate	155	118	176	182	128	121	96
Range of speech rate	95	100	72	118	135	109	98
Emotions discriminated ($P < 0.05$)	all except sad, neutral	all except disgust, fear	All except happy, sad	all pairs	all except neutral, happy	all except surprise, disgust	all except surprise, Fear

Anger was expressed with a significantly high speech rates. Disgust was characterized by the lowest values of minimum, maximum and mean speech rates. Neutral also has typical mean speech rates above 200 words per minute.

Results of ANOVA indicate no significant discrimination amongst the mean duration of fear, disgust and surprise and amongst mean speech rates for neutral, happy and sadness. This in turn indicates possible confusions in the accurate recognition of these emotions, solely based on the speech rates. Anger was the best discriminated from the rest of the emotions. Both anger as well as neutral could be easily discriminated from fear, surprise or disgust, solely based on the speech rates, since the mean speech rates were significantly different in these cases ($P < 0.001$). The minimum, maximum and average Hindi speech rates in words per minute are presented in Figure 4.9. The important speech rate statistics were found to be independent of the valence of emotions.

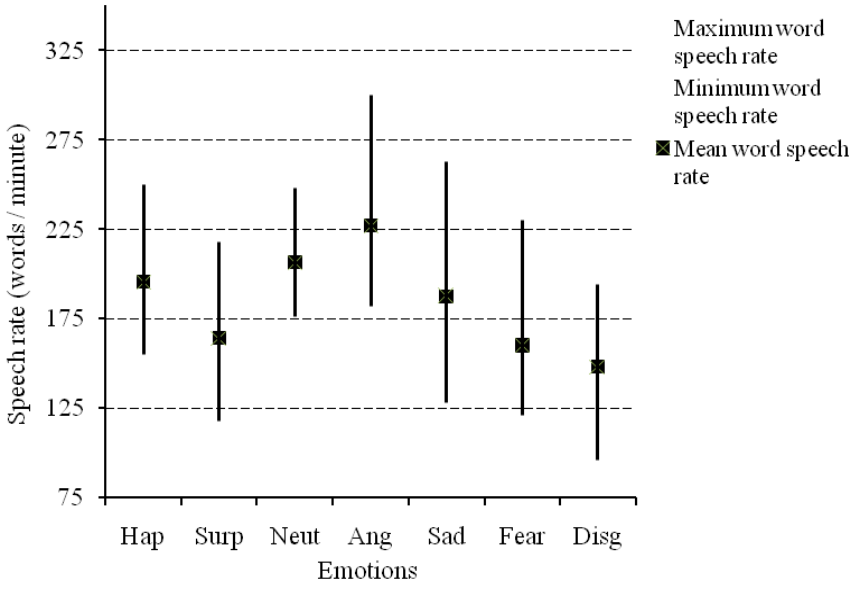


Figure 4.9: Minimum, mean and maximum speech rates in Hindi

Statistical analysis of speech rates was followed by classification of emotions. Table 4.18 presents the emotion classification rates obtained with the KMeans, Naïve Bayes, KNN and the ANN classifiers.

Table 4.18: Consolidated speech rate based SER rates for Hindi

Emotions	SER rates in percentage for the various classifiers			
	K Means	KNN	NB	ANN
Happy	35.7	60	100	50
Surprise	28.6	25	25	100
Neutral	28.6	28.6	71.4	66.7
Anger	57.1	33.3	37.5	100
Sad	21.4	14.3	14.3	50
Fear	21.4	42.9	28.6	50
Disgust	21.4	16.7	66.7	50
Average	30.6	31.5	49.1	66.7

With the 60% cut off in SER rate, all emotions except sadness and fear are recognized. The maximum SER rate for both sadness and fear was 50% only, by the ANN classifier. All instances of happiness, surprise and anger could be recognized. Neutral and happiness were recognized by more than one classifier. The results of classification agree with the statistical discrimination results of ANOVA given in Table 4.17. The best results were obtained with the NB and ANN classifier. The obtained SER rates were found to be valence independent.

4.3.6. Malayalam Word speech rate analysis

The word speech rates of native speakers of Malayalam were investigated. Results of the statistical discrimination are provided in Table 4.19 along with the summary statistics of speech rates.

Table 4.19: Summary statistics of ANOVA of word speech rates in Malayalam

Statistical Parameters	Speech rate in words / minute for various emotions						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
Mean word rate	154.06	142.04	139.75	171	117.08	125.46	120.5
SD	19.12	29.15	27.79	23.94	22.71	20.86	15.63
SEM	4.51	5.95	5.67	4.89	4.64	4.26	3.19
Maximum word rate	191	189	209	209	146	167	154
Minimum word rate	103	93	94	117	76	96	94
Range	88	96	115	92	70	71	60
Emotions discriminated ($P < 0.01$)	all except neutral, surprise	all except neutral, happy	all except happy, surprise	all pairs	all except fear, disgust	all except disgust, sad	all except fear, sad

The analysis of speech rate measurements indicated anger as having the highest speech rate. The second highest mean speech rate was for happiness while the second lowest mean speech rate was for disgust. Statistical analysis using ANOVA indicated no significant difference amongst the mean duration of neutral, happy and surprise as well as amongst fear, disgust and sadness.

Anger was the best discriminated from all other emotions. Figure 4.10 shows the mean minimum maximum and range of speech rate for each emotion.

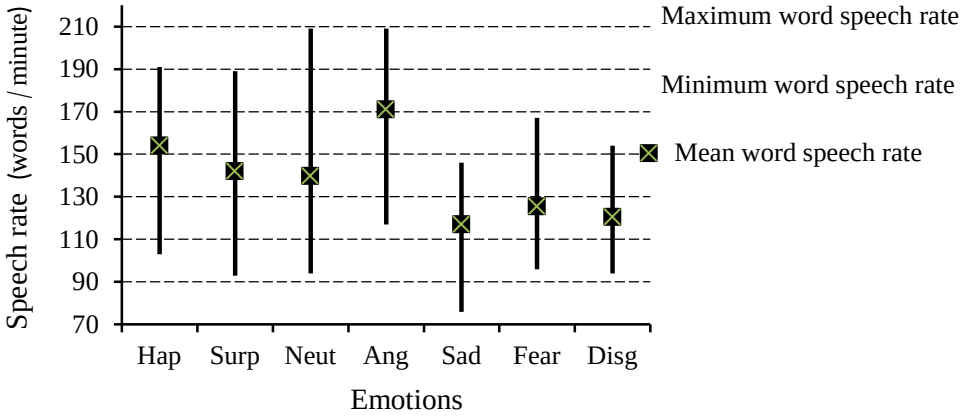


Figure 4.10: Minimum, mean and maximum speech rates in Malayalam

Even though surprise and neutral had very close values of speech rates, the relatively large difference in their maximum speech rate values could be the basis for discriminating these two emotions. The emotion classification results from the four classifiers are consolidated in Table 4.20.

Table 4.20: Consolidated speech rate based SER rates for Malayalam

Emotions	SER rates in percentage for the various classifiers			
	KMeans	KNN	NB	ANN
Happy	55.6	22.2	22.2	75
Surprise	20.8	25	58.3	50
Neutral	33.3	16.7	50	75
Anger	33.3	36.4	72.7	100
Sad	25	18.2	36.4	100
Fear	20.8	36.4	18.2	50
Disgust	33.3	36.4	36.4	50
Average	31.7	27.3	42.0	71.4

With the 60% arbitrary cut off in classification rate, happiness, neutral, anger and sadness have been well recognized by the ANN classifier, with 100% SER rate for anger and sadness. Surprise, fear and disgust are poorly recognized. Anger was the best recognized across the various classifiers, in agreement with the results of ANOVA given in Table 4.19. The obtained SER rates were found to be valence independent.

4.3.7. Summary of Speech rate analysis across English, Hindi, and Malayalam

The salient features of the speech rate analysis in the three languages are as follows. The correlation matrix of speech rates for the three languages pairs and for seven emotions showed maximum correlation between English and Hindi speech rates with correlation coefficient, $\gamma = 0.6235$. Positive correlation among the speech rates in these three languages was obtained for surprise only.

The mean speech rate profiles across English, Hindi and Malayalam, for the seven emotions are presented in Figure 4.11.

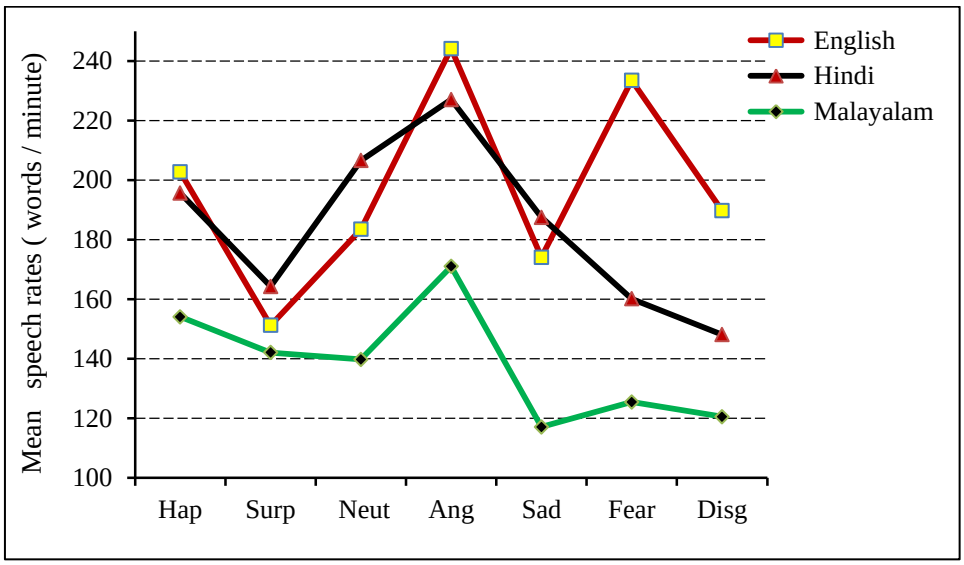


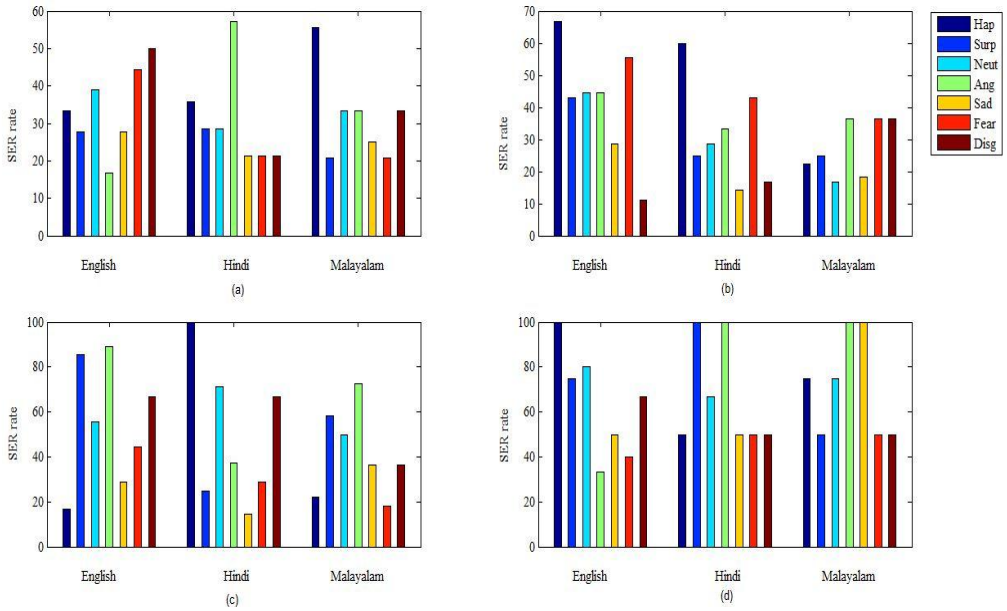
Figure 4.11: Emotion specific mean speech rates in English, Hindi and Malayalam

Irrespective of emotions, the least speech rate was for Malayalam. Speech rates among the three languages converged for surprise, and were the

most spread out for fear. Irrespective of the language, the highest word speech rates were for angry utterances. Therefore, duration or its speech rate measure is a good feature for the identification of anger.

Figure 4.12 helps to assess possible SER performances for the various subsets of emotion classes, prior to their actual classification. For instance, in English as well as Malayalam, a class 3 SER problem most probable of giving high SER rates would be the classification of anger, disgust, and surprise, since their mean speech rates are spread out. Similarly in Hindi, high classification accuracies could be expected when discriminating anger, happiness and fear than when discriminating anger, surprise and fear.

In all the three languages, the various statistical measures of duration and word rates were found to be independent of the valence of emotions. The complete picture of speech rate based classification of the seven emotions in English, Hindi and Malayalam, using the four different classifiers is given in Figure 4.13.



(a) KMeans classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier.

Figure 4.12: Comparison of Speech rate based SER in English, Hindi and Malayalam

Based on Figure 4.13, it can be concluded that all instances of happiness in English and Hindi; surprise in Hindi; anger in Hindi and Malayalam and sadness in Malayalam were recognised. Further, surprise, neutral and anger for English were recognized with accuracy of 80% or above. Speech rate was the most effective in the recognition of anger, for all three languages which is in agreement with the statistical discrimination results of ANOVA given in Tables 4.14, 4.17 and 4.19. Fear was universally (in all three languages) the least recognized on the basis of word speech rates, which was as expected, from the poor statistical discrimination for fear. The word speech rate based SER rates were valence independent.

The best speech rate based SER rates obtained in English, Hindi and Malayalam, for each of the seven emotions are presented in Table 4.21.

Table 4.21: Best speech rate based SER rates in English, Hindi and Malayalam

Emotions	SER rates in percentage for various languages		
	English	Hindi	Malayalam
Happy	100	100	75
Surprise	85.7	100	58.3
Neutral	80	71.4	75
Anger	88.9	100	100
Sad	50	50	100
Fear	55.6	50	50
Disgust	66.7	66.7	50
Average	75.27	76.87	72.61

From all the results presented in this section, it is concluded that emotions are better recognized for English and Hindi, than for Malayalam, on the basis of speech rates of suprasegmental utterances.

4.4. Pitch based SER

This section discusses pitch based SER at the segmental level, as well as at the suprasegmental level (with single words and multiword utterances). The investigations included pitch contours too as they provide a visual, global picture of pitch over the entire duration of an utterance. Knowledge of the characteristic

features of pitch contours of various emotions is of great use in speech emotion recognition, emotional speech recognition as well as emotional speech synthesis.

Therefore typical pitch contour characteristics for each of the seven emotions have been identified, by grouping contours of known emotional content.

4.4.1 Pitch based English SER

The pitch values of various vowel utterances were statistically analyzed for different emotions. The utterance wise mean pitch and pitch range for the seven emotions are presented in Table 4.22.

Table 4.22: Utterance specific mean pitch and pitch range for each emotion

Emotions	Mean Pitch values in Hertz for different emotions and utterances						
	a	**e**	**i**	**o**	**u**	**Pitch Range**	**Mean Pitch for each emotion (Hertz)**
Happy	278.78	299.29	295.86	305.19	289.9	26.39	293.81
Surprise	347.89	304.07	318.57	345.88	323.9	43.82	328.06
Neutral	221.14	241.08	214.03	231.91	219.5	27.05	225.52
Anger	268.04	308.04	267.96	290.36	269.7	17.68	280.82
Sad	215.42	268.65	234.49	236.85	239.0	53.13	238.87
Fear	285.15	311.8	293.55	260.37	286.4	51.46	287.46
Disgust	171.91	212.54	204.88	215.72	226.9	54.96	206.39
Range	175.98	99.26	113.69	130.16	104.4	-	-

The least pitch for all vowels except u was for disgust. The highest pitch was for surprise, except for eh. The least pitch range for vowels was for anger, whereas the maximum pitch range was for disgust. Positive valence emotions had higher average pitch than neutral and the negative valence emotions. All emotion pairs except sad-neutral and anger-fear were statistically discriminated by ANOVA ($P < 0.01$).

The important pitch based features identified for each emotion were as follows:

- Happiness had the second highest pitch value.

- Surprise had the highest pitch value aross the different vowel sounds, followed by happy and fear.
- Neutral had the second least pitch value and low pitch range.
- Anger was found to be the most utterance independent emotion and had the least pitch range.
- Sadness was expressed with a wide spread of pitch values, over different utterances. Sadness was characterized by utterance dependent pitch values.
- Fear had the highest pitch value among the negative valence emotions. Both mean pitch value and pitch range were high. Maximum difference in mean pitch was between e and o.
- Disgust had the least mean pitch value and the largest pitch range.

The mean pitches for single worded utterances as well as multi worded utterances are given in Tables 4.23 and 4.24 respectively.

Table 4.23: Mean pitch of single words in English, Hindi and Malayalam

Language	Mean pitch in Hertz for single words for various emotions						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
English	370	277	247	238	236	235	208
Hindi	272	304	267	272	239	244	246
Malayalam	278	317	252	302	275	250	233

In each language, the emotion with the highest average pitch was found to be the same for both single words as well as multi worded utterances**.** Irrespective of the utterance length, the highest average mean pitch was for happy in English, and for surprise in Malayalam and Hindi.

Table 4.24: Mean pitch of multi worded utterances in English, Hindi and Malayalam

Language	Mean pitch in Hertz of multi worded utterances for various emotions						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
English	302	297	224	268	221	225	231
Hindi	308	328	253	270	220	228	233
Malayalam	288	307	237	290	277	260	213

In English, the least mean pitch was for disgust and sadness, for single and multi worded utterances respectively; the highest mean pitch was for happiness irrespective of utterance length. The least pitch range for short multiword utterances was for fear, with anger and disgust above it. The highest pitch range was for happiness followed by surprise and neutral.

In Hindi, irrespective of the utterance length, the lowest average of the mean pitch was for sadness; the highest mean pitch was for surprise. In Malayalam, irrespective of the utterance length, the lowest value of mean pitch was for disgust; the highest mean pitch value was for surprise.

The statistical discriminations between various emotions in English, Hindi and Malayalam, at the suprasegmental level are presented in Table 4.25.

Table 4.25: Consolidated list of emotions that were discriminated by the ANOVA of mean pitch values

Emotions	Emotions discriminated in each of the three languages (P < 0.05)		
	* English	Hindi	Malayalam
Happy	all pairs	all except surprise	all pairs
Surprise	all pairs	all except happy	all pairs
Neutral	happy, surprise	all except sadness	all pairs
Anger	happy, surprise, sad, disgust	all pairs	all pairs
Sad	happy, surprise, anger, fear	all except neutral	all pairs
Fear	happy, surprise, sad	all except disgust	all pairs
Disgust	happy, surprise, anger	all except fear	all pairs

*P < 0.01 for English

Thus all emotions were statistically well discriminated for suprasegmental utterances in Malayalam, though at a low significance level of P < 0.05 only. At the suprasegmental level in English, only happiness and surprise were statistically discriminated from the rest of the emotions.

The results of pitch based emotion classification of the segmental English utterances are given in Table 4.26. At the segmental level, 100% classification accuracy was obtained for surprise, neutral and disgust. With the classification rate fixed arbitrarily at 60%, all emotions except fear was recognized.

Table 4.26: Consolidated Segmental pitch based SER rates

Emotions	SER rates in percentage for various classifiers				
	FCM	K Means	KNN	NB	ANN
Happy	48	52	80	70	60
Surprise	60	52	93.3	80	100
Neutral	68	68	50	75	100
Anger	40	28	83.3	66.7	66.7
Sad	32	60	50	91.7	66.7
Fear	32	8	22.2	25	50
Disgust	44	32	76.9	76.9	100
Average	46.29	42.86	65.1	69.32	77.63

Fear was recognized at 50% only, which agrees with the ANOVA results of Table 4.25, that indicated confusion between anger and fear on the basis of segmental pitch. The best results were obtained with the ANN classifier with average SER of 77.73 %. The SER rates were found to be valence independent. Next, the pitch based classification of suprasegmental utterances are presented in Table 4.27.

Happiness, anger and fear were classified with accuracy above 80%. Neutral and disgust too were classified. Surprise was not recognized satisfactorily. Utterances at the suprasegmental level were found to be less efficient than segmental utterances, in the pitch based classification of emotions. Results indicated that the length of the analysis unit could be made as small as a vowel, for surprise, neutral and disgust, since higher SER rates were obtained for these emotions at the segmental level compared to the suprasegmental level.

Table 4.27: Consolidated pitch based SER rates for Suprasegmental English utterances

Emotions	SER rates in percentage for various classifiers				
	FCM	K Means	KNN	NB	ANN
Happy	40.2	38.9	76.9	76.9	89.5
Surprise	18.6	22.2	46.2	38.5	51.5
Neutral	70.8	72.2	33.3	6.7	30.95
Anger	10	5.6	81.3	31.3	48.9
Sad	43.7	50	64.7	41.9	35.3
Fear	14	16.7	81.2	94	35.1
Disgust	13	16.7	68.8	19	46.2
Average	30.04	31.76	64.63	44.04	48.21

In the case of fear, better emotion recognition was obtained with increased size of the analysis unit. The best SER rates were obtained with the KNN classifier giving an average of 64.63%. The obtained SER rates were found to be valence independent. The SER rates agreed with the results of ANOVA given in Table 4.25, except for surprise.

4.4.2. Pitch based Hindi SER

The pitch values of Hindi utterances were statistically analyzed. As indicated in Table 4.25, anger was statistically discriminated from the rest of the emotions. Each of the other six emotions could be discriminated from five emotions. The classification results are given in Table 4.28. .

Table 4.28: Consolidated pitch based SER rates for Hindi

Emotions	SER rates in percentage for various classifiers			
	KMeans	KNN	NB	ANN
Happy	35.3	30.4	47.1	57.1
Surprise	12	35	50	55.6
Neutral	57.8	44.4	83.3	55.6
Anger	63	53.8	100	100
Sad	25	13.3	13.3	10
Fear	29.2	27.8	72.2	63.6
Disgust	50	27.8	27.8	77.8
Average	38.9	33.2	56.2	58.5

No strong valence dependency could be observed for the SER rates. 100% classification accuracy was obtained for anger in agreement with the results of ANOVA given in Table 4.25. The pitch based speech emotion recognition in Hindi was not effective for happiness, surprise and sadness since these failed to recognized with the 60% SER rate cut off.

4.4.3 Pitch based Malayalam SER

Statistical discrimination of the mean pitch values have been presented in Table 4.25. The results of classifications are given in Table 4.29.

Table 4.29: Consolidated pitch based SER rates for Malayalam

Emotions	SER rates in percentage for various classifiers			
	KMeans	KNN	NB	ANN
Happy	24.4	46.3	40	11.1
Surprise	12.2	24.4	7. 3	50
Neutral	56.5	26.8	36.5	50
Anger	36.6	31.7	50.5	57.1
Sad	25	46.3	34.1	87.9
Fear	14.3	43.9	5.6	57.1
Disgust	25	60	34.1	100
Average	27.71	39.91	29.73	59.03

All instances of disgust were classified. The SER rates were valence dependent, with better performance for negative valence emotions. Investigations were done on SER with the various combinations of these three features and the best results were obtained with the complete feature set, as presented in detail in the following section.

4.5 Complete Prosodic Feature Set based SER

This section presents the results of the ANN classification based on the entire prosodic feature set comprising intensity, pitch and speech rate (at the suprasegmental level) in each of the three languages, as well as for the segmental utterances. The reason for preferring the ANN classifier has already been discussed in Section 3.7. Besides, results discussed in this chapter and presented in Table 4.30, indicate comparatively higher SER rates with ANN classifier for the listed features, with the average SER rate (across seven emotions) indicated against it.

Table 4.30: Consolidated higher overall SER rates obtained by the ANN classifier

Languages	Features that gave best results by the ANN classifier (overall SER in percentage)		
English	Intensity - 74.8%;	Speech rate - 63.6%.	-
Hindi	Intensity - 51.9%;	Speech rate - 66.7% ;	Pitch - 58.5%
Malayalam	Intensity - 74.64%;	Speech rate - 71.4% ;	Pitch - 59%

Thus based on majority, best performance vote, the ANN classifier has been chosen for the final classification with the complete prosodic feature set.

Table 4.31 helps to assess the universality in prosodic feature based SER across English, Hindi, and Malayalam. Surprise, anger and disgust were universally recognized across English, Hindi and Malayalam based on these three prosodic features.

Table 4.31: Complete Prosodic feature set based SER rates by ANN classifier

Emotions	SER rates in percentage for three languages		
	English	Hindi	Malayalam
Happy	75	100	75
Surprise	100	100	100
Neutral	66.7	50	100
Anger	100	100	100
Sad	100	50	83.3
Fear	100	33.3	100
Disgust	100	100	100
Average	91.67	76.19	94.04

The SER rates with the complete prosodic feature set shows remarkable improvement compared to the SER rate based on individual features, especially for English and Malayalam. Whereas the overall average SER rates obtained were 91.67% and 94.04% for English and Malayalam respectively, it was 76.19% for Hindi. In Hindi, the maximum SER rate was only 50% for fear and neutral, based on the complete prosodic feature set.

Table 4.32 presents the results of similar ANN classification of the prosodic feature set extracted from segmental utterances in English.

Table 4.32: SER rates of segmental utterances based on complete prosodic feature set for English

Emotions	Segmental SER in percentage
Happy	100
Surprise	100
Neutral	100
Anger	71.4
Sad	100
Fear	100
Disgust	100
Average	95.91

Higher overall average SER was obtained at the segmental level than at the suprasegmental level. The obtained results were further validated with

feature values from new wave files for each of the seven emotions. 10% validation was done. With a few false hits for surprise and neutral, and misses for anger and sad, the overall SER accuracy with the validation set was 99.3%. Thus the prosodic feature based SER at the segmental level in English was found to be more efficient than at the suprasegmental level.

4.6. Pitch contour based SER

This section presents the findings of an exploratory data analysis of the pitch contours of utterances. Data was visualised and certain general conclusions about the emotional speech could be drawn from the pitch contours. This was possible since the analysis was based on very short utterances with high perceptual rating with respect to the emotional content. Further, the choice of female speakers within the specified age group has served to minimise variances in the pitch profile within any specific emotion class. Pitch contours were grouped in seven different emotion classes. Pitch contours were found to vary with choice of speaker, utterance, emotion. Even so, pitch contour representatives as well as the characteristic features of each emotion for segmental utterances were identified and presented in Figure 4.13, Figure 4.14 and Table 4.33 respectively. Figure 4.13 indicates pitch breaks due to the nature of the utterance.

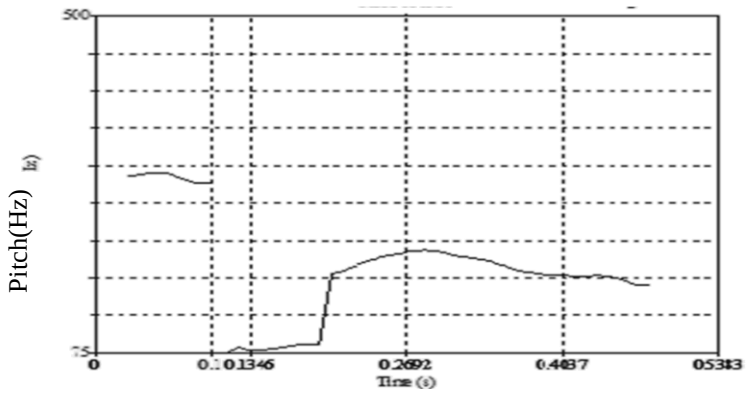


Figure 4.13: Pitch contour of "oh" uttered in disgust

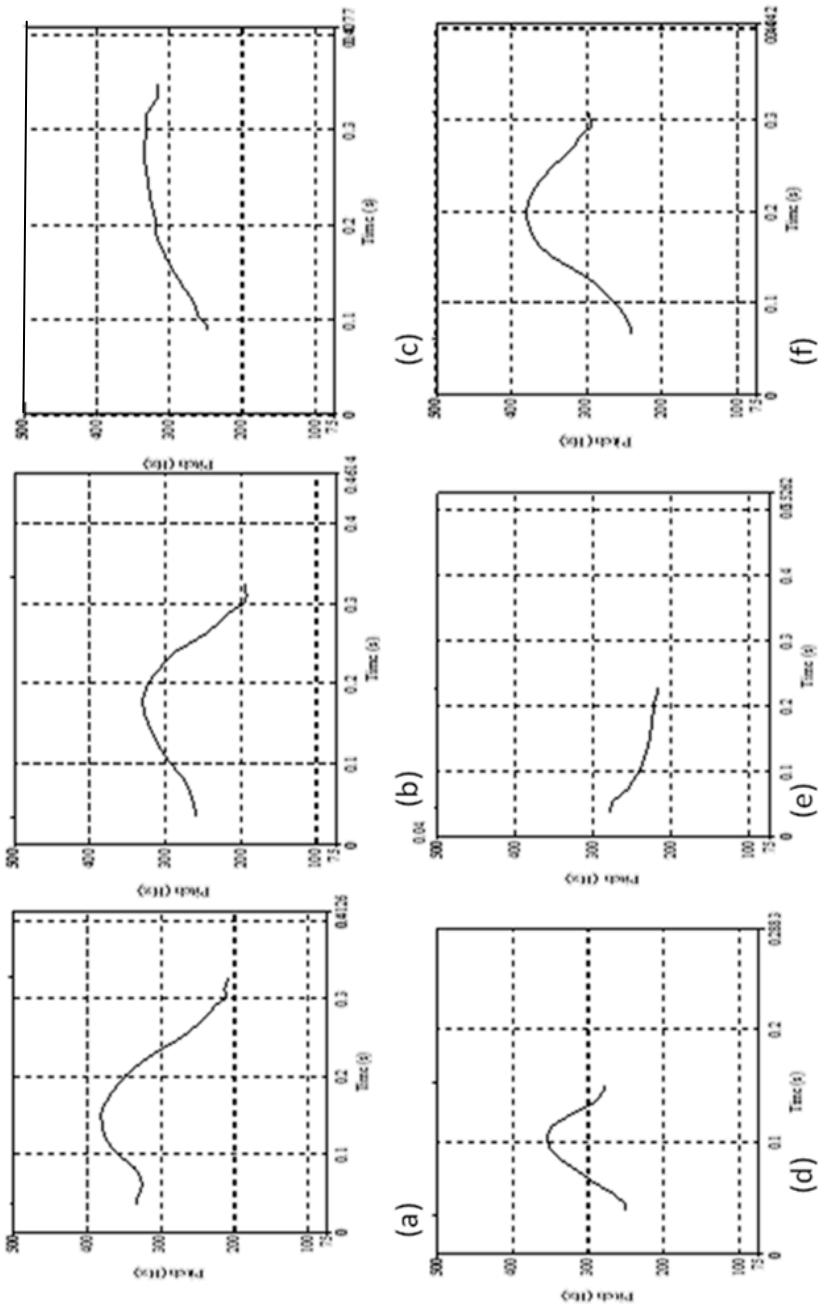


Figure 4.14: Typical pitch contours of "oh" under (a) happiness, (b) surprise, (c) neutral, (d)anger, (e) sad and (f) fear.

Table 4.33: Characteristic features identified at the segmental level for various emotions

Emotions	Characteristic features observed
Happy, surprise (positive valence)	Similar, right skewed, pitch contours; higher average pitch values for surprise.
Neutral, sad	Least pitch variations across utterance.
Anger	Rise to a peak followed by either a decrease or levelling out of the pitch values; left skewed , small utterance duration
Sad	Constancy of pitch over a major part of the utterance
Fear	Pitch increase to a peak, followed by slight decrease. Left skewed.
Disgust	Mostly with pitch breaks and longer utterance durations

SER was done by visually matching test pitch contours of segmental utterances with the class representatives. The observations were been validated with new sample values. The pitch contours of short independent utterances in English, Hindi and Malayalam were analyzed in a similar manner. The pitch contours of similar stand alone, short suprasegmental utterances under each of the seven emotions and for all three languages are presented in Figure 4.15 to Figure 4.19. The specific utterances are given in the legends. Despite differences in the actual utterances investigated, similarity was observed in the overall pitch profiles for disgust in the three languages. Happy utterances in English had larger pitch range than those under surprise.

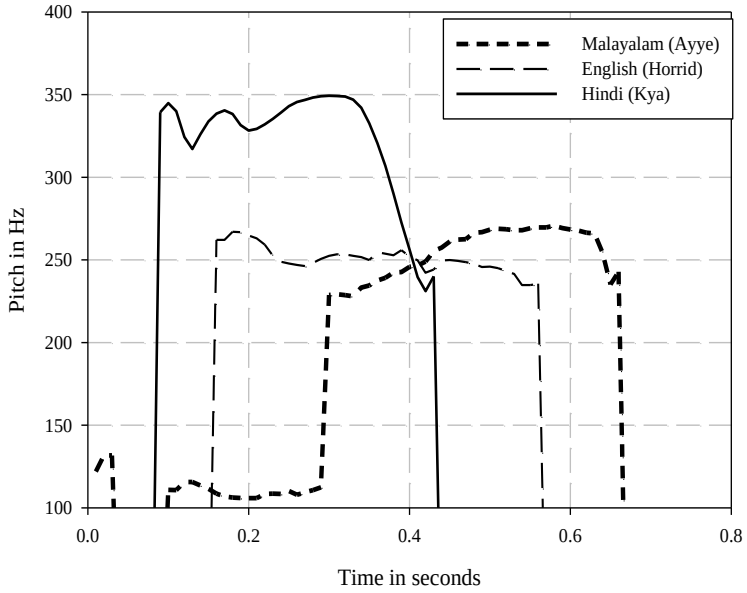


Figure 4.15: Typical pitch contours of disgust in the three languages

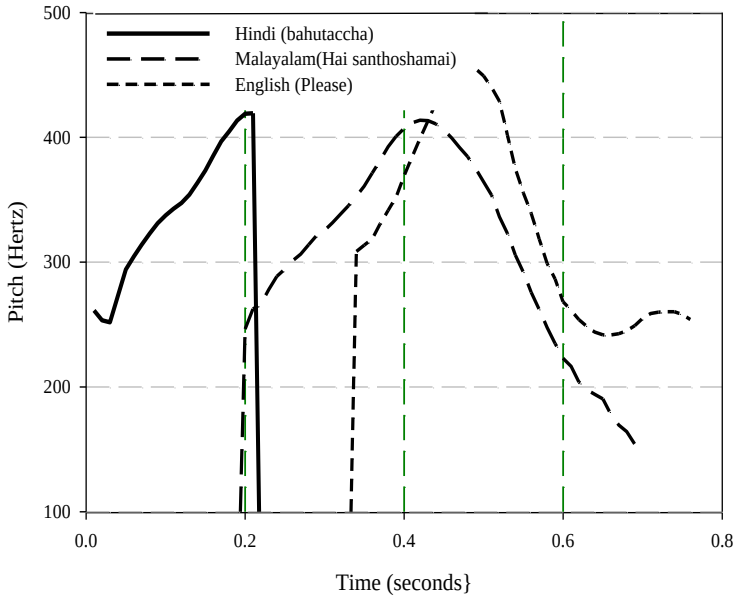


Figure 4.16: Pitch contour of happiness

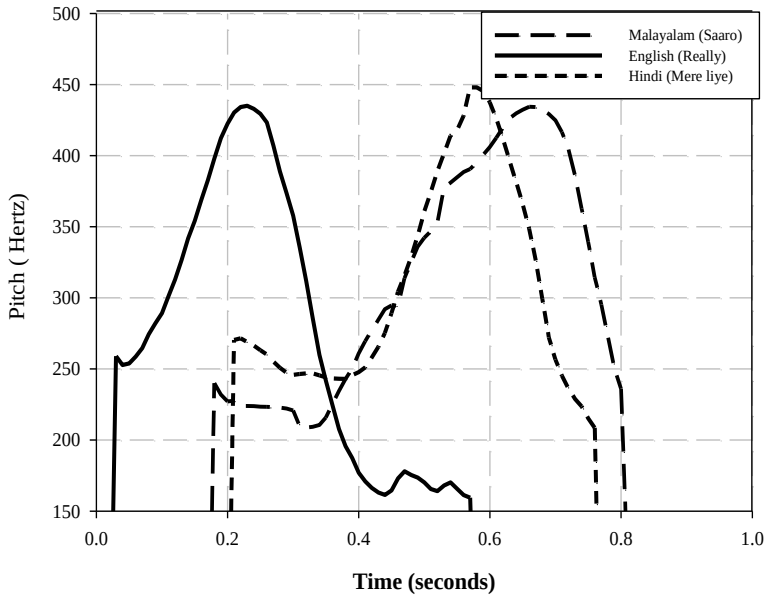


Figure 4.17: Pitch contours of surprise

The trends of prosody contours include discriminating information about the emotions. However very few efforts to describe the shape of these contours in a

systematic manner can be found in literature [17]. This study has examined contours at both the segmental and suprasegmental levels in English and also examined the feature variations among English Hindi and Malayalam.

Some of the most salient characteristic features identified from the suprasegmental utterances in various emotions, are presented in Table 4.34.

Table 4.34: Characteristic features of pitch contours in English, Hindi and Malayalam

Emotions	Features identified
Happy	Similar to pitch contour of surprise, but sharper decline in pitch contours of Hindi utterances.
Surprise	Identical maximum pitch values in English, Hindi, and Malayalam. Similar pitch contours in 3 languages; both a smooth rise to and decline from a peak.
Neutral	Least pitch values in Malayalam.
Anger	Language dependent pitch contours.
Sad	Pitch decreases towards the end of the utterance. Least values in English. Similar across the three languages.
Fear	Pitch breaks in Hindi, Malayalam. Language dependent.
Disgust	Similar pitch range for English, Malayalam.

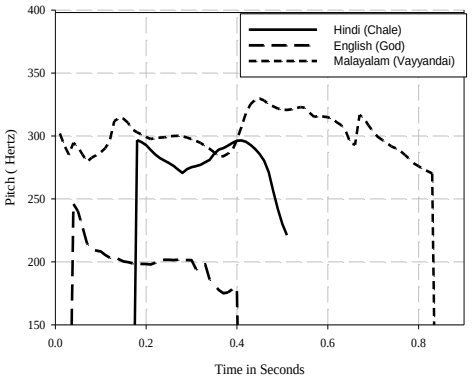


Figure 4.18: Pitch contours of sadness

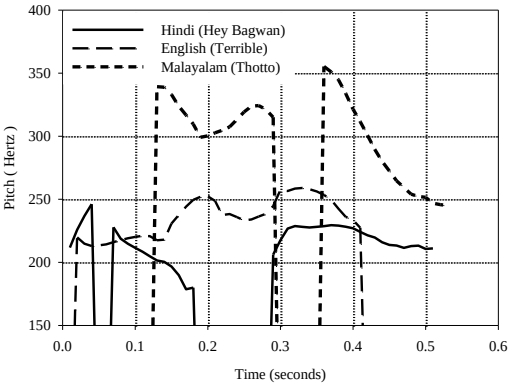


Figure 4.19: Pitch contours of Fear

Across the three languages pitch profiles are similar for happiness, surprise, sadness and disgust. Maximum language dependence or deviation in pitch profiles was noted for fear. Similarity in pitch profiles are noted across segmental utterances and suprasegmental utterance in English for happiness, surprise and sadness. Since the emotion recognition approach used in this

investigation is one of visual matching, it is simple and easy once the template or specimen contours are obtained.

4.7. Comparisons with the state of the art

Relevant, available results for SER based on prosodic features are presented to facilitate comparison of the obtained results with the state of the art. Classification rates of 57.1% for female speakers by a Bayes classifier and Gaussian Probability density functions, using ten features including energy of the DES database have been reported by Ververidis [77] in 2004, for a five emotion class SER experiment. The highest classification rate of 66 % obtained by GMM, for male samples, has been reported for large feature sets including other acoustic parameters, besides intensity [78].

Intensity based SER accuracies of 48.16% and 39.61% on the EMO-DB and the DES have been reported by Schuler et al. in 2005, for German and Danish respectively [79]. Classification accuracy of up to 64% has been obtained by Polzin and Waibel [130] for acted type anger, sad and neutral emotion based on verbal, non verbal and intensity based prosodic features. In the speaker independent mode, average recognition rate of 57.57% has been reported by Kim et al. in 2009, with gender and emotion classification [90]. In 2010, maximum SER rates of 62.6% and 60.9%, by Naïve Bayes and KNN classifiers respectively, were reported for frequency and energy features in female speech [64].

The SER rates obtained in this work for English for disgust, anger and sad are the highest reported result in any class-7 emotion recognition problem. In this respect, the present experimental results, especially those reported for elicited emotions in English and Malayalam solely based on intensity values are significant.

Even though no similar study of emotion recognition based solely on duration measures has been conducted especially in Hindi and Malayalam, certain relevant results from the available literature are presented. Whereas experimental results of Nwe et al [76] demonstrated an even spread of the utterance durations for the basic emotions in Burmese and Mandarin, results of this experimental study show significant differences in duration measures among certain of the seven emotion classes, though not between all classes. Anger has the smallest vowel duration as well as the highest speech rates in English, Hindi and Malayalam.

Schuller et al. 2005 estimated durations of pauses and syllable of the public corpus, EMODB and DES (consisting of anger, joy, sadness, and surprise plus neutrality in acted emotions). The recognition accuracies obtained based on duration was 27.46% for the DES and 19.08% for EMO-DB respectively [79].

The best overall intensity SER rates obtained in this work for the class-7 emotion recognition problem are much higher, with 57.86 % based on vowel duration; and 63.6% (ANN) 66.7% (ANN) and 71.4% (ANN) based on the speech rate accuracies for English, Hindi and Malayalam respectively. Universality in intensity based speech emotion was noted only for sadness which was the best recognized (at 100%) across English, Hindi and Malayalam.

Besides, the highest overall classification accuracies obtained for English, Hindi and Malayalam, in this seven class, female, speaker independent, SER problem, were 95.91% (at segmental level), 76.19% and 94.04% respectively, by the ANN classification of the complete prosodic feature set.

In the light of the earlier findings cited, the results obtained in this experimental study on the prosodic features for seven emotions in three languages can be considered significant.

4.8. Chapter Summary

This chapter has discussed the experimental results of the prosodic feature set based speech emotion recognition for English, Hindi and Malayalam. The SER performances for each feature (as well as their various combinations) have been evaluated with multiple classifiers. Results reaffirm the popularity of prosodic features for SER in the Indian context too. Experimental results in English prove the existence of prosody at the segmental level itself, despite the description of prosody as a suprasegmental phenomenon. In English, anger was very well classified based solely on the segmental intensity. Happiness, surprise, neutral, and disgust were well classified based on segmental duration. All emotions except fear, were recognized well based on the segmental pitch. The SER rates were found to be valence independent at the segmental level.

Among the three languages, intensity based emotion classification was found to be the most effective for English, as all emotions, except surprise and neutral were classified with 100% accuracy. It was the least effective for Hindi, since high classification accuracy was obtained for sadness only. For Malayalam 100%, recognition of neutral and sadness were obtained with the ANN classifier. The ANN classifier gave the best overall classification accuracies. The obtained SER rates mostly validated the statistical discriminations of emotions based on ANOVA of the various feature values.

In English, more emotions were recognized based on the syllable rate. Word rate and syllable rate based SER rates showed valence dependency. At the suprasegmental level speech rate based SER was the most effective for English and Hindi, than for Malayalam.

In English, irrespective of the utterance size, the pitch based SER rate for each emotion was above 80%. In Hindi, anger was the best classified

whereas in Malayalam, sadness and disgust were well recognized, based on pitch. Pitch based emotion classification was found to be language dependent; being the most effective for English. Characteristic features were identified from the pitch contours of segmental and suprasegmental utterances for applying rule based emotion classification.

Universality was observed only in the recognition of surprise, anger and disgust in all three languages. The SER rates obtained in English and Malayalam, based on the complete feature set are comparable with human perception ratings given in Section 3.4 of this thesis. But in the case of Hindi, the experimentally obtained SER rate is lesser than rated by humans. Therefore, this calls for investigations with other features such as jitter and shimmer which are variations in prosodic features. The following chapter therefore discusses the results of such investigations.

Chapter 5

SPEECH EMOTION RECOGNITION BASED ON JITTER AND SHIMMER

Contents

This chapter investigates speech emotion recognition in English, Hindi and Malayalam using micro perturbations in pitch, called jitter, as well as very small variations in intensity, called shimmer. Since it is difficult to bring about such minute variations in intensity and pitch, artificially, without actually experiencing the emotions, jitter and shimmer are proposed as features for speech emotion recognition. Therefore, more than certain other observable prosodic features, which can be acted, jitter and shimmer are expected to reflect true emotions only. The investigations are carried out separately, at the segmental level and suprasegmental level, based on jitter, shimmer and their combination (at the suprasegmental level). Performance comparisons demonstrate the effectiveness of jitter and shimmer in speech emotion recognition. Finally, universality in emotion recognition across the three languages is assessed for each emotion.

5.1. Introduction

Jitter has been extensively used in clinical diagnosis more than applications for speech recognition [131]. Jitter and shimmer have sparingly been used along with many other features for emotion detection [132], [133]. However, there are no reports available, for jitter and shimmer based SER for Indian English, Hindi and Malayalam as proposed in this investigation. This

chapter presents the results of the statistical analysis of jitter and shimmer values, followed by the results of SER based on the KMeans, KNN, NB and the ANN classifiers. This work intended to investigate the scope of improving the prosodic feature based SER discussed in Chapter 4, since certain emotions such as fear in Hindi and Malayalam, were not recognised well, based on certain individual, prosodic features. Besides, in Hindi, neutral, sadness and fear were not recognised based on the complete prosodic feature set.

5.2. Jitter based SER in English

This section presents the results of statistical analysis and classification of the jitter of both segmental as well as suprasegmental utterances in English. Similar analyses were done for Hindi and Malayalam

5.2.1. Jitter Analysis in English at the Segmental Level

The features were extracted using the Praat voice analysis software which recommends jitter measurements on sustained vowels. Therefore segmental level investigations in jitter based SER for English, were done with vowel utterances. Table 5.1 gives the summary statistics of vowel jitter

Table 5.1: Summary statistics of ANOVA of Jitter of Segmental English utterances

Statistical parameters of vowel jitter	**Decimal equivalent of the percentage local jitter for various emotions**						
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean Jitter	0.022	0.0236	0.0215	0.0269	0.0152	0.0296	0.0214
SD	0.0052	0.008	0.0076	0.0096	0.007	0.0135	0.0088
Maximum Jitter	0.031	0.039	0.0387	0.0437	0.035	0.0657	0.038
Minimum Jitter	0.01	0.0097	0.0127	0.01	0.0057	0.011	0.007
Emotions discriminated P < 0.001	neutral, anger, sad, fear, disgust	anger, sad, fear	happy, anger, sad, fear, disgust	happy, surprise, neutral, sad, disgust	happy, surprise, neutral, anger, fear	happy, surprise, neutral, sad, disgust	happy, neutral, anger, fear

The highest mean jitter as well as the maximum jitter were for fear. The least value of mean jitter was for sadness. The highest SD from the mean jitter was for fear, while the lowest SD was for happiness.

Since the investigations at the segmental level were based on five types of vowel utterances, the best choice of segmental utterances were those that maximized the difference in mean jitter across the various emotions. The utterance specific jitter values were as presented in Figure 5.1. The best choice of segmental utterances for SER based on mean jitter, can be made by examination of Figure 5.1, so as to ensure better classification accuracy. For instance, a class-3 SER problem for recognizing happiness, anger and fear, could result in higher SER rates, when based on the jitter values of 'i' or 'u', since the mean jitter for each of these utterances, are spaced wide apart for these three emotions. This approach of selecting favourable utterances for SER, is feasible in this work, which is concerned with the recognition of emotions of utterances of known semantic content.

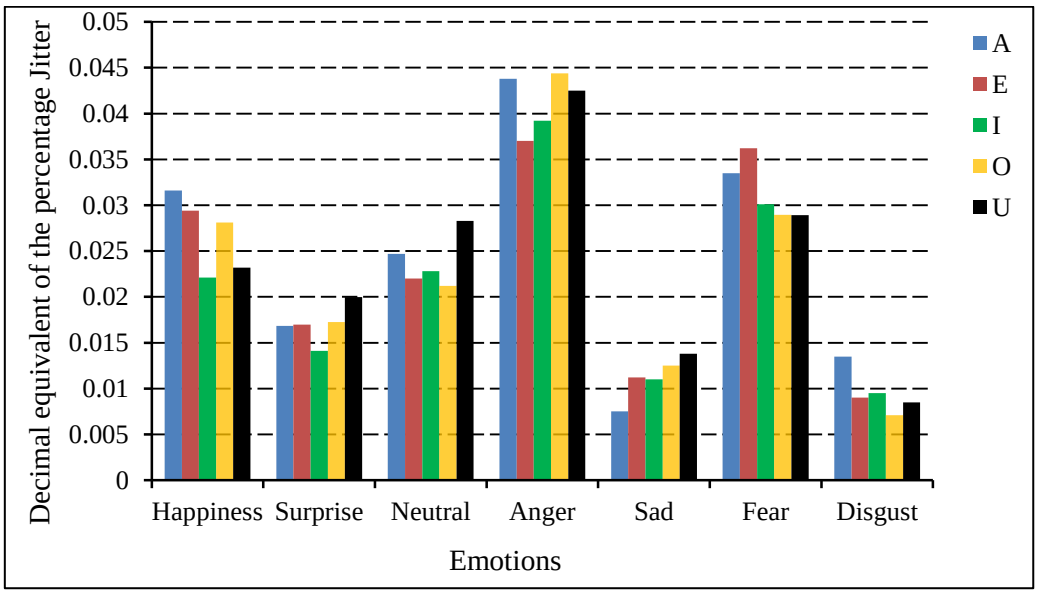


Figure 5.1: Vowel specific jitter values for different emotions

One way standard ANOVA was done on the jitter values and the emotion pairs with significant statistical difference in jitter values were found out as given in Table 5.1. Happiness, neutral, anger, sadness and fear were discriminated better than surprise and disgust. The vowel utterances for each emotion were chosen appropriately, to ensure maximum statistical discrimination. These vowel jitter values were given to each of the four classifiers - KMeans, KNN, NB and the ANN. Table 5.2 gives the consolidated classification rates, emotion wise, and classifier wise.

Table 5.2: Consolidated Segmental Jitter based SER rates.

SER rates in percentage for the various classifiers				
Emotions	**Kmeans**	**KNN**	**NB**	**ANN**
Happy	100	40	50	74
Surprise	89	25	37.5	66.7
Neutral	67	30	10	50
Anger	33	30.8	15.4	20
Sad	78	50	60	60
Fear	67	38.5	23.1	100
Disgust	67	12.5	10	50
Average	71.57	32.4	29.4	60.1

The highest average SER rate was for sadness, followed by happiness and fear. 100% recognition rates were obtained for happiness and fear. Disgust had the least average SER. The ANN and the KMeans classifier gave better classification accuracies than the KNN and NB classifiers. The SER rates for happiness, surprise, neutral, sad, fear and disgust were above the cut off (arbitrarily chosen as 60%). These results, especially of the average SER for each emotion, agree with those of ANOVA given in Table 5.2, except for anger. The obtained SER rates were found to be valence independent.

5.2.2. Jitter analysis in English at the Suprasegmental level

The average of the mean jitter of suprasegmental utterances in English, along with their standard deviations, are as plotted in Figure 5.2. The average jitter was the highest for anger, whereas it was the least for fear. The SD of the

mean jitter was the largest for sadness and the least for fear.

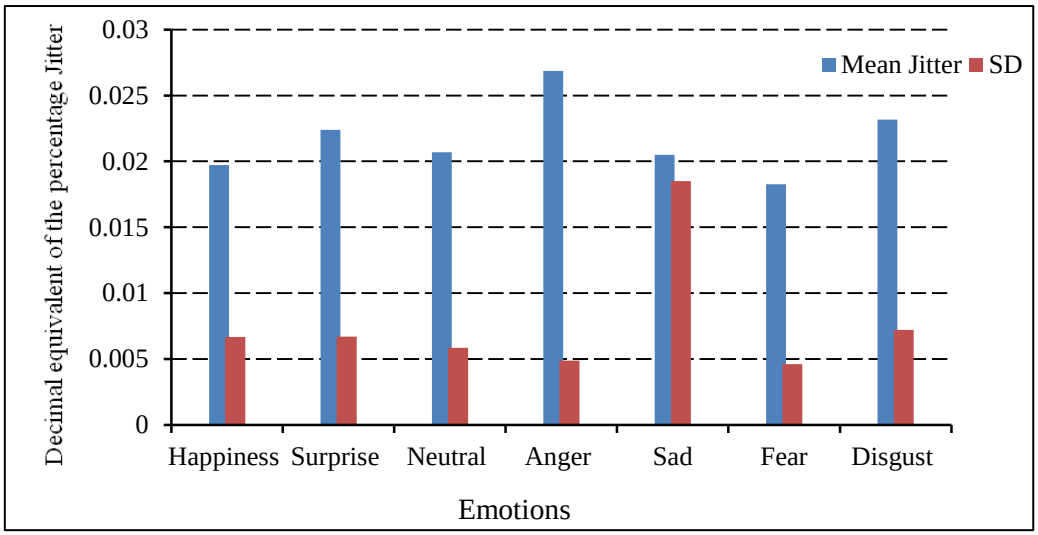


Figure 5.2 Mean and standard deviations of the jitter of Suprasegmental utterances

Table 5.3 summarizes the statistical discrimination of emotions based on the jitter of suprasegmental utterances in English.

Table 5.3: Jitter based statistical discrimination of emotions in English

Emotions	Emotions best discriminated by ANOVA of jitter with P < 0.05
Happy	Anger
Surprise	-
Neutral	Anger
Anger	Happy, neutral, sad, fear
Sad	Anger
Fear	Anger
Disgust	-

Anger was the best discriminated. Surprise and disgust were not at all discriminated from the rest of the emotions. Happiness, neutral, sadness and fear were discriminated at the same level. All the discriminated emotions were done so, with a low level of significance.

The jitter based classification accuracies obtained with the four classifiers are given in Table 5.4.

Table 5.4: Consolidated Suprasegmental Jitter based SER rates in English

SER rates in percentage for the various classifiers				
Emotions	**Kmeans**	**KNN**	**NB**	**ANN**
Happiness	19.04	20.0	60	66.7
Surprise	38.10	50.0	10	50
Neutral	14.8	20.0	60	50
Anger	42.80	36.40	80	66.7
Sad	14.8	20.00	40	50
Fear	33.30	27.30	36.4	66.7
Disgust	38.10	36.40	54.5	50
Average	28.71	30.01	48.7	57.15

Anger was the only emotion with high (80%) classification accuracy. Surprise, sadness and disgust were not recognized, at the SER cut off rate of 60%. The classification results obtained by various classifiers agreed with the results of the statistical discrimination results of ANOVA given in Table 5.3. The NB and ANN classifiers gave the best SER rates. The SER rates were found to be independent of the valence of the emotions.

5.2.3. Comparison of Jitter Based SER rates at the Segmental and Suprasegmental Levels in English.

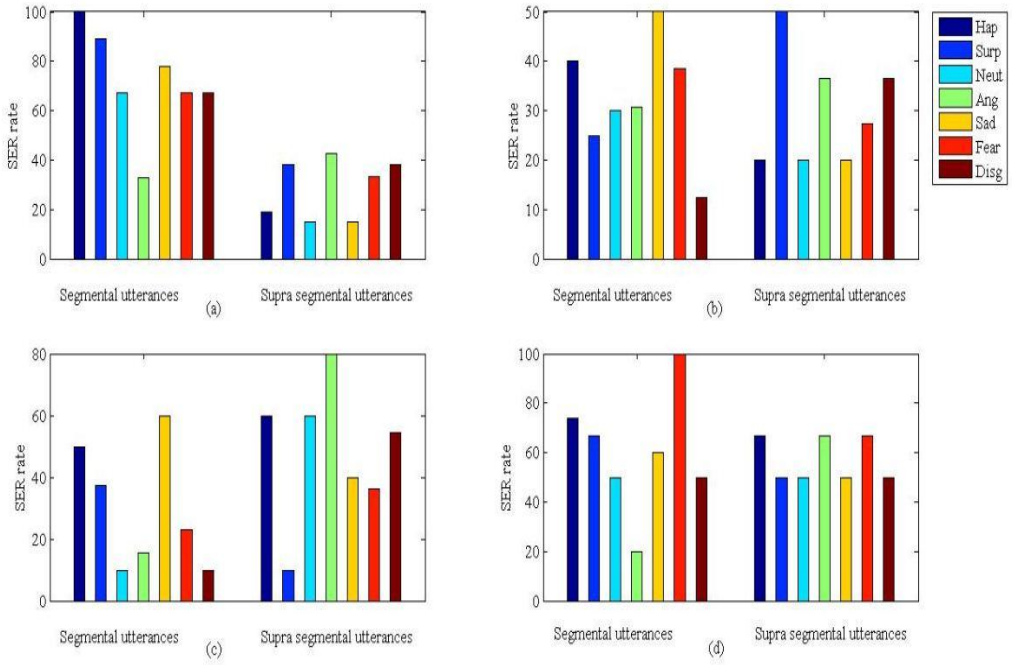
Figure 5.3 gives a comprehensive picture of the performance of each of the four classifiers, at the segmental and suprasegmental level in English, for the various emotions.

The following inferences could be drawn from Figure 5.3.and Tables 5.2 and 5.4. At the segmental level,

- All emotions except anger were better recognized, than at the suprasegmental level.
- Happiness, surprise and fear were the best recognized.
- Higher overall classification accuracies were obtained.
- The KMeans and the ANN classifier gave good results.

At the suprasegmental level,

- Anger was the best recognised in agreement with Anova results given in Table 5.3.
- The NB classifier gave the best results



(a) KMeans classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier

Figure 5.3: Comparison of jitter based SER rates at the Segmental and Suprasegmental levels in English for various emotions

Jitter based SER at the suprasegmental level was found to be complementary to that at the segmental. The emotions with the highest average jitter were among the best classified. At both the analysis levels, the classification accuracies were independent of the valence of emotions.

5.2.4. Jitter based Hindi SER

The jitter values of suprasegmental Hindi utterances were statistically analyzed for the seven emotions and the mean jitter along with the standard

deviation are as illustrated in Figure 5.4.

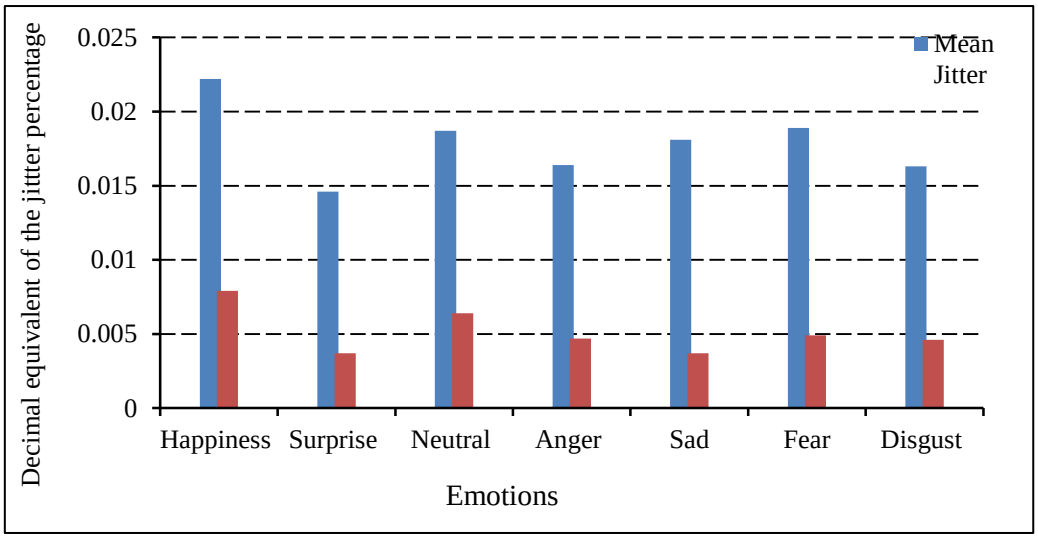


Figure 5.4: The mean jitter and standard deviations of jitter for Hindi

The highest mean jitter was for happiness whereas, the lowest mean jitter was for surprise. The mean values as well as the standard deviations were not valence dependent. The statistical discrimination for different emotion pairs are consolidated in Table 5.5. Happiness and surprise, which were the emotions with the highest and lowest mean jitter respectively, were the most discriminated. Sadness and neutral were less discriminated.

Table 5.5: Jitter based Statistical discrimination of emotions for Hindi utterances

Emotions	Emotions discriminated by repeated MANOVA with P < 0.05
Happy	Surprise, neutral, anger, sad, fear, disgust
Surprise	Happy, neutral, anger, sad, fear, disgust
Neutral	Happy, surprise, anger, disgust
Anger	Happy, surprise, neutral, sad, fear
Sad	Happy, surprise, anger, disgust
Fear	Happy, surprise, anger, sad, disgust
Disgust	Happy, surprise, neutral, sad, fear,

Table 5.6 presents the jitter based classification accuracies of suprasegmental utterances in Hindi.

Table 5.6: Consolidated Suprasegmental Jitter based SER rates in Hindi

SER rates in percentage for the various classifiers				
Emotions	**Kmeans**	**KNN**	**NB**	**ANN**
Happy	29.16	54.50	54.5	50
Surprise	41.70	41.70	66.7	50
Neutral	4.20	25.00	41.7	66.7
Anger	29.16	16.70	33.3	33.3
Sad	20.80	8.30	25	75
Fear	29.16	25.00	25	85.7
Disgust	25.00	16.70	50	80
Average	25.60	26.84	42.31	62.9

The highest average classification rates across the four classifiers was for surprise, which agreed with the ANOVA results given in Table 5.5. Fear and disgust were well classified. Happiness and anger could not be recognised with 60% cutoff in SER rate. The obtained SER rates were found to be valence independent. The ANN classifier gave the best SER rates.

5.2.5. Jitter Based Malayalam SER

The Figure 5.5 shows the mean and standard deviation of the jitter of suprasegmental utterances in Malayalam.

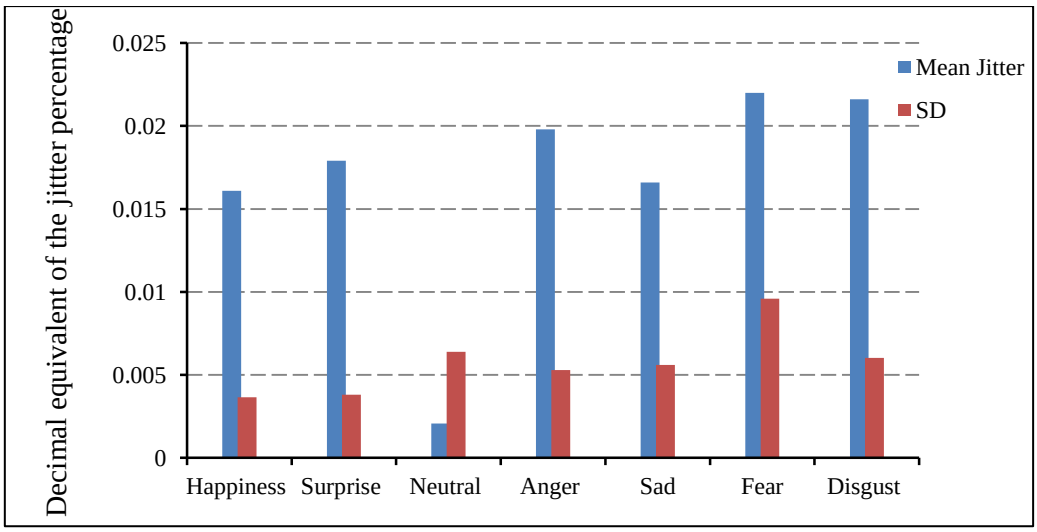


Figure 5.5: The mean jitter and standard deviations of jitter for Malayalam

The least jitter was for neutral, whereas the highest was for fear. The mean jitter values were independent of the valence of the emotions. The standard deviation was the highest for fear. Table 5.7 indicates the emotion pairs that could be discriminated statistically.

Table 5.7: Statistical discriminations for various emotions

Emotions	Emotions discriminated by ANOVA of the jitter of Malayalam utterances with P < 0.05
Happy	Neutral, anger, fear, disgust
Surprise	Neutral*, anger, fear, disgust
Neutral	Happy, surprise, sad
Anger	Happy, surprise, sad, fear, disgust
Sad	Neutral, anger, fear, disgust
Fear	Happy, surprise, sad, anger
Disgust	Happy, surprise, anger, sad,

* P < 0.01

For Malayalam utterances, anger was the best discriminated statistically. All emotions, except for the surprise - neutral emotion pair were discriminated at a low level (P < 0.05).

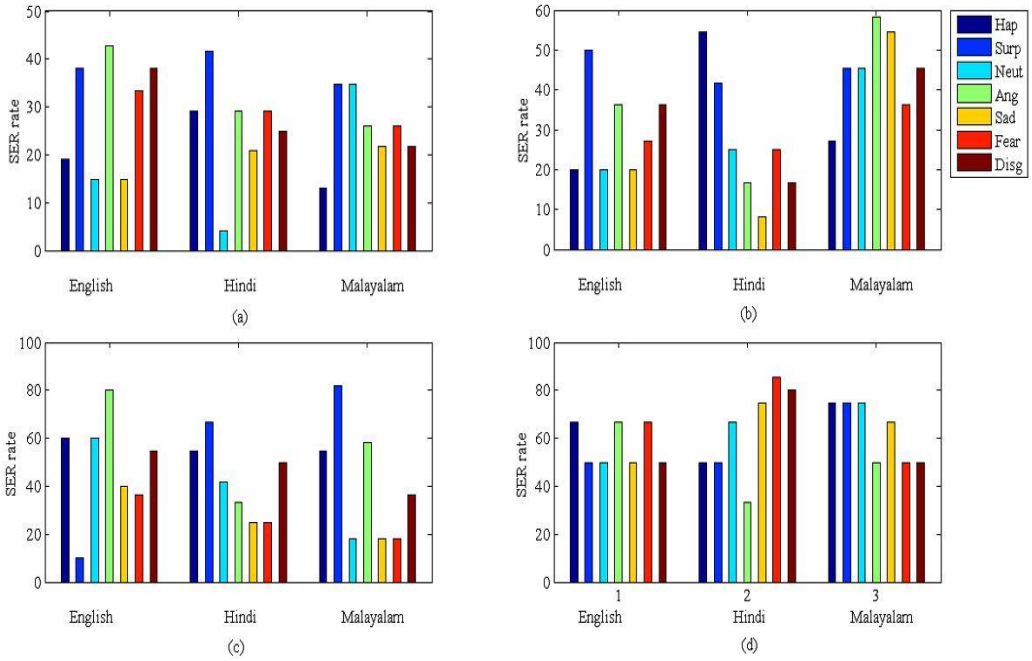
The speech emotion recognition accuracies are given in Table 5.8. Surprise was classified with accuracy above 80% and happiness too was well recognised at 75%. The maximum classification accuracy for anger, fear and disgust was less than 60%. Therefore the SER rates showed slight valence dependency since positive valence emotions were classified better than the negative valence emotions.

Table 5.8: Consolidated Suprasegmental Jitter based SER rates for Malayalam

Emotions	SER rates in percentage for various classifiers				
	K Means	KNN	NB	ANN	Average
Happy	13.04	27.30	54.5	75	42.46
Surprise	34.80	45.45	81.8	75	59.26
Neutral	34.80	45.45	18.2	75	43.36
Anger	26.08	58.30	58.3	50	48.17
Sad	21.70	54.50	18.2	66.7	40.28
Fear	26.08	36.40	18.2	50	32.67
Disgust	21.70	45.45	36.4	50	38.39
Average	25.47	44.69	40.8	63.1	43.52

5.2.6. Comparison of jitter based SER in English, Hindi and Malayalam

The consolidated recognition rates for the seven emotions across the three languages, at the suprasegmental level are presented in Figure 5.6.



(a) KMeans classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier

Figure 5.6: Comparison of SER rates based on Jitter of suprasegmental utterances in English, Hindi and Malayalam for various emotions

At the suprasegmental level in English, anger was the best recognized, whereas sadness and surprise were the least recognized. The highest recognition rates in Hindi were for fear and disgust, whereas the least were for anger and happiness. In Malayalam, the SER rates were high for surprise and low for fear and disgust. There was no universality in the recognition of emotions across English, Hindi and Malayalam, based solely on the jitter of suprasegmental utterances.

Jitter based speech emotion recognition was the most effective for English (at the segmental level), and the least effective for Malayalam. There

were no instances of 100% classification accuracy for any emotion in any language, based on the jitter of suprasegmental utterances. On the whole, the classification accuracies were independent of the valence of emotions except in the case of Malayalam where the positive valence emotions were recognised better than the negative valence emotions.

5.3. Shimmer based SER in English

Shimmer based investigations at the segmental level in English were followed by investigations at the suprasegmental level in all three languages.

5.3.1. Shimmer of Segmental Utterances

The segmental shimmer values in decibels were extracted using Praat and tabulated. Table 5.9 summarizes the results of the statistical analysis on vowel shimmer. The statistical discriminations among the emotions are also evaluated.

Table 5.9: Summary statistics of ANOVA of the Shimmer of Segmental English utterances

Statistical parameters of shimmer	Segmental Shimmer in dB for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean Shimmer	0.81295	0.86838	0.7209	0.9320	0.6421	0.90669	0.73742
Maximum Shimmer	1.33867	1.34633	1.0663	1.156	1.059	1.35567	1.14133
Minimum Shimmer	0.35867	0.44467	0.472	0.6393	0.358	0.59567	0.5133
Range	0.98	0.90166	0.5943	0.5167	0.701	0.76	0.628
Emotions discriminated with $P < 0.05$	all except fear	all except anger, fear	all	all except surprise, fear	all	all except happy, surprise, anger	All

The mean shimmer values were not valence dependent. The highest mean shimmer was for anger and the least was for sadness. Neutral had the smallest range for shimmer, and happiness had the highest range. Each of the emotions – neutral, sad and disgust were statistically discriminated from the rest

of the emotions. Table 5.10 gives the emotion wise and classifier wise vowel shimmer based SER rates.

Table 5.10: Consolidated segmental shimmer based SER rates in Engish

Segmental shimmer based SER rates in percentage for various classifiers				
Emotions	**KMeans**	**KNN**	**NB**	**ANN**
Happy	12.5	36.4	18.2	33.3
Surprise	33.3	25	66.7	66.7
Neutral	20.8	16.7	41.7	50
Anger	16.7	25	58.3	37.5
Sad	45.8	50	58.3	100
Fear	37.5	54.5	54.5	66.7
Disgust	41.7	66.7	41.7	50
Average	29.76	39.19	48.49	57.74

All emotions other than happiness, neutral and anger were recognized. The average recognition rate was the highest for sadness followed by disgust, in agreement with the results of ANOVA given in Table 5.9. The obtained SER rates were valence independent.

5.3.2. Shimmer based SER at the Suprasegmental level in English

The most relevant results of statistical analysis are presented in Table 5.11

Table 5.11: Summary statistics of ANOVA of Shimmer for Suprasegmental English utterances

Statistical parameters of shimmer	**Decimal equivalent of the percentage local shimmer**						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean shimmer	0.06566	0.08233	0.08095	0.08941	0.06764	0.06793	0.0863
SD of shimmer	0.019	0.0179	0.02647	0.04001	0.016	0.0137	0.03189
Emotions discriminated with $P < 0.05$	all except sad	all except neutral, fear	happy, sad	all except neutral, fear	all except happy	happy, sad	all except neutral, fear

Statistically, happiness and sadness could be discriminated from five other emotions. The highest value of the average shimmer was for anger, whereas the lowest average shimmer was for happiness. The largest standard deviation was for anger, and the least standard deviation was for fear. The obtained classification accuracies are given in Table 5.12.

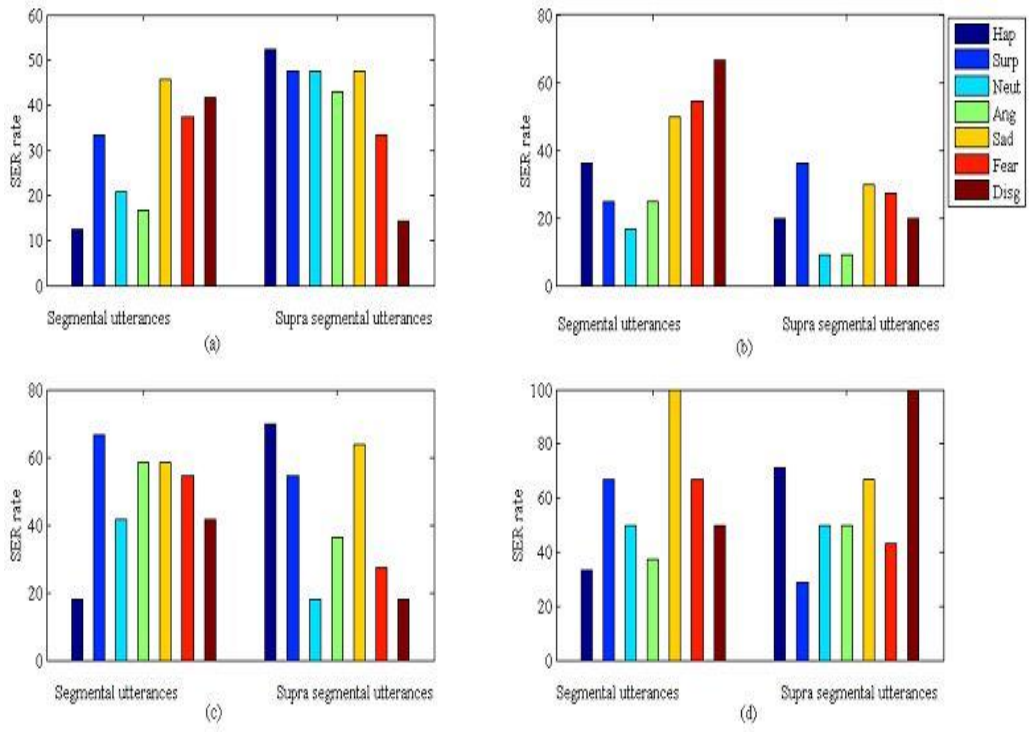
Table 5.12 Consolidated Shimmer based SER rates of Suprasegmental English utterances

Emotions	Shimmer based SER rates in percentage			
	K Means	KNN	NB	ANN
Happy	52.4	20	70	71.4
Surprise	47.6	36.36	54.5	28.6
Neutral	47.6	9.1	18.2	50
Anger	42.86	9.1	36.4	50
Sad	47.6	30	63.6	66.7
Fear	33.3	27.27	27.3	42.9
Disgust	14.29	20	18.2	100
Average	40.81	21.69	41.17	58.51

All instances of disgust were classified. Happiness and sadness were classified by both the NB and the ANN classifier in accordance with the results of ANOVA given in Table 5.1. The obtained SER rates were found to be valence independent.

5.3.3. Comparison of Shimmer based SER rates at the Segmental and Suprasegmental level

Figure 5.7 presents the emotion wise, classifier wise, shimmer based recognition accuracies for segmental and suprasegmental utterances. Further the inferences drawn from the results of shimmer based SER at both levels and presented in Table 5.8 and 5.10 are as follows.



(a) K Means classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier

Figure 5.7: Comparison of shimmer based SER rates at the segmental and suprasegmental levels in English.

At the 60% cut off in SER rate, neutral and anger failed to be recognised at both the segmental and suprasegmental levels based on the utterance shimmer. Sadness and disgust were classified well at both the levels. All instances of disgust were classified correctly at the suprasegmental level by the ANN classifier, which gave higher SER rates compared to the other classifiers. The best classification accuracies were obtained with the ANN classifier. The obtained SER rates were valence independent at both the segmental and suprasegmental levels.

5.3.4. Shimmer based SER in Hindi

The summary statistics of ANOVA of Hindi utterances are presented in Table 5.13 along with the statistical discriminations, for each emotion. Both

surprise and disgust could be discriminated from the rest of the emotions. These had respectively the lowest and second highest mean shimmer values. Sadness was the least discriminated.

Table 5.13: Summary statistics of ANOVA of the Shimmer for Hindi utterances.

Statistical parameters of shimmer	Decimal equivalent of the percentage local shimmer						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
Mean shimmer	0.0864	0.0662	0.0778	0.0775	0.0816	0.0965	0.0907
SD of shimmer	0.0276	0.0128	0.018	0.021	0.0167	0.0196	0.023
Emotions discriminated by repeated MANOVA with $P < 0.05$	surprise, neutral anger	all emotions	all except anger, sad	all except neutral, sad	happy, neutral, anger	all except happy	all emotions

The classification accuracies are given in Table 5.14. The highest average recognition rate was for surprise, by both the Naïve Bayes and the ANN classifiers. Classification rate above 80% was obtained only for surprise and disgust. Happiness and fear were poorly classified, in agreement with the statistical analysis given in Table 5.13.

Table 5.14: Consolidated shimmer based SER rates for Hindi

Emotions	SER rates in percentage for the various classifiers			
	K Means	KNN	NB	ANN
Happy	30.40	27.30	9	50
Surprise	21.70	75.00	83.3	83.3
Neutral	34.80	50.00	58.3	66.7
Anger	21.70	41.70	41.7	54
Sad	30.40	58.30	58.3	50
Fear	4.30	22.20	11.1	50
Disgust	21.70	25.00	33.3	83.3
Average	23.57	42.79	42.15	62.47

5.3.5. Shimmer based Malayalam SER

The summary statistics of average shimmer values for Malayalam utterances are as given in Table 5.15.

Table 5.15: Summary statistics of Shimmer of Malayalam utterances

Statistical parameters of shimmer	Decimal equivalent of the percentage local shimmer						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Mean	0.0788	0.0765	0.0919	0.0797	0.0735	0.0976	0.1009
SD	0.0119	0.0101	0.024	0.0111	0.0185	0.043	0.0254
Emotions discriminated with $P < 0.01$	surprise, anger, sad	happy, anger, sad	fear, disgust	happy, surprise, sad	happy, surprise, anger	neutral, disgust	neutral, fear

The highest shimmer was for disgust, whereas the least was for sad. Neutral, anger, fear and disgust were the better discriminated statistically. The classification accuracies given in Table 5.1 indicate comparatively high average recognition rates for anger and fear.

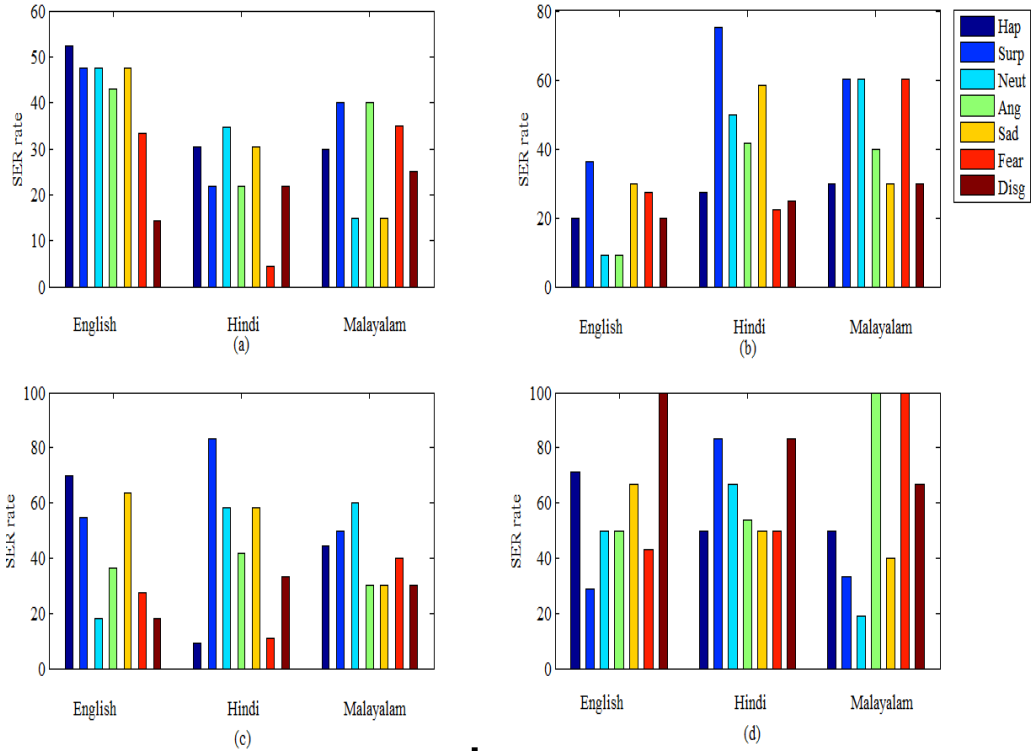
Table 5.16: Consolidated shimmer based SER rates for Malayalam

Emotions	SER rates in percentage for the various classifiers			
	K Means	**KNN**	**NB**	**ANN**
Happy	30	30	44.4	50
Surprise	40	60	50	33.3
Neutral	15	60	60	18.8
Anger	40	40	30	100
Sad	15	30	30	40
Fear	35	60	40	100
Disgust	25	30	30	66.7
Average	28.57	44.29	40.6	58.4

All instances of anger and fear were classified correctly by the ANN classifier. With the cutoff of 60%, surprise, neutral and disgust too, were well classified. All the classification results agreed with the results of statistical discrimination of emotions by ANOVA given in Table 5.15. Further, the emotion with the lowest shimmer was the least classified. The SER rates were found to be valence independent.

5.3.6. Comparison of Shimmer based SER rates in English, Hindi and Malayalam.

Figure 5.8 provides ready comparison of the shimmer based SER rates of the various emotions in English, Hindi and Malayalam, by the KMeans, KNN, NB and the ANN classifiers.



a) K Means classifier (b) KNN classifier (c) Naïve Bayes (d) ANN classifier

Figure 5.8: Comparison of Shimmer based SER at suprasegmental level for English, Hindi and Malayalam.

At 80% cut off, correct classification was obtained, only for disgust in English, surprise and disgust in Hindi, and anger and fear in Malayalam. In all cases, the classification results agree with those of ANOVA. None of the emotions were equally classified across the three languages. The ANN classifier gave the best classification results. Even at a reduced cut off of 60%, surprise, neutral, anger, and fear in English; happy, anger, sad and fear in Hindi; happy and sad in Malayalam were not classified correctly.

5.4. Jitter and Shimmer based English SER

The investigations were repeated for the combination of jitter and shimmer to explore their scope for improved classification accuracies in comparison with the performance based on individual features.

5.4.1. Jitter and Shimmer of English utterances

The ANN classification of the combined values of jitter and shimmer of suprasegmental English utterances, resulted in 100% classification accuracy of neutral, fear and disgust. The optimum network had 55 neurons in the hidden layer and overall accuracy of 64.7%. Results show fear and disgust to have several false hits. Anger failed to be classified on the basis of the combined features whereas; it was classified at 100%, based on jitter alone. The confusion matrix of the classification accuracies are given in Table 5.17

Table.5.17 Confusion matrix of the classification accuracies for English SER, based on jitter and shimmer

Classification accuracies in percentage for Jitter and Shimmer							
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	66.7	0	0	33.3	0	0	0
Surprise	0	20	20	0	20	20	20
Neutral	0	0	100	0	0	0	0
Anger	0	0	0	0	0	100	0
Sad	0	0	0	0	66.7	0	33.3
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	0	0	100

5.4.2. Jitter and Shimmer based SER of Hindi utterances

Based on both jitter and shimmer, 100% classification accuracies were obtained for happiness, sadness, fear and disgust, with an ANN having 45 neurons in the hidden layer. The confusion matrix is given in Table 5.18.

Table 5.18: Confusion matrix of the classification accuracies for Hindi SER, based on jitter and shimmer

Classification accuracies in percentage for Jitter and Shimmer							
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	100	0	0	0	0	0	0
Surprise	33.30	66.7	0	0	0	0	0
Neutral	0	0	50	50	0	0	0
Anger	0	0	0	66.7	33.3	0	0
Sad	0	0	0	0	100	0	0
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	0	0	100

The overall classification accuracy is 83.34%. All emotions other than neutral have been well recognized with combination of jitter and shimmer.

5.4.3. Jitter and Shimmer of Malayalam utterances

The confusion matrix of the classification accuracies based on the combination of jitter and shimmer of Malayalam utterances are presented in Table 5.19. The optimum network for classification of both jitter and shimmer had 45 neurons in the hidden layer. It gave an overall accuracy of 68.1%.

Table 5.19: Confusion matrix of the classification accuracies for Malayalam, SER based on jitter and shimmer

Classification accuracies in percentage of Jitter and Shimmer						
Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
100	0	0	0	0	0	0
20	10	0	60	0	0	10
0	0	66.7	0	0	0	33.3
0	50	0	50	0	0	0
0	0	0	25	50	0	25
0	0	0	0	0	100	0
0	0	0	0	0	0	100

All instances of happiness, fear and disgust were correctly recognized. Whereas surprise was well recognized on the basis jitter alone (as given in Table 5.8), the combination of jitter with shimmer led to very poor recognition of surprise. Sadness too had comparatively less recognition accuracy as shown in the confusion matrix of Table 5.19.

5.5. Performance Summary

The best classification rates obtained for each emotion and in each language with jitter, shimmer and their combination, at the suprasegmental level, across the KMeans, KNN, NB and the ANN classifier, are consolidated in Table 5.20.

Table 5.20: Performance Summary of SER with jitter or shimmer and their combination for English, Hindi and Malayalam

Best emotion classification rates based on jitter , shimmer for English, Hindi and Malayalam									
Emotions	**Jitter**			**Shimmer**			**Jitter and Shimmer**		
	***Eng**	***Hin**	***Mal**	**Eng**	**Hin**	**Mal**	**Eng**	**Hin**	**Mal**
Happy	66.7	54.5	75	71.4	50	50	66.7	100	100
Surprise	50	66.7	81.8	54.5	83.3	60	20	66.7	10
Neutral	60	66.7	75	50	66.7	60	100	50	66.7
Anger	80	33.3	58.3	50	54	100	10	66.7	50
Sad	50	75	66.7	66.7	58.3	40	66.7	100	50
Fear	66.7	85.7	50	42.9	50	100	100	100	100
Disgust	54.5	80	50	100	83.3	66.7	100	100	100
Average	61.13	65.99	65.26	62.21	63.66	68.1	64.77	83.34	66.67

*Eng- English;*Hin-Hindi; *Mal- Malayalam

The ANN classification of the combined jitter and shimmer values resulted in 100% classification accuracy, for fear and disgust in English, Hindi and Malayalam. This result was obtained with an ANN having 20 neurons in the hidden layer. This further validated the classification results, based on individual features of jitter or shimmer, which indicate the possibility of universality in SER for these three languages. The average SER rates showed the maximum improvement for Hindi. The SER rates were valence independent. In Hindi neutral was recognised at 66.7% whereas sadness and fear were recognised at 100%, thereby proving the efficiency of jitter and shimmer over the prosodic features in the recognition of these emotions.

5.6. Performance Comparisons

There are no available reports for the use of either jitter or shimmer, or their combination, as the sole feature for speech emotion recognition, especially in English, Hindi and Malayalam. The number of works in which these have been used in other languages, along with scores of other features, are also very few and as given in Table 5.21 for performance comparison.

Table 5.21: Comparison with other relevant works in speech emotion recognition

Reference	Jang and Kwon [132]	Hendy and Farag [133]	Adopted approach
Language	Korean	German	English / Hindi / Malayalam
Gender	Male and Female	Male and Female	Female
Emotions	Happy, surprise, neutral, anger, sad fear, bored	Happy, neutral, anger, sad fear, bored, disgust	Happy, surprise, neutral, anger, sad, fear, disgust
Acted / elicited	Elicited	simulated	Elicited
Features	Jitter, shimmer, log energy, pitch, formants	More than 100 other features along with perturbation factor and perturbation quotient of jitter, shimmer.	Jitter, Shimmer
Classifiers	SVM	Probablistic Neural network	KMeans, KNN, NB and ANN
Overall Recognition accuracy	58.6%	49% ; with raw data	English - 64.77 % ; Hindi - 83.3% ; Malayalam - 68.1%, with raw data

Comparison of the obtained results with the cited results indicate jitter and shimmer based SER to be the most effective for Hindi.

5.7. Chapter Summary

The effectiveness of jitter and shimmer for SER in English, Hindi and Malayalam was investigated and evaluated, using statistical analysis and multiple classifiers. Additionally, in English, investigations on jitter and / or shimmer based SER were carried out at the segmental level. The obtained

classification rates demonstrated that jitter and shimmer of suprasegmental utterances, are successful in classifying all instances of fear and disgust in all the three languages. Due to the unprecedented nature of these investigations, specifically in Hindi and Malayalam there are no publicly available results for direct performance comparison. At the 60% cut off in SER rate, all emotions could be recognized based on either jitter or shimmer, or their combination (at the segmental or suprasegmental levels). High SER rates were obtained for cases where the data belonging to the different emotion classes were verified to be statistically different, with high level of significance. These experiments on female speech database have demonstrated that jitter and shimmer give excellent results in speech emotion recognition, as complementary features of the prosodic parameters investigated in Chapter 4.

Chapter 6

FORMANT AND BANDWIDTH BASED CLASSIFICATION OF UTTERANCES AND EMOTIONS

Contents

SER methods generally use a large number of features and considerable signal processing effort. This chapter investigates the use of a minimum number of formants and bandwidths in an efficient, yet simple approach to classify neutral and six basic emotions in English, Hindi and Malayalam. For each language, the best vocal tract features among formants and bandwidths are identified by the KMeans, KNN and NB classification of individual features, followed by the ANN classification of the optimum features. In English, the performance of formant based SER at the segmental level are compared with those at the suprasegmental level. The effect of reduction in the number of emotion classes, on the emotion classification accuracy are studied, for each language. The manifestation of universality in the vocal expressions of emotions across the three languages, are also investigated. This chapter further reports the results of the classification of segmental utterances comprising vowels, on the basis of first four formants. Quantitative information regarding such formant based Speech Recognition (SR) at the segmental level, and the identification of emotions suitable for segmental SR have not been reported so far. Statistical analysis of formants and bandwidths was carried out, as a part of these investigations. Further, SER for Malayalam has been modeled using Decision trees and logistic regression.

6.1. Introduction

Investigations in SER based on prosodic features and their variations, namely jitter and shimmer (as detailed in Chapters 4 and 5 respectively), motivated this work on spectral features, so as to finally compare the performances of time domain and spectral features, for simple and efficient SER.

6.1.1. Choice of Formants and Bandwidths as Spectral Features for SER

Formants are the resonant frequencies of the vocal tract and are called so by speech scientists since these resonances tend to "form" the overall spectrum of speech. Each formant is characterized by its center frequency and its bandwidth. It has been reported that speakers during stress or under depression do not articulate voiced sounds with the same effort as in the neutral emotional state [134]. A strong dependency of these spectral characteristics (formants) on phonemes, and therefore on the phonetic content of an utterance has also been verified [135]. Vowel phonemes are distinguished primarily by the location of the first three formant frequencies [16]. Therefore, formants are sensitive to both emotions and utterances. These observations set the directions for this investigation in emotional speech based on the first four formants and their respective bandwidths.

6.1.2. Formant and Bandwidth based SER

The objective of this investigation was to identify and use the optimum vocal tract features among formants and their bandwidths to achieve efficient speech emotion recognition, with minimum algorithmic complexity. Investigations in formant and bandwidth based emotion classification at the segmental level were followed by similar investigations at the suprasegmental

level in English, Hindi and Malayalam. SER with formants, especially at the segmental level is challenging due to the sensitivity of formant values not only to the type of utterance, but also to the type of emotion. To reduce such cross sensitivity of formant values to utterances, detailed investigations were done with the formant values of vowel utterances in different emotions so as to identify those emotions and utterances that would give the best SER rates. These investigations are doubly significant since it also helped to identify formants and emotions most suitable for segmental SR, apart from the principal aim of segmental level SER with optimum spectral features.

6.2. Formant based SR at the Segmental level

Investigations were carried out at the segmental level to verify whether the five vowel / diphthong utterances a, e, i, o, u in the Indian English accent, could be discriminated solely based on the location of their first four formants. The effect of emotional content of the utterances on formant based classification / recognition of these vowel sounds was also studied. Statistical analysis as well the NB and KNN classification of each of the four formants, were used to identify both emotions as well as formants with the best vowel discrimination ability. For proper utterance discrimination based on formant values, first of all, the variations in the formant values due to differences in emotions had to be minimized, for any specific class of formant and utterance. The first four formant values of each English vowel sound were extracted and tabulated for each of the seven different emotions. The distributions of the first four formants of the five vowels under surprise are as plotted in Figure 6.1.

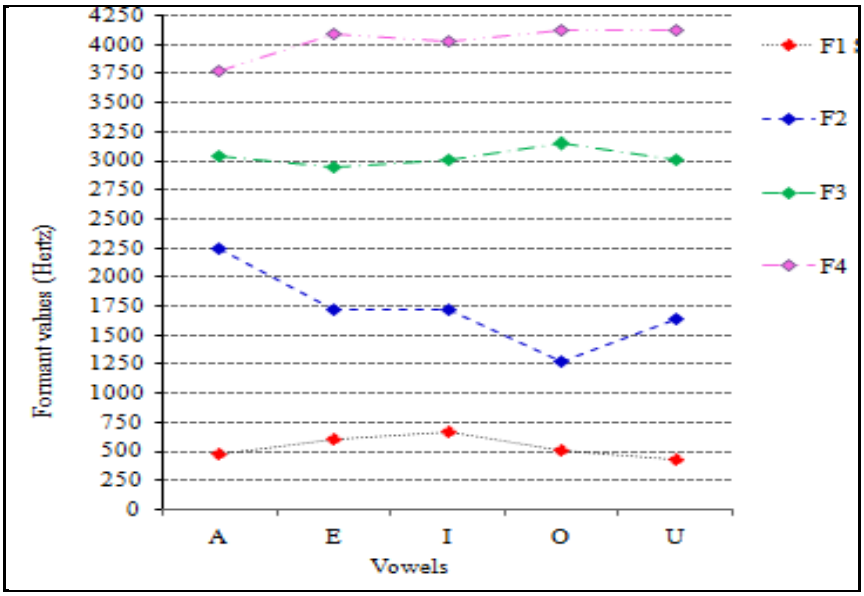


Figure 6.1: Location of the average values of the first four formants of segmental utterances under surprise

This figure helps to make the right choice of formants so as to obtain the maximum discrimination among the various segmental utterances. For instance, Figure 6.1 illustrates that for surprise, 'a' is the best discriminated from the other vowel sounds on the basis of the second formant values. Similarly, the vowel utterances 'e' and 'i' are best discriminated on the basis of F1 values. So also 'i' and 'o' were best discriminated on the basis of F2.

6.2.1. Statistical Analysis of Vowel Formants

Statistical analysis by ANOVA was used to identify significant differences in formant values among these five different vowel utterances, under each emotion. Consolidated results of the emotion specific, statistical discrimination of vowels on the basis of the first two formants are given in Table 6.1. The P values are indicated by appropriate asterisk ratings as illustrated in Chapter 3. The investigation aimed to achieve utterance discrimination with

minimum number of formants. Hence the individual contributions of F1and F2 are distinctly represented in two different colours for easy understanding.

Table 6.1: Emotion specific ANOVA based discrimination of vowels for F1, F2

Utterance pairs	Significance levels for utterance discrimination under various emotions						
	Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
a-e	***	***	***	***	***	*	*
a-i	***	***	***	***	***	***	***
a-o	***	***	***	***	***	***	***
a-u	***	***	***	***	***	*	***
e-i	**ns**	***	***	*	*	*	*
e-o	***	***	***	***	**ns**	*	**
e-u	***	***	***	***	***	***	***
i-o	***	***	***	***	***	**	***
i-u	***	***	***	***	***	***	***
o-u	***	***	***	***	***	**	***

Contributions of formants: *- F1; * - F2;
ns - no significant difference in formant values based on F1, F2, F3, F4

Statistically, utterances were the best discriminated for surprise, neutral and anger. The utterance pairs ‘e - i’ and ‘e - o’ under happiness and sadness respectively, could not be statistically discriminated on the basis of F3 or F4 also. The asterisk ratings of P values in the Table 6.1 indicate that the segmental, vowel utterances were the least discriminated under fear.

6.2.2. Classification of the Vowel Formants

Classification of segmental utterances on the basis of the various vowel formants was done separately by the NB and the KNN classifiers. The vowel sounds with the highest classification rates were considered as the best recognized. The NB classification results are given in Tables 6.2 and 6.3. The cut off rate for accurate recognition was arbitrarily fixed at 60% (as for SER) and classification accuracies equal to or above cut off were considered to be correctly recognized. Out of the thirty five cases (corresponding to five utterances in seven emotions), twenty two cases were recognized on the basis of F1, and twenty three cases on the basis of F2.

Table 6.2: NB classification of vowel formants based on F1, F2

Emotions	Vowel formant classification accuracies in percentage for various emotions									
	F1					F2				
	A	e	i	o	u	A	e	i	o	U
Happy	0	0	100	60	40	80	80	60	100	100
Surprise	100	20	80	40	60	100	60	40	100	100
Neutral	80	80	80	0	80	60	0	20	40	40
Anger	60	40	100	20	100	80	0	40	100	80
Sad	100	40	80	60	100	60	40	20	80	60
Fear	40	75	60	40	80	80	0	60	60	40
Disgust	40	80	80	80	40	80	0	80	100	60
Average	60	47.9	82.9	42.9	71.4	77	25.7	45.7	82.9	68.6

Thus the individual contributions of F1 and F2, to utterance discrimination were 62.9% and 65.7% respectively. However, the additional contribution of F2 to the overall recognition of segmental utterances was only 28.57%. The contributions of the third and fourth formants to utterance classification can be inferred from Table 6.3.

Table 6.3: NB classification of vowel formants based on F3, F4

Vowel formant classification accuracies in percentage for various emotions										
Formants→	F3					F4				
Vowels → / Emotions	a	e	i	o	u	a	e	i	o	U
Happy	60	25	60	80	20	0	0	20	60	20
Surprise	60	20	40	0	20	100	40	80	80	20
Neutral	80	20	0	20	80	0	20	0	80	60
Anger	60	20	20	20	40	60	80	0	60	20
Sad	40	60	40	20	0	60	0	0	80	80
Fear	80	0	20	20	40	20	20	40	80	0
Disgust	20	40	0	40	100	0	20	0	60	20

The individual contributions of F3 and F4, to utterance discrimination were 28.57 % and 40% respectively. The additional contribution of F3 and F4 to the overall recognition of segmental utterances was 2.9% and 5.7% respectively. Table 6.4 presents a consolidated summary of the contribution of individual formants to the classification of segmental utterances.

Table 6.4: Consolidated vowel recognition rates by the NB classifier based on F1, F2, F3 and F4

Emotions	Percentage recognition of vowels with the NB classifier				
	a	e	i	o	U
Happy	80	80	100	60	100
Surprise	100	60	80	100	60
Neutral	80	80	80	80	80
Anger	60	80	100	100	100
Sad	100	60	80	60	100
Fear	80	75	60	60	80
Disgust	80	80	80	80	60
Average	82.85	73.57	82.86	77.14	88.57

Contributions of formants by colour: red - F1; violet - F2; green - F3; blue - F4

Based on Table 6.4, it can be concluded that,

- At the 60% cut off in SR rate, the correct classification of all the five vowels could not be achieved by any single formant.
- The best recognized vowel was "u".
- Emotions most favorable for utterance discrimination were anger followed by happy and disgust. The utterances were the least recognized for fear.
- F1 supplemented by F2 was sufficient to discriminate all vowel utterances, with a lower SR cut off of 40%.

Classifications of vowel formants were repeated using the KNN classifier, in order to validate the results obtained by the NB classifier, for utterance discrimination. The consolidated KNN classification rates are given in Table 6.5.

Table 6.5: Consolidated vowel recognition rates by the KNN classifier based on F1, F2, F3 and F4

Emotions	Percentage recognition of vowels with the KNN classifier				
	a	e	i	o	U
Happy	60	80	100	100	60
Surprise	60	60	100	60	60
Neutral	80	80	100	60	80
Anger	60	100	60	40*	100
Sad	80	60	60	80	100
Fear	80	60	60	60	100
Disgust	60	60	60	80	100
Average	68.57	71.43	77.14	73.33	85.71

Contributions of formants by colour: - red- F1; violet - F2; green - F3; blue –F4

* Angry utterances of "o" could be recognized at a maximum of 40% only.

Based on the vowel formant classification rates given in Table 6.5, it was concluded that,

- At the 60% cut off in classification rate, only 68.57% of the total utterances were correctly classified with F1.
- F2 contributed to the recognition of another 11.4%.
- The remaining utterances were classified by equal contributions from F3 and F4, of 8.57% each.
- The best recognised utterance was "u".

The emotions most favourable for utterance discrimination were happy, neutral and anger.

6.2.3. Consolidated Summary of Segmental Level SR

This section summarizes the findings of the investigations on utterance discrimination at the segmental level using vowel formants. Table 6.6 presents a consolidated list of the most favorable emotions for classifying various vowel utterances by the Naïve Bayes and KNN classifiers.

Table 6.6: Emotions most favourable for vowel classification by each formant class

Formants	Vowels	Emotions giving the best utterance recognition
F1	a	Surprise (N*), sad, neutral (K*,N)
	e	Neutral (K, N), anger (K), disgust (N),
	i	Anger, sad, disgust (N), happy, surprise, neutral (K, N)
	o	Disgust (K, N)
	u	Neutral (N), anger, sad, fear (K, N), disgust (K)
F2	a	Surprise, happy (K, N), anger, fear, disgust (N)
	e	Happy (K, N), neutral (K)
	i	Happy, neutral (K), disgust (N)
	o	Happy, surprise (K, N), neutral (K)
	u	Surprise, Neutral (K), happy, anger, sad, disgust (N)
F3	a	Neutral (K), fear (K, N)
	o	Happy (K), anger (N)
	u	Neutral (K, N), disgust (N)
F4	a	Surprise (K, N), neutral (K)
	e	Anger (N)
	i	Surprise (K, N), fear (K)
	o	Sad (K, N), surprise, neutral, fear (N)
	u	Fear (K), sad (N)

N*- Naive Bayes classifier; K*- KNN classifier

The information provided in the Table 6.6 was used in the segmental SER based on vowel formants, since the cross sensitivity of formant values to utterance differences, could be reduced by avoiding the identified utterance - emotion combinations (presented in Table 6.6). Results of the statistical analyses and classification of vowel formants showed that the SR at the segmental level was emotion dependent. Both the NB and the KNN classifier failed to classify certain vowel utterances, under certain emotions, based solely on F3 or F4. Irrespective of classifiers, "u" was the best classified utterance, whereas "e" and "a" were less accurately classified. Happiness and anger were the most favorable emotions for the segmental SR of stand-alone vowels. Fear was the least favorable for the recognition of vowels. F3 and F4 were necessary for vowel discrimination, even though their contributions were less compared to that of F1 and F2. F1 gave poor overall recognition rate for the mid-back, tense rounded vowel "o", as well as for "e". Based on F2, the diphthong "i" as well as "e" were poorly recognised. F3 gave poor recognition rates for "e". F4 gave better recognition rate for "o" than for "e". Thus this investigation in formant based utterance discrimination, helped to identify utterances most favorable for emotion recognition based on vowel formants. The identified formant vowel combinations could therefore be used in segmental level SER. The results of this investigation on vowel discrimination have increased significance owing to the stand alone nature of vowels / diphthongs in contrast with other phones that are uttered in a sequence.

6.3. Formant based SER at the Segmental Level

This section discusses SER at the segmental level based on the four formants - F1, F2, F3, F4, of five English vowels namely, a, e, i, o, u. The selection of vowel formants to form the optimum, minimal feature set for SER, was made based on the results of classification by the KMeans, NB and KNN classifiers, for individual formant classes. Formants with the maximum SER

rates for the various emotions were chosen as optimum, and the rest were considered irrelevant and avoided in the final ANN classification.

Similar investigations were done on vowel bandwidths. But the very poor SER rates obtained, led to the conclusion that vowel bandwidths are not effective for segmental level SER in English.

6.3.1. Statistical analysis at the segmental level

The average values of the first four formants for the seven emotions were as given in Table 6.7.

Table 6.7: Mean values of the various vowel formants for the seven emotions

Formants	Mean formant values in Hertz for vowels for various emotions						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
F1	575.78	582.64	542.34	603.86	533.14	585.07	598.8
F2	1756.2	1776.99	1800.15	1844.18	1877.77	1789.39	1857.64
F3	2970.75	3045.39	3006.75	2994.56	3070.88	3048.26	3089.74
F4	4054.97	4069.73	4098.89	4095.19	4189.29	4095.10	4175.43

The least formant values were for sadness for F1; happiness for F2, F3 and F4. The highest formant values were for anger for F1; sadness for F2 and F4, and disgust for F3. The formant values were independent of the valence of the emotions.

The summary of the statistical discrimination of emotions based on segmental formants is presented in Table 6.8. The results of ANOVA indicated the statistical discriminations (at three levels of significance) of the formant values between the various pairs of emotion classes. Within any specific emotion, the formant values varied with the specific vowel utterances as discussed in the previous sections of this chapter. Statistical analyses were therefore carried out to identify significant differences in formant values for the five different vowel utterances within the same emotion class. Finally as much as possible, those formant values which reflected differences in emotions (and not differences in utterances) were used for emotion classification.

Table 6.8: Summary statistics of ANOVA of formants of Segmental English utterances

Statistical discrimination of vowel formants of different emotions with P < 0.01		
Formants	**Best discriminated**	**Least discriminated**
F1	Anger, sad, disgust	Neutral
F2	*Happy, neutral, fear	Surprise
F3	Sadness, disgust	All other emotions
F4	Surprise, happy	Rest of the emotions

* P < 0.001

Following were the main inferences of the statistical analysis for the various formant classes:

F1 - Anger, sadness and disgust were better discriminated than the rest of the emotions considered.

F2 - Happiness could be distinguished from surprise, even though at a low level of significance. Statistically, there was no significant difference within the emotion pairs - anger, disgust and surprise, sadness. Neutral and fear were the best discriminated.

F3 - Sadness was the best discriminated from all other emotions. Surprise was better discriminated from the rest of the emotions.

F4 - Compared to the first three formants, F4 showed poor statistical discrimination among the seven emotion classes. Summarizing the results of ANOVA, all the seven emotions were discriminated well across the four formants, though at different levels of significance.

6.3.2. The Optimum Feature set for Segmental SER

Separate, formant wise (F1 to F4) classifications of emotions by three base classifiers namely Kmeans, NB and KNN were done in order to identify the formants that gave the best results for SER. Table 6.9 gives the consolidated results of classification by the KMeans, NB and the KNN classifiers. With the single formant, baseline recognition rate fixed arbitrarily at 20%, each of the four formants recognized almost the same number of cases; though for different

emotions and with different classifiers. Therefore it was inferred that the contribution of each of these four formants was significant, and so all the four formants were included in the optimum feature set for the final ANN classification at the segmental level.

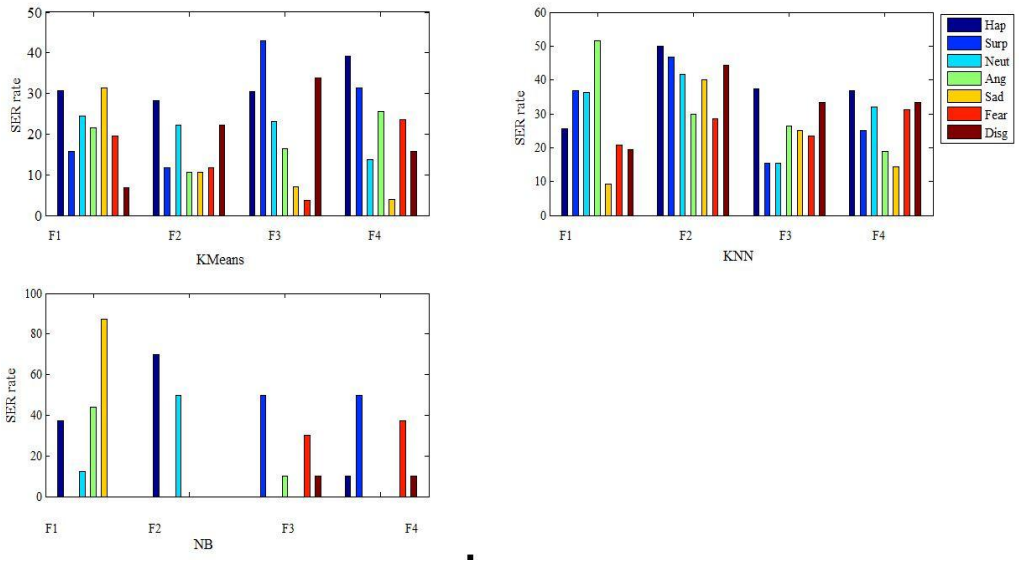
Table 6.9: Consolidated Vowel formant based SER rates by the KMeans, KNN and NB (base) classifiers

Vowel Formants	Classifier	Formant and classifier wise SER rates in percentage for various emotions						
		Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
	K Means	30.7	15.8	24.5	21.6	31.3	19.5	6.9
F1	NB	37.5	0	12.5	43.8	87.5	0	0
	KNN	25.7	36.7	36.4	51.5	9.3	20.7	19.4
	K Means	28.2	11.7	22.3	10.6	10.6	11.7	22.3
F2	NB	70	0	50	0	0	0	0
	KNN	50	46.7	41.7	30	40	28.6	44.4
	K Means	30.4	42.9	23.2	16.5	7.1	3.6	33.9
F3	NB	0	50	0	10	0	30	10
	KNN	37.5	15.4	15.4	26.3	25	23.5	33.3
	K Means	39.2	31.4	13.7	25.5	3.9	23.5	15.7
F4	NB	10	50	0	0	0	37.5	10
	KNN	36.8	25	32	19	14.3	31.3	33.3

The formant wise, performance of each of the three classifiers for each emotion, was as illustrated in Figure 6.2.

Based on the Figure 6.2 and information given in Tables 6.8 and 6.9, it was concluded that,

- Among the three classifiers, the KNN classifier gave the best emotion recognition based on individual vowel formant values and could recognize all the emotions, even though at different rates. This was followed by KMeans classifier, which could also recognize all emotions. The NB classifier failed to recognize all emotions based on any specific formant class.



a) K Means classifier (b) KNN classifier (c) Naive Bayes classifier

Figure 6.2: Comparison of SER rates of the three base classifiers

Further from the formant wise, SER performances it was inferred that,

- On the basis of F1, anger and sadness could be classified, in agreement with the results of ANOVA given in Table 6.8. F1 was not recommended for the detection of surprise, fear and disgust as the average SER rate for these emotions were below the cutoff rate.
- On the basis of F2, neutral and disgust were the best classified, followed by happiness. The rest of the emotions were recognized with much lower accuracy. Results of ANOVA too indicated good statistical discrimination for happiness and neutral.
- Based on F3, the highest emotion classification accuracy by the KMeans method was for surprise, followed by disgust.
- Based on F4, surprise and fear had good SER rates. Results of ANOVA had already indicated the ability of F4 to discriminate surprise from the rest of the emotions.

Summarizing the overall performance of these three classifiers based on the single formant values, it was found that neutral and the positive valence emotions namely, happiness and surprise were better classified based on vowel formants. Each formant class contributed to the recognition of one or more emotions. Therefore all the formants were included in the optimal feature set for formant based SER.

6.3.3. ANN Classification for Segmental SER

The classification results of the three base classifiers had indicated that all four formants contributed significantly to SER. This was further validated by ANN classification of the individual features as well as their various combinations A two-layer feed-forward network, with sigmoid input and output neurons and 55 neurons in its hidden layer was used for the optimum feature set based SER. Table 6.10 presents the confusion matrix of recognition accuracies for the various emotions.

Table 6.10: Confusion matrix of ANN classification accuracies based on the first four formants

Emotions	Confusion matrix of classification accuracies in percentage						
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	100	0	0	0	0	0	0
Surprise	0	100	0	0	0	0	0
Neutral	0	0	100	0	0	0	0
Anger	3.4	0	0	96.6	0	0	0
Sad	0	0	3.6	0	96.4	0	0
Fear	0	0	0	3.8	11.5	84.6	0
Disgust	0	4.2	0	4.2	0	0	91.7

Based on the ANN classification results, it was found that the overall emotion classification accuracy was 95.6 %. All instances of both the positive valence emotions and the neutral could be recognised on the basis of the vowel formants. Anger and sadness were recognised at the same level. Fear was the least recognized. The results of the final ANN classification of the first four formant values of the vowels, validates the consolidated findings of ANOVA

(given in Table 6.8), as well as the results of classifications by the other three classifiers (given in Table 6.9) based on the individual formant values.

6.3.4. Conclusion of Segmental SER

An efficient system for the detection of seven emotions in English, at the segmental level itself, using only formants of stand-alone vowels, has been investigated. The classification results by the KMeans, NB and KNN classifiers for various emotions, based on the individual formant values agreed with the results of the statistical analysis. The specific SER rates obtained by this approach varied with the type of emotion, order of the formant, and the classifier used. Very high SER rates were obtained with the final ANN classifier especially for happiness, surprise and neutral (100% classification accuracy). This approach has advantages of obvious saving in time and effort, and can therefore be adapted to implement a truly real time SER system. The next section presents spectral feature based SER for suprasegmental utterances.

6.4. Formant and Bandwidth based SER for English

In certain situations where it is difficult to obtain the speech samples of isolated, stand-alone vowels, (without co-articulation effects), SER in English would invariably be based on suprasegmental utterances. This section discusses SER of suprasegmental English utterances, on the basis of the statistical analysis and classification of their formants and bandwidths. The obtained results are finally compared with those of other works on similar features in other languages, and with results based on other features in the same language in Section 6.8.

6.4.1. Statistical analysis of Suprasegmental English Utterances

The mean values of the formants of suprasegmental English utterances are given in Table 6.11.

Table 6.11: Mean values of formants of Suprasegmental English utterances

Mean values of the formants in Hertz for various emotions							
Formants	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
F1	567.46	553.57	541.58	634.04	506.44	581.7	619.2
F2	1818.8	1684.29	1868.3	1880.9	1777.33	1949.41	1791.92
F3	2869.2	2738.53	2860.0	2831.1	2932.6	2970.7	3647.4
F4	3948.1	3904.93	3939.9	3807.9	3976.6	3989.9	3931.7

The emotion with the highest formant value varied with the formant class. The highest formant value was for anger based on F1; for fear based on F2 and F4 and for disgust based on F3. The mean values of the bandwidths are given in Table 6.12.

Table 6.12: Mean values of bandwidths of Suprasegmental English utterances

Mean values of the bandwidths in Hertz for various emotions							
Bandwidths	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
B1	103.284	100.726	90.3812	116.178	198.735	154.717	119.822
B2	379.756	439.488	297.196	564.098	283.145	626.331	669.276
B3	381.079	628.371	509.716	401.834	579.679	564.724	575.455
B4	218.192	233.718	277.691	189.213	212.535	270.853	552.336

The bandwidths of the second and fourth formants were the highest for disgust. The formant and bandwidth values were not valence dependent. The results of investigations on the statistical discrimination of various emotions, based on each formant or bandwidth using ANOVA are presented in Table 6.13.

Table 6.13: Statistical discrimination of emotions based on the formants and bandwidths of suprasegmental English utterances

Statistical discrimination of emotions for suprasegmental English utterances (P < 0.01)		
Features	**Most discriminated emotions**	**Least discriminated emotions**
F1	Neutral, sad	Anger
F2	Surprise	Happy, anger, sad, disgust
F3	Surprise, disgust	Happy, neutral, anger
F4	Anger	Happy, surprise, neutral, sad, fear, disgust
B1	Sad, fear	Surprise, neutral, anger
B2	-	All emotions
B3	Anger	Happy, surprise, neutral
B4	Disgust	Happy, surprise, neutral, sad, anger, fear

It was found that the emotions with the least formant values were some of the most discriminated, statistically.

6.4.2. The Optimum feature set for Suprasegmental English

The features to be included in the optimum, minimal feature set were selected on the basis of their individual SER rates obtained by the Kmeans, KNN and the NB classifiers as detailed in Section 6.3.2 for segmental level SER. Table 6.14 presents the consolidated, graded SER rates for each emotion across the 3 different classifiers (with each of the 8 features,) as belonging to one of the four classes - very good, good, fair and poor. Grading was used for the easy comparison of SER rates, for the subsequent feature selection, since the primary concern of this investigation is the selection of features to the minimal set, rather than a detailed comparison of classifier performances. However, it was observed that the KNN classifier identified all the seven emotions on the basis of the formant and bandwidth features and gave the highest overall SER rate. The NB classifier performed poorly in this class seven emotion recognition problem in English.

Table 6.14: Consolidated graded SER rates with formants and bandwidths of suprasegmental English utterances

Consolidated graded SER rates of the formants and bandwidths								
Emotions	F1	F2	F3	F4	B1	B2	B3	B4
Happy	G	P	F	P	F	F	G	G
Surprise	F	G	G	G	F	G	F	F
Neutral	P	F	F	F	G	F	F	F
Anger	√	F	P	G	F	F	√	F
Sad	F	F	F	G	G	F	F	F
Fear	G	√	G	F	G	P	F	F
Disgust	F	F	G	F	G	P	P	F

Key: very good (√) - RR > 50; good (G) - 50 > RR > 35; Fair (F) - RR < 35; Poor (P) RR < 15

Table 6.14 indicates that across the various features and emotions, there were only three cases of very good classification rates.

Rejection of B2 and B3 from the final feature set: Across the three classifiers and eight features, the overall recognition rate was the least with B2 and it gave very poor recognition rates for disgust and fear. The emotions best

recognized with B3 were, happy, anger and surprise, all of which were better recognized by B4, F1, F2, F3 and F4. Disgust was poorly recognized on the basis of B3. Since the few emotions that were recognized using B2 and B3, could be better recognized by the other six features (as evident from Table 6.14), B2 and B3 were not used in the final ANN classification. Therefore the optimum, minimal feature set was constituted by the six spectral features, F1, F2, F3, F4, B1 and B4. This was further checked by the SER rates of the ANN classifier for the individual spectral features and their various combinations.

6.4.3. ANN classification for Suprasegmental English SER

The ANN SER rates were found out for each of the individual formants and the two bandwidths as well as for their various combinations. Investigations with the formant group gave the best results in terms of the least percentage error, for the combination of all the four formants. Separate classifications based on the two bandwidths and their various combinations, resulted in significantly less classification rates than those obtained with formants. Subsequently, classification was done with various combinations of formants and bandwidths. A slight improvement in performance was noted for a combination of formants with B1 and B4 over that based on formants alone. The salient results are presented in Table 6.15. The performance of the classifier was noted by changing the network size that is, for different number of neurons in the hidden layer. The optimum number was found to be 60 neurons as it yielded the least percentage error on the test data.

Table 6.15: Performance of Spectral features for SER in English

Features	Feature Identity	Percentage accuracy
Formants	F1 to F4	84.14
Bandwidths	B1	31.43
	B4	40
Formants and bandwidths	F1 to F4 and B1, B2	85.28

Even though the overall recognition accuracies obtained on the basis of the four formants were slightly less than that obtained with the minimal feature set, formant based classification was investigated in detail, for performance comparisons with the segmental level formant based SER. The confusion matrix of formant based recognition accuracies are as given in Table 5.16.

Table 6.16: Confusion matrix of formant based emotion classification of suprasegmental English utterances

Formant based emotion classification accuracies in percentage							
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	67.4	0	9.3	4.7	13.95	4.7	0
Surprise	0	100	0	0	0	0	0
Neutral	9.4	0	84.4	0	0	6.3	0
Anger	0	0	5.6	86.1	0	0	8.3
Sad	0	0	0	0	90.5	4.8	4.8
Fear	0	0	0	0	9.7	80	11.4
Disgust	3.2	0	0	0	16.1	0	80.6

The performance of this formant based classifier can be assessed in terms of misses and false hits respectively; both of which have to be nil or minimum for the best classifier. From Table 6.16, it can be concluded that there were no misses or false hits for surprise. From the perspective of misses, better SER performances were for surprise and sadness, than for happiness, fear and disgust. The maximum number of false hits was for sadness. The overall accuracy for this formant based neural network classifier was 84.14%, with maximum recognition accuracies of 100% for surprise and 90.5% for sad. Table 6.16 suggests that better SER performance can be obtained for SER problems, with reduced number of emotion classes. The results of such investigations with varied size of SER problem classes, are presented in Table 6.17.

Table 6.17: ANN performance measures for formant based English SER for various problem classes

Problem class and description	**Number of neurons**	**Percentage accuracy**
2 -positive and negative	45	100
3-positive, neutral and negative	60	97.14
4- surprise, neutral, anger, sad,	60	95
5- surprise, neutral, anger, sad, fear	60	92
6-happy, surprise, neutral, anger, sad, fear	60	90
7- neutral and six basic emotions	60	84.14

The obtained results indicated steady increase in the percentage error values with increase in the number of classes in the emotion recognition problem. The error was seen to increase linearly. But the rate of increase changed beyond a certain number of classes.

The output of the classifier for the minimal feature set of all formants and B1and B4 is presented in the confusion matrix given in Table 6.18

Table 6.18: Confusion matrix of the classification accuracies based on all formants, B1 and B4 for suprasegmental English utterances

Formant and bandwidth based emotion classification accuracies in percentage							
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	90.3	0	6.5	0	3.3	0	0
Surprise	5.2	84.5	10.3	0	0	0	0
Neutral	9.1	0	81.8	6.1	0	3	0
Anger	0	0	0	79.5	0	2.5	17.9
Sad	0	0	0	0	90.3	0	9.6
Fear	0	0	0	0	7.9	81.6	10.5
Disgust	0	0	0	0	5	0	95

There were no false hits for surprise, whereas disgust had the highest recognition rate as well as the maximum number of false hits. The least recognition rate was for anger and it was confused mostly with disgust. The overall SER rate of this neural network classifier for English (using all formants, B1 and B4) was 86.14% , with a maximum recognition accuracy of 95% for disgust, and 90.3% for sad and happy emotions**.**

Table 6.19 illustrates the improvement in SER performance obtained by reducing the number of emotion classes.

Table 6.19: ANN performance for formant and bandwidth based English SER for various problem classes

Problem class and description	**Number of neurons**	**Percentage accuracy**
2 positive and negative	45	100
3 positive, neutral and negative	60	97.14
4 surprise, neutral, anger, sad,	60	95
5 surprise, neutral, anger, sad, fear	60	92
6 happy, surprise, neutral, anger, sad, fear	60	90
7 neutral and six basic emotions	60	86.14

6.4.4. Comparisons of Formant based SER between Segmental and Suprasegmental Levels

This sections compares the results of formant based, English SER at the segmental level with that at the suprasegmental level. Comparisons with other relevant works are also given in Section 6.8.

Based on the comparisons of the SER rates of the three base classifiers between these two levels, as given in Tables 6.9 and 6.14,Table 6.20 recommends the specific feature as well as the effective analysis level for each emotion. Better classification of emotions was obtained at the suprasegmental than at the segmental level for fear.

Table 6.20: Best feature and analysis level for formant based SER in English

Emotions	Feature Identity	Analysis level
Happy	F2	V
Surprise	F2 / F3 / F4	V/ S*
Neutral	F2	V
Anger	F1	V/S
Sad	F1	V
Fear	F2	S
Disgust	F2	V

* V- vowel; S - suprasegmental

In these investigations emotion classification rates of 95.6% and 84.14% were obtained for formant based, ANN classification at the segmental and suprasegmental levels respectively. Detailed comparison of the ANN classification results (given in Tables 6.10 and 6.16) of segmental and suprasegmental utterances indicated higher classification accuracies for each emotion, at the segmental level. Therefore, in view of the above results, it is concluded that for English, SER is more effective at the segmental level than at the suprasegmental level.

6.5. Formant and Bandwidth based Hindi SER

This section investigates SER for Hindi, using the first four formants and their bandwidths following the steps explained in Section 6.4 for

suprasegmental English utterances. First, the various results of the statistical analysis are ANOVA are presented. This is followed by the classification results of the three base classifiers used for selecting optimum features for the subsequent ANN classification.

6.5.1. Statistical Analysis of Formants and Bandwidths for Hindi

The mean values of the various formants of suprasegmental Hindi utterances are presented in Table 6.21.

Table 6.21: Mean values of the various formants of suprasegmental Hindi utterances

Mean values of the formants in Hertz for various emotions							
Formants	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
F1	599	541.6	523	638.6	581.7	539	559
F2	1840.45	1820.19	1941.27	1606.28	1900.77	1651.58	1925.27
F3	2906	2804	2854	2920	3041	2809	2845
F4	3988.5	3825	3930	3899.9	4089.8	3906.5	3832.6

Sadness had high values in all the four formant classes. The formant values are however not valence dependent. The mean bandwidths are presented in Table 6.22.

Table 6.22: Mean values of bandwidths of suprasegmental Hindi utterances

Mean values of the formant bandwidths in Hertz for various emotions							
Bandwidths	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
B1	186.12	120.87	418.93	319.77	251.13	161.84	213.12
B2	474.39	237.51	405.12	526.44	385.33	218.97	275.09
B3	622.52	509.92	398.74	886.00	551.15	597.57	442.31
B4	465.83	473.47	668.08	1067.49	764.61	456.80	432.12

The second, third and fourth formant bandwidths were the largest for anger. Two of the formant bandwidth values were very close, for most emotions. Table 6.23 summarizes the results of statistical discrimination by ANOVA of the four formants and their bandwidths, for the various emotion pairs.

Table 6.23: Summary statistics of ANOVA of formant and bandwidths in Hindi, with emotion discrimination

Statistical discrimination of emotions for Hindi (P < 0.01)		
Features	**Most discriminated emotions**	**Least discriminated emotions**
F1	Anger, happiness, sadness	Fear, surprise
F2	Anger, fear	Disgust
F3	Sadness	Happiness, surprise, anger, sad, fear, disgust
F4	Sadness, happiness	Neutral, anger, fear
B1	Surprise, anger	Sadness
B2	Happiness	Sadness
B3	Happiness, anger, sad and fear	Disgust
B4	Anger	Happiness

Statistical analysis of each formant and its bandwidth indicated anger as the most discriminated from all other emotions, followed by happiness and sadness. Fear and disgust were statistically the least discriminated, based on formants and their bandwidths.

6.5.2. The Optimum feature set for Hindi SER

The features to be included in the optimum, minimal feature set were selected on the basis of their individual SER rates obtained by the Kmeans, KNN and the NB classifiers. Table 6.24 presents the consolidated, graded SER rates across the three different classifiers, for each emotion (under each of the 8 features) as belonging to one of the four classes - very good, good, fair and poor.

The summary of the classification results by the Kmeans, NB and KNN classifiers are as follows:

The K Means classifier gave very poor recognition rates across the various formant and bandwidth features for emotions like sad, surprise, disgust. The highest classification rate obtained was 53.13%, for sad and fear for B1. The NB classifier could not recognize all emotions, based on individual formants / bandwidths. The highest classification rate of 81.8% was obtained for anger, based on F1. The kNN classifier recognized all emotions based on each feature, and gave higher overall classification accuracy than the NB and k-

means classifiers. The highest SER rate obtained with this classifier was 84.6% for happy emotion, based on B1.

Table 6.24: Consolidated graded SER rates with formants and bandwidths of suprasegmental Hindi utterances

Consolidated emotion classification grades for each feature								
Emotions	**F1**	**F2**	**F3**	**F4**	**B1**	**B2**	**B3**	**B4**
Happy	G	G	F	F	G	√	F	G
Surprise	F	F	P	F	F	F	F	F
Neutral	P	G	F	F	F	F	F	F
Anger	√	√	G	F	F	F	F	G
Sad	F	P	G	G	F	P	P	F
Fear	G	F	F	F	√	G	G	F
Disgust	F	G	F	P	G	G	√	F

Key : Very good (√) - RR > 50; Good (G) 50 > RR > 35; Fair (F) - RR < 35; Poor (P) RR < 15

Good overall recognition rates were obtained with F1, F2, B1, B2, and B3. F3 as well as F4 contributed to the recognition of sadness. With B4, moderate recognition rates were obtained for each emotion. These observations of the effectiveness of each feature in the recognition of the various emotions, supported the inclusion of all formants and bandwidth in the optimal feature set for SER in Hindi.

6.5.3. ANN classification for Hindi SER

Details of the final classification by a two layer feed forward back propagation artificial neural network are given in Table 6.25. The formants and bandwidths were first classified in separate groups and the resulting accuracies were noted. The ANN classification was repeated for different combinations of features and the best results obtained are presented in Table 6.25.

Table 6.25: Performance of Spectral features for Hindi SER

Feature identity	Percentage accuracy
F1 to F4	94.29
B1 to B4	83
F1 to F4 & B1 to B4	97.14

The ANN classification of the combined group of formant bandwidth values gave the maximum classification rate of 97.14%, indicating all the formants as well as bandwidths in Hindi, to be very good markers of emotion. Table 6.26 presents the confusion matrix of formant and bandwidth based emotion classification accuracies for suprasegmental Hindi utterances, as percentage belonging to each class.

Table 6.26: Confusion matrix of the classification accuracies for formant and bandwidth based Hindi SER

Classification accuracies in percentage for various emotions							
	Happy	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	100	0	0	0	0	0	0
Surprise	0	100	0	0	0	0	0
Neutral	0	0	100	0	0	0	0
Anger	0	0	0	80	0	0	10
Sad	0	0	0	0	100	0	0
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	0	0	100

The ANN classifier gave 100% accuracy for all emotions except anger. Figure 6.3 shows the variation in the percentage error on the test data, for five different sizes of ANN.

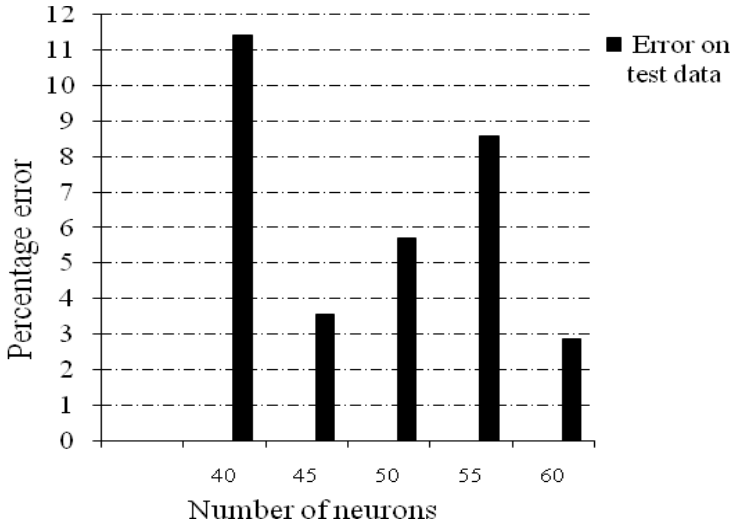


Figure 6.3: Percentage error for various network sizes

Table 6.27 gives the ANN performance measures in terms of its classification accuracy, for the spectral feature set based Hindi speech emotion recognition, for various problem classes.

Table 6.27: ANN performances for formant and bandwidth based Hindi SER for various problem classes

Problem class and description	Number of neurons	Percentage accuracy
2 - positive and negative	20	100
3 - positive, neutral and negative	60	100
4 - happy, surprise, sad, fear	60	100
5 - happy, surprise, neutral, sad , fear	30	100
6 - happy, surprise, neutral, sad, fear ang	30	100
7 - neutral and six basic emotions	60	97.14

Fully recognized emotion groups of size ranging from two to six were identified. Not all combinations of emotions were recognized fully. This implies that other features are needed to give complete recognition in those cases. Thus improvement in performance of the ANN was obtained by reducing the number of emotion classes from seven.

6.5.4. Conclusion of Spectral feature based Hindi SER

An efficient, though simple approach for class seven SER in Hindi, using minimal feature set of formants and bandwidths was implemented. The recognition rate obtained with this approach was 97.14%, which is the highest reported so far, for Hindi, supporting the use of the proposed minimal feature set for Hindi SER. The classification results of the three base classifiers were validated by the performance of ANN classifier and by the results of the human perception tests given in Table 3.4. The obtained SER rates are comparable with that of human SER. The SER rate increased with decrease in the number of emotion classes, from seven.

6.6. Formant bandwidth based Malayalam SER

The investigations for Malayalam SER were along the same lines as detailed for English and Hindi, in sections 6.4 and 6.5 respectively. The average

values of the first four formants of Malayalam utterances are as given in Table 6.28.

6.6.1. Statistical Analysis of formants and bandwidths for Malayalam

The mean values of the various formants are given in Table 6.28

Table 6.28: Mean values of the various formants of suprasegmental Malayalam utterances

Mean values of the formants in Hertz for various emotions							
Formants	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
F1	654.50	577.52	645.9	555.343	579.561	507.49	638.47
F2	1741.61	1630.03	1749.59	1767.33	1841.68	1876.55	1833.02
F3	2818.72	2781.03	2718.49	2744.46	2928.04	2887.33	2909.93
F4	3938.45	3778.41	3665.70	3642.58	3746.31	3874.97	3758.19

The formant values varied across the different emotions. The highest values for F1 and F4 were for happiness. It was for fear, based on F2; and for disgust, based on F3. The least F1 was for fear and the least F2 was for surprise. The least values for F3 and F4 were for neutral and anger respectively.

The mean values of the first four bandwidths for the seven emotions are given in Table 6.29.

Table 6.29: Mean values of the bandwidths of Malayalam suprasegmental utterances.

Mean values of the formants in Hertz for various emotions							
Bandwidths	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
B1	155.90	117.10	208.40	162.70	102.80	125.50	144.60
B2	272.30	296.60	239.30	325.50	171.20	272.00	211.50
B3	402.5	383.57	458.24	439.72	325.35	396.79	392.80
B4	294.95	529.80	699.68	511.99	430.30	435.29	587.33

The least values of the first, second and third bandwidths were for sadness. The least value of B4 was for happiness. Except for B2, the highest bandwidth value was for neutral. The highest value for B2 was for anger. The bandwidth values were not valence dependent. The statistical discrimination of formants and bandwidths of Malayalam utterances are given in Table 6.30.

Table 6.30: Summary statistics of ANOVA of formants and bandwidths of Malayalam utterances

Statistical discrimination of emotions with P < 0.01		
Features	**Most discriminated emotions**	**Least discriminated emotions**
F1	Anger, fear	Happy, neutral, disgust
F2	Surprise, fear	Happy, neutral, anger
F3	Sad, happy	Surprise, neutral, anger
F4	Neutral, fear	Anger,
B1	Neutral	Fear, disgust
B2	Neutral, anger, sad	Happy, surprise, fear
B3	Sad	Happy
B4	Happy	Anger, neutral, surprise, disgust

6.6.2. The Optimum Feature set for the SER in Malayalam

The KMeans classifier gave the highest SER rate of 66.7% with bandwidths B4 and B3, for happy and fear. This classifier recognized all emotions, but at a lesser rate, based on each of the eight individual features values. The NB classifier failed to recognize certain emotions at the expense of high RR of certain other emotions. Among formants, the highest classification accuracies were obtained with F4. Surprise was the most recognized. Anger was the worst recognized.

The kNN classifier classified all emotions, based on each feature, and gave the highest overall classification accuracy. The highest recognition rate obtained was 80% for neutral, based on F2 values. In the bandwidth based classification, the kNN classifier gave 100% recognition rates for sad and fear based on B1, leading to the inclusion of the latter in the feature group for the final ANN based classification of emotional speech.

The consolidated, graded average classification rates are given in Table 6.31. The first three bandwidths give very good classification rates for several of the seven emotions. F1 and B2 give very good recognition rates. Very good were obtained for each emotion on the basis of the various individual features other than F4. For Malayalam, the fourth formant was only moderate performer

in speech emotion recognition. Since there was no instance of poor recognition on the basis F4, it was included with other seven features in the final ANN classification.

Table 6.31: Consolidated graded SER rates with formants and bandwidths of suprasegmental Malayalam utterances

Graded SER rates based on formants and bandwidths								
Emotions	**F1**	**F2**	**F3**	**F4**	**B1**	**B2**	**B3**	**B4**
Happy	√	G	G	F	G	F	F	√
Surprise	√	F	F	F	F	√	F	G
Neutral	G	√	F	G	F	G	F	F
Anger	F	F	F	F	√	√	√	G
Sad	G	G	G	F	√	√	√	F
Fear	F	F	√	G	√	F	F	P
Disgust	√	G	P	F	√	G	G	G

Key: very good (√) - RR > 50; good (G) - 50 ≥ RR > 35;
Fair (F) - 35 ≥ RR > 15; Poor (P) RR ≤ 15

6.6.3. ANN Classification for Malayalam SER

Table 6.32 gives results of ANN classification based on combinations of formants alone, bandwidths alone and both formants as well as bandwidths.

Table 6.32: Spectral features giving the best SER rates in Malayalam

Features	Feature identity	Percentage accuracy
Formants	F1 to F4	78.35
BW	B2	40
	B1	34.29
Formants and BW	F1 to F4 & B1 to B4	86.76
	F1 to F4 & B1, B2	84.85

In Malayalam, the highest overall classification accuracy of 86.76%, was obtained for a feature set comprising all four formants and their bandwidths. Such a high emotion classification accuracy based on these vocal tract features, has not been reported earlier. The confusion matrix of the classification accuracies (as percentage in each class) based on all four formants and their bandwidths is given in Table 6.33.

Table 6.33: Confusion matrix of classification accuracies based on all formants and bandwidths in Malayalam

Classification accuracies in percentage for various emotions							
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	78.6	0	0	0	0	0	21.4
Surprise	0	94.1	0	0	2.94	2.94	0
Neutral	16.7	0	73.8	0	2.4	2.4	4.8
Anger	2.6	2.6	5.2	84.6	2.6	2.6	0
Sad	10	0	0	0	90	0	0
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	10.3	3.4	86.2

The 100% recognition of fear in this class-7 emotion recognition problem is significant. Surprise, anger, sad and disgust too were classified at accuracies above 80%. Table 5.34 gives the ANN performances for various number of emotion classes. With reference to the percentage error in the classification of different numbers of emotion classes for Malayalam, 100% accuracy in classification was observed for a class 2 SER of positive and negative emotions. With the addition of neutral samples to the group, however the maximum percentage of overall recognition accuracy of the class three problems decreased to 91.43%, due to the possible confusion in the formant - bandwidth values of neutral with those of certain other emotions from both groups. The results were obtained using an optimum network with sixty neurons in the hidden layer

Table 6.34: ANN Performance for formant and bandwidth based Malayalam SER

Problem class	Description	Number of neurons	Percentage accuracy
2	positive and negative	45	100
3	positive, neutral and negative	60	91.43
3	Surprise, neutral, fear	45	100
4	Surprise, neutral, anger, sad	55	100
5	surprise, neutral, sad, anger, fear	55	100
6	happy, surprise, neutral, anger, sad, fear	60	96.67
7	neutral and six basic emotions	60	86.76

However, for the specific combination of surprise neutral and fear, 100% was obtained as the highest accuracy for a class-3 SER problem. The highest

recognition accuracy obtained here, for a class four problem was 100% whereas available reports show a maximum of accuracy of 72.1% only, with a different feature set [136]. For the class-5 SER problem the highest accuracy of 100% obtained only by avoiding happiness and disgust has to be viewed from the perspective of the confusion matrix for the class-7 SER problem, which clearly indicates several misses and false hits for happiness and disgust. For class-6 SER problem, the overall accuracy obtained was 96.67%.

6.7. Universality in Formant and Bandwidth based SER across English, Hindi and Malayalam

The confusion matrices of Tables 6.18, 6.26 and 6.33 suggest universality in formant and bandwidth based SER in English, Hindi and Malayalam. Therefore all formants and bandwidths were given to an ANN classifier. The best classification results obtained for an ANN network with 35 neurons in the hidden layer, are presented in the confusion matrix of Table 6.35.

Table 6.35: Confusion matrix of classification accuracies of formant and bandwidth based SER across English, Hindi and Malayalam

	Classification accuracies in percentage for various emotions						
Emotions	**Happy**	**Surprise**	**Neutral**	**Anger**	**Sad**	**Fear**	**Disgust**
Happy	100	0	0	0	0	0	0
Surprise	0	75	0	0	0	0	25
Neutral	0	0	100	0	0	0	0
Anger	0	0	0	100	0	0	0
Sad	10	0	0	0	66.7	11.1	22.2
Fear	0	0	0	0	0	100	0
Disgust	0	0	0	0	0	0	100

All instances of happiness, neutral, anger, fear and disgust were recognised by this ANN classifier. Therefore, it is concluded that there is universality in recognition of these five emotions, across these three languages. The overall accuracy obtained was 91.67%.

6.8. Models for SER in Malayalam

This section presents the results of modeling SER in Malayalam using Decision Trees and Logistic Regression, with formants and bandwidths as features.

Scheme for modelling using binary trees and logistic regression

The focus of this thesis is the classification of seven emotions. However, in the course of building intuitive models for SER the following binary classifications were considered as relevant:

i. Classification of speech samples as emotional versus neutral

ii. Classification of emotional speech as positive valence versus negative valence

iii. Classification of positive valence emotional speech as happy versus surprise.

Additionally, the following cases too were investigated:

iv. Classification of speech samples as belonging to any of the three - neutral, positive valence or negative valence emotional class.

v. Classification of negative valence emotions as belonging to any of the four- anger, sad, fear or disgust.

vi. Finally, the single stage classification of emotions as happy, surprise, neutral, anger, sad, fear or disgust.

The schematic of the intuitive modelling for classifications under case (i) to (iii) and (v) listed above, are as given in Figure 6.4.

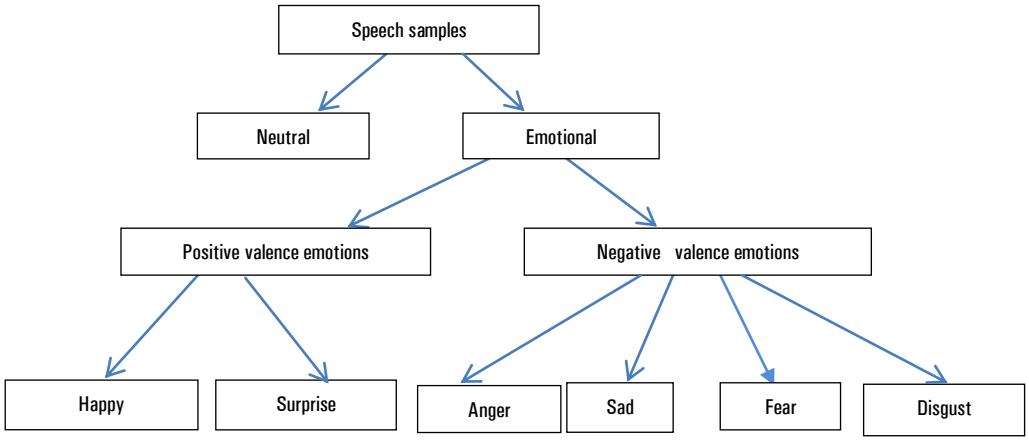


Figure 6.4: Schematic of the modeling for certain cases of SER in Malayalam

Decision Trees - Results for the binary classifiers

The approach adopted in this thesis for modeling SER in Malayalam has been given in Section 3.9. As an illustration of the modeling and consequent validation of the model, case (i) (viz. the classification of speech samples as neutral or emotional), is considered first.

The predictor importance was assessed for each classifier in order to arrive at a possible subset of important predictors. The results (coefficient values) obtained for two ensembles of 50 trees and 100 trees respectively, for the classification of neutral versus emotional samples, are given in Table 6.36.

Table 6.36: Coefficients indicating predictor importance

Number of trees in ensemble	Coefficient values indicating predictor importance							
	F1	**F2**	**F3**	**F4**	**B1**	**B2**	**B3**	**B4**
50	0.0012	0.0019	0.0047	0.0080	0.0011	0.0019	0.0008	0
100	0.0008	0.0015	0.0036	0.0061	0.0009	0.0013	0.0008	0.0004

The entries in the above table indicate that the eighth feature x8 which is in fact the fourth bandwidth B4 is the least important predictor with regard to the neutral versus emotional, binary classification. However, for the purpose of verification and comparison, classification was done both for the full set as well as with a subset of seven features; excluding B4. The results were validated

several times by resubstitution as well as cross validation. The training set was used in resubstitution. Cross validation was done by partitioning dataset into folds and estimating the accuracy on each fold.

The textual description of a pruned classification tree used for the classification of neutral and emotional speech samples was obtained as follows:

Textual Description of Pruned Decision tree

1 if x3<2887.69 then node 2 elseif x3>=2887.69 then node 3 else 1
2 if x4<4081.5 then node 4 elseif x4>=4081.5 then node 5 else 2
3 if x2<1992.39 then node 6 elseif x2>=1992.39 then node 7 else 1
4 if x3<2805.97 then node 8 elseif x3>=2805.97 then node 9 else 2
5 class = 2
6 if x4<3973.17 then node 10 elseif x4>=3973.17 then node 11 else 2
7 if x7<887.335 then node 12 elseif x7>=887.335 then node 13 else 1
8 class = 2
9 if x3<2881.81 then node 14 elseif x3>=2881.81 then node 15 else 1
10 if x7<171.171 then node 16 elseif x7>=171.171 then node 17 else 1
11 if x6<230.643 then node 18 elseif x6>=230.643 then node 19 else 2
12 if x4<4307.48 then node 20 elseif x4>=4307.48 then node 21 else 1
13 class = 2
14 if x5<307.234 then node 22 elseif x5>=307.234 then node 23 else 1
15 class = 2
16 class = 2
17 if x1<653.622 then node 24 elseif x1>=653.622 then node 25 else 1
18 if x2<1573.76 then node 26 elseif x2>=1573.76 then node 27 else 2
19 if x7<397.252 then node 28 elseif x7>=397.252 then node 29 else 1
20 if x1<533.505 then node 30 elseif x1>=533.505 then node 31 else 1
21 if x1<635.396 then node 32 elseif x1>=635.396 then node 33 else 2
22 class = 1
23 class = 2
24 if x1<540.627 then node 34 elseif x1>=540.627 then node 35 else 1
25 class = 1
26 class = 1
27 if x7<153.281 then node 36 elseif x7>=153.281 then node 37 else 2
28 if x5<200.04 then node 38 elseif x5>=200.04 then node 39 else 2
29 if x6<511.885 then node 40 elseif x6>=511.885 then node 41 else 1
30 if x7<189.503 then node 42 elseif x7>=189.503 then node 43 else 1

```
31  class = 1
32  class = 1
33  class = 2
34  class = 1
35  if x2<1675.79 then node 44 elseif x2>=1675.79 then node 45 else 1
36  class = 1
37  class = 2
38  if x2<1909.53 then node 46 elseif x2>=1909.53 then node 47 else 2
39  if x3<3072.19 then node 48 elseif x3>=3072.19 then node 49 else 1
40  class = 1
41  class = 2
42  class = 2
43  class = 1
44  class = 2
45  if x5<276.758 then node 50 elseif x5>=276.758 then node 51 else 1
46  if x7<196.714 then node 52 elseif x7>=196.714 then node 53 else 2
47  class = 2
48  class = 1
49  class = 2
50  class = 1
51  class = 2
52  class = 2
53  if x6<399.608 then node 54 elseif x6>=399.608 then node 55 else 1
54  class = 1
55  class = 2
```

Similar descriptions of the unpruned decision tree for the same problem are given in Appendix C. It can be understood that the eighth predictor is not included in this model, as it was already found to be unimportant. The node numbers are given on the extreme left. Nodes 5, 8, 15, etc. are the terminal nodes giving the category labels as either 1 or 2 corresponding to neutral or emotional speech. Node 1 is the root of the tree and nodes 2, 3, 4 etc. are the internal nodes where predictor values are checked. The unpruned decision tree model presented in Appendix C has more number of nodes to perform similar functions. The graphical description of a pruned classification tree is given in Figure 6.5.

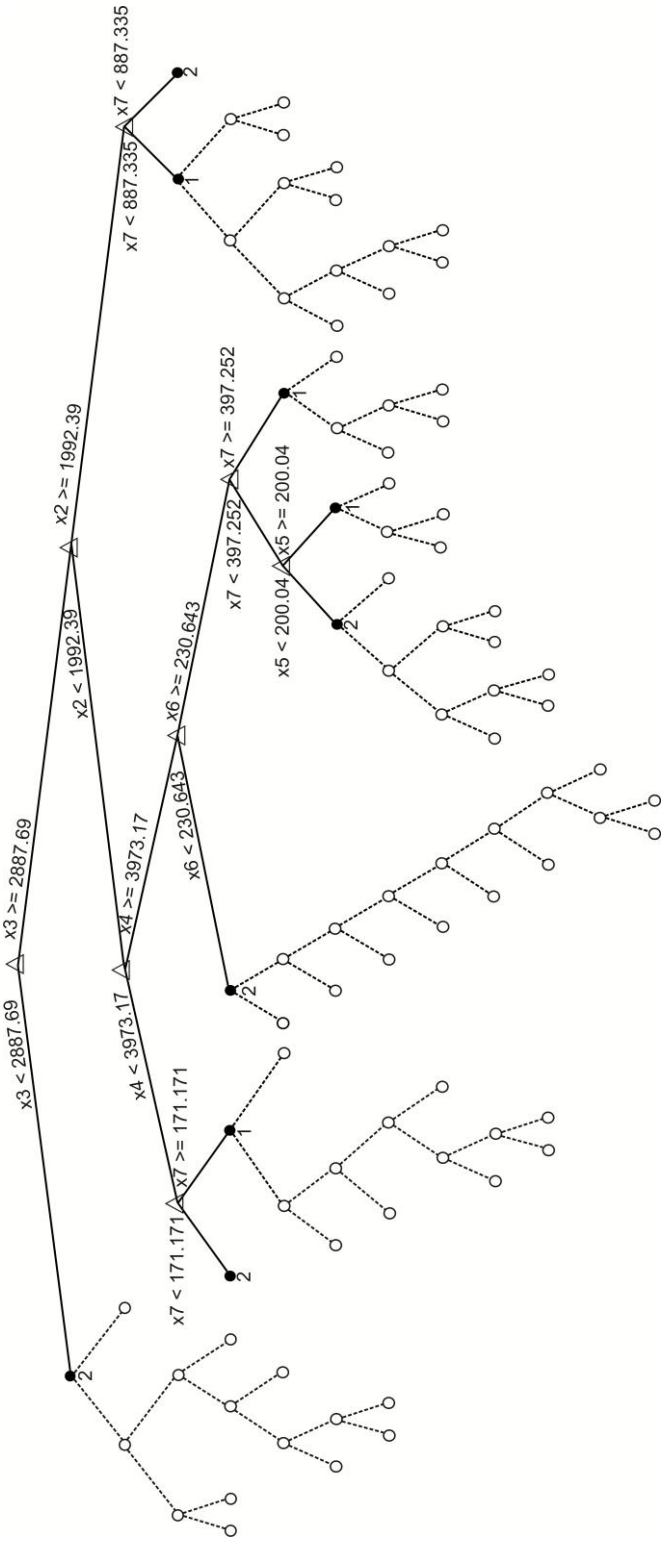


Figure 6.5: Schematic of a pruned decision tree

The pruned tree shown in Figure 6.5 clearly indicates that the initial classification in this tree has been performed based on the values of the third formant (x3). The subsequent classifications have been based on the second formant (x2), fourth formant (x4) and third bandwidth (x7). Similarly the graphical description for an unpruned decision tree is given in Appendix C. Whereas the unpruned tree has 93 nodes this pruned tree has only 17 functional nodes.

The average SER accuracy obtained on the train class was 89.44%.

The confusion matrix obtained for this classification is given in Table 6.37, for the test class.

Table 6.37: Confusion matrix for neutral/emotional Speech classification of test class

Percentage accuracies for SER		
Output /Input	**Neutral**	**Emotional**
Neutral	85	15
Emotional	16.11	83.89

The average SER accuracy obtained on the test class was 84.45%.

The classification models for case (ii) and case (iii) were arrived at, by the same approach and these models too were validated. The classification results of the three binary classifiers are summarised in Table 6.38.

Table 6.38: Summary of Results of various binary classifications

Percentage accuracies for SER						
Case	**Class**	**Description**	**8 Features**		**7 Features**	
			Resub*	**Cross Val#**	**Resub**	**Cross Val**
(i)	2	Neutral versus Emotional	88.61	82.68	89.44	84.45
(ii)	2	Positive valence versus Negative valence	90.80	87.76	91.51	87.67
(iii)	2	Happy versus Surprise	92.69	84.19	93.63	87.31

***Resubstitution; # Cross Validation**

It is found that very good decision tree models in binary classification were obtained for case ii and case iii. Except for the classification based on the valence

of emotions, higher accuracy was obtained with the subset of seven predictors rather than with the complete set (of 8 features) validating the determination of predictor importance for binary logistic regression. The classification results of the three multiclass classifiers are summarised in Table 6.39.

Table 6.39: Summary of Results of various multiclass classifications

Classification accuracy in percentage					
Case	Description	8 Features		7 Features	
		Resub*	Cross Val#	Resub	Cross Val
(iv)	Neutral, positive and negative valence	87.19	85.06	86.43	84.62
(v)	Anger, sad, fear and disgust	82.01	66.80	79.81	75.96
(vi)	Happy, surprise, neutral anger, sad, fear and disgust	77.86	71.78	75.24	70.59

***Resubstitution; # Cross Validation**

In all the three cases cited in table 6.38, higher classification accuracies were obtained by resubstitution rather than by cross validation, both on the complete feature set as well as on the reduced one with 7 features. The classification accuracy decreased with increase in the number of classes. However, the high emotion classification accuracies obtained in almost all cases validate the decision tree model.

In summary, the tree has easy interpretability in terms of logical expressions. The decision of any specific test pattern is interpreted as a conjunction of decisions along the path to its leaf node. It helps one to understand what all attribute values / ranges lead to a decision and also provides clear interpretation of the categories in terms of logical expressions of attributes and is therefore truly, a white box model.

Logistic Regression: Binary logistic regression uses data from a variety of predictors to model the probability of each of the two possible outcomes. Hence it can be appropriately applied to the three cases of binary classifications listed in the beginning of this section. Binary logistic regression can therefore be used to understand and predict (i) the presence of emotion (ii) valence of emotion

(iii) happiness / surprise from positive valence emotional speech; based on the first four formants F1-F4, along with their corresponding bandwidths B1-B4.

Case (i): Beginning with case (i) stated above, the objective was to arrive at an appropriate regression equation that could predict the probability of a speech sample as belonging to either the neutral or emotional group based on predictor values. This also took care of including only those predictor values that contributed to the response, in a statistically significant way. The feature or predictor value and responses were checked to satisfy the assumptions mentioned in Section 3.9. First a regression model was fit for the entire training data. A portion of this set was replaced by new samples for prediction. Prediction of probable outcomes based on feature values of new samples was done using the obtained regression equation. The binary logistic regression equation represented by Y' is an algebraic representation of the regression line and is used to describe the relationship between the response and predictor variables.

Considering the probability that an event occurs as P(1),

$$P(1) = \exp(Y') / (1 + \exp(Y')); \quad (6.1)$$

Stepwise logistic regression for case (i) eliminated the fourth bandwidth (eighth predictor) upon evaluating its contribution as insignificant. Subsequently, the regression equation for the binomial logistic regression model was obtained as follows:

$$Y_s' = -4.07 - 0.00393\ F1 + 0.001854\ F2 + 0.00323\ F3 - 0.001990\ F4 + 0.00489\ B1 + 0.003452\ B2 - 0.000995\ B3 \quad (6.2)$$

The various aspects related to the interpretation of results of regression have been discussed in Section 3.9. Accordingly it can be deduced that the third formant and the first bandwidth contribute positively, to the probability of occurrence of the event, namely the presence of emotion.

For the sake of performance comparison, another regression model including all predictors was obtained as,

$$Y' = 4.20 + 0.00380\ F1 - 0.001839\ F2 - 0.00328\ F3 + 0.001971\ F4 - 0.00487\ B1 - 0.003482\ B2 + 0.000987\ B3 + 0.000305\ B4 \quad (6.3)$$

Case (ii): The stepwise regression equation obtained for the prediction of positive and negative valence emotions based on the most relevant features was,

$$Y_s' = 17.58 - 0.001403\ F2 - 0.004967\ F3 - 0.001418 B1 + 0.001138\ B2 - 0.000795\ B3 \quad (6.4)$$

The regression equation including the 8 predictors for case (ii) was obtained as follows:

$$Y' = 19.03 - 0.001172\ F1 - 0.001314\ F2 - 0.004697\ F3 - 0.000441\ F4 - 0.001486\ B1 + 0.001132\ B2 - 0.000814\ B3 + 0.000240\ B4 \quad (6.5)$$

Case (iii): The stepwise regression equation for the prediction of positive valence emotions as happy or surprise was obtained as follows:

$$Y_s' = 16.13 - 0.00246F3 - 0.00208F4 - 0.002774B2 \quad (6.6)$$

But the HL statistic was indicative of a poor fit for this model with few features.

The equation for the regression model including all the predictors is given as,

$$Y' = 15.67 + 0.00439\ F1 - 0.00260\ F2 - 0.00153\ F3 - 0.00200\ F4 - 0.00004\ B1 - 0.00290\ B2 + 0.00033\ B3 - 0.000848\ B4 \quad (6.7)$$

Analyses of these three distinct cases of binary logistic regression yielded variance inflation factors close to 1, which indicates that the predictors are not correlated. The classification accuracies and Homer Lemeshow statistic

obtained for cases (i) to (iii) using the ordinary as well as stepwise regression are given in Table 6.40.

Table 6.40: Prediction accuracies (in percentage) for the various cases of binary logistic regression

Case	Class	Description	Prediction accuracy using 8 Features (%)			Prediction accuracy with Features got by Stepwise method (%)		
			Train	Test	*HLS	Train	Test	*HLS
(i)	2	Neutral versus Emotional	70	68.06	0.692	66.9	66.85	0.692
(ii)	2	Positive valence versus Negative valence	69.22	68.13	0.881	70.87	69.49	0.631
(iii)	2	Happy versus Surprise	73	70	0.765	68	67.6	0.452

***Hosmer Lemeshow Statistic**

In summary, binomial logistic regression was run to understand the contributions of the first four formants and their respective bandwidths in these three cases of binary classifications. The Hosmer-Lemeshow test showed that the model fitted the data well especially with the entire feature set. The regression equation obtained using Minitab was used to make predictions about response (the dependent variable) based on values of the independent variables. Table 6.40 shows that the highest accuracy was obtained for case (iii), on both the test as well as train set and based on all predictors.

Multiclass SER

Case (iv): Classification of speech as neutral/ positive valence/negative valence

The Response Information displays the number of observations that fall into each of the response categories (neutral, positive and negative). Neutral emotion was chosen as the reference event and assigned the response value 3. Positive and negative valence emotions were assigned values 2 and 1 corresponding to events 2 and 1 respectively. Here Nominal Logistic Regression was performed for the Response Information versus F1, F2, F3, F4, B1, B2, B3, B4. The first set of estimated logits, labeled Logit (1), are the parameter estimates of the change in logits of negative valence emotions relative to the reference event, neutral. The coefficients along with their p values are presented in Table

6.41. Considering the co-efficients for both logits, the positive values associated with F2, B1 and B2 indicate that these features favour the occurrence of negative valence emotions over neutral. F1 coefficient is the estimated change in the logit with unit increase in F1, with other features held a constant. The p-values of 0.517 and 0.534 for F4 and B4, respectively (for logit (1)), are more than the default acceptable alpha-level. This indicates that there is sufficient evidence to conclude that a change in these features did not affect the choice of occurrence of positive valence emotions over neutral. The second set estimated logits, labeled Logit (2), are the parameter estimates of the change in logits of positive valence emotion relative to the reference event, neutral. The obtained classification accuracy was 68.83%.

Table 6.41: Logistic Regression Table showing constants and coefficients for both logits

Constant/	Coefficients		P values	
Predictor	Logit1: (2/3)	Logit 2: (1/3)	Logit 1: (2 /3)	Logit 2: (1/3)
Constant	-0.602863	-7.22399	0.843	0.020
F1	-0.001785	0.0033297	0.196	0.016
F2	0.0007288	0.0027125	0.223	0.000
F3	-0.0010223	0.0021198	0.245	0.022
F4	0.0003916	-0.0008753	0.517	0.147
B1	0.0050674	0.0057252	0.000	0.000
B2	0.0045349	0.0027366	0.000	0.000
B3	-0.0008686	-0.0007845	0.050	0.073
B4	-0.0002492	-0.0001761	0.534	0.659

The p value for the Pearson Goodness of Fit test was 0.719 for a Chi-square statistic of 1034.9; with 1062 degrees of freedom. This indicates that there is evidence to suggest that the regression model for neutral, negative and positive valence emotions fits the data well.

Case (v): Classification of negative valence speech as anger /sad /fear /disgust

In this case, disgust was chosen as the reference event and assigned the response value 4. Anger, sad and fear were assigned values 1, 2 and 3 respectively. Table 6.42 presents the coefficient values for the three logits.

Table 6.42: Logistic Regression Table showing constants and coefficients for three logits

Constant/ Predictor	Coefficients			P values		
	Logit1: Fear/disgust (3/4)	**Logit2: Sad/disgust (2/4)**	**Logit3: anger/disgust (1/4)**	**Logit1: Fear/disgust (3/4)**	**Logit2: Sad/disgust (2/4)**	**Logit3: anger/disgust (1/4)**
Constant	13.3951	57.7865	50.8699	0.258	0.000	0.000
F1	0.0008421	-0.0034889	-0.0034895	0.835	0.517	0.492
F2	0.0028467	0.0055478	-0.0034467	0.293	0.083	0.296
F3	-0.0005221	-0.0068552	-0.0092964	0.862	0.044	0.004
F4	-0.0033792	-0.0110046	-0.0023614	0.177	0.000	0.394
B1	-0.0078347	-0.0075737	-0.0092694	0.001	0.032	0.002
B2	-0.0013985	0.0019511	-0.0049060	0.311	0.243	0.009
B3	-0.0003011	0.025495	0.0028536	0.832	0.143	0.073
B4	-0.0019334	-0.0004048	-0.0029143	0.120	0.763	0.073

From the entries of the Table it can be interpreted that F1 does not affect the choice of occurrence of anger / sad / fear over the occurrence of disgust.

The p value for the Pearson goodness-of-fit test was 0.967 with Chi – square statistic of 287.25, with 333 degrees of freedom. With the eight features for the SER of four negative valence emotions, the Pearson test has p-value far greater than 0.05 indicating that the regression model fits the data adequately. The highest average classification accuracy was 65%, with all features.

Case (vi): Classification of emotions as happy, surprise, neutral, anger, sad, fear and disgust

Neutral emotion was chosen as the reference event and assigned the response value 7. The other emotions happy, surprise anger, sad, fear and disgust, were assigned values 1-6, in that order. Hence their corresponding logits are as follows:

Logit 1: (Disgust / Neutral); Logit 2: (Fear / Neutral); Logit 3: (Sadness / Neutral); Logit 4: (Anger / Neutral); Logit 5: (Surprise / Neutral); Logit 6: (Happiness / Neutral)

Tables 6.43 to 6.45 present the coefficient values for the six logits.

Table 6.43: Logistic Regression Table showing constants and coefficients for the first and second logits

Constant/ Predictor	Coefficients		P values	
	Logit1: (6/7)	Logit2: (5/7)	Logit1: (6 /7)	Logit 2: (5/7)
constant	-43.8163	-21.1766	0.003	0.051
F1	0.0014753	0.0021251	0.754	0.622
F2	0.0077139	0.0086786	0.003	0. 010
F3	0.0026497	0.0028532	0.412	0.292
F4	0.0043564	-0.0007863	0.077	0.646
B1	0.0155208	0.0077531	0.000	0.026
B2	0.0018788	-0.0000390	0.285	0.980
B3	-0.0021766	-0.0023632	0.152	0.072
B4	-0.0038051	-0.0050937	0.012	0.000

Table 6.44: Logistic Regression Table showing constants and coefficients for the third and fourth logits

Constant/ Predictor	Coefficients		P values	
	Logit3: (4/7)	Logit 4: (3/7)	Logit3: (4 /7)	Logit 4: (3/7)
constant	32.4868	47.9444	0.004	0.000
F1	-0.0095486	-0.0072485	0.064	0.191
F2	0.0114285	0.0029350	0.000	0.184
F3	-0.0063387	-0.0116414	0.013	0.000
F4	-0.0075553	-0.0033780	0.000	0.060
B1	0.0084902	0.0080945	0.027	0.029
B2	0.0028897	-0.0032241	0.071	0.084
B3	0.0002888	0.0019880	0.828	0.117
B4	-0.0026915	-0.0036620	0.057	0.013

Table 6.45: Logistic Regression Table showing constants and coefficients for the fifth and sixth logits

Constant/ Predictor	Coefficients		P values	
	Logit5: (2/7)	**Logit6: (1/7)**	**Logit5:(2 /7)**	**Logit6: (1/7)**
constant	-0.826584	36.3043	0.938	0.000
F1	-0.0025993	-0.0035687	0.576	0.469
F2	0.0088334	0.0068858	0.000	0.001
F3	-0.0100503	-0.0094724	0.000	0.000
F4	0.0036539	-0.0040588	0.075	0.017
B1	0.0071078	0.0043388	0.050	0.241
B2	0.0016681	0.0016157	0.286	0.341
B3	-0.0029259	-0.0006953	0.037	0.578
B4	-0.0012226	-0.0035304	0.360	0.11

The entries in the Tables 6.43 to 6.45 indicate F1 and B2 have lesser role in the occurrence of other emotions with respect to the occurrence of neutral. The p value for the Pearson deviance goodness-of-fit test was 0.998 with Chi-square statistic of 1075.54; with 1212 degrees of freedom.

The highest average classification accuracy was 60.47%, with all features. Sadness and disgust were the best recognized, at 73.3% and 70% respectively.

Upon comparison it is seen that the modeling by decision tree or logistic regression helped to identify the important predictors for the SER. The classification accuracies / performances by either of the modeling techniques were comparable, for each of the six classes.

6.9. Comparison with the State of the art

The results of the investigation are not directly comparable with state of the art results reported elsewhere in the literature on SER. This is because, even for SER in English itself or based exclusively on vowels, the exact database, feature sets and classifiers differ strikingly. Nevertheless, certain qualitative

comparisons between the current results and the state-of-the-art are presented for the segmental and supra segmental SER.

In the case of segmental level SER, Schuller et al. have reported a speaker independent SER rate of 88.6% on the Berlin speech database [137]. Hassan and Damper [102] have reported 79.5% on the Berlin database and 80.1% on the Serbian database using the 3DEC method developed by them. In view of these cited recognition accuracies; we logically conclude that the vowel formant based approach described here, which achieves 95.6% SER rate, is superior in accuracy .

The highest emotion classification accuracy reported for elicited emotions is 80.8%, using hundreds of other features and deep neural network based generalized discriminant analysis technique. The investigations had been reported for a bench marked emotional speech database [138].

Very few studies have been reported for Indian languages. An overall accuracy of 87.9% has been reported for Hindi SER using MFCC [139]. An average emotion recognition rate around 81% for female speech utterances has been reported for MFCC and a GMM classifier [140]. The highest overall accuracy obtained in this work with the optimum feature set of all four formants and bandwidths was 97.14%. The formant feature group used in this work gave a recognition rate of 94.29% with the ANN classifier. A maximum recognition rate of 63.73% and 41.79% only, have been reported on the formant features of the emotional speech database EMO-DB and the DES respectively, with SVM classifiers [79]. A recent review has cited the average recognition rate for happy, anger, neutral and sad as 85.7%, obtained with acoustic prosodic and semantic label information [141].

6.10. Chapter Summary

This chapter has discussed the results of investigations on emotion and utterance discrimination, using eight vocal tract features comprising the first four formants and their respective bandwidths. Results of segmental level classification of utterances and emotions were discussed. Segmental level SR was found to be the most effective for anger and happiness and "u" was the best recognized of the five vowel sounds. Formant based SER at the segmental level resulted in 100% accuracy for happiness, surprise and neutral and was more effective than at the supra segmental level. Minimal feature sets of formants and bandwidths have been identified for simple, yet efficient class- seven SER in English, Hindi and Malayalam. The emotion classification rates obtained with such a minimal set of features were higher than those reported in these and other languages with larger feature sets. The classification rates mostly agreed with results of statistical discrimination. On the whole, higher SER rates were obtained in this investigation for spectral features than those obtained in Chapters 4 and 5. This further validates the adopted approach of using three base classifiers to identify the best features followed by the final ANN classification of the optimum, minimal feature set. For all three languages, the SER rate increased with decrease in the number of emotion classes, from seven. Universality in SER across the three languages has been verified for happiness, neutral, anger, fear and disgust.

The SER for Malayalam has been modeled better by using decision tree than using logistic regression, since the validation results indicate comparatively higher emotion recognition accuracies in the former case.

Chapter 7

CONCLUSION AND FUTURE WORK

Contents

This concluding chapter of this thesis is organized as follows. The specific contributions of this research are listed. This is followed by the assessment of the SER performance in terms of the proposed research objectives for the three languages, by the three feature sets, and at both suprasegmental level and segmental level (for English). This thesis finally concludes giving suggestions for future work from the perspective of the various results already obtained in these investigations.

7.1. Specific Contributions of this Research

This section of the chapter reiterates the principal contributions of this research. The most efficient feature / feature set for the SER in each of the three languages has been identified.

- There are no available reports of investigations involving SER in English, Hindi and Malayalam, with comparison of results thereof.
- Exclusive emotional speech databases in the three languages had been developed for this research and their emotional content was evaluated using acoustic perception tests.
- The investigations in speech had been conducted on elicited emotions as opposed to the prevalent use of acted speech for SER.

- Appropriate non-neutral content was used in this research for recording of emotions other than the neutral, which helped to obtain emotions with the right activation (whereas, most available reports of SER are based on emotional speech of neutral semantic content). This has been possible by designing the speech database after taking into account various social, psychological and linguistic aspects in the speech of educated, urban, females (as explained in Section 3.3 of Chapter 3).
- Manual segmentation had been employed for obtaining linguistically accurate utterance boundaries especially for the segmental level SER.
- Statistical analysis of the feature values (acoustic correlates) of emotional speech in the six basic emotions and the neutral, were done for the mean pitch, speech rate, duration, intensity, local jitter, shimmer and first four formant values along with their bandwidths. The statistical discrimination of these emotions was investigated using ANOVA. On the whole, the obtained SER rates agreed with the results of the ANOVA, showing higher SER rates for the best discriminated emotions. Segmental and suprasegmental pitch contours too were analysed for identifying characteristic features for the rule based classification of emotions by visual inspection of test pitch contours.
- The individual contribution of each of the above mentioned features to SER in the three languages was assessed and the best features for SER in each language were identified. The spectral features outperformed both prosodic features as well as their variations (comprising jitter and shimmer) in English Hindi and Malayalam, for SER.

The following facts were concluded from the evaluation of the valence dependency of SER rates for each feature and in each language: The obtained

SER rates were independent of the valence of emotions for suprasegmental utterances in English and Hindi.

- SER rates for Malayalam showed valence dependency (though to a small extent only) with pitch and jitter features.
- At the segmental level, valence dependency of SER rates was observed for SER based on F3, with positive valence emotions having slightly higher SER rates than the negative valence emotions. Also the final ANN classification results based on the optimum set of all the four formants revealed better performance for the positive valence emotions.

➢ Investigations regarding universality in the manifestation of any specific emotion, across English, Hindi and Malayalam gave positive results for the following cases:

- Surprise, anger and disgust, based on prosodic features.
- Fear and disgust, based on jitter and shimmer.
- Happiness, neutral, anger, fear and disgust, based on spectral features.

➢ Therefore it is concluded that disgust is universally recognised based on the features considered in this investigation.

➢ The SER for English was conducted at both the segmental and suprasegmental levels. This was followed by emotion wise and feature wise performance comparisons between the two levels. The following features of segmental level utterances outperformed those at the suprasegmental level in terms of the word based SER rates, for the listed emotions.

- Duration - surprise and disgust,
- Pitch - surprise, neutral, anger, sad, disgust,

- Jitter - all except anger,
- Shimmer - surprise, sad and fear,
- Formants - all emotions, except surprise (for which the suprasegmental SER rate was 100%).

These higher SER accuracies obtained at the segmental serve to reduce the complexity of the SER system. These further indicate the success of this approach to SER using minimal inputs, comprising small analysis units as well the use of a minimal feature set.

In the case of suprasegmental English utterances, 100% SER was achieved for the following features with,

- Intensity for all negative valence emotions and happiness
- Speech rate for happiness,
- Shimmer for disgust
- Formants for surprise.

The feature set of all four formants was sufficient for the classification of the seven emotions from suprasegmental English utterances (at the 60% SER rate cut off).

The features that gave 100% SER rates for Hindi utterances were:

- Intensity for sadness,
- Speech rate for happiness, surprise and anger.
- Pitch for anger
- Formants and bandwidths for all emotions except anger.

Therefore in the case of suprasegmental emotional utterances in Hindi, the spectral features consisting of the first four formants along with their bandwidths were sufficient to accurately classify all the seven emotions

considered in this work. Likewise the features that gave 100% SER rates for suprasegmental Malayalam utterances were:

- Speech rate for anger and sadness
- Pitch for disgust,
- Shimmer for anger and fear.
- Four formants and their bandwidths for fear.

Besides these, results of investigations with spectral features showed these to be sufficient for the accurate classification of all the emotions for suprasegmental Malayalam utterances.

- This research had investigated the role of formants in speech recognition at the segmental level, under various emotions. The results of experiments conducted with stand-alone vowel utterances indicate happiness, surprise, anger and disgust, apart from the neutral to be the most suitable for speech recognition. Fear and sadness were found to be the least favorable. "a", "i" and "u" were the best recognized.
- The specific contributions of this research work would be incomplete without a brief assessment of the various classifiers used in this work. The FCM, KMeans, KNN, Naive Bayes and ANN classifiers were used in this research to investigate speech emotion recognition in different languages. These classifiers served to assess the obtained SER rates, in terms of the consistency of emotion recognition. Assessment of the overall contribution to SER shows that the ANN classifier has given the best SER rates, across languages, features, analysis levels (segmental and suprasegmental levels) and emotions.
- The SER problem for Malayalam has been modelled appropriately using Decision trees and logistic regression, taking into account the emotional

content and valence of emotions. The modeling was carried out using formants and bandwidths. In the classification of seven emotions by logistic regression, even though the model fitted the data well (P value for Pearson test was 0.998), the obtained classification / prediction accuracy was not very high.

- The decision tree provided a white box model for the SER in Malayalam indicating clearly the order of selection of features along with their ranges in the course of classification. The models were validated by resubstitution and cross validation. In this thesis work, high accuracies were obtained when the decision tree classifier was used for the three binary classifications (cases (i) to (iii)) as well as multiclass classifications of neutral, positive and negative valence emotions.

7.2. Suggestions for Future Work

The effectiveness of the approach adopted in this research work for the analysis and classification of female, elicited emotional speech in the Indian context has been clearly demonstrated through the various obtained results presented in Chapters 4-6. This section is a pointer to the various issues that can be addressed in future research in SER.

- This research had been restricted to Ekman's basic set of six emotions and the neutral, considering these adequate to cover the normal ambit of emotions in everyday life. Therefore, implementation of a similar study for the extended set of basic emotions namely; contentment, embarrassment, excitement, relief, satisfaction, shame and guilt could be a possible direction for future research.
- The SER with emotional speech of males and transgenders can be investigated for these or other languages.

- In addition to the three languages considered for this research on SER, well designed and developed speech databases of other popular Indian languages like Tamil, Telugu and Kannada could also be included for a more comprehensive study.
- Only one popular dialect of Malayalam and Hindi, as used by the educated, urban middle-class has been considered for the speech database. The other dialects have not been looked into, due to limitations of time and since it would also result in a more extensive database. Hence, another objective for future work could be to study the extent of variation of emotional expressions across dialects within any particular language itself.
- Speech emotion recognition could be pursued with various other features and classifiers, other than those used in this investigation. The SER can be designed to incorporate sophisticated optimization techniques, and also aim at real time identification of emotions.
- The SER in other languages such as Hindi, English, etc. may be modelled using Decision Trees, Logistic regression or other methods.

REFERENCES

[1] Cowie, R. E., Douglas-Cowie, Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J.G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32-80.

[2] Frijda, N.H. (1986). *The Emotions*. Cambridge University Press: Cambridge.

[3] Candland, D. (1977). The persistent problems of emotion. In D. Candland, Fell, J., Keen, E., Leshner, A., Tarpy, R., and Plutchik, R., (Eds.,) *Emotion* (pp.2-84). Monterey, C.A: Brooks – Cole.

[4] Ekman, P., Ricard Sorenson, E., Friesen, W.V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, New Series, 164 (3875): 86-88.

[5] Ling He (2010). "Stress and Emotion Recognition in Natural Speech in the Work and Family Environments" Ph.D. Thesis. School of Electrical and Computer Engineering Science, RMIT University. Australia.

[6] Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40: 227-256.

[7] Neviarouskaya, A., Prendinger, H., Ishizuka, M. (2010). EmoHeart: Conveying emotions in second life based on affect sensing from text. *Advances in Human-Computer Interaction*, 2010:13 pages.

[8] Sommers, S.M. (2006). Evaluating voice-based measures for detecting deception. *The Journal of Credibility Assessment and Witness Psychology*, 7(2): 99-107.

[9] Sloman, A., and Croucher, M. (1981). Why robots will have emotions In *Proc. of the 7th International Joint Conference on AI*, IJCAI, 1981, pp. 197-202.

[10] Thayer, R.E. (1989). *The Biopsychology of Mood and Arousal*. Oxford University Press: New York.

[11] Oakley, K., and Jenkins, J.M. (1996). *Understanding emotions.* MA: Blackwell Publishers: Cambridge.

[12] Luca, J., and Tarricone, P. (2001). Does emotional intelligence affect successful teamwork? Meeting at the Crossroads. In *Proc.of the ASCILITE*, 2001, pp. 367-376.

[13] Vogt, T., Andr´e, E., and Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. *Affect and Emotion in Human-Computer Interaction.*

Lecture Notes in Computer Science. C. Peter and R. Beale (Eds.). 4868, pp. 75-91. Springer.

[14] Agrawal, S.S. (2011). Emotions in Hindi speech - analysis, perception and recognition, In *Proc. Oriental COCOSDA*, 2011, pp. 7-13.

[15] Roach, P. (1997). *English Phonetics and Phonology*. Second edition. Cambridge University Press: Cambridge.

[16] Douglas O' Shaughnessy, D.O'. (2001). *Speech Communication: Human and Machine*. Universities Press India Private Limited: India.

[17] Ververidis, D.and Kotropoulos, C. (2006) Emotional speech recognition: resources, features, and methods. *Speech Communication*, 48(9):1162- 1181.

[18] Silva, D.G., Oliveira, L.C., and Andrea, M. (2009). Jitter estimation algorithms for detection of pathological voices. *EURASIP Journal on Advances in Signal Processing*, 2009:9 pages.

[19] Brockmann, M., Drinnan , M.J, Storck C., Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task, *Journal of Voice*, 25: 44-53.

[20] Oliver, R.L. (1997) *Satisfaction: A behavioural perspective on the consumer*. First Edition, Mc Graw Hill Education: New York.

[21] Cornelius, R. R. (2000). Theoretical Approaches to Emotion, In *Proc. of the ISCA ITRW on Speech and Emotion*, 2000, pp. 2755-2758.

[22] Ekman, P. (1999). Facial Expressions. Chapter 16, in *The Handbook of Cognition and Emotion.* Eds. Dalgleish, T. and Power, M.J., John Wiley and Sons Ltd, New York. pp. 301-320.

[23] Lazarus, R.S. (1982).Thoughts on the relations between emotion and cognition. *American Psychologist*, 37(9):1019-1024.

[24] Averill, James R. (1983). Studies on Anger and Aggression. Implications for Theories of Emotion. *American Psychologist*, 38(11):1145- 1160.

[25] Derry, S.J. (1999). A Fish called peer learning: *Searching for common themes*, A.M .O' Donnell and A. King (Eds.).

[26] Mc Mahon, M. (1999). Social constructivism and the World Wide Web - A Paradigm for Learning, In *Proc. ASCILITE*, 1999.

http://www.curtin. edu.au:80/conference/ASCILITE97/papers

[27] Busso, C. and Narayanan, S.S. (2007). Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8): 2331-2347.

[28] Calvo, A.R. and D'Mello, S. (2010). Affect Detection: An inter disciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1): 18 - 37.

[29] Plutchik, R. (2001). The Nature of Emotions. *American Scientist*, 89(4): 344-350.

[30] Jack, E.R., Garrod, G.B. and Schyns. P.G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, 24(2):187-192.

[31] Pantic, M. Rothkrantz, and L. J. M. (2003). toward an affect-sensitive multimodal human-computer interaction, in P*roc. of the IEEE,* 91(9) pp. 1370-1390.

[32] Swati Patra (2004). Role of Emotional Intelligence in Educational management, *Journal of Indian Education*, 2004: 98-104.

[33] Mayer, J.D., Salovey, P., David R., Caruso D.R., Sitarenios. G. (2001). Emotional Intelligence as Standard Intelligence. *Emotion* 1(3): 232-242.

[34] Kappas, A. (2010). Smile when you read this, whether you like it or not: conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, 1(1): 38-41.

[35] Izard, E.C. (2010).The many meanings / aspects of emotion: definitions, functions, activation, and regulation, *Emotion Review*, 2(4): 363–370.

[36] Zachar, P. (2010). Has there been conceptual progress in the science of emotion*?, Emotion Review*. 2 (4): 381–382.

[37] Espinosa, H.P. Garcia, J.O. ; Pineda, L.V. (2010). Features selection for primitives estimation on emotional speech, In *Proc. of the IEEE ICASSP* 2010, pp. 5138 -5141.

[38] Lee, S. (2007). "Discrete emotion and motivation: relative activation in appetitive and aversive motivational system as a function of anger, sadness, fear, and joy embedded in the content of televised information campaigns". Ph.D. Dissertation, Indiana University. USA 154 pages.

[39] Grimm, M., Kroschel, K., Narayanan, S. (2008). The Vera Am Mittag German Audio-Visual Emotional Speech Database, In *Proc. of the ICME*. pp. 865-868.

[40] Sundberg, J., Patel, S. Bjorkner, E., and Scherer, K.R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, 2(3):162-174.

[41] Fontaine, J., Scherer, K.R., Roesch, E.B., and Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050-1057.

[42] Pallier, C., Bosch, L. and Sebasti´an-Gall´es, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition* 64 (3): B9 - B17.

[43] Giri, V.N. (2004). *Gender role in communication style*, IIT Kharagpur, Concept publishing House: New Delhi.

[44] Tannen, D. (1986). *That's not what I meant! How conversational style makes or breaks relationships*. Ballantine books: NewYork.

[45] Yrizarry, N., Matsumoto, D. and Wilson-Cohn, C. (1998). American-japanese differences in multiscalar intensity ratings of universal facial expressions of emotion. *Motivation and Emotion*, 22(4): 315-327.

[46] Matsumoto, D. (2006). Are cultural differences in emotion regulation mediated by personality traits. *Journal of Cross-cultural psychology* 37(4):421- 437.

[47] Álvarez, A., Cearreta, I., López, J.M., Arruti, A., Lazkano, E., Sierra, B., Garay, N. (2007). Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech, In Proc. *of the 4th international conference on Non linear speech processing*, NOLISP 2007, Paris, May 2007, pp.71-74.

[48] Banziger, T., Scherer, .K.R. (2005).The role of intonation in emotional expressions. *Speech Communication,* 46(3-4): 252 – 267.

[49] Yacoub, S., Simske, S., Lin, X., Burns,J. (2003). Recognition of emotions in interactive voice response systems, *In Proc. of 8th European Conference on Speech Communication and Technology, Eurospeech 2003*, Geneva, September 2003, pp. 729-732.

[50] Ramakrishnan, S. & Ibrahiem M.M. El Emary. (2011). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*. 46 (3):191-193.

[51] Lee, C. M. and Narayanan, S.S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*.13 (2): 293-303.

[52] Bowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1535): 3515–3525.

[53] Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1171-1178.

[54] Schuller, B., Seppi, D., Batliner, A., Maier,A., and Stefan Steidl, (2007). Towards more reality in the recognition of emotional speech. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007*. 4:941-944,

[55] Weninger, F., Bjorn Schuller, B., Batliner, A., Stefan Steidl, and Dino Seppi, D. (2011). Recognition of non prototypical emotions in reverberated and noisy speech by non-negative matrix factorization. *EURASIP Journal on Advances in Signal Processing,* 2011, Article ID 838790, 16 pages. DOI:10.1155/2011/838790.

[56] Hansen, J.H.L., Bou - Ghazale, SE., Sarikaya R., Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. *Eurospeech*. 97(4):1743-1746. .

[57] Johnstone, T., Van Reekum, C.M., Kathryn Hird, Kirsner, K., Scherer, K.R. (2005). Affective Speech Elicited With a Computer Game. *Emotion* 5(4): 513-518.

[58] McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schro¨der, M. (2007). The SEMAINE Database: annotated multimodal records of emotionally colored conversations between a person and a limited agent, *IEEE Transactions on Affective Computing.* 6(1): 5-17.

[59] Ververidis D. and Kotropoulos, C. (2003). Review of emotional speech databases, in *the Proc.of the 9th Panhellenic Conference on Informatics (PCI),* Greece, November 2003, pp. 560-574.

[60] Aggarwal, R.K. and Dave, M. (2013). Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system. *Telecommunication Systems*, 52(3): 1457-1466.

[61] Campbell, N. (2005). Developments in corpus-based speech synthesis: approaching natural conversational speech. *IEICE Transactions on Information and System.* 88(3):376-383.

[62] Ng, R.W.M., Lee, T., Chi Leung, C., Ma, B., Li, H. (2009). Analysis and selection of prosodic features for Asian language recognition. *International Journal on Asian Language Processing.* 19(4):139-152.

[63] Hirschberg, J., Liscombe, J., and Venditti, J. (2003). Experiments in Emotional Speech, In *Proc. of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo. April 13-16, 7 pages. ISCA. http://www.isca-speech.org/archive. 2003.

[64] Zhongzhe Xiao & Emmanuel Dellandrea and Weibei Dou & Liming Chen. (2010). Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools and Applications.* 46(1):119-145.

[65] Petrushin, V.A. (2000). Emotion recognition in speech signal: Experimental study, development, and application, In *the Proc.of the ICSLP* 2000, pp. 222- 225.

[66] Cai, L., Jiang, C., Wang, Z., Zhao, L., and Zou, C. (2003). A method combining the global and time series structure features for emotion recognition in speech, *In Proc. of the International Conference on Neural Networks and Signal Processing*, 2:904-907.

[67] Shami, M., and Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classifications of emotions in speech. *Speech Communication*, 49: 201-212.

[68] Altun, H and Polat, G. (2009).Boosting selection of speech related features to improve performance of multiclass SVMs in emotion detection. *Expert Systems with Applications*, 36: 8197-8203.

[69] Karlsson, T. Banziger, J. Dankovicov, T. Johnstone, J. Lindberg, H. Melin, Nolan, F., K. Scherer. (2000). Speaker verification with elicited speaking styles in the VeriVox project. *Speech Communication*, 31:pp.121-129. 2000.

[70] Theune, M., Meijs, K., Heylen, D., and Ordelman, R. (2006). Generating Expressive Speech for Story telling Applications. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4): 1137-1144.

[71] Rouas, J.L. (2007). Automatic Prosodic Variations Modeling for Language and Dialect Discrimination. *IEEE Trans. Audio, Speech, and Language Processing*, 15(6):1904-1911.

[72] Mary, L. and Yegnanarayana, B. (2008). Extraction and Representation of Prosodic Features for Language and Speaker Recognition. *Speech Communication*. 50(10):782-796.

[73] Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

[74] Li, Xi, Tao, Jidong Johnson, M.T., Soltis, S., Savage,A., Kirsten M. Leong, Newman, J.D. (2007). Stress and emotion classification using jitter and shimmer features. In *Proc. of IEEE, ICASSP 2007*. IV - 1081 -1084.

[75] Farrús, M., Hernando, J., Ejarque,P. (2007). Jitter and shimmer measurements for speaker recognition. *interspeech, 2007*, pp. 778 - 781.

[76] Nwe, T.L., Foo, S.W., Liyanage C. De Silva, L.C. (2014). Speech emotion recognition using hidden Markov models. *Speech Communication.* 41:603–623.

[77] Ververidis, D., Kotropoulos C. (2004). Automatic speech classification to five emotional states based on gender information, In *Proc. of 12th European Signal Processing Conference*. Austria, pp. 341–344.

[78] Ververidis, D., Kotropoulos, C. (2005). Emotional speech classification using Gaussian Mixture models and the Sequential Floating forward selection algorithm, In *IEEE International Conference on Multimedia and Expo, ICME*, 2005. pp. 1500 - 1503.

[79] Schuller, B., Arsi, D., and Wallhoff, F., Lang, M. and Rigoll, G. (2005). Bioanalog acoustic emotion recognition by genetic feature generation based on low-level descriptors In *the International Conference on computer as a tool. EUROCON 2005*, Belgrade. 2:1292 – 1295.

[80] Athanaselis, T., Bakamidis, S., Dologlou, I. (2005). Automatic recognition of emotionally coloured speech. *World Academy of Science, Engineering and Technology*, 12: 484-487.

[81] Ramamohan, S.and Dandapat, S. (2006). Sinusoidal model-based analysis and classification of stressed speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3): 737 -746.

[82] Navas, E., Hernáez, I. and Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS, *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1117-1128.

[83] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao (2007). manifolds based emotion recognition in speech. *Computational Linguistics and Chinese Language Processing*, 12 (1): 49-64.

[84] Neiberg, D., Elenius, K. (2008). Automatic recognition of anger in spontaneous speech, *In Proc.of the ISCA. Interspeech 2008*, Australia, pp. 2755-2758.

[85] Wang, Y. and Guan, L. (2008). Recognizing Human Emotional State from Audiovisual Signals. *IEEE Transactions on Multimedia*, 10(5): 936-946.

[86] Bozkurt, E., Engin Erzin, Ci˘gdem Ero˘glu Erdem, Erde, A.T. (2009). Improving automatic emotion recognition from speech signals, Interspeech 2009. *Emotion Challenge*, 4 pages.

[87] Park, J.S., Kim, J.H., Yung-Hwan. (2009). Feature Vector Classification based Speech Emotion Recognition for Service Robots. *IEEE Transactions on Consumer Electronics*, 55(3):1590-1596.

[88] Eye, F., Wollmer, M., and Schuller, B. (2009). OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit, In *Proc. of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. *ACII 2009*, pp.1–6.

[89] Busso, C., Lee, S and Narayan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):582-596.

[90] Kim, E.H, Hyun, K.H., Kim, S.H., Kwak, Y.K. (2009). Improved emotion recognition with a novel speaker-independent feature. *IEEE / ASME Transactions on Mechatronics*, 14(3):317-325.

[91] Gharavian, D., Sheikhan, M., Janipour, M. (2010). Pitch in emotional speech and emotional speech recognition using pitch frequency *Majlesi Journal of Electrical Engineering*, 4(1):19-24.

[92] Gharavian, D., Sheikhan, M. (2010). Emotion recognition and emotion spotting improvement using formant-related features. *Majlesi Journal of Electrical Engineering*, 4(4):1-8.

[93] Stefan Steidl, Anton Batliner, Dino Seppi, and Bjorn Schuller. (2010). On the impact of children's emotional speech on acoustic and language models, *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, Article ID 783954, 14 pages. DOI:10.1155/2010/783954.2010.

[94] Yoon, W.J and Park, K.S. (2011). Building robust emotion recognition system on heterogeneous speech databases. *IEEE Transactions on Consumer Electronics*, 57(2):747 -750.

[95] Caponetti, L., Buscicchio, C.A, and Castellano, G. (2011). Biologically inspired emotion recognition from speech. *EURASIP Journal on Advances in Signal Processing*, 2011: 24, 10 pages.

[96] Mower, *E.*, Maja J Mataric´, Narayanan, S. (2011). A Framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech and Language Processing*. 5:1057 - 1107.

[97] Chung Wu, H and Wei.B.L. (2012). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on affective computing*, 2(1):- 21.

[98] Ntalampiras, S. and Fakotakis, N. (2012). Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*, 3(1):116-125.

[99] Koolagudi, S.G., Sreenivasa Rao, K. (2012). Emotion recognition from speech: a review. *Int. Journal of Speech Technology* 15:99-117. DOI 10.1007/s10772-011-9125-1. 2012.

[100] Yun, S., and Yoo, C.D. (2012). Loss-scaled large-margin Gaussian mixture models for speech emotion classification. *IEEE transactions on Audio, Speech, and Language Processing*, 20(2): 585-598.

[101] Feraru, S.M. (2012). Speech Emotion Analysis in the Romanian Language Economy Transdisciplinarity. *Cognition*, 15(1): 267-272,

[102] Hassan, A. and Damper, R.I. (2012). Classification of emotional speech using 3DEC hierarchical classifier. *Speech Communication*, 54(7) 903-916.

[103] Attabi. Y. and Dumouchel, P. (2013). Anchor Models for Emotion Recognition from Speech. *IEEE Transactions on Affective Computing*, 4(3): 290 -390.

[104] Fujisaki, H., and Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics.*16(1):73-77.

[105] Iriondo, I., Planet, S., Socor´o, J.C., Francesc Al´ıas. (2007). Objective and subjective evaluation of an expressive speech corpus, *In Proc. of the 4th international conference on Non linear speech processing. NOLISP* 2007. Paris. May2007, pp. 15-18.

[106] Dowdy, S., Wearden, S., Chilko, D. (2004). *Statistics for Research* John Wiley and sons: New Jersey:

[107] Ekman. P. (1982). *Emotions in the Human Face*, Cambridge University Press.

[108] Daniel Jones. (2004). *Cambridge English pronouncing dictionary*. Cambridge University Press: Cambridge.

[109] Bern, S.L. (1981). Gender Schema Theory: A Cognitive account of sex typing. *Psychological review*, 88: 354 -364.

[110] Gumperz, J. J. and Herasimchuk, E. (1982). Intonation and meaning in conversation, language and communication., 2: 123–131.

[111] Chateau, N., Maffiala, V., Ehrette, T., d'Alessandro, C. (2002). Modeling the emotional quality of speech in a telecommunication context, In *Proc. of the 2002 International Conference on Auditory Display*, Japan, July 2002.

[112] Karlsson, T. Banziger, J. Dankovicov, T. Johnstone, J. Lindberg, H. Melin, F. Nolan K. Scherer (2000). Speaker verification with elicited speaking styles in the VeriVox project. *Speech Communication* 31:121-130.

[113] Boersma, P.P and Weenink, D., "Praat: doing phonetics by Computer (Version 4.6.09)," 2005 [Computer program], http://www.Praat.org/.

[114] Rabiner, L.R. (1977). On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(1): 24-33.

[115] Boersma, P. (1993). Accurate short –term analysis of the fundamental frequency and the Harmonics to nois ratio of a sampled sound." In *IFA Proceedings*, University of Amsterdam, 17: 97-110.

[116] Lieberman, P., Michaels, S.B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America*, 34: 922-927.

[117] Snell, C.R. and Milinazzo, F. (1993). Formant Location from LPC Analysis data. *IEEE Transactions on Speech and Audio Processing* 1(2).

[118] Deller Jr., J.R., Proakis, J.G., Hansen, J.H. (1993). *Discrete-Time Processing of Speech Signals.* New York: Macmillan Publishing Company.

[119] Andrew Gelman. (2005). Analysis of variance – why it is more important than ever. *The Annals of statistics*, 33(1):1-53.

[120] Talar M. Hopyan, Misakyan, Karen A. Gordon, Maureen Dennis and Blake C. Papsin. (2009). Recognition of affective speech prosody and Facial affect in deaf children with unilateral Right cochlear implants. *Child Neuro Psychology Press,* Taylor & Francis Group, 15: 136 – 146.

[121] Nayak, G.S., Davide, O. Puttamadappa C. (2010). Classification of bio optical signals using KMeans clustering for detection of skin Pathology. *International Journal of Computer Applications,* l (2): 91-96.

[122] Gose, E., Johnsonbaugh, R. and Jost, S. (2000). *Pattern Recognition and Image Analysis.* Prentice Hall India: New Delhi.

[123] Lee C.M., and Narayanan, S. (2003) Emotion recognition using a data-driven fuzzy inference system, In *the Proc. of the European Conference on Speech Communication and Technology*, pp. 157-160.

[124] Lior Rokach, Oded Maimon, (2014). *Data Mining with Decision Trees Theory and Applications,* 2nd Edition. Series in Machine Perception and Artificial Intelligence: Volume 81, ISBN: 978-981-4590-07-5

[125] Masami Akamine and Jitendra Ajmera, (2012), Decision tree-based acoustic models for speech recognition *EURASIP Journal on Audio, Speech, and Music Processing* 2012. **DOI:** 10.1186/1687-4722-2012-10.

[126] Duda O.R., Stork D.G., (2001): *Pattern Classification*. 2nd edition. John Wiley & Sons, , Hoboken, NJ, USA

[127] Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com).

[128] Agresti, A., (2007). *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[129] Murray, I., Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion *Journal of the Acoustical Society of America*. 93(2), pp. 1097–1108.12

[130] Polzin T, Waibel A (2000). Emotion-sensitive human- computer interface. In *the Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 201 – 206, Newcastle, Northern Ireland.

[131] Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and pathological larynges. *Journal of the Acoustic Society America* 35(3).

[132] Jang, K.D and Kwon, O.W. (2006). Speech Emotion Recognition for affective human-robot interaction, *SPECOM* 2006, St. Petersburg, 25-29 June 2006, pp. 419-422.

[133] Hendy, N. A. and Farag, H. (2013). Emotion Recognition Using Neural Network: A Comparative Study. *World Academy of Science, Engineering and Technology*. 75:1149-1155.

[134] Tolkmitt, F. J., Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12 (3):302–313.

[135] Schuller, B., Rigoll, G. and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector Machine - belief network architecture. *International Conference on Acoustics, Speech, and Signal Processing- ICASSP* 2004, pp. 577-580.

[136] Firoz Shah, A. Raji Sukumar, and Babu Anto, P. (2010). Discrete wavelet transforms and artificial neural networks for speech emotion recognition. *International Journal of Computer Theory and Engineering*, 2(3):319-321.

[137] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G.,and Wendemuth, A. (2009). Acoustic Emotion Recognition: A Benchmark Comparison of Performances". *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU 2009)*, Merano, Italy, pp. 552–557.

[138] Stuhlsatz, A., Meyer, C., Eyben, F. Zielke, T., Meier G., Schuller, B. (2011). Deep Neural networks for acoustic emotion recognition: raising the benchmarks In *Proc. of ICASSP 2011*. pp.5688-5692.

[139] Anurag Jain, Nupur Prakash, S.S. Agrawal (2011). Evaluation of MFCC for Emotion Identification in Hindi Speech, In *Proc. of IEEE 3rd International Conference on Communication Software and Networks* (*ICCSN*) May 2011, pp. 189-193.

[140] G. Koolagudi, Maity, S., Anilkumar V., Chakravarti S., Sreenivasa Rao. (2009). IIT KGP-SESC: Hindi speech data base for emotion analysis In *Proc. International Conference on Information, Communication and Computing*, 40: pp.485-492.

[141] Lanjewar, R.B. and D. S. Chaudhari (2013). speech emotion recognition: a review. *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), 2(4), March 2013.

Appendix A

SEGMENTAL UTTERANCES

Happy	Surprise	Neutral	Anger	Sad	Fear	Disgust
Aye (address)	Aye (What a surprise!)	Aye	Aye (reprimand)	Aye (no hope)	Aye	Aye (ugh!)
Eh? (prompt)	Eh?	Eh	Eh? (rough)	Eh…	Eh?	Eh?!
I	I?	I	I (growl)	I	I (hesitant)	I
Oh (please come in)	Oh! (That's great)	Oh (great thing!)	Oh	Oh	Oh (bridge is falling!)	Oh! (Stinking)
You	You?	You	You (are responsible)	You (pathetic)	You(fear)	You?

Appendix B

SUPRA SEGMENTAL UTTERANCES IN HINDI AND MALAYALAM

Emotion Specific Hindi Utterances

Emotion	Specific Utterances in Hindi
खुशी	अन्दर आओ
आश्चर्य	कितना सुन्दर मकान!
निर्विकार	हो सकता है
गुस्सा	चुप रहो
दुख	बहुत बुरी बात है
डर	मुझे डर लग रहा है
घृणा	खाना खराब है

Emotion Specific Malayalam Utterances

Emotion	Specific Utterances in Malayalam
സന്തോഷം	എത്രനാളായി കണ്ടിട്ട്
അതിശയം	എന്തൊരുഭംഗി
നിർവ്വികാരം	അതെവിടെ വച്ചോളൂ
ദേഷ്യം	എനിക്ക് വേറെ പണിയുണ്ട്
സങ്കടം	അതൊരപകടമായിരുന്നു
പേടി	എനിക്ക് പേടിയാകുന്നു
അറപ്പ്	എന്തൊരു വൃത്തികേട്

Utterances Common to all Emotions

	Hindi	Malayalam
1.	अन्दर आना	ശരി
2.	करेगा क्या	തീർച്ചയായും
3.	ज़रूर	നോക്കട്ടെ
4.	जल्दी में है क्या	തിരക്കാണോ
5.	हे ईश्वर प्रभु	എന്റെ ദൈവമേ
6.	शुक्रिया	ഇതോ

Appendix C

DESCRIPTIONS OF UNPRUNED DECISION TREES FOR A BINARY CLASSIFICATION

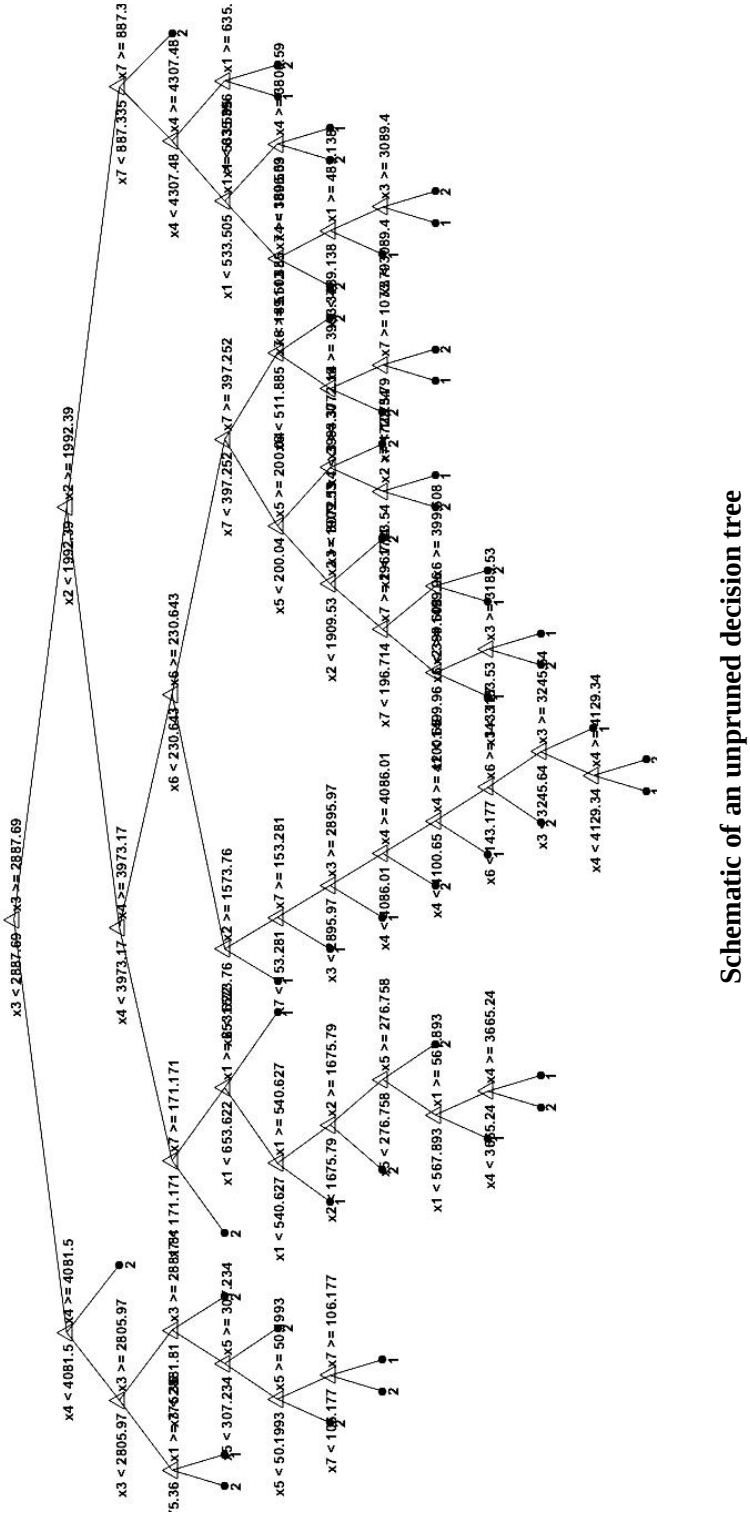
The textual description of an unpruned decision tree used for the classification of neutral and emotional speech was obtained as follows:

Textual description of unpruned decision tree for binary classification

```
 1  if x3<2887.69 then node 2 elseif x3>=2887.69 then node 3 else 1
 2  if x4<4081.5 then node 4 elseif x4>=4081.5 then node 5 else 2
 3  if x2<1992.39 then node 6 elseif x2>=1992.39 then node 7 else 1
 4  if x3<2805.97 then node 8 elseif x3>=2805.97 then node 9 else 2
 5 class = 2
 6  if x4<3973.17 then node 10 elseif x4>=3973.17 then node 11 else 2
 7  if x7<887.335 then node 12 elseif x7>=887.335 then node 13 else 1
 8  if x1<775.36 then node 14 elseif x1>=775.36 then node 15 else 2
 9  if x3<2881.81 then node 16 elseif x3>=2881.81 then node 17 else 1
10  if x7<171.171 then node 18 elseif x7>=171.171 then node 19 else 1
11  if x6<230.643 then node 20 elseif x6>=230.643 then node 21 else 2
12  if x4<4307.48 then node 22 elseif x4>=4307.48 then node 23 else 1
13  class = 2
14  class = 2
15  class = 1
16  if x5<307.234 then node 24 elseif x5>=307.234 then node 25 else 1
17  class = 2
18  class = 2
19  if x1<653.622 then node 26 elseif x1>=653.622 then node 27 else 1
20  if x2<1573.76 then node 28 elseif x2>=1573.76 then node 29 else 2
21  if x7<397.252 then node 30 elseif x7>=397.252 then node 31 else 1
22  if x1<533.505 then node 32 elseif x1>=533.505 then node 33 else 1
23  if x1<635.396 then node 34 elseif x1>=635.396 then node 35 else 2
24  if x5<50.1993 then node 36 elseif x5>=50.1993 then node 37 else 1
25  class = 2
```

```
26  if x1<540.627 then node 38 elseif x1>=540.627 then node 39 else 1
27  class = 1
28  class = 1
29  if x7<153.281 then node 40 elseif x7>=153.281 then node 41 else 2
30  if x5<200.04 then node 42 elseif x5>=200.04 then node 43 else 2
31  if x6<511.885 then node 44 elseif x6>=511.885 then node 45 else 1
32  if x7<189.503 then node 46 elseif x7>=189.503 then node 47 else 1
33  if x4<3806.59 then node 48 elseif x4>=3806.59 then node 49 else 1
34  class = 1
35  class = 2
36  class = 2
37  if x7<106.177 then node 50 elseif x7>=106.177 then node 51 else 1
38  class = 1
39  if x2<1675.79 then node 52 elseif x2>=1675.79 then node 53 else 1
40  class = 1
41  if x3<2895.97 then node 54 elseif x3>=2895.97 then node 55 else 2
42  if x2<1909.53 then node 56 elseif x2>=1909.53 then node 57 else 2
43  if x3<3072.19 then node 58 elseif x3>=3072.19 then node 59 else 1
44  if x4<3993.37 then node 60 elseif x4>=3993.37 then node 61 else 1
45  class = 2
46  class = 2
47  if x1<489.138 then node 62 elseif x1>=489.138 then node 63 else 1
48  class = 2
49  class = 1
50  class = 2
51  class = 1
52  class = 2
53  if x5<276.758 then node 64 elseif x5>=276.758 then node 65 else 1
54  class = 1
55  if x4<4086.01 then node 66 elseif x4>=4086.01 then node 67 else 2
56  if x7<196.714 then node 68 elseif x7>=196.714 then node 69 else 2
57  class = 2
58  if x2<1723.54 then node 70 elseif x2>=1723.54 then node 71 else 1
59  class = 2
60  class = 2
61  if x7<1073.79 then node 72 elseif x7>=1073.79 then node 73 else 1
62  class = 1
```

63 if x3<3089.4 then node 74 elseif x3>=3089.4 then node 75 else 1
64 if x1<567.893 then node 76 elseif x1>=567.893 then node 77 else 1
65 class = 2
66 class = 2
67 if x4<4100.65 then node 78 elseif x4>=4100.65 then node 79 else 2
68 if x2<1499.96 then node 80 elseif x2>=1499.96 then node 81 else 2
69 if x6<399.608 then node 82 elseif x6>=399.608 then node 83 else 1
70 class = 2
71 class = 1
72 class = 1
73 class = 2
74 class = 1
75 class = 2
76 class = 1
77 if x4<3665.24 then node 84 elseif x4>=3665.24 then node 85 else 1
78 class = 1
79 if x6<143.177 then node 86 elseif x6>=143.177 then node 87 else 2
80 class = 1
81 if x3<3183.53 then node 88 elseif x3>=3183.53 then node 89 else 2
82 class = 1
83 class = 2
84 class = 2
85 class = 1
86 class = 2
87 if x3<3245.64 then node 90 elseif x3>=3245.64 then node 91 else 2
88 class = 2
89 class = 1
90 if x4<4129.34 then node 92 elseif x4>=4129.34 then node 93 else 2
91 class = 1
92 class = 1
93 class = 2



Schematic of an unpruned decision tree

PUBLICATIONS

This section presents the list of publications of the author, based on this thesis.

Journals

1) Agnes Jacob and P. Mythili. (2013). Upgrading the performance of speech emotion recognition at the segmental level. *Journal of Computer Engineering, International Organisation of Scientific Research*. 15(3): 48-52.

2) Agnes Jacob and P. Mythili. (2013). Minimal feature set based classification of emotional speech. *International Journal of Scientific and Engineering Research*. 4(11): 46-51.

3) Agnes Jacob and P. Mythili. (2013). An Improved Hindi Speech Emotion Recognition System. *International Journal of Innovative Technology and Exploring Engineering*.3 (6): 25-29.

Proceedings of International Conferences

1) Agnes Jacob and P. Mythili. (2010) Development and Evaluation of emotional speech databases. In the *Proceedings of the 4th International Conference on Intelligent Systems and Control* (ISCO2010), Coimbatore, February 2010.

2) Agnes Jacob and P. Mythili. (2010). Empowerment of Women through assessment of Vocal Expressions of Emotions. In the *Proceedings of the International Conference on Women's Education for Empowerment* (ICW2010). Puducherry, February 2010

3) Agnes Jacob and P. Mythili. (2012). A socio friendly approach to the analysis of emotive speech. In the *SciVerse Science Direct; Procedia Engineering* 30 (2012) 577–583. Elsevier. *Proceedings of the International Conference on Communication Technology and System Design.* Coimbatore, December 2011.

4) Agnes Jacob and P. Mythili. (2011). Jitter measurements for performance enhancement in the service sector. *In the Proceedings of the* Annual *IEEE India Conference (INDICON)* December 2011, Hyderabad, pp. 1-4 DOI: 10.1109/INDCON.2011.6139496.

5) Agnes Jacob and P. Mythili. (2015). Prosodic feature based speech emotion recognition at segmental and suprasegmental levels. *In IEEE International Conference on Signal Processing, Informatics, Communication and Energy systems (SPICES), Kozhikode 2015, pp.1-5* DOI: 10.1109/ SPICES.2015.7091377.